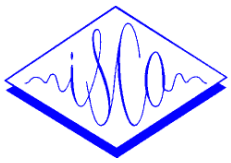


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

INTEGRATING TDNN-BASED DIPHONE RECOGNITION WITH TABLE-DRIVEN MORPHOLOGY PARSING FOR UNDERSTANDING OF SPOKEN KOREAN

Kyunghee Kim**, Geunbae Lee**, Jong-Hyeok Lee**, and Hong Jeong*

Department of Computer Science**, Department of Electronic
& Electrical Engineering*, POSTECH, KOREA

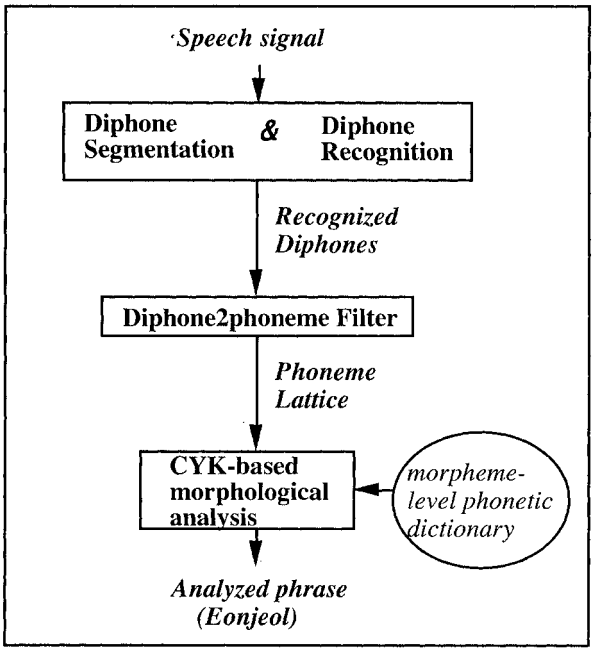
ABSTRACT

In this paper, we propose a spoken Korean morphological analysis model extensible to large vocabulary continuous speech recognition. This model consists of a diphone recognizer, a diphone2phoneme filter and a CYK-morphological analyzer. Two-level hierarchical TDNNs (time-delayed neural networks) recognize Korean diphones which are transformed into a phoneme lattice (a set of phoneme candidates hypothesized by a speech recognition module) by a diphone2phoneme filter. The morphological analyzer parses the phoneme lattice with the phonological changes handling and produces the morphology-segmented Korean words (called Eojeols). Using the TDNN diphone speech recognizer, we obtained 95.2% of 17 Korean vowel recognition and 93.7% of 72 diphone recognition. The speaker-dependent and continuous Eojeol recognition experiments using the current model show that the morphological analysis for spoken Korean can be achieved for medium sized vocabularies with 90.6% of success rate.

1. INTRODUCTION

Spoken language understanding system requires an integrated speech and natural language processing. To integrate speech recognition with natural language processing, the final speech recognition results have to be morphologically analyzed words (called Eojeols in Korean). A Korean Eojeol usually consists of more than one morphemes. To recognize large spoken Eojeols, the recognition unit has to be smaller than the word, such as phoneme. However phoneme-level recognition and its post-processing is very difficult especially in Korean because of the lack of distinctive features and characteristics of the diphthongs. Many rising diphthongs are very similar to mono-vowels, which makes it difficult to distinguish between them [1][2]. Moreover, different positioned same consonants (first consonant and final consonant in a syllable) make a recognition difficult, too [3]. In the case of Japanese that has less phonemes than Korean, the phoneme-based recognition rate is very high. But the experiment results for words show that the postprocessing of phoneme-level recognition is still difficult [4]. Therefore, instead of direct phoneme-level recognition, we adopted a new Korean Eojeol recognition unit - diphone with 4 different types. To recognize the diphones, we use the well-known TDNN model that has excellent phoneme spotting capabilities [5]. The recognized diphones are transformed into a phoneme lattice later using a simple rearrangement scheme. We suggest a new post-processing method based on the phoneme lattice for continuously spoken Korean Eojeol. The method analyzes the recognized symbols and reconstructs the original morphemes regardless of the Korean phonological phenomena such as assimilation, dissimilation, contraction, and insertion.

