Investigation on Rescoring Using Minimum Verification Error (MVE) Detectors

Qiang Fu and Biing-Hwang Juang

School of Electrical and Computer Engineering,
Georgia Institute of Technology
{qfu,juang}ece.gatech.edu

Abstract

Discriminative training, especially Minimum Verification Error (MVE) method plays an important role in the detection-based ASR. Recently, discriminative training also has been shown to be effective in large vocabulary continuous speech recognition [1]. In this paper, we propose a rescoring framework to show the improvement by fusing MVE-trained detectors with a conventional recognizer. The recognizer performs regular Viterbi decoding, generating possible recognition candidates with corresponding likelihood in a fashion of either N-best lists or word graphs. Detectors trained under MVE criterion form and conduct hypothesis testing for all test tokens to accomplish additional scores. A number of linear or non-linear rescoring methods are then presented to combine these two groups of scores. The experiments were conducted on the TIMIT database, and the results indicates that combining based on word graphs outperforms the one on N-best lists in the final accuracy. This rescoring framework explores possible ways to combine other independent knowledge sources with a conventional recognizer. Further more, it can guide the future research of the pure detection-based ASR techniques.

1. Introduction

Although we have witnessed fast development of automatic speech recognition (ASR) techniques for decades, the framework of state-of-the-art speech recognition systems is still known to be too rigid to incorporate new knowledge or information. These techniques are in general task specific with a fixed system construct which does not allow alternation to adapt to new applications without totally re-designing the entire system. Moreover, mis-matched design scenarios such as out-of-vocabulary words or different training and testing conditions will lead to severe performance degradation.

Detection-based ASR is an alternative paradigm [2]. It conducts a bottom-up hypothesis testing framework based on the Neyman-Pearson lemma. This framework endows the detection-based ASR its flexibility to combine different knowledge sources and the ability to fuse lower level information into higher level hypotheses, and at the same time to neglect superfluous input. We have already seen encouraging results in [3] .

Discriminative training methods such as Minimum Verification Error (MVE) training [4] are data-driven approaches that aim at minimizing an empirical estimate of the test error. They have been extensively applied to event verification applications, such as speaker verification. In [4], we have studied the performance of MVE on various broad phonetic class detection tasks. It is a solid manifestation of the effectiveness of the MVE modeling method in the detector design in detection-based ASR.

In this paper, we introduce a rescoring framework by combining the scores of a conventional decoder and those computed by MVE-trained detectors. Unlike other rescoring methods such as [5][6], no other "knowledge-based" front-end feature is utilized to accomplish additional information. We are trying to fuse two or more relatively independent "inference" measures, which are computed on the same feature space, to enhance the system performance. Normally a conventional recognizer organizes decoded candidates in two forms: an N-best list or a word graph. The comparison experiments indicate that word graph is a more applicable structure for knowledge incorporation because it contains a richer search space. Three rescoring approaches are studied in this paper on labels transcribed by taxonomical phoneme sets studied in [4].

One issue in this rescoring scheme is that we don't alter anything in the conventional Viterbi decoding procedure. It means the segmentation information provided by the decoder, which is not highly trustworthy, is somehow kept valuating. However, it is well known that the segmentation information is critical in speech recognition. Therefore this limitation would impact the final improvement. Embedding the detectors into path searching and pruning of the recognition is a viable solution to create a real hybrid system that the detectors are fully exploited [3]. Furthermore, it is hard to predict which specific rescoring method performs better in a particular task though we are able to rank them in this paper. It is to be noted that rescoring is to enhance the decoded result of a conventional recognizer which has its own optimality considerations; any additional optimality claim that involves other sources of knowledge would not be the objective here. We are trying to explore a global framework under which people can combine the conventional decoding results and other independence information sources such as hypothesis testing from detectors in a flexible course. Figure

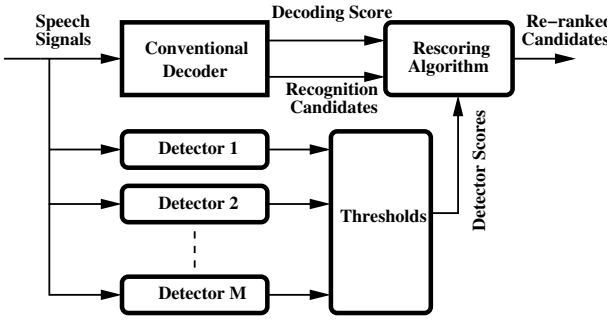

Figure 1: *The rescoring diagram using MVE-trained detectors.*

1 depicts this diagram. We can replace any approach in the "Rescoring Algorithm" box, adjust the structure of recognition candidates, and tune thresholds up to any particular tasks.

