Using Confidence Vector in Multi-Stage Speech Recognition

Hyungbae Jeon, Kyuwoong Hwang, Hoon Chung, Seunghi Kim, Jun Park, Yunkeun Lee
Speech/Language Information Research Center
Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea
{hbjeon, kyuwoong, hchung, seunghi, junpark, yklee}@etri.re.kr

Abstract

This paper presents a new method of using confidence vector as an intermediate input feature for the multi-stage based speech recognition. The multi-stage based speech recognition is a method to reduce the computational complexity of the decoding procedure and thus accomplish faster speech recognition. In the multi-stage speech recognition, however, the error of previous stage is transferred to the next stage. This tends to cause the deterioration of the overall performance. We focus on improving the accuracy by introducing confidence vector instead of phoneme which typically used as an intermediate feature between the acoustic decoder and the lexical decoder, the first two stages in the multi-stage speech recognition. The experimental results show up to 16.4% error reduction rate(ERR) of word accuracy for 220k Korean Point-of-Interest (POI) domain and 29.6% ERR of word accuracy for hotel reservation dialog domain.

1 Introduction

Recently, demand of fast and small size speech recognition engine is increasing. One example is the mobile automatic speech translation device. To implement the speech translation function, we need to integrate the speech recognition, the text translation, and the speech synthesis modules all together into a single device. Another example is recognition of Point-of-Interests (POIs) on navigation device. The number of POIs used in a commercial navigation system is about several hundreds of thousands to one million. To make a fast and more efficient speech recognition engine, many computationally efficient methods were reported and multi-stage speech recognition is one of competitive approaches [1][2].

The multi-stage speech recognition completes recognition procedure through sequential two stage decoding, the acoustic decoding and the lexical decoding. The optional rescoring stage can follow the lexical decoding, depending on the application. In the acoustic decoding stage, we find the phoneme sequence which has the maximum likelihood for a sequence of input acoustic feature vector. In the lexical decoding, we recover the optimal word sequence whose phone sequence has a minimal edit distance from an input phoneme sequence from the acoustic decoding stage.

Comparing to the one-stage speech recognition, the multi-stage speech recognition can significantly reduce computation load. This is because the lexical decoding stage in the multi-stage speech recognition is repeated for every phoneme in the input phoneme sequence, while the lexical decoding stage in the one-stage speech recognition is repeated for typically every ten milliseconds. Since the duration of a phoneme is up to several hundred milliseconds, this computational saving is quite big, especially for cases with a large lexical search space. However, the multi-stage speech recognition has a shortcoming that the overall performance is highly affected by accuracy level of the phoneme recognizer. Moreover, the performance of the phoneme recognizer is generally poor since it depends only on the acoustic feature without using any higher level of information like a language model.

In this paper, we propose a new method of using a confidence vector as an input feature for lexical decoding stage to improve performance of multi-stage speech recognition system. For a phoneme segment, we introduce a confidence vector in which each element is defined as the likelihood of each phoneme for the given segment. And, we use this confidence vector as the input feature for the lexical decoder rather than just the phoneme sequence as in the conventional multi-stage speech recognition. This means that more information is transferred from acoustic decoder to lexical decoder.

This paper is organized as follows. At first, we present the structure of multi-stage speech recognition in section 2. In section 3, we describe the proposed method that uses confidence vector as an input feature of lexical decoding. In section 4, we explain the experiment environment and results. Finally, we summarize our work in section 5.

2 Multi-Stage Speech Recognition

Multi-stage speech recognition system is composed of acoustic decoder, lexical decoder and optional n-best rescoring [2][3].

The purpose of acoustic decoding is to convert an input acoustic feature sequence into a corresponding phoneme sequence as correctly as possible while minimizing the consumption of computational power. In most cases, CDHMM-based automatic phone recognizer is used for acoustic decoding. In this work, we also used automatic phone recognizer for the acoustic decoding, in which CDHMMs corresponding to 46 Korean phonemes including silence are used and a finite state network representing Korean syllable composition structure is used for pronunciation model.

The purpose of the lexical decoding is to recover an optimal word sequence from input phoneme sequence given as a result of acoustic decoding. Acoustic decoder may generate three types of phone errors, substitution, insertion and deletion for a reference phoneme due to inaccurate acoustic modeling, surrounding noises and so on. We model these error patterns in the form of probabilistic edit cost. For each speech utterance in training DB, the lexical decoder aligns phoneme sequence from the acoustic decoder with the phoneme sequence of reference words by dynamic time warping (DTW). By accumulating the alignment results for all the utterances in training DB, we obtain the substitution probability of each phoneme. While decoding, the cost at each node is derived from the substitution probability of each phoneme.