2. A SPOKEN KOREAN ANALYSIS MODEL



[Figure 1.Spoken Korean morphological analysis model]

There are three major components in the system, as illustrated in Figure 1. The first component transforms the speech signal into a symbol description, i.e., a diphone string. Using Energy, ZCR and time information, a segmenter finds the location of the endpoints of Eonjeol (a Korean phrase which has no pause during the pronunciation). A group of TDNNs, organized into two-level hierarchies, recognize the patterns. The TDNN for vowel group identification operates using low and middle frequency speech signals and the sub-TDNNs for diphone recognition use all frequency speech signals. The second component converts a recognized diphone string into a

phoneme lattice. The final component produces morphologically analyzed Eojeols in an Eonjeol. The morphological analysis is performed by consulting a morpheme-level phonetic dictionary, checking the connectivity between morphemes, and reconstructing the original morphemes regardless of the phonological changes and the irregular conjugations.

3. DIPHONE RECOGNITION

group name	phoneme kind	phoneme number	diphone number
V	vowel	21	21
CV	consonant	18	378
	vowel	21	
VC	vowel	21	147
	consonant +	7	
CC	consonant +	7	126
	consonant	18	

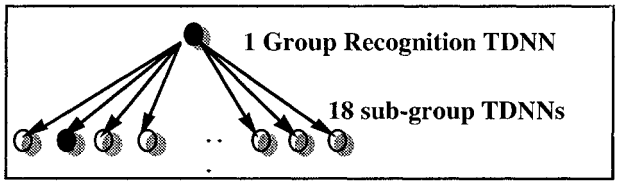
[Table 1. Korean diphone groups (+:second consonant)]

Korean is a syllable-based phonetic language, i.e., Korean syllable is the basic unit of pronunciation and has its unique sound. But syllable-based speech recognition has many problems like phoneme-level recognition. Recognizing more than 2500 Korean syllables frequently used is not easy. The different time durations of syllables make signal segmentation difficult. Fortunately, a Korean syllable is a combination of first consonant, vowel, and final consonant. In a few cases, two consonants are tightly coupled when pronounced. Table 1 is a Korean diphone group for the model. Diphone-based recognition has several advantages: 1) Existing diphone number is smaller than the existing syllables (about 350 diphones are extracted from 2350 syllables), 2) Its post-processing is easier than that of phoneme, and 3) Segmentation range decision of diphones is easier than that of phonemes.

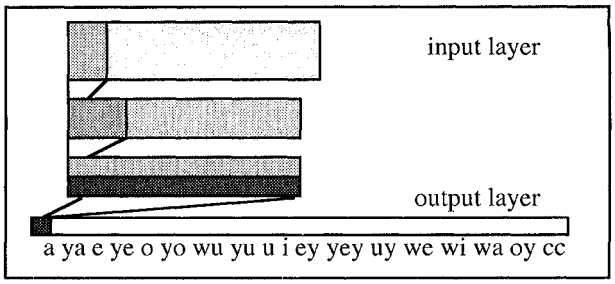
3.1 Korean diphone segmentation

A diphone segmenter checks whether each feature vector is a diphone while it scans the segmented Eonjeol. In scanning, the segmenter makes use of the information : 1) the time when the Energy is increased or decreased across the threshold, and 2) the time when the magnitude of signal wave is increased or decreased across the threshold.

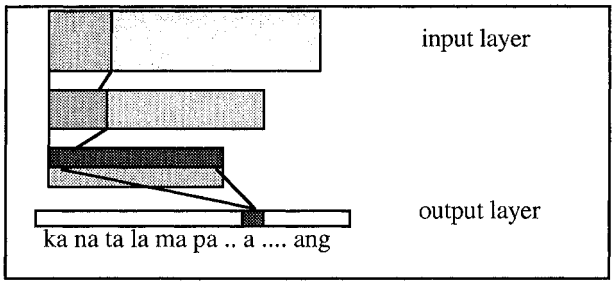
3.2 Korean diphone recognition by TDNNs



[Figure 2. Diphone recognizing hierarchical TDNNs]



