IMPROVING SPEECH RECOGNIZER BY BROADER ACOUSTIC-PHONETIC GROUP CLASSIFICATION

Youngjoo Suh, Kyuwoong Hwang, Oh-Wook Kwon, and Jun Park

ETRI, 161 Kajong-Dong, Yusong-Gu, Taejon, Korea

E-mail: {yjsuh, hkw, owkwon, junpark}@etri.re.kr

ABSTRACT

We propose a new approach to improve the performance of speech recognizers by utilizing acoustic-phonetic knowledge sources. We use the unvoiced, voiced, and silence (UVS) group information of the input speech signal in the conventional speech recognizer. We extract the UVS information by using a recurrent neural network (RNN), generate a rule-based score, and then add the score representing the UVS information to the conventional spectral feature-driven score in the search module. Experimental results showed that the approach reduces 9% of errors in a 5,000-word Korean spontaneous speech recognition domain.

1. INTRODUCTION

It is very important to select good input features for high performance speech recognition [1]. Most of the current speech recognizers adopt hidden Markov model (HMM) as their method. Their input feature vectors are mostly based on spectral analysis-driven parameters such as perceptually linear prediction (PLP)-cepstrum, mel-cepstrum, or filter-bank output, [2], [3]. Although the speech recognizer based on spectral analysis feature shows considerable performance, it still needs to improve further to satisfy user requirements from real-world applications.

When we examined recognition results of the spectral feature-based speech recognizer, we found notable amounts of the UVS misclassification. This undesirable phenomenon is due to the fact that the spectral feature-based speech recognizer is primarily designed to discriminate very confusing classes such as phones or subphone-like units and not intended to classify more obvious classes like the UVS phone groups. Therefore we can improve the performance of the speech recognizer if we classify the UVS group of speech signals more accurately. From this observation, we introduce a new approach that utilizes the UVS group information of speech signals to improve the spectral feature-based speech recognizers.

This paper is organized as follows. In section 2, we first represent the algorithm of our method. We then represent the experimental procedure and evaluation results with discussions in section 3. Finally, we conclude our works in section 4.

2. THE PROPOSED APPROACH

Our approach is composed of two parts: extraction of the UVS information from speech signal using the RNN and integration of the UVS and the spectral information in the search procedure of speech recognition.

2.1. UVS Group Information Extraction

We use a recurrent neural network (RNN) to extract the UVS information from the input speech signal for each frame. In the RNN approach, we first select features for the RNN input nodes. We then train the RNN-based UVS information extractor using hand-labeled speech material. The performance of the RNN is compared with that of a multilayer perceptron (MLP).

Feature selection

The features for the input of the UVS information extraction are carefully chosen to efficiently represent the characteristics of each UVS group and extracted from every speech frame. In the features, frame energy and level crossing rate are widely used to classify UVS groups. We added three kinds of frequency band energy ratios to improve the UVS information extractor. These features are represented as follows.

1. Frame energy

$$ENERGY(k) = 10 * \log_{10}(\sum_{n=0}^{N-1} x(n)^2 + 1) \qquad (1)$$

where k represents the analysis frame and N is the length of the analysis frame and $x(n)$ is input speech signals without preemphasis.

2. Level crossing rate

$$LCR(k) = \sum_{n=0}^{N-1} \mathrm{sgn}(n) \qquad (2\text{-}1)$$

$$\begin{aligned} \mathrm{sgn}(n) &= 1 \quad \text{iff } [(x(n)-lcr\_level)*(x(n+1)-lcr\_level)] < 0 \\ &= 0 \quad \text{otherwise} \qquad (2\text{-}2) \end{aligned}$$

where lcr_level represents the level defined in the level crossing rate and sgn(n) becomes 1 if the speech signal crosses the predefined level. In our case, this value is set as three times the average value of the positive samples from the 100 ms silence region.

