Integrating connectionist, statistical and symbolic approaches for continuous spoken Korean processing*

Geunbae Lee, Jong-Hyeok Lee, Kyubong Park, Byung-Chang Kim

Department of Computer Science & Engineering
Pohang University of Science & Technology
San 31, Hoja-Dong, Pohang, 790-784, Korea
gblee@vision.postech.ac.kr

ABSTRACT

This paper presents a multi-strategic and hybrid approach for large-scale integrated speech and natural language processing, employing connectionist, statistical and symbolic techniques. The developed spoken Korean processing engine (SKOPE) integrates connectionist TDNN-based phoneme recognition technique with statistical Viterbi-based lexical decoding and symbolic morphological/phonological analysis techniques. The modular large-scale TDNNs are organized to recognize all 41 Korean phonemes using 10 component networks combined through 3 glue networks. In performance phase, continuously shifted TDNN outputs are integrated with HMM-based Viterbi decoding using a tree-structured lexicon. The Viterbi beam search is integrated with Korean morphotactics and phonological modeling, and produces a morpheme-graph for high-level parsing module. Currently, SKOPE shows average 76.2% phoneme spotting performance for all 41 Korean phonemes (including silence) from continuous speech signals and exhibits average 92.6% morpheme spotting performance from erroneous TDNN outputs after morphological analysis. Other extensive experiments verify that the multi-strategic approaches are promising for complex integrated speech and natural language processing, and the approaches can be extended to other morphologically-complex agglutinative languages such as Japanese.

1. Introduction

This paper presents a statistical and symbolic hybrid technique for speech and natural language integration on top of connectionist phoneme recognition. Recent researches on connectionist speech recognition [2] and statistical language processing [1] can complement standard HMM (hidden markov model) approaches and conventional symbolic approaches in speech and natural language processing. Integrated speech and language processing techniques need not depend on the characteristics of specific languages, but it is inevitable that the techniques also get some influences from target languages. Our speech and language integration efforts focus on a class of languages called agglutinative languages, especially on Korean. In these languages, a word consists of clearly separable morphemes, usually single free-morpheme and several functional-morphemes, and the functional morphemes play important roles in a grammar. For these agglutinative languages, we need a morphologically-conditioned integration technique between speech and natural language. The language processing should begin with the morphological level, not with the syntactic level, and the speech recognition must generate morphemes, not words, as interface units. In this regard, our speech and natural language integration technique, called V-morph (Viterbi morphological analysis), fully considers the characteristics of agglutinative languages.

2. TDNN-based continuous phoneme recognition

For large-vocabulary continuous speech recognition, we selected all 41 Korean phonemes for direct recognition targets[1]. All 41 Korean phonemes are divided into consonant and vowel groups, and each group is divided again according to their acoustic-phonetic characteristics. For example, each consonant group is divided into 6 classes according to the co-articulation manners, and each vowel group into 4 classes according to the formant characteristics. For each phoneme class, we develop a separate TDNN (time-delayed neural network) [5] to recognize the 3 or 4 phonemes in the class, so altogether we need 10 TDNNs for the whole phoneme recognition in Korean. These 10 TDNNs are used together with 3 other glue networks, each for differentiating consonants from vowels, and for selecting the class in the consonant group and the class in the vowel group. The total 13 TDNNs form a large-phonemic TDNN [3, 6] by adding one more extra layer before the output layer. Figure 1 shows a structure of the large-phonemic TDNN for all 41 Korean phoneme recognition.

*This research was partly supported by KOSEF grant no. 941-0900-084-2 and is being supported by Ministry of Information and Telecommunications information super-highway application projects no. 95-122

[1] 40 phonemes plus one silence symbol

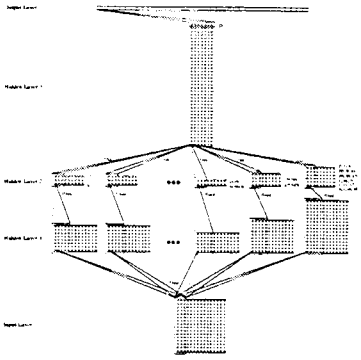

Figure 1: Large-phonemic TDNN architecture for Korean phoneme recognition

3. The V-morph speech and language integration

Figure 2 shows the V-morph (Viterbi morphological analysis) speech and language integration architecture. The core com-

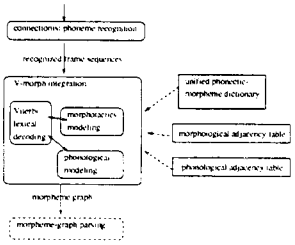

Figure 2: The V-morph speech and natural language integration: ovals designate processing modules and boxes represent common linguistic resources. The morpheme-graph parsing module is also implemented and integrated, but not described in this paper.

ponent of the V-morph integration is Viterbi-based lexical decoding which interacts with morphotactics and phonological modeling components. The Viterbi-based lexical decoding is performed on the recognized frame sequences from the large phonemic TDNNs. The large phonemic TDNN module completes acoustic-phonetic processing and transduces speech signals into error-contained phoneme sequences in an activation vector for each 10 msec frame. The result of V-morph integration is a morpheme graph that contains all the spotted morphemes in the input sentence.