[Figure 2-1. Group recognition TDNN]



[Figure 2-2. /a/ sub-group TDNN]

Figure 2 shows the diphone recognizing hierarchical TDNN architecture. A group recognition TDNN has 17 vowel group nodes and 1 cc group node as outputs (Figure 2-1). The cc group has diphones without a vowel component. Each sub-TDNN has different number of target nodes. Sub-TDNNs individually recognize the diphones in each group. The training time and the recognition time can be reduced by parallel training of each sub-TDNN. Figure 2-2 is a /a/ sub-group recognition TDNN structure. The process of diphone segmentation and spotting is repeated on the speech signals.

4. DIPHONE TO PHONEME FILTERING

The diphone2phoneme filter transforms the recognized diphone string into a phoneme lattice using the following 4 translation steps (Figure 3) :

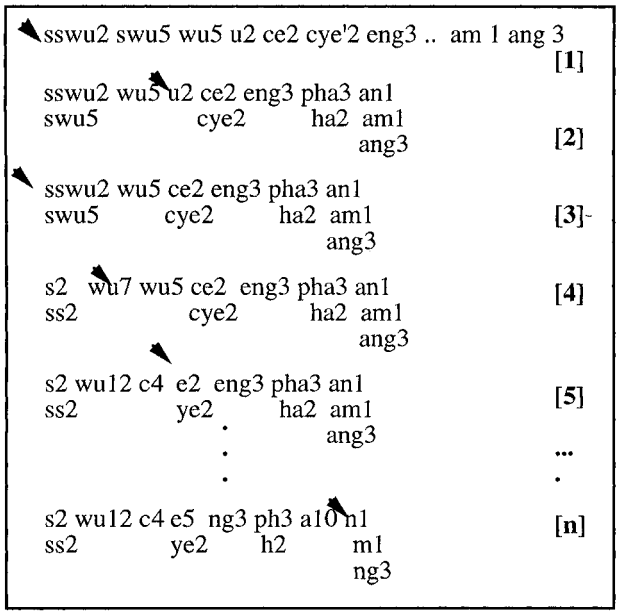
(1) Diphone Grouping : The recognized diphone string is converted to a grouped diphone lattice using the segmentation information and the properties of diphones. Each diphone group consists of the same type diphones with their frequency count ([1] to [2] in Fig.3).

(2) Time Alignment : The group whose count is less than the defined threshold is discarded. This process prevents the insertion errors happened in the segmentation process. (e.g. /u 2/ is deleted, [2] to [3] in Fig.3).

(3) Diphone Splitting : Each diphone is split into one or two phonemes with counts. ([3] to [4] in Fig.3).

(4) Phoneme Merging : When two continuous diphone groups are to be combined, two connected phoneme groups are merged into one group according to their similarities ([4] to [5] in Fig.3).

Steps (3) and (4) are repeated until the filter produces a complete phoneme lattice ([n] in Fig.3).



[Figure 3. Diphone2phoneme filtering " sucenghan" (meaning : updated)]

5. KOREAN MORPHOLOGICAL ANALYSIS

The CYK-morphological parser analyzes the acquired phoneme lattice using a morpheme-level phonetic dictionary. The parser handles the phonological changes both within the morpheme and across the morpheme boundary by modeling the Korean phonology rules. We modified the well-known CYK algorithm [6] to process the phoneme-lattice input in two ways : checking of the morpheme connectivity between Eojeols, and handling of the phonological changes.

5.1 Morpheme-level phonetic dictionary and connectivity checking

phonetic transcription header		morpheme	
sswu		swu	
left C.I.	right C.I.	left P.I	right P.I
bound noun	bound noun	Ps2ss	P-wu

[Figure 4-1. Morpheme-level phonetic dictionary, C.I : Connectivity Information, P.I : Phoneme-level connectivity Information]