This paper is organized as following: we will briefly review the theory of MVE in the next section. Rescoring methods will be introduced in section 3. Experiments and results are presented in section 4. Finally, we conclude the paper in section 5.

2. Minimum Verification Error (MVE) Training

Analogous to MCE [7], the essence of MVE [4] is to directly minimize the total detection errors. In detection problems, there are two different kinds of errors: type I error (missing) and type II error (false alarm). Viewed from a classification problem perspective, there are two misclassification measures respectively. Assume there are M classes and K training tokens in the training set. For any training token labeled in the ith class, a type I error (miss) may result when applied to the detector of the ith class, and possibly $M-1$ type II errors (false alarm) when applied it to detectors for all the other classes. The type I misclassification measure for an incoming training token $\mathbf{O^i}$ labeled in the ith class can be formulated as

$$d_I = -g_t^i(\mathbf{O^i}|\Theta_t^i) + g_a^i(\mathbf{O^i}|\Theta_a^i) + \gamma_i \tag{1}$$

where $g_t = \frac{1}{T} LR_t(\mathbf{O^i}|\Theta_t^i)$ is the normalized log likelihood of the target model for the ith class. T is the number of frames in the incoming token. $g_a = \frac{1}{T} LR_a(\mathbf{O^i}|\Theta_a^i)$ is the normalized log likelihood of the anti-model for the ith class. Θ_t and Θ_a are respectively parameter sets of the target and the anti models. γ_i is the decision threshold for class i.

At the same time, the type II misclassification measure of the jth class for an incoming training token $\mathbf{O^i}$ labeled in the ith class is

$$d_{II}^j(\mathbf{O^i}|\Theta^j) = +g_t^j(\mathbf{O^i}|\Theta_t^j) - g_a^j(\mathbf{O^i}|\Theta_a^j) + \gamma_j \tag{2}$$

$$j = 1, 2, \ldots, M,\ j \neq i$$

The two misclassification measures can be embedded into smoothed loss functions written as

$$l_I^i(d_I^i) = \frac{1}{1 + \exp\{-\alpha_i d_I^i\}} \tag{3}$$

and

$$l_{II}^j(d_{II}^j) = \frac{1}{1 + \exp\{-\alpha_j d_{II}^j\}} \tag{4}$$

$$j = 1, 2, \ldots, M,\ j \neq i$$

Finally, the empirical loss for a training set $\{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_K\}$ is given by

$$L_{total}(\tilde{\Theta}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{M} l_{total}^i(\mathbf{O}_k|\Theta^i) 1(\mathbf{O}_k \in class\ i) \tag{5}$$

where the parameter set $\tilde{\Theta}$ is defined by $\tilde{\Theta} = \{\Theta_t^i, \Theta_a^i\}$, $i = 1, 2, \ldots, M$. The composite error estimation function $l_{total}^i(\mathbf{O}_k|\Theta^i)$ is a combination of type I and type II errors.

$$\begin{aligned} l_{total}^i(\mathbf{O}_k|\Theta^i) &= PE_I l_I^i(\mathbf{O}_k|\Theta^i) \\ &+ PE_{II} \sum_{j=1, j\neq i}^{M} l_{II}^j(\mathbf{O}_k|\Theta^j) \end{aligned} \tag{6}$$

PE_I and PE_{II} are penalty weights for type I and type II errors. The minimization of L_{total} can be done through the generalized probabilistic descent (GPD) method [7] w.r.t. all parameters.

3. RESCORING METHODS

We investigated three rescoring methods to combine the likelihood and scores generated from the conventional decoder and MVE-trained detectors. Suppose there are M classes of sub-word units, hence there are M corresponding detectors that each of them consists of a target model and an anti-model. For a segment that is decoded as the ith class with log likelihood $S_{decode}^{(i)}$, its jth ($j = 1, 2, \ldots, M$) detector scores are $S_{tgt}^{(j)}$ and $S_{anti}^{(j)}$, respectively. Namely, the likelihood ratio for the jth detector is $ratio^{(j)} = S_{tgt}^{(j)} - S_{anti}^{(j)}$. We call the score for the test segment belonging class i after combination $S_{new}^{(i)}$.

