A Two-stage Recognition Approach Based on Error Pattern Hypotheses for Connected Digit Recognition

*Oh Wook Kwon and *Chong Kwan Un

Abstract

In this paper, a two-stage recognition approach based on error pattern hypotheses is proposed to reduce errors of a connected digit recognizer. In the approach, a conventional recognizer is first used to produce N-best candidate strings, and then error patterns are hypothesized by examining the candidate strings. For substitution error pattern hypotheses, error-pattern-dependent classifiers having more discriminative power than the first-stage classifier are used; and for insertion and deletion errors, word duration and energy contour information are exploited to discriminate confusing pairs. Simulation results showed that the proposed approach achieves 15% decrease in word error rate for speaker-independent Korean connected digit recognition when a hidden Markov model-based recognizer is used for the first-stage classifier.

Ⅰ. Introduction

Recently many speech recognition systems have used hybrid approaches such as hidden Markov models (HMM's) combined with neural networks and two-stage recognition strategies to improve recognition accuracies. In particular, the two-stage approach has often been adopted to improve the performance of isolated word recognizers based on dynamic time warping [1] and HMM's [2, 3]. In the two-stage approach, the first-stage classifier performs coarse classification of input patterns and generates several candidates, and then the second-stage classifier performs detailed classification by using more sophisticated recognizers or using new features discriminating the candidates better. Some systems combine likelihoods of the two classifiers using different weights on each state to enhance recognition accuracies further.

Examining our preliminary experimental results on frequent error patterns in a Korean connected digit recognizer based on HMM's, we have found that a small number of error patterns contribute to most recognition errors. This phenomenon is common in connected digit recognizers, regardless of languages. To reduce errors in such a situation, we extend the two-stage approach to connected digit recognition and propose to use a new classifier for each error pattern to discriminate frequently confusing word pairs. The new classifier is designed to have more discrimination power by estimating its parameters using the speech segments that have generated the corresponding error pat tern. Different features are used for different error pattern types(substitution, insertion, or deletion error patterns).

Ⅱ. A Two-stage Recognition Approach Based on Error Pattern Hypotheses

A block diagram of the two-stage recognition system based on error pattern hypotheses using the N-best paradigm [4] is shown in Fig. 1. The first-stage recognizer produces N-best candidate strings, and an error pattern hypoth-esizer generates possible error patterns by examining the N-best candidate strings. In an error-pattern-dependent second-stage classifier, the likelihood of each candidate pattern is calculated and combined with the likelihood in the first-stage classifier. Finally, a candidate string having the largest likelihood is chosen as a final recognition result.

2.1 Error pattern hypothesizer

In the training mode, we find error patterns between the correct and the recognized strings and estimate parameters of the second-stage classifier for each error pattern. In the recognition mode, we calculated the error patterns between the best candidate and the N-th ($N \geq 2$) best strings.

While word-level string matching using the dynamic programming technique [5] generally produces correct

*Department of Electrical Engineering at KAIST

Manuscript Received June 24, 1996.

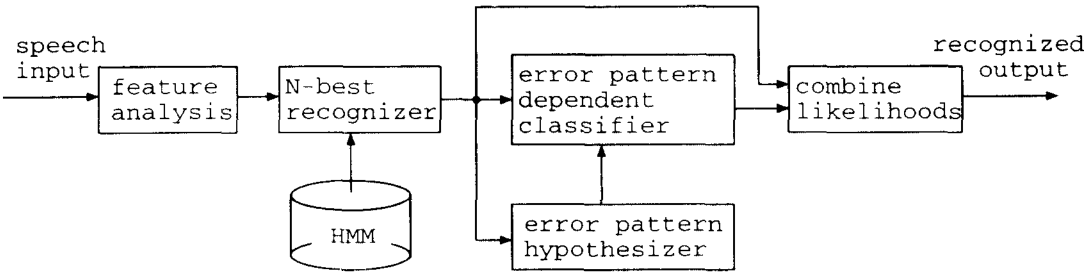

Figure 1. A two-stage recognition system.

numbers of substitution, insertion, and deletion errors in calculating recognition accuracies, it can not identify error pattern types when insertion or deletion errors have occurred. For example, when "a b c" is recognized as "a c", there can be three possible error patterns according to segmentation information. That is, the speech segment "a b" can be misrecognized as the segment "a", "b c" misrecognized as "c", or "a b c" misrecognized as "a c". The segmentation information for the candidate strings obtained in the first-stage recognizer is exploited ot identify the error pattern among possible patterns.

