Confidence Score Based Unsupervised Incremental Adaptation for OOV Words Detection

Wei Chu, Xi Xiao, and Jia Liu

Department of Electronic Engineering,
Tsinghua University,
Beijing 100084, China
chuwei@stu.ee.tsinghua.edu.cn

Abstract. This paper presents a novel approach of distinguishing in-vocabulary (IV) words and out-of-vocabulary (OOV) words by using confidence score-based unsupervised incremental adaptation. The unsupervised adaptation uses Viterbi decode results which have high confidence scores to adjust new acoustic models. The adjusted acoustic models can award IV words and punish OOV words in confidence score, thus obtain the goal of separating IV and OOV words. Our Automatic Speech Recognition Laboratory has developed a Speech Recognition Developer Kit (SRDK) which serves as a baseline system for different speech recognition tasks. Experiments conducted on the SRDK system have proved that this method can achieve a rise over 41% in OOV words detection rate (from 68% to 96%) at the same cost of a false alarm (taken IV words as OOV words) rate of 10%. This method also obtains a rise over 11% in correct acceptance rate (from 88% to 98%) at the same cost of a false acceptance rate of 20%.

1 Introduction

Nowadays speech recognition system can perform quite well on isolated words recognition if only providing IV word utterances as input and a vocabulary which is not very large. But the situation gets worse as the appearance of OOV words. In real world, OOV words input problem should not be overlooked, because the recognizer is faced with the OOV words spoken by users all the time.

Confidence score is utilized to evaluate the reliability of recognition results by S. Cox [1]. Later on, many approaches of calculating confidence score are introduced. T. J. Hazen has done prominent work in summarizing and devising confidence scores in word-level and utterance-level [2]. But for our practical short isolated words recognition, it is hard to distinguish IV words from OOV words in confidence score domain.

One major reason is that the acoustic models used in SRDK can not generate confidence scores which are separable for IV and OOV words. In this paper, a confidence score-based unsupervised incremental adaptation method is used to adjust the acoustic models used in SRDK system.

During the adaptation, we first send adaptation data including IV and OOV words into SRDK system, then use the Viterbi decode results of the recognizer which have high confidence scores to guide the model adaptation. A Threshold for confidence

score is set in order to ensure that almost all the words used for adaptation are correctly recognized IV words.

Because the adaptation data are limited, we adopt maximum likelihood linear regression (MLLR) + maximum a posteriori (MAP) adaptation method. Our experiments have proved this unsupervised adaptation procedure can greatly improve the later performance of OOV words detection.

2 Word Confidence Scoring

In this OOV words detection task, two classical but proved to be efficient confidence scores are employed. For computational reasons, we adopt a two-pass search strategy in which a semi-syllable based confidence score [3] is calculated in the first-pass, and a filler model based confidence score [4] is calculated in the second-pass. Finally, we combine the two confidence scores into a single dimensional confidence score through a simple linear discrimination method.

2.1 Semi-syllable Based Likelihood Ratio

$C_{ss}(X, W_0)$ is the semi-syllable based confidence score of the best hypothesis W_0 when the observed vector sequence is X which focuses on likelihood ratio:

$$C_{ss}(X, W_0) = \frac{P(X|W_0)}{P(X)}. \quad (1)$$

Consider the states alignment of the observed vectors, we can express $P(X|W_0)$ as:

$$P(X|W_0) = P(X_1, X_2, \dots, X_m | h_1, h_2, \dots, h_m), \quad (2)$$

where h_i is the semi-syllable alignment of the observed vector sequence X_i , The corresponding relationship between h_i and X_i is determined during Viterbi match. Assuming the observed vectors X_i are independent of each other, we have:

$$P(X) = \prod_{i=1}^m P(X_i). \quad (3)$$

Furthermore, we assume that semi-syllables h_i are independent of each other. We represent the conditional probability $P(X|W_0)$ as:

$$P(X|W_0) = P(X_1, X_2, \dots, X_m | h_1, h_2, \dots, h_m) = \prod_{i=1}^m P(X_i | h_i). \quad (4)$$

So, we get

$$CS_{ss}(X, W_0) = \prod_{i=1}^m \frac{P(X_i | h_i)}{P(X_i)}. \quad (5)$$

For each segmented observed vector X_i ,

$$P(X_i) = \sum_{j=1}^M P(X_i | h_j). \quad (6)$$

In our system based on semi-syllable, if h_j is matched as a consonant (or vowel), M represents the amount of all the consonants (or vowels). Consequently, $CS_{log_ss}(X, W_0)$ in the log domain can be described as:

$$CS_{log_ss}(X, W_0) = \log P(W_0 | X) = \sum_{i=1}^m [\log P(X_i | h_i) - \log \sum_{j=1}^M P(X_i | h_j)]. \quad (7)$$

2.2 Filler Model Based Likelihood Ratio

$CS_{fl}(X, W_0)$ is the filler model based confidence score of the best hypothesis W_0 when the observed vector sequence is X .

$$CS_{fl}(X, W_0) = \frac{P(X | W_0)}{P(X | H_{Filler})}. \quad (8)$$

In our system, online garbage model $H_{Online_Garbage}$ is considered to work as filler model H_{Filler} .