The first method is called *Naive-Adding (NA)*. From its name we can know that it is a quite naive score combination scheme. In this approach, the new score of each segment being decoded as the ith class is

$$S_{new}^{(i)} = S_{decode}^{(i)} - S_{anti}^{(i)} + ratio^{(i)} \tag{7}$$

The reason for subtracting $S_{anti}^{(i)}$ is to scale the decoding score into a relatively close dynamic range with the likelihood ratio. This procedure is also taken in the following two methods.

The second method is named *Competitive-Rescoring (CR)*. In this approach, we define a new "competitive" score $S_c^{(i)}$.

$$S_c^{(i)} = ratio^{(i)} - \log\{\frac{1}{M-1}\sum_{j\neq i}^{M}\exp(\eta \cdot ratio^{(j)})\}^{1/\eta} \quad (8)$$

and

$$S_{new}^{(i)} = S_{decode}^{(i)} - S_{anti}^{(i)} + S_c^{(i)} \quad (9)$$

In the first method only the likelihood ratio from underlying class of detectors are used for rescoring. But in this case, we first compute a distance measure between the claimed class to a geometric average of the other competitive classes. This quantity $S_c^{(i)}$ is similar to the "misclassification measure" function d in MCE training [7] but using the corresponding detectors' likelihood ratio and there is a sign difference.

The third method is called *Remodeled Posterior Probability (RPP)*. Borrowing from the idea of the recognition word graph, we formed a pseudo-graph for each phoneme segment. We can consider the detection results of the total M detectors are M extra pathes for the testing speech segment. A remodeled posterior probability of the claimed class i is defined as the ratio of two scores. The score on the numerator is the scaled decoding score of claimed class i plus the likelihood ratio of the detector for class i. The score on the denominator is the sum of the numerator score and all the other detection scores. i.e,

$$S_{new}^{(i)} = \frac{\exp(S_{decode}^{(i)} - S_{anti}^{(i)}) + \exp(ratio^{(i)})}{\exp(S_{decode}^{(i)} - S_{anti}^{(i)}) + \sum_{j=1}^{M}\exp(ratio^{(j)})} \quad (10)$$

4. Experiments and results

The experiments were conducted on the TIMIT database. The training set has 3,696 utterances and the test set has 1,344 utterances (the utterances for speaker adaptation are ignored). The acoustic model of the baseline decoder consists of 48 CI phones defined in [8]. Each phone is modeled by a 3-state HMM, with each state represented by 16 Gaussian mixtures. The model parameters are trained by embedded Baum-Welch algorithm using 39 dimensional feature vectors with 12MFCC,12Δ, 12Δ^2 and 3 log energy values.

Three taxonomical phonetic category detectors are defined and trained following the same way of MVE training in [4]. These categories include 6 classes (stops, vowels, nasals, fricatives, silence, and others. see[9]), 14 classes (see[10]), and 48 classes phonemes respectively.

The entire test set first went through a Viterbi decoding process to generate N-best lists and word graphes with a simple word-loop language model. Note that the term "N-best" here refers to the N utterances with the highest word string likelihood. A forced-alignment is conducted on each testing utterance to acquire phone boundaries. We mapped the 48-phone transcription into 6-phone and 14-phone labels to check the capability of rescoring algorithms in different scenarios. The system performance reaches its upper bound when selecting the candidate in the N-best list or word graph which best matches the reference phone transcription. In this paper, we approximate the value of this upper bound to be the one we get in a 100-best list. Hence, to evaluate a rescoring algorithm, the relative accuracy improvement is defined by the ratio of the absolute improvement over the offset between the upper bound accuracy and the baseline accuracy. Table 1 shows phone recognition accuracy of the baseline decoder and upper bound of 100-best list for transcriptions in three kinds of phoneme sets.

Table 1: *Baseline decoder accuracy and upper bounds.*

Acc(%)	6class	14class	48class
Baseline	75.44	63.61	55.33
Upper bound	80.84	70.85	62.08

4.1. N-best lists rescoring

In our experiments, we assume $N = 100$. All three rescoring algorithms are conducted for each segment of phonemes. Each phoneme will have new scores as described in the last section. We add all scores in an utterance together to compute a new total score of the string. The N-best list is reranked based on the new total score.