3. Differential level crossing rate

$$DLCR(k) = \sum_{n\ 0}^{N-1} \mathrm{sgn}(n) \qquad (3\text{-}1)$$

$$\mathrm{sgn}(n) = 1 \text{ iff } [(dx(n)-dlcr\_level)*(dx(n+1)-dlcr\_level)]<0$$

= 0 otherwise

$dx(n) = x(n) - x(n-1)$ (3-2)

where dlcr_level is obtained by the same method used in 2.

4. Voiced band energy ratio 1 (low)

VBER1(k) = log energy in the region between 180 Hz and 1000 Hz – log energy in the region between 4000 Hz and 8000 Hz

This feature is selected from the fact the first formant of voiced sounds usually lies between 180-1000 Hz.

5. Voiced band energy ratio 2 (high)

VBER2(k) = log energy in the region between 1000 Hz and 2300 Hz – log energy in the region between 4000 Hz and 8000 Hz

We select this feature to reflect the second formant energy of voiced sounds.

6. Unvoiced band energy ratio

UBER(k) = log energy in the region between 4000 Hz and 8000 Hz – log energy in the region between 180 Hz and 4000 Hz

This feature value is large for unvoiced sounds. The features in 4-6 are used to accurately classify between voiced and unvoiced frames. We apply the fast Fourier transform algorithm to calculate the spectral-domain features defined in 4-6.

The neural network-based UVS information extraction

The structures of the MLP and the RNN to extract UVS information are shown in Figure 1 and 2, respectively.

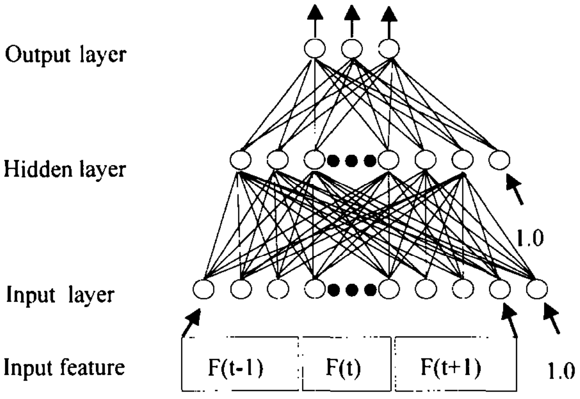

Figure 1: Multilayer perceptron-based method .

As shown in Figure 2, the structure of the RNN is very similar to that of the MLP except that the input layer has three additional nodes to feed the delayed outputs of the output layer recurrently. The input layer of the RNN is composed of 21 nodes, 18 nodes for features extracted from 3 consecutive frames and 3 nodes for one frame-delayed RNN outputs. The number of fiidden nodes in hidden layer_is chosen from the experimental results. Three output nodes in output layer represent the probability of each UVS group.

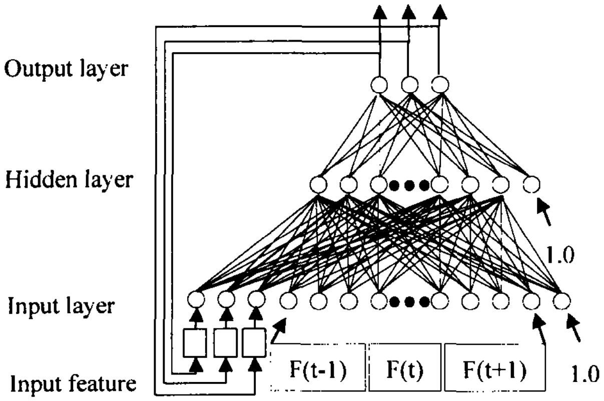

Figure 2: Recurrent neural network-based method.

2.2. Integration of the UVS and the Spectral Information

To integrate the UVS and the spectral information, we first establish a set of rules to derive proper UVS score from the UVS feature parameters, i.e., the RNN output values. The three output values from the RNN output layer are applied to a set of rules to derive the score representing the final UVS information of the analysis frame. We determine the rules as follows. All HMM models are mapped to their corresponding UVS classes in advance. Then the score for each HMM model becomes relatively high when the class indicated by the UVS information matches to a class that the HMM model is included. These rules consist of two main steps. First, we choose the proper threshold value for each UVS class to determine the analysis frame is included one of 3 UVS class obviously or ambiguously. Second, for each UVS group, we derive a score function that compares RNN output value with the threshold value. The score function is a weighted sum of the output signals of the RNN output and the three weight values are determined heuristically. When only the output node with the maximum value is used, recognition results become worse.