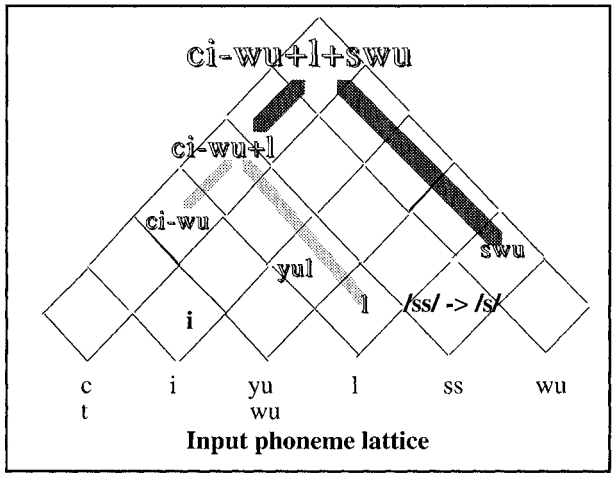
trie indexing	phoneme lattice
... s ... o wu ... ... n l ... c ...	ss wu t eng h a n s c ph m ng

[Figure 4-2. Trie indexing of transcription header]

A morpheme-level phonetic dictionary has a morpheme-level entry with a phonetic transcription header (Figure 4.1). The separate morphotactic information table (called connectivity table) is also provided for the morpheme connectivity checking. To process a morphological combination, the morphotactic information is divided into a left and right connectivity information (e.g. morpheme class : bound noun, left connectivity class : adnominalizing suffix, right connectivity class : verb). The phoneme-level connectivity information is also provided with a left and right ones (e.g. Ps2ss means that the phoneme /ss/ is from phoneme /s/ and P-wu means no change with phoneme /wu/). We use the trie indexing for looking up the varying size phonetic transcription header in the dictionary. The trie indexing eliminates the unnecessary search of several phoneme strings (e.g. all strings beginning with "swut" are pruned in figure 4-2).

5.2 Handling of phonological changes

The phonological changes which happen within the morpheme can be easily handled by the phonetic transcription headers in the dictionary. The phonological changes which happen across-morpheme boundary should be handled with the phoneme-level connectivity information table during the CYK analysis. When a phonological rule is applied, the original phoneme is suggested for the dictionary access (e.g., recovering consonant dissimilation : /sswu/ -> /swu/ in Figure 5). The parser goes on analyzing the string with the suggested original phoneme. Figure 5 shows the CYK morphological analysis with the phonological changes handling.



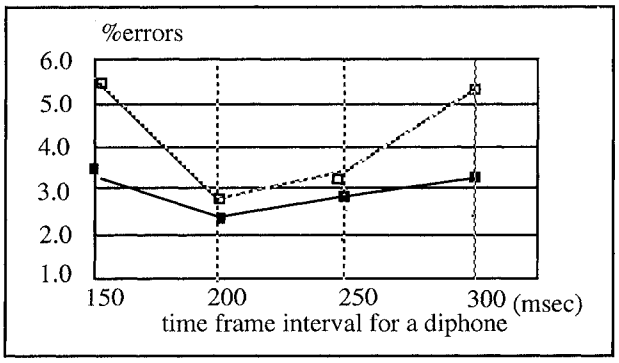
[Figure 5. The morphological anaysis of *"ci-wu+l+swu"* (removable)]

6. EXPERIMENT RESULTS

6.1 Time-shift Invariance of Diphones

We generated 2400 diphone samples for the 12 diphones (ko, kyo, kwu, kyu, to, tyo, twu, tyu, po, pyo, pwu, pyu). Eojeols were recorded in a normal laboratory environment with an average S/N ratio of 12dB. Speech data was sampled at 16kHz, and hamming windowed. From this windowed data, 512-point DTFTs were computed at 5 msec intervals. The DTFTs were used to generate the 16 Melscale coefficient spectra at 10 msec intervals [5]. These spectra were normalized to produce suitable input levels for the TDNNs. We used hyperbolic arc tangent error function in the weight updating [7]. The training was carried out using the fast back-propagation

algorithm. We updated the weights after a small number of iterations [8]. The patterns in the two tests are set the same in order to compare the "no shift" case and the "shift" case. Figure 6 shows that the diphone recognition has the time-invariance property of TDNN and suggests the optimal time interval value near 200 msec.