To find the error pattern hypotheses, we assume that a string composed of arrays of word identifications (ID), segment-start and segment-and points. The procedure is as follows. Conventional word-level string matching between two strings is performed. Then, null segments are inserted in the appropriate positions to have the same number of segments in the two strings. For each segment of a string, the word ID of the segment is compared with the string-matched word ID of the segment of the other string. If they are different, segments whose starting points are similar within the tolerance range, T, in the backward direction and segments whose and points are similar in the forward direction are searched. Next, word ID's of the segments between the start and end points in the correct string and the recognized string are concatenated. Then, the concatenated word ID's constitute an error pattern. Deletion error patterns related with monophonemic digits are always hypothesized in the recognition mode.

The overall error pattern hypothesizing algorithm is as follows.

```
**typedef struct** {
  $w()$: array of word ID's of segments
  *start*(): array of start points of segments
  *end*(): array of end points of segments
} *StringType*
```

```
**procedure Err Pat Hypothesizer** (*StringType S, StringType T,*
  *Error Pattern List EP*)
perform word-level string matching between $S$ and $T$

**for** $i = 0$ **to** number of segments in $S$
  **if** $S.\,w(i)$ is a monophonemic word
    **and** hypothesizer is in recognition mode
    $P = S.\,w(i)$
    add $PP \rightarrow P$ to $EP$
  **else if** $S.\,w(i) \neq T.\,w(i)$
    $b = i$
    **while** $|S.\,start(b) - T.\,start(b)| > T$
      **or** $S.\,w(b) = null$ **or** $T.\,w(b) = null$
      $b = b - 1$
    **end**
    $e = i$
    **while** $|S.\,end(e) - T.\,end(e)| > T$
      **or** $S.\,w(e) = null$ **or** $T.\,w(e) = null$
      $e = e + 1$
    **end**
    $P$ = Concatenation of $S.\,w(b) .. S.\,w(e)$ excluding $null$
    $Q$ = Concatenation of $T.\,w(b) .. T.\,w(e)$ excluding $null$
    add $P \rightarrow Q$ to $EP$
  **endif**
**end**
**return** $EP$
**end** *Err Pat Hypothesizer*
```

2.2 Error-pattern-dependent classifier

Error-pattern-dependent discriminative classifiers are used in the second stage. They may be discriminatively trained HMM-based recognizers or neuarl networks such as multilayer perceptrons, which are known to have high discrim ination power. In the training mode, parameters of the second-stage classifier for an error pattern is estimated based on speech segments that have generated the error pattern. In the recognition mode, the classifier calculates likelihoods of N-best candidate string using the classifier parameters for the error pattern hypotheses obtained between the best candidiate and the N-th ($N \geq 2$) best strings.

The feastures used in the second-stage classifiers may be the same as or different from those used in the first-stage classifier according to error pattern types. In particular, for insertion and deletion errors, duration and energy information in addition to spectral features are useful as the features. In this paper, for substitution error patterns, the same features and classifier parameters estimated by the generalized probabilistic descent (GPD) method are used; and for insertion and deletion error patterns, word duration and energy contour are used as the features. These features have been shown to reduce segmentation errors in both training and recognition phases [6].

2.3 Word duration and energy contour modeling

A word duration is modeled by a string-boundary-dependent Gaussian distribution. String-boundary dependency means that word durations at the start and end of a string are modeled separately. To discriminate between two hypothesized word clusters that have generated insertion or deletion errors, a word-cluster duration is also modeled by a string-boundary-dependent Gaussian distribution.