After calculation of the UVS score, the final score for each HMM model is obtained by combining the two scores, i.e., the UVS score and the conventional spectral score, with an appropriate weighting factor. The score obtained at this stage is used as a new score in the search of the speech recognition. The scores for all active HMM models at given search instant are computed by this method. This rule-based score combining approach saves the size of speech recognizer significantly compared with that of adding the UVS features to the conventional spectral feature parameters. The speech recognizer adopting this proposed algorithm is represented in Figure 3.

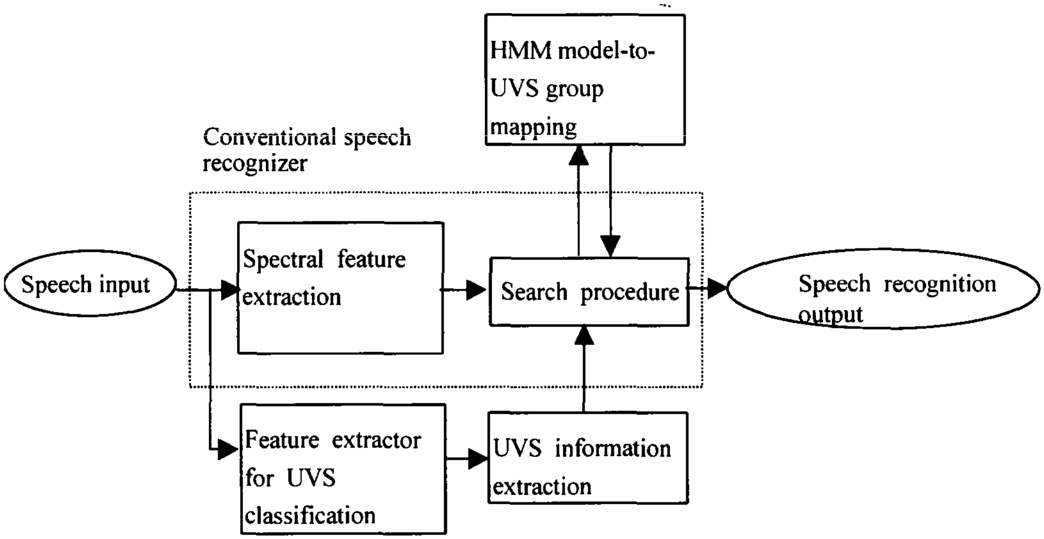
HMM model-to-UVS group mapping
Conventional speech recognizer
Speech input
Spectral feature extraction
Search procedure
Speech recognition output
Feature extractor for UVS classification
UVS information extraction

Figure 3: The structure of the UVS information-incorporated speech recognizer.

3. EXPERIMENTAL RESULTS

3.1. UVS Group Classification

To train and evaluate the RNN-based UVS group classification, we used 44 Korean spontaneous speech sentences uttered by 11 male and 6 female speakers. We labeled this speech database manually. With the speech data and the labeling data, we extracted input data and target data to train and evaluate RNN. We adopted the leave-one-out method as an evaluation method because of the limited amount of database. In this method, we used 40 sentences as learning data and the remaining 4 sentences as evaluation data. The error back propagation algorithm was used to the RNN learning algorithm.

Table 1 represents the results of the UVS classification by the two methods. In this table, the performance of the RNN is 92.9% and 91.7% for the closed and the open test, respectively and better than that of the MLP by 17.4% and 8.8% error reduction rate, respectively. The portion of unvoiced, voiced, silence region was 13%, 61%, and 26%, respectively. The corresponding classification accuracy was 61.6%, 99.0%, and 90.7%, respectively. This means that many classification errors are still remaining at unvoiced or transient regions. This performance seems to be relatively low compared to already reported methods. However, we think this result is due to the worse target domain, that is, the spontaneous speech database.