[Figure 6. Average error rates of segmented time frame (solid line) vs. time frame with maximum 40 msec temporal shift (gray line)]

6.2 Performance of Diphone Recognition

This experiment is to show that the diphone can improve the recognition rate of Korean vowels regardless of many diphthongs. We set 150 msec time range for the phoneme and 200msec for the diphone. Unlike in the phoneme, the recognition rate didn't decrease in the diphone when the number of targets with similar features increased (Table 2). Moreover, the unit with more than one features can be recognized at the high rate in the diphone recognition.

unit of recognition	number of targets	number of samples	recognition rate
phoneme	9	1080	94.06 %
diphone	9	1080	95.42 %
phoneme	17	2040	89.80 %
diphone	17	2040	95.27 %

[Table 2. Diphone recognition vs. phoneme recognition]

6.3 Performance of Eojeol Recognition

Eojeol class	number of Eojeol	recognition rate
trained Eojeol	66	92.4 %
new Eojeol	296	90.2 %
total mixed Eojeol	362	90.6 %

[Table 3. Eojeol recognition results]

In order to test the ability of full Eojeol recognition, a middle-vocabulary experiment was carried out. The task is speaker-dependent, continuous Eojeol recognition which produce the phoneme lattice from the speech signals, including the correct phoneme sequence in the Eojeol. Then the morphological analyzer can correctly produce the segmented Eojeols from the phoneme lattice. We extracted 72 diphones from 66 different Eojeols. The training was performed with the 72 diphones (each 66 Eojeols pronounced 15 times). We then collect 296 new Eojeols which include the trained 72 diphones. We test on the total 362 Eojeols. Table 3 shows the recognition results for the trained Eojeols, new Eojeols, and mixed Eojeols.
The high recognition rate of a non-trained new Eojeols show that the diphone-based speech recognition easily extends the size of the vocabulary.

7. CONCLUSIONS

We present a new spoken Korean recognition model in this paper. We implemented a prototype of the model and conducted experiments for the performance test. Our model has the following contributions to the Korean spoken language processing : (1) The new Eojeol unit, diphone, makes it easy to determine the segmentation range and the unit decision. (2) Korean phonological changes can be easily handled during the morphological analysis in a declarative way. (3) The continuous spoken Eojeol segmentation can be possible using the morphotactic information table. (4) The two-level TDNNs reduce the training and recognition time and make expansion of target Eojeols easy. (5) The diphone-level recognition can recognize many non-trained new Eojeols. In sum, we developed a tightly coupled morphological processing architecture in a connectionist large-vocabulary continuous speech recognition system. And we hope our spoken language morphological analysis model will play a major role to connect speech recognition field to natural language processing field.

REFERENCES

[1] S.G.Kim, *Phonetics*. JeungEum Pub., 1983 (in Korean).
[2] Y.U.Kwon and H.Y.Chung, "Analysis of Korean Phonemes Using Multi-Dimensional Scaling Method", . *Journal of Korea Electronics Engineering*, Vol.29, pp 868-876, 1992 (in Korean).
[3] D.G. Kim, C.G.Jung and H.Jeong, "Time-Delay Neural Networks for Korean Phoneme Recognition", *Journal of Korean Information Science Society*, Vol.18, No.3, pp.360-373, 1991 (in Korean).
[4] Hidefumi SAWAI, "TDNN-LR Continuous Speech Recognition System Using Adaptive Incremental TDNN Training", IEEE, Proceedings of ICASSP-91, 1991
[5] A.Waibel and T.Hanazawa, "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Transactions on Acoustics Speech and signal Processing, Vol.37, 1989.*
[6] AU.Aho and J.D.Ullman, *The theory of parsing, translation and compiling, vol. 1: parsing,* Prentice-Hall, 1972.
[7] Scott E.Fahlman, "Faster-Learning Variations on Back-Propagation : An Empirical Study", *Proc. of the 1988 Connectionist Models summer School,* 1988.
[8] P.Haffner, A.Waibel and H.Sawai and K.Shikano, "Fast Back-Propagation Learning Methods for Large Phonemic Neural Networks", *European Conference on Speech Communication and Technology,* 1989.