Energy contour information is represented by valley depth of a segment computed as areas between energy contour and its convex hull (the region A in Fig. 2) and is modeled by a Gaussian distribution. By assuming that points on the frame energy contour, segment boundaries and the frame axis corresponding to a segment constitute a set of points, the convex hull of the segment can be obtained by a well-known method such as the package-wrapping method [7]. The deeper valley has the energy contour, the more likely the corresponding speech segment consists of two words. This feature is useful in reducing deletion errors in monophonemic digits, decause energy-normalized spectral characteristics of two consecutive utterances of a monophonemic digit are nearly the same as those of one utterance while frame energies change as time elapses (e.g., the cases where "o o" is misrecognized as "o", or "i i" misrecognized as "i" in Korean connected digit recognizers).

III. Task and Database

To investigate the performance of the proposed approach in speech recognition, the speech material used in this experiment was a Korean connected digit database produced by 140 speakers (90 males and 50 females) in a quiet room. Each speaker pronounced 40 digit strings which varied from three to seven digits and then erroneous utterances were discarded by listening. Words from 93 speakers (60 males and 33 females) were used as training data, and those from the other 47 as test data. The speech signal was sampled at 16 kHz and segmented into 30 msec frames with each frame advancing every 10 msec. Each frame was parameterized by a 26-dimensional feature vector consisting of (1) 12 linear predictive coding (LPC) derived liftered cepstral coefficients and energy and (2) their corresponding time derivatives.

All Korean digits are monosyllabic. These are transcribed in Table 1. Table 2 shows our preliminary experimental results on frequent error patterns in a

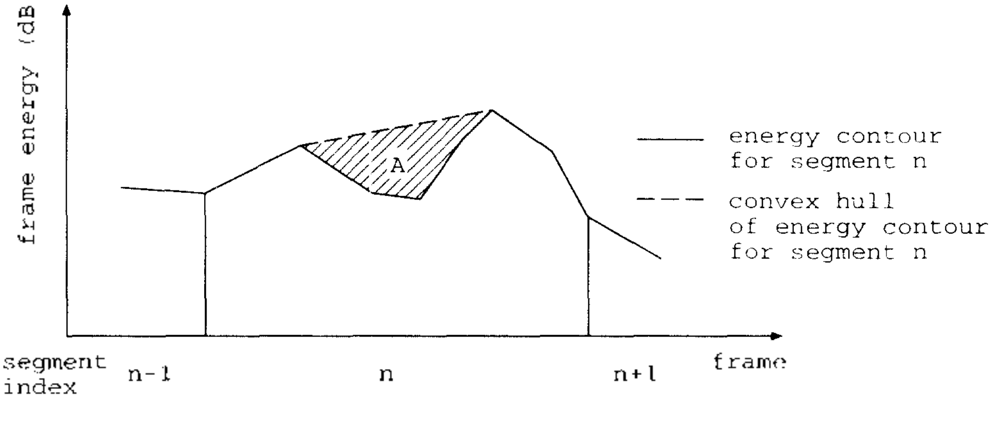
frame energy (dB)
A
energy contour for segment n
convex hull of energy contour for segment n
segment index
n-1
n
n+1
frame

Figure 2. An energy contour-based feature.

Korean connected digit recognizer. In Table 2, 10 of 170 error patterns contribute about 40% of total errors in case of an HMM recognizer trained by the maximum likelihood criterion. This fact justifies using error-pattern-dependent classifiers in connected digit recognition.

Table 1. 11 Korean digits used in experiments

digit	1	2	3	4	5	6	7	8	9	0	oh
trans.	il	i	sam	sa	o	yuk	chil	pal	ku	yeong	kong

Table 2. Frequent recognition error patterns

ML training	GPD training
il ←→ i (67)	il ←→ i (37)
o ←→ ku (38)	o ←→ ku (32)
ii ←→ i (30)	ii ←→ i(32)
sam ←→ sa (28)	oo ←→ o (20)
ku o ←→ ku (21)	ku ←→ kong (16)
ku ←→ kong (19)	ku o ←→ ku (11)
il ←→ i il (19)	chil il ←→ il (11)
oo ←→ o (19)	yuk o ←→ yuk ku (10)
chil ←→ chil il (13)	yuk ←→ yeong (9)
yeong ←→ i yeong (11)	il ←→ i il (9)

Note: "a" ←→ "b" (*c*) denotes a substitution error pattern and "a" → "b" (*c*) denotes an insertion or deletion errgr pattern occurring *c* times.