Table 2 display the performance of all three rescoring approaches on the 100-best list for 6-phone, 14-phone, and 48-phone transcriptions. We can see that in 6-phone and 14-phone cases, all rescoring methods achieve considerable improvement. Moreover, method 1 (Naive-adding) has the least performance boosting and method 3 (Remodeled Posterior-Probability) obtains the most gain. It is not surprising since method 1 is the most naive approach among those three while the method 3 tries to find a candidate with maximum value of a remodeled posterior probability, which bears relationship to Bayes risk.

In the case of 48-phone transcription, the relative improvement is not as high as the former two. There is even a performance degradation when using NA. One of the reasons could be that the accuracy of detectors is not as good as that of 6-phone and 14-phone. Further more, a combination score may be sensitive to the dynamic range of the likelihood, whose numeric behavior may be erratic in a naive method.

4.2. Word graph rescoring

The word graph pruning criterion in the experiments are set in a way that only 3 pathes can exist at the same time. In one graph, each node represents a time point and each arc represents a word. We forced-aligned each arc (word) for phoneme boundaries and applied all three algorithms accordingly to calculate new scores for each phone in this word.

Table 2: *N-best rescoring performance.*

phoneme class	Acc(%)	method1 (NA)	method2 (CR)	method3 (RPP)
6class	Baseline	75.44	75.44	75.44
	Upperbound	80.84	80.84	80.84
	Rescored	76.36	76.38	80.00
	Relative	**17.04**	**17.40**	**84.44**
14class	Baseline	63.61	63.61	63.61
	Upperbound	70.85	70.85	70.85
	Rescored	65.38	67.27	68.45
	Relative	**24.45**	**50.55**	**61.88**
48class	Baseline	55.33	55.33	55.33
	Upperbound	62.08	62.08	62.08
	Rescored	55.04	55.61	55.91
	Relative	**-4.30**	**4.15**	**8.59**

The new score for a word is the sum of all new scores of the phones belonging to this word. A best word sequence is then selected from the underlying word graph based on new word scores.

We only conduct experiments on the 48-phone case. We can see that RPP still outperforms the other two methods. The other observation is that the rescoring performance with word graphs is higher than the ones with N-best lists. It is expected since a word graph represents much larger search space than a N-best list.

Table 3: *Word graph rescoring performance for the 48-phone transcription.*

Acc(%)	method1 (NA)	method2 (CR)	method3 (RPP)
Baseline	55.33	55.33	55.33
Upperbound	62.08	62.08	62.08
Rescored	55.30	56.02	56.10
Relative	**-0.44**	**10.22**	**11.4**

5. Conclusions

We have proposed a rescoring framework for combining relatively independent information sources. In this paper, we specifically rescored decoding likelihood using MVE-trained detectors. Three different rescoring methods have been introduced and the experiment results show that creating a pseudo-phone graph and recomputing the posterior probability accomplishes the best performance enhancement. Meanwhile, two structures for re-ranking decoding candidates, the N-best list and the word graph, are investigated. As expected, the word graph outperforms N-best list in our rescoring tasks since it represents a richer search space. In this paper, MVE training shows promising results in helping the conventional ASR techniques. However, detectors can only apply incremental impact on the final results since inaccurate segmentation information is kept during rescoring. To overcome this limit, We will migrate to complete detection-based ASR in the future.

6. Acknowledgement

This work is supported in part by a grant from AT&T.

7. References

[1] W. Macherey, L. Haferkamp, R. Schluter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Interspeech-2005*, Lisbon, Portugal, Sep. 2005, pp. 2133–2136.

[2] C.-H. Lee and B.-H. Juang, "A new detection paradigm for collaborative automatic speech recgnition and understanding," in *SWIM-2004*, Maui, Hawaii, Jan. 2004.

[3] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 558–568, Nov. 1998.

[4] Q. Fu and B. H. Juang, "Segment-based phonetic class detection using minimum verification error (mve) training," in *Interspeech-2005*, Lisbon, Portugal, Sep. 2005.

[5] J. Li, Y. Tsao, and C.-H. Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," in *ICASSP-05*, Philadelphia, Pennsivania, March 2005.

[6] B. Launay, O. Siohan, A. Surendan, and C.-H. Lee, "Towards knowledge-based features for hmm based large vacabulary automatic speech recognition," *icassp*, pp. 817–820, 2002.

[7] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.

[8] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[9] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[10] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, NY, 1999.