3 Confidence Vector Based Approach

In this paper, we introduce a phoneme confidence vector as an input feature for lexical decoding. This confidence vector based approach comprises four stages, phoneme segmentation, confidence vector extraction, lexical decoding, n-best rescoring. Figure 1 illustrates a block diagram of the multi-stage speech recognition using confidence vector.

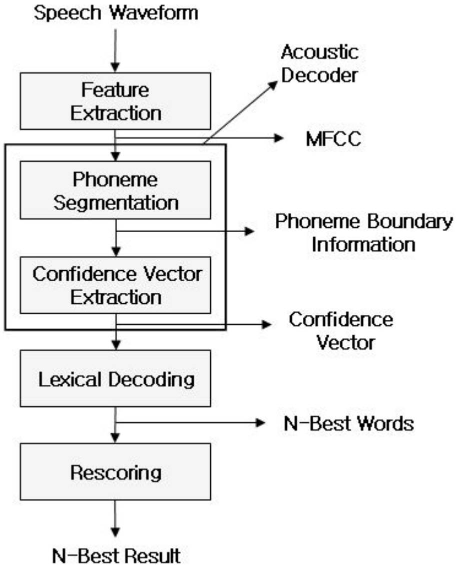

Figure 1. Block diagram of multi-stage speech recognition using confidence vector.

Firstly, the automatic phoneme recognizer finds optimal phoneme boundary information for an input utterance and then every acoustic likelihood corresponding to each of 46 Korean phoneme units is calculated for each segment, according to the phone segmentation result. Secondly, we define the confidence value for each phoneme at a given phoneme segment as follows,

$$confidence(i) = \frac{likelihood(i)}{\sum_{j=1}^{N} likelihood(j)} \qquad (1)$$

where confidence(i) means confidence value of i'th phoneme and N is the number of phonemes.

Because lexical decoding is a sort of dynamic programming, we need to define a local match cost to weight distance between two phonemes. For confidence vector approach, we defined lexical model as mean of confidence vectors of each phoneme. The cost of DTW is defined with mean of confidence vectors of reference phoneme and confidence vectors of input speech segment. We use three cost definitions that represent distance between two confidence vectors.

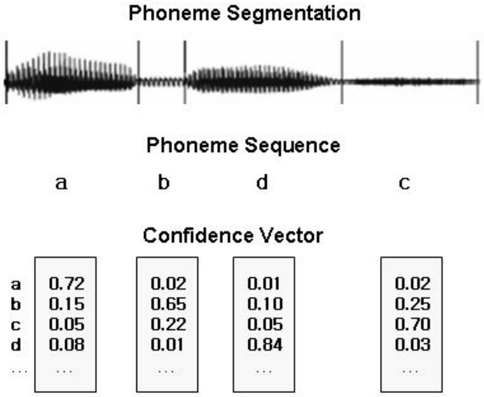

Figure 2. An input feature for lexical decoding, phoneme sequence and confidence vectors

3.1 Kullback-Leibler Divergence

Kullback-Leibler divergence measures a difference between two probability distributions. If we regard the confidence vector as a probability distribution, we can define the difference as cost. Equation 2 is a cost definition by Kullback-Leibler divergence for reference phoneme 'p'.

$$\mathrm{cos}t(f \mid p) = \sum_{i=1}^{N} \left(f(i) \times \log \frac{f(i)}{m_p(i)} \right) \quad (2)$$

In equation 2, f(i) is confidence value of i'th phoneme and mp(i) is i'th value of mean vector of phoneme 'p'.

3.2 Weighted Sum of Confidence

The second definition of cost is weighted sum of confidence. As weight values, mean of confidence vector that was previously trained is used. This weighted sum of confidence corresponds to substitution probability with consideration of prior probability distribution. It takes –log to make cost value from probability. Equation 3 is a cost definition by weighted sum of confidence.

$$\mathrm{cos}t(f \mid p) = -\log \left(\sum_{i=1}^{N} \left(m_p(i) \times f(i) \right) \right) \quad (3)$$

In equation 3, weight term can be considered as loss term of minimum risk decoding [4]. If the weight value is loss quantity of decision to phoneme 'p' with confidence f(i), decision for minimum cost is identical to decision for minimum risk.