Ⅳ. Simulation Results

In the first-stage classifier, each digit was modeley by a 9-state left-to-right HMM without skip transitions. Observation densities were modeled by mixture Gaussian pdf's with the number of mixtures for each state varying from 1 to 4. Two kinds of the first-stage classifier were trained and tested using the maximum likelihood estimation (MLE) and GPD methods. We generated 3 candidate strings with the unknown length decoding constraint using a frame-synchronous search algorithm [8]. For substitution errors, we used second-stage classifiers with HMM structures and features the same as the first-stage classifier and estimated their parameters by the GPD method. The difference from the first-stage classifier is that only the speech segments having generated an error pattern are used to estimate parameters of the second-stage classifier for the error pattern. For insertion and deletion errors, word durations and energy contour for a segment were modeled by Gaussian distributions. The energy component below 1 kHz was used for energy contour modeling. The tolerance range in the error pattern hypothesizer was set to 3 frames. The likelihoods of the two classifiers were combined with equal weighting.

Thress kinds of experiments were performed to justify the proposed two-stage approach. In Exp²riment A, the second-stage classifiers were used to reduce only substitution errors; in Experiment B to reduce only insertion and deletion errors; and in Experiment C to remove all three types of errors. In all experiments, only error patterns within 20 most frequent ones were hypothesized.

Tables 3 and 4 show recognition results when the HMM-based first-stage classifier was trained by the MLE and GPD methods, respectively. The recognition accuracies of the baseline systems were obtained without using the second-stage classifieres. The values in the parentheses indicate string accuracy. Simulation results show that the proposed approach consistently improves recognition accuracies with different training methods and HMM parameters. The proposed approach achieves 15% decrenase in word error rate when an MLE-trained recognizer with 4 mixtures is used for the first-stage classifier. Decrease in error rates of a system with the MLE-trained HMM recognizer used as the first-stage classifier is larger than that of the GPD-trained system and it is mostly due to the decrease in substitution errors. This is because most substitution errors were effectively removed by virtue of the discriminative nature of the GPD-trained HMM recognizer.

Table 3. Word and string accuracies (%) when an MLE-trained HMM recognizer is used as the first-stage classifier.

Number of mixtures	Baseline system	Experiment A	Experiment B	Experiment C
1	92.12(67.09)	92.70(69.23)	92.53(68.83)	93.11(70.91)
2	92.25(67.61)	93.10(70.73)	92.63(68.19)	93.64(73.11)
4	93.17(70.85)	93.80(73.28)	93.47(72.24)	94.17(74.96)

Note: The number in parenthesis indicates string accuracy.

Ⅴ. Conclusions

In this paper, a two-stage recognition approach based on error pattern hypotheses was proposed to reduce errors of a connected digit recognizer. In the approach, a conventional HMM-based recognizer was first used to produce N-best candidate strings, and then error patterns were hypothesized by examining the candidate strings. For substitution error pattern hypotheses, a new classifier having more discriminative power than the first-stage classifier was used for each error pattern. For insertion and deletion errors, word-cluster durations and valley depth in energy contour of a segment have been modeled by Gaussian distributions to discriminate confusing pairs. Simulation results showed that the proposed approach achieves 15% decrease in word error rate for speaker-independent Korean connected digit recognition when an HMM-based recognizer is used as the first-stage classifier.

Table 4. Word and string accuracies (%) when a GPD-trained HMM recognizer is used as the first-stage classifier.