Based on this result of the RNN, we adopt the RNN-based UVS information extractor in improving the performance of the speech recognizer. This result is also consistent with the previous results in UVS classification [4], [5].

3.2. Integration of UVS Information into Speech Recognizers

Speech database and speech recognizer

In the experiments about the improvement of speech recognizer by the use of the UVS information, we use 283 Korean spontaneous speech sentences consisting of 3012 words as the evaluation speech database. The baseline speech recognizer is developed for 5,000-word Korean spontaneous speech domain and shows 79.6% word accuracy. This speech recognizer uses 32 dimensional linear discriminative analysis (LDA) outputs transformed from the 39 dimensional PLP, delta PLP, and delta-delta PLP coefficients as its spectral feature parameters.

Experimental results and discussion

We integrated the proposed method to the baseline speech recognizer to examine the improvement by the addition of the UVS information. First we appended the three output values of the RNN output nodes or 18 features to the spectral feature vectors, and achieved recognition accuracy of 76.5% and 78.5%, respectively. Another attempts to augment feature vectors did not give any performance improvements. This result is because the recognizer can not efficiently extract the UVS information from the mixed features. We also tried to adopt the UVS information score in lattice rescoring did not yield any improvements.

We then incorporated the UVS information into the search module directly. The recognizer yielded the best recognition accuracy of 80.9% when the weight for the UVS information score was 8. After correcting unvoiced errors in the silence region, the recognizer achieved 81.4% word recognition accuracy, which means 9% error reduction rate. Incorporation of the UVS information made little increase of recognition time by less than 10%.

From this result, we know that there is still notable amount of misclassification in the UVS region. We thus presume that more attention to improve the performance of the UVS classification may helpful in obtaining higher accuracy of the speech recognizer.

Tests (leave-one-out)	Accuracy [%]			
	MLP		RNN	
	Learning	Evaluation	Learning	Evaluation
1	91.3	92.4	92.3	93.4
2	91.3	92.0	92.5	93.4
3	91.3	91.5	92.6	91.9
4	91.8	91.8	93.3	92.5
5	91.5	84.2	93.7	85.8
6	91.3	94.5	92.9	94.6
7	91.3	94.1	93.2	95.0
Average	91.4	91.7	92.9	92.5
E. R. R.	-	-	17.4	9.6

Table 1: The accuracy of UVS classification with the MLP and RNN classifiers.

4. CONCLUSION

Most of the current speech recognizers use the spectral analysis-driven feature parameters as their input feature. Although this kind of feature has relatively good capability of discriminating between very confusing phonetic classes, it shows some erroneous results in the viewpoint of the UVS classification. We thus proposed a method of improving the performance by utilizing the UVS information in the conventional HMM-based speech recognizer. We compared two approaches to integrate the UVS information: appending the UVS features to the existing feature vectors and adopting them in the search module of the speech recognizer directly. The proposed method shows about 9 % error reduction rate in the large vocabulary Korean spontaneous speech recognition domain when the UVS information is integrated in the search module. Also, this performance improvement is achieved with a relatively small computational cost.

5. ACKNOWLEDGEMENTS

This research was funded by the Korea Ministry of Information and Communications.

6. REFERENCES

1. Y. Lee and K. Hwang, "Selecting good speech features for recognition," *ETRI Journal*, vol. 18, no. 1, pp. 29-40, 1996.
2. P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young, "The Development of the 1994 HTK Large Vocabulary Speech Recognition System," *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, pp. 104-109, January 1995.
3. R. Roth, L. Gillick, J. Orloff, and F. Scattone, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer," *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, pp. 116-120, January 1995.
4. R. P. Lippmann, "Review of Neural Networks for Speech Recognition," *Readings in Speech Recognition*, pp. 374-392, Morgan Kaufmann Publishers, San Mateo, 1990.
5. T. Fukada, S. Aveline, M. Schuster, and Y. Sagisaka, "Segment Boundary Estimation using Recurrent Neural Networks," *Proceedings of EuroSpeech'97*, vol. 5, pp. 2839-2842, Sep. 1997.