3.3 Smoothing of Confidence

Even though we use weighted sum of confidence instead of substitution probability of best phone sequence, the cost is highly sensitive to the result of acoustic decoder. To prevent extreme decision of phoneme recognizer, we add smoothing parameters, α and β, like equation 4. α and β are assigned to be less then 1 and this gives effect to emphasize small values of probability.

$$\mathrm{cos}t(f \mid p) = -\log \left(\sum_{i=1}^{N} \left(m_p(i)^{\alpha} \times f(i)^{\beta} \right) \right) \quad (4)$$

4 Experiment Results

4.1 Experiment on POI domain

To evaluate the performance of the proposed algorithm, we performed experiments on Korean POI task domain which is an isolate word recognition composed of 220k POI entries. Test set comprises with 7 speakers, each speaker makes 99 utterances, so total test set is 693 utterances.

The speech signal was sampled at 16kHz, and the frame length was 20ms with 10ms shift. Each speech frame was parameterized as a 39-dimentional feature vector containing 12 Mel-Frequency Cepstral Coefficients (MFCCs), C0 energy, their delta and delta-delta coefficients. In order to evaluate the robustness of the proposed algorithm on acoustically, lexically mismatched condition, we took the experiments in two different test environment where N-fold test means the test is done in lexically matched condition and open test in acoustically and lexically mismatched condition.

N-fold Test

In order to reflect phoneme error patterns to the lexical decoding, we used the phoneme result of 6 speaker's utterance to train lexical model. We performed test for remained one speaker data, and we did 7 times with different test speakers. Each case training data of lexical model is 594 utterances and test data is 99 utterances. Lexical model is substitution probability or mean of confidence vector for reference phoneme.

Because training data of lexical model has the same vocabulary with test data and the same channel condition, training of lexical model in N-fold test can train error pattern of same vocabulary and environment condition. Table 1 shows the result of N-fold test experiment. Conventional multi-stage speech recognition method uses phoneme sequence as an input feature for lexical decoding, and its performance is the lowest among the rest methods. Because the phoneme accuracy of acoustic decoder is about 60~70% for this open test set, the performance of conventional method is degraded. If we use confidence vector as input feature for lexical decoding, the performance is increased by almost 10%. This performance enhancement proves that the confidence vector contains more useful information at lexical decoding. The cost definition with weighted sum of confidence has a similar performance with Kullback-Leibler divergence that is a well known distance measure. Especially with smoothing of confidence we get the best performance, but smoothing makes the search network enlarge and slow down a little because smoothing reduces the differences between confidence vectors.

Lexical Feature	Cost Definition	N-best Word Recognition Rate (%)		
		1	5	10
Phoneme Sequence	Substitution Prob.	64.5	75.6	78.6
Confidence Vector	KL Divergence	74.5	88.7	91.8
	Weighted sum of Confidence	74.5	88.5	92.4
	Smoothing of Confidence	78.8	90.9	93.8

Table 1. Word Recognition Rate of N-fold test

Lexical Model Training with Open set

In N-fold test, environment condition and error pattern of certain vocabulary is trained although we do not intend to do so. For an open test, we trained lexical model using word DB. We used ETRI phonetically optimized word (POW) DB which has 92k utterances.

Table 2 shows the results of open test experiments. If we use phoneme sequence as input feature for lexical decoder, performance is better than N-fold test. Because training DB for lexical model is enlarged to 92k, substitution probability might be modeled correctly. But confidence vector approach shows some performance degradations from N-fold test. In N-fold test, the performance is highly affected by the training of error pattern of certain vocabulary. In open test, KL divergence has better performance than cost definition by weighted sum of confidence in n-best aspect. And smoothing method has the best performance.

Lexical Feature	Cost Definition	N-best Word Recognition Rate (%)		
		1	5	10
Phoneme Sequence	Substitution Prob.	68.4	83.6	86.0
Confidence Vector	KL Divergence	71.4	86.6	90.9
	Weighted sum of Confidence	71.3	83.8	87.2
	Smoothing of Confidence	73.6	88.0	91.9

Table 2. Word Recognition Rate of Open test

N-best Rescoring

N-best rescoring stage does one-pass search among N-best result words. We used 500-best result for N-best rescoring. Table 3 shows the result of N-best rescoring. In the N-best rescoring test we used the lexical model from POW DB.

500-best performance of conventional multi-stage speech recognition system is 97.3%, and when we use confidence vector as input feature for lexical decoding with Kullback-Leibler divergence cost, our 500-best performance is 98.8% and 98.9% when we use weighted sum of confidence vector as cost. When we use smoothing method, 500-best performance is 99.1%. Since the performance of N-best rescoring is highly dependent with 500-best performance, conventional multi-stage method has the worst rescoring performance. Because 500-best performance of weighted sum of confidence and smoothing method are similar, the performance of N-best rescoring is almost same for these two cost definitions. If we use conventional one-pass decoder, word recognition rate is 82.9%. By N-best rescoring we can achieves the same level of performance with the one-pass decoder.