3.1. Unified phonetic morpheme dictionary

The V-morph speech and language integration heavily depends on a common dictionary, called unified phonetic-morpheme dictionary (UPM dictionary). The UPM dictionary can be searched using a phonetic transcription of each morpheme, and includes morphological and phonological information for each header. Figure 3 shows an example entry *nun* in the UPM dictionary. Single phonetic header can be

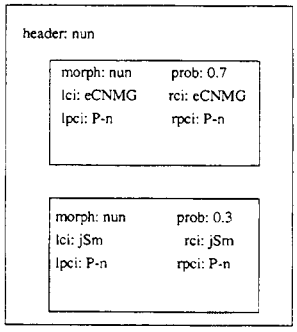

Figure 3: An example entry *nun* in the UPM dictionary. Yale romanization will be used for Korean alphabet in this paper.

associated with several different morphemes, and the dictionary fully reflects one-to-many mappings between the phonetic header and its corresponding morphemes. For example, in Korean, the phonetic sequence *nun* can be associated with two different morphemes with categories of adnominal verb-ending and auxiliary particle. In figure 3, the first box under the header *nun* represents the adnominal verb-ending *nun*, where morph designates lexical form in orthographic spelling, and prob represents a prior probability of the morpheme for the given phonetic sequence header. The lci and rci represent left and right morphological category, and the lpci and rpci is for left and right phonological category. The category values eCNMG and P-n will be explained shortly in the next sections. The reason we encode both left and right morphological category is that we sometimes need to directly encode idiomatic expressions[2] in which left and right morphological category may be different. For single morpheme encoding, the left and the right morphological categories are always the same. The phonological categories show a legal combination of sound changes when two morphemes are combined. The UPM dictionary is an essential component of V-morph integration and the morpheme-graph construction, so the dictionary must be organized so as to guarantee fast and efficient access. For the fast access, the phonetic header is organized into a tree structure, and each path in the tree forms a HMM (hidden markov model) of the corresponding morpheme.

3.2. morphotactics and phonological modeling of Korean

Morphotactics modeling requires morphological categorization of each morpheme. We maintain about 400 fine grained morphological categories for Korean. The morphological category is represented using a hierarchical symbol sequence.

[2]For example, *l-swu-iss* is an idiomatic expression meaning "be able to", and the lci must be a kind of adnominal verb-ending while rci must be a kind of predicate particle.

For example, the lci (rci) value "eCNMG" in figure 3 means specialized verb-ending category, organized with symbols: e (verb-ending) C (making complex sentences) N (making embedded sentences) M (noun phrase embedding) G (adnominalization). On the top-level, there are 15 most coarse morphological categories: noun, pronoun, number, adnominal, adverbial, exclamation, verb, adjective, particle (noun-ending), verb-ending, auxiliary predicate, predicate particle, prefix, suffix, and special-symbol. These coarse categories are refined step by step into several levels to make up final 400 morphological categories encoded in the dictionary. Among them, the noun, verb, and verb-endings are most finely sub-categorized since these categories represent very rich classes in Korean. Morphotactics modeling enforces ordering constraints on the morphemes. For the morpheme ordering constraint modeling, we check connectabilities of two adjacent morphemes in an *eojeol* (Korean word)[3]. In order to check the connectability, we use a morphological adjacency table, in which, for each morphological category, the connectable (morphotactically correct) morphological categories are all listed. Using morphological category in the dictionary and morphological adjacency table, we can model the ordering constraints between morphemes. Phonological modeling is also essential for speech and language integration since phonetic spellings are sometimes different from orthographic spellings in Korean. Phoneme recognition module only emits phonetic transcriptions of input speech, but high-level parsers require orthographic transcriptions as their inputs. Moreover, phonological changes in Korean sometimes occur with irregular conjugations (inflections), and mostly occur between morpheme boundaries when two morphemes are combined to form an *eojeol*. We adopt declarative strategy in which all the possible phonological changes are categorized and encoded in the dictionary. We consulted Korean standard pronunciation rules[4], and collected over 50 phonological changes between morphemes. These phonological changes are categorized and encoded using the special symbols. In figure 3, the lpci and rpci value "P-n" is such a symbol and designates no phonological change in the n sound for this morpheme. Other symbols are like "Pk2kk", meaning that k sound is changed to kk sound (glottalization phenomenon). The standard Korean phonological phenomena include consonant normalization, glottalization, consonant assimilation, consonant contraction, palatalization, and so on. Together with the phonological change categories, a phonological adjacency table is used to designate the contexts of phonological changes, which is much similar to the morphological adjacency table in its role.

[3]Linguistically, *eojeol* is a spacing unit of Korean orthography, and it corresponds to a word or a phrase in English. It is much similar to Japanese Bunsetsu.