In the back-end, N -best hypotheses are listed out. Besides the best hypothesis W_0 , the left $N-1$ hypotheses are called online garbage. $P(X | H_{Online_Garbage})$ is as follows:

$$P(X | H_{Online_Garbage}) = \frac{1}{N-1} \sum_{i=1}^{N-1} P(X | W_i). \quad (9)$$

So the normalized confidence score $CS_{log_fl}(X, W_0)$ in the log domain is expressed as follows:

$$CS_{log_fl}(X, W_0) = \frac{1}{n_x} \left\{ \log P(X | W_0) - \log \left[\frac{1}{N-1} \sum_{i=1}^N P(X | W_i) \right] \right\}, \quad (10)$$

where n_x represents the frame numbers of word W_0 , and makes words with different frame numbers comparable in confidence score domain.

2.3 Confidence Score Combination

Since the two confidence scores are of different information, better performance will be achieved while combining them together.

For computational reasons, Fisher linear discrimination is used to find the projection vector $\mathbf{p}^T$, and generate a linear discriminating plane between the IV and OOV words. Now we obtain a single dimensional confidence score $CS_{single}(X, W_0)$:

$$CS_{single}(X, W_0) = \mathbf{p}^T CS_{multi}(X, W_0) = [\alpha \quad \beta] \begin{bmatrix} CS_{log_ss}(W_0) & CS_{log_fl}(W_0) \end{bmatrix}^T. \quad (11)$$

3 Unsupervised Incremental Adaptation

After finishing calculating confidence score, we find that IV and OOV words can not be separated easily in confidence score domain. It is hard for us to detect OOV words by confidence score. The main reason for this phenomenon is that the initial acoustic models are trained with generic speech data. This initial acoustic models can perform speaker-independent speech recognition tasks quite well when providing only IV word utterances as input. But when adding OOV word utterances into the input sequence, these acoustic models can not generate separable confidence scores for IV and OOV words. For this reason, we have an idea that we can use specific IV word utterances to adjust suitable acoustic models. The acoustic model is a specific context-dependent phoneme HMM to our OOV word detect task. D. Wang has used this confidence score-based unsupervised adaptation method to improve the performance of speech recognition [5]. Our experiments have proved that these acoustic models after adaptation can also award IV words and punish OOV words in confidence score domain.

Because we use IV word utterances to perform unsupervised incremental adaptation, it is possible that wrongly recognized results will degrade the model parameters accuracy. Our strategy is to select only correctly recognized IV word utterances with high confidence scores for the adaptation.

3.1 MAP Adaptation

In MAP adaptation, the following formulas are used in each step of re-estimation, for each Gaussian pdf [6]:

$$\mu = \frac{\frac{N_{prior}}{N_{Init}} \sum_{i \in Init} x_i + \sum_{j \in Adapt} w(x_j) x_j}{N_{prior} + \sum_{j \in Adapt} w(x_j)}, \quad (12)$$

$$\sigma^2 = \frac{\frac{N_{prior}}{N_{Init}} \sum_{i \in Init} x_i^2 + \sum_{j \in Adapt} w(x_j) x_j^2}{N_{prior} + \sum_{j \in Adapt} w(x_j)} - \mu^2, \quad (13)$$

where N_{prior} is a control parameter of the adaptation process. The fewer N_{prior} is, the more adaptation utterances are taken into account with respect to prior data. $w(x_j)$ is a weighting factor to determine in what way the utterances should be used in the adaptation process. In our system, We adopt a strict strategy for $w(x_j)$:

$$\begin{cases} w(x_j) = 1 & \text{if } CS_{single}(x_j, W_0) > Th \\ w(x_j) = 0 & \text{if } CS_{single}(x_j, W_0) \leq Th \end{cases} \quad (14)$$

In our experiment we find that when the confidence scores of recognition results exceed a certain threshold, all the Viterbi decoder output is right. Only utterances with confidence score above Th can be used for adaptation in order to ensure that wrong Viterbi decode results will not perform negative effect on model parameters.

3.2 MLLR Adaptation

MLLR adaptation [7] is suitable when the amount of adaptation data is small or limited. MLLR adaptation performs faster than MAP adaptation when given the same amount of adaptation data. For each Gaussian pdf, μ_{ik} is transformed by using the following formula:

$$\tilde{\mu}_{ik} = A_c \mu_{ik} + b_c, \quad (15)$$

where A_c is a regression matrix and b_c is an additive bias vector associated with some broad class c , which can be either a broad phone class or a set of tied Markov states. We also only utilize those utterances with confidence scores over Th in MLLR adaptation, just as in our MAP adaptation.

4 Experiment Results

To show the effectiveness of the proposed method, we conduct experiments on our SRDK system. The initial acoustic models of the SRDK system are trained from approximately 100 hour speech data.