Lexical Feature	Cost Definition	N-best Word Recognition Rate (%)		
		1	5	10
Phoneme Sequence	Substitution Prob.	81.2	92.4	94.2
Confidence Vector	KL Divergence	82.0	94.2	95.5
	Weighted sum of Confidence	83.2	94.5	95.9
	Smoothing of Confidence	82.0	94.5	95.8

Table 3. Word Recognition Rate of N-best Rescoring

Speed Improvement

By dividing search procedure into acoustic and lexical parts, we can have the advantage of speed. In the acoustic decoder we used context-dependent model and a real-time factor of acoustic decoder is 0.03 on 3G Hz Dual-core CPU machine. Table 4 shows the recognition speed of multi-stage speech recognition system. Multi-stage approach is about 4-times faster than conventional one-pass decoder. If we use confidence vector as input feature for lexical decoding, we need additional amount of computation to extract confidence values. But the total speed is faster than conventional phoneme sequence feature. Since confidence vector feature makes more discriminative cost in lexical decoding, the total number of active nodes of search network is decreased and we can avoid increasing time.

Decoder/Lexical Feature		Real Time Factor
One-pass		1.16
Multi-pass	Phoneme Sequence	0.29
	Confidence Vector	0.27

Table 4. Speed of multi-stage speech recognition system

4.2 Experiment on hotel reservation domain

We applied the proposed method to the dialog speech recognition. The acoustic decoder is same as in the POI domain experiment. The lexical decoder searches the finite state network with class as its node which covers the 25,000 sentence templates in the hotel reservation domain. The test speech data is composed of 2,161 sentences uttered by 50 speakers, which is covered by the sentence templates. Table 5 shows that the use of the confidence vector is shown to be very effective, showing 29.6% error reduction rate to the conventional multi-stage method. Also, the proposed approach is shown to be competitive to the one-stage speech recognition while the execution speed is improved more than five times.

Decoder/Lexical Feature		Real Time Factor	Word Accuracy(%)
One-Pass		1.17	95.90
Multi-pass	Substitution Prob.	0.33	95.71
	KL Divergence	-	96.18
	Weighted sum of Confidence	0.22	96.39
	Smoothing of Confidence	-	96.98

Table 5. Word accuracy and speed of multi-stage speech recognition for hotel reservation domain

5 Conclusion

In this paper, we proposed a new method of using the confidence vector as an input feature for lexical decoding in the multi-stage speech recognition. We also introduced diverse cost definitions to measure the difference between confidence vector and reference vector. By transferring more information in the form of confidence vector to the lexical decoding, we can achieve better performance than the conventional method. The experiment results show up to 16.4% ERR of word accuracy for 220k Korean Point-of-Interest (POI) domain and 29.6% ERR of word accuracy for hotel reservation dialog domain.

Also, comparing to the one-pass decoder, the proposed method is shown to be far faster while maintaining competitive performance. Thus, this method is appropriate to build an embedded speech recognition engine on a mobile device.

As a future research, we will investigate the possibility of integrating the confidence vector approach with a phoneme lattice as an input feature for lexical decoding. A plain phoneme lattice can convey more information to lexical decoding, but mmay increase the computational complexity without any performance gain. We expect its advantage is maximized with combining the confidence vector approach.

References

[1] Victor Zue, James Glass, David Goodine, Michael Phillips, and Stephanie Seneff. The SUMMIT Speech Recognition System: Phonological Modeling and Lexical Access. In Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 49-52, 1990.

[2] Hoon Chung, Ikjoo Chung. Memory efficient and fast speech recognition system for low resource mobile devices. In IEEE Transactions on Consumer Electronics, Volume: 52, Issue: 3, pages 792 – 796, 2006.

[3] Hyungbae Jeon, Kyuwoong Hwang, Hoon Chung, Seunghi Kim, Jun Park, and Yunkeun Lee. Multi-stage Speech Recognition for POI. In Proc. Conference of Korean Society of Phonetic Sciences and Speech Technology, pp. 131-134, 2007.

[4] Vaibhava Goel, Shankar Kuman, and William Byrne. Confidence Based Lattice Segmentation and Minimum Bayes-Risk Decoding. In Proc. Conference of Eurospeech, 2001.