[4]Korean orthography and pronunciation rules distributed by Ministry of Education, Korea

3.3. morpheme-graph construction in V-morph

Morpheme-graph construction utilizes Viterbi-based lexical decoding integrated with morphotactics and phonological modeling. A morpheme graph is a graph with morphemes as nodes and morphotactic and phonological constraints as links. Based on continuous connectionist phoneme recognition results which are sequence of activation vectors with each 10 msec speech frame having one activation vector, Viterbi beam search decodes the vector sequences into connectable morpheme sequences in a morpheme graph. During the beam search, the morphological and phonological connectability play a low-level morpheme-pair language model that reduces the search space. The Viterbi beam search is performed on phonetic headers in the UPM dictionary using the activation vectors from the large phonemic TDNNs as inputs, where the phonetic headers are tree-structured HMMs (hidden markov models) with phone duration modeling. Figure 4 shows a tree-structured HMM model for some of the entries in the dictionary. Each node represents a single phone HMM model with the transition probabilities for minimum 3 and maximum 8 phone duration modeling [4]. For the phone emission probability for each state, we use the

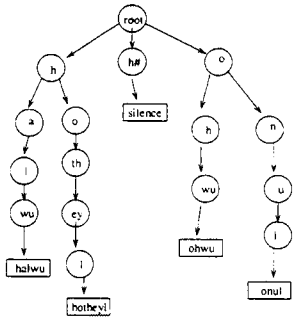

Figure 4: Tree-structured HMM dictionary for morphemes *ho-theyl* (hotel), *ha-lwu* (single day), *o-nul* (today), *o-hwu* (afternoon). The h# designates a silence (pause) in the sentence.

following formula:

$$p(speech \mid state) = \sum_{phone} p(speech \mid phone)p(phone \mid state) \quad (1)$$

where $p(speech \mid state)$ is a probability of speech signal given HMM state and plays a HMM emission probability for the Viterbi search. In equation 1, probability $p(speech \mid phone)$ comes from the activation vectors divided by the prior probability $p(phone)$ according to the Bayesian formalism. Probability $p(phone \mid state)$ is acquired by confusion matrix which averages the activation vectors out within a phone boundary. In this way, we can adjust the Viterbi search to cope with the connectionist continuous phoneme recognizer characteristics of degrading performance with a frame in a phone boundary. Using the transition and emission probabilities defined,

morph	correct	insert	delete	sub
2380	2204 (92.6%)	415	26	139

Table 1: The V-morph performance for 326 sentences (2380 morphemes).

the Viterbi beam search is performed frame-synchronously on the tree-structured HMM dictionary.

4. Experiment results

We collected 326 sentences with 2380 morphemes and obtained 76.2% of average phoneme spotting performance for all 41 Korean phonemes using the large phonemic TDNNs. Since the system deals with the continuous speech, there are no phone boundaries specified in the output activation-vector sequences. Therefore, single whole activation-vector sequence represents one sentence. The V-morph system segments these vector sequences into phonemes and morphemes, and finally constructs a morpheme graph for the sentence. We constructed 1000 morpheme UPM dictionary, and using the phonetic header, also made a tree-structured HMM for all 1000 morphemes. Table 1 shows the performance of the V-morph integration architecture. In the table, "correct" means the number of correctly spotted morphemes in the morpheme graph, while insert, delete, and sub designate the number of inserted, deleted, and substituted morphemes when compared to the correct morpheme graphs. The 92.6% continuous morpheme spotting performance is remarkable since V-morph does not employ any high-level language model except morphological and phonological constraints. So we can achieve general morphological and phonological processing for spoken agglutinative language processing without sacrificing the morpheme spotting performance. When the high-level linguistic models are embedded with the search, the performance will be much increased.

5. Concluding remarks

This paper presents a connectionist, statistical and symbolic hybrid technique for speech and natural language integration, called V-morph, for morphologically complex agglutinative languages. The V-morph technique utilizes statistical and symbolic approaches on top of a connectionist continuous phoneme recognition engine. We suggested a UPM dictionary for efficient speech and language integration, and implemented a Viterbi-based lexical decoding scheme using the dictionary. The V-morph presents a morpheme graph as an interface unit for general morphological and phonological processing. We also present declarative morphotactics and phonological modeling for Korean to be integrated into Viterbi-based search to construct the morpheme graph. Using large phonemic TDNN-based experiments, V-morph shows the possibility of morpheme-level speech and natural language integration for agglutinative languages with good morpheme spotting performance. However, more efficient beam search needs to be implemented to meet the real-time spoken language processing requirements. We are continuously improving the V-morph search technique and also working on the application of the V-morph to Korean/Japanese speech-to-speech translation system.

6. REFERENCES

1. E. Charniak. *Statistical language learning.* MIT press, 1994.
2. R. Lippmann. Review of research on neural networks for speech recognition. *Neural Computation*, 1, 1989.
3. M. Miyatake, H. Sawai, Y. Minami, and K. Shikano. Integrated training for spotting Japanese phonemes using large phonemic time-delay neural networks. In *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1990.
4. J. Tebelskis. Speech recognition using neural networks. Technical Report CMU-CS-95-142, Ph.D Thesis, School of Computer Science, CMU, 1995.
5. A. Waibel, T. Hanaazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328 – 339, 1989.
6. A. Waibel, H. Sawai, and K. Shikano. Consonant recognition by modular construction of large phonemic time delay neural networks. In *Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1989.