The vocabulary size of IV words is 200. Because our OOV words detection is expected to perform well in adverse situation, it is assumed that OOV word utterances occupy 50% of the total utterances. 3000 IV words and 3000 OOV word utterances are prepared as input utterances.

We use one third of the total input utterances for MLLR+MAP unsupervised incremental adaptation. Left 4000 utterances including IV and OOV word utterances are taken as OOV words detection test set.

In MLLR+MAP adaptation, to find an optimum value for Th , we compare OOV words detection performance under different Th . The results are depicted in Figure 1. The work point refers to the OOV words detection rate at the point where OOV words detection rate + false alarm rate = 1. Given a Th , each work point represents the best work condition under this Th . We want to mention that all our following experiments are conducted under this optimum Th (4×10^5).

The original OOV words detection point is 82.5% before our unsupervised adaptation. When the Th is higher than the optimum Th (4×10^5), the amount of utterances which used in adaptation decreases, and the work point falls, but always over the initial work point. When the Th is lower than the optimum Th , the performance of OOV words detection falls greatly. It is mainly because the incorrectly recognized words have performed negative effects on the unsupervised adaptation. We observed that when the input data used in adaptation contain a few OOV word utterances, the work point after adaptation is still higher than the initial work point.

Fig. 1. Change in work points of OOV words detection depending on the changes in threshold of confidence score after MLLR+MAP adaptation

Figure 2 and Figure 3 below illustrate that IV and OOV words become easier to separate in confidence score domain after adaptation. During the incremental adaptation procedure, the acoustic model parameters are gradually adjusted according formulas (12), (13) and (14), thus can generate higher confidence scores for the later coming utterances used in adaptation. So it is also feasible to gradually lower Th as the adaptation procedure goes gradually. But we want to perform a robust unsupervised incremental adaptation, thus a fixed Th is used during the adaptation procedure to prevent the possible underlying instabilities which may perform negative effects on OOV word detection.

In Figure 4, false alarm rate in Figure 4 refers to the rate of false acceptance of IV word as OOV words. Figure 4 shows that the proposed method has achieved a rise

Fig. 2. Confidence score distribution of IV and OOV words before adaptation

Fig. 3. Confidence score distribution of IV and OOV words after adaptation. Under $Th = 4 \times 10^5$

Fig. 4. OOV words detection performance before adaptation and after adaptation. Under $Th = 4 \times 10^5$.

over 41% in OOV words detection rate (from 68% to 96%) at the same cost of a false alarm rate of 10%.

In Figure 5, false acceptance rate refers to the percentage of wrongly recognized words which are accepted. The correct acceptance rate refers to the percentage of correctly recognized words which are accepted. It is essential for us to examine the recognition performance of the adapted models. Figure 5 depicted that we can achieve a rise in correct acceptance rate (from 88% to 98%) at a false acceptance rate of 20%, when the input data are composed of 50% IV word utterances and 50% OOV word utterances. But when the input data are all IV word utterances, we observe degradation in correct acceptance rate (from 88% to 68%) at a false acceptance rate of

20%. The main reason is that the adjusted acoustic models is task-oriented (best fit for 50% IV + 50% OOV), and its performance relies greatly on the proportions of IV and OOV word utterances in the input data.

Fig. 5. ROC curve of recognition with rejection before adaptation and after adaptation. Under $Th = 4 \times 10^5$.

5 Conclusions

This paper presented a new method for improving the OOV word detection rate by using unsupervised incremental adaptation based on confidence score. The effectiveness of this method has been proved by experiments on our SRDK system. Our future work will include applying this idea not only to the acoustic models, but also to the language model of a real-time human-machine interactive system in which input utterances are composed of isolated words and sentences. It is important to find a balance point between OOV words detection rate and recognition rate in practical usage according to the real feelings of the users.

Acknowledgements

The research is supported by National Natural Science Funds of China (Item No. 60572083). The authors would like to thank everyone who contributed to building and improving the SRDK system.

References

1. Cox, S. and Rose, R.: Confidence measures for the SWITCHBOARD database. In Proceedings of ICASSP 1996. Atlanta. (1996) 511-514
2. Hazen, T. J., Seneff, S. and Polifroni, J.: Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*. 16(1). (2002) 49-67

3. Sankar, A. and Wu, S.-L.: Utterance verification based on statistics of phone-level confidence scores. In Proceedings of ICASSP 2003. Menlo Park. (2003) 584–587
4. Boite, J., Boulard, H., D'hoore, B. and Haesen, M.: A new approach towards keyword spotting. In Proceedings of Eurospeech 93. Berlin. (1993) 1273-1276
5. Wang, D., and Narayanan, S. S.: A confidence-score based unsupervised map adaptation for speech recognition. In Proceedings of 36th Conference on Signal, Systems and Computers. Pacific Grove. (2002) 222–226
6. Charlet, D.: Confidence-measure-driven unsupervised incremental adaptation for HMM-based speech recognition. In Proceedings of ICASSP 2001. Salt Lake City. (2001) 357–360
7. Leggetter, C. J. and Woodland, P. C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*. Vol. 9, No. 2. (1995) 171-185