Number of mixtures	Baseline system	Experiment A	Experiment B	Experiment C
1	93.99(74.03)	94.20(74.78)	94.44(76.17)	94 66(76.92)
2	94.79(76.92)	94.91(77.39)	95.20(78.48)	95.32(79.12)
4	95.11(78.31)	95.20(78.77)	95.44(79.76)	95.55(80.28)

Note: The number in parenthesis indicates string accuracy.

References

1. L. R. Rabiner and J. G. Wilpon, "A two-pass pattern-recognition approach to isolated word recognition", *Bell Syst. Tech. J.*, Vol. 50, No. 5, pp. 739-766, May 1981.
2. E. A. Martin, "R. P. Lippmann, and D. B. Paul, "Two-stage discriminant analysis for improved isolated-word recognition", *Proc. ICASSP 87*, pp. 709-712, April 1987.
3. K. -Y. Su and C. -H. Lee, "Speech recognition using weighted HMM and subspace projection algorithm", *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 1, Part 1, pp. 69-79, Jan. 1994.
4. R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, and G. Zavaliagkos, "New uses for the N-best sentence hypotheses within the BYBLOS speech recognition system", *Proc. ICASSP 92*, pp. I.1-I.4, March 1992.
5. J. Picone, K. M. Goudie-Marshall, G. R. Doddington, and W. Fisher, "Automatic text alignment for speech system evaluation", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. 34, No. 4, pp. 780-784, Aug. 1986.
6. V. Gupta, M. Lenning, P. Mermelstein, P. Kenny, P. F. Seitz, and D. O'Shaughnessy, "Use of minimum duration and energy contour for phonemes to improve large vocabulary isolated-word recognition", *Computer Speech and Language* (1992) 6, pp. 345-359.
7. R. Sedgewick, *Algorithms in C*, Addison-Wesley, Reading, Massachusetts, pp. 359-372, 1990.
8. C. -H. Lee and L. R. Rabiner, "A frame synchronous network search algorithm for connected word recognition", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. 37, No. 11, pp. 1649-1658, Nov. 1989.

▲Oh Wook Kwon

Oh Wook Kwon received the B.S. degree in electronic engineering from Seoul National University, and the M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology(KAIST) in 1986 and 1988, respectively.

Since 1988 he has been with Electronics and Telecommunications Insitute, where in is doing research in the area of continuous speech recognition. Since 1992 he is a Ph.D. Candidate in electrical engineering at KAIST.

▲Chong Kwan Un

Chong Kwan Un was born in Seoul, Korea. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Delaware, Netwark, in 1964, 1996, and 1969, respectively.

From 1969 to 1973, he was an Assistant Professor of Electrical Engineering at the University of Maine, Portland, where he taught communications and did research on synchronization problems. In May 1973 he joined the staff of the Telecommunication Sciences Center, SRI International, Menlo Park, CA, where he did research on voice digitization and bandwidth compression systems. Since June 1977 he has been with Korea Advanced Institute of Science and Technology(KAIST), where he is a Professon of Electrical Engineering, teaching and doing research in the areas of digital communications and signal processing. So far, he has supervised 59 Ph.D. and more than 110 M.S. graduates. He has authored or coauthored over 250 papers on B-ISDN, protocol design and analysis, very high-speed packet communication systems, speech coding and processing, adaptive signal processing and data communications. Also, he hold seven patents granted. From February 1982 to June 1983 he served as Dean of Engineering at KAIST.

Dr. Un is a Fellow of IEEE and a Fellow of the Korean Academy of Sciences and Engineering. Also, he is a Founding Member of the National Academy of Engineering of Korea. He received a number of awards, including the 1976 Leonard G. Abraham Prize Paper Award from the IEEE Communications Society, the National Order of Merits from the Government of Korea, and Achievement Awards from the Korea Institute of Telematics and Electronics, the Korea Institute of Communication Sciences, and the Acoustical Society of Korea (ASK), Highest Achievement Award from KAIST. He was President of the ASK from 1988 to 1989. He is a member of Tau Beta Pi and Eta Kappa Nu.