A STUDY ON RESCORING USING HMM-BASED DETECTORS FOR CONTINUOUS SPEECH RECOGNITION

Qiang Fu, Biing-Hwang Juang

{qfu,juang}@ece.gatech.edu
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT

This paper presents an investigation of the rescoring performance using hidden Markov model (HMM) based attribute detectors. The minimum verification error (MVE) criterion is employed to enhance the reliability of the detectors in continuous speech recognition. The HMM-based detectors are applied on the possible recognition candidates, which are generated from the conventional decoder and organized in phone/word graphs. We focus on the study of rescoring performance with the detectors trained on the tokens produced by the decoder but labeled in broad phonetic categories rather than the phonetic identities. Various training criteria and knowledge fusion methods are investigated under various semantic level rescoring scenarios. This research demonstrates various possibilities of embedding auxiliary information into the current automatic speech recognition (ASR) framework for improved results. It also represents an intermediate step towards the construction of a true detection-based ASR paradigm [1].

Index Terms: rescoring, MVE, phone/word graph , detection-based ASR

1. INTRODUCTION

It is well known that the state-of-the-art speech recognition framework faces challenges in incorporating new knowledge or information. The current techniques are inflexible and task-specific and in general do not allow adaption to new applications without a substantial system adjustment. Furthermore, any mismatch between the training and test environments such as out-of-vocabulary words or different noise conditions would cause a serious performance degradation. Detection-based ASR is an alternative paradigm [1]. It conducts a bottom-up hypothesis testing framework based on the detection theory. This framework is flexible in its ability to combine different knowledge sources and the to fuse lower level information into higher level hypotheses, while neglecting superfluous inputs. We have already seen encouraging results in [2, 3, 4, 5, 6].

We have witnessed many related research in regard to the detector design methodology and information integration approaches [3, 4, 7, 8]. We are reminded by the previous exploration that before building a real and complete detection-based system, it is helpful to incrementally investigate the effect of combining detectors and conventional decoders. Thus a rescoring system, a hybrid of attribute detectors and a conventional decoder, is the research objective here. In [7], a frame-based detector was introduced and "knowledge-based" front-end features are utilized to accomplish enhanced recognition accuracy. A segment-based rescoring system was reported in [4], showing preliminary improvements; it exploits a set of HMM-based detectors to help a conventional recognizer in reaching the final decision. Two or more relatively independent "inference" measures were integrated in the same observation space.

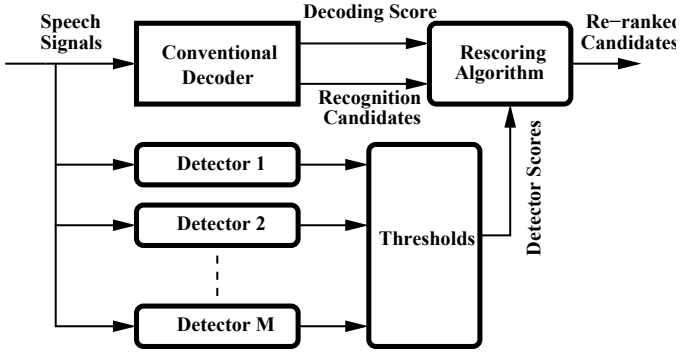

Fig. 1. The rescoring diagram using HMM-based detectors.

In this paper, we follow the research in [4] and present a more extensive study of the rescoring performance of HMM-based detectors. Fig.1 depicts the general framework of our rescoring strategy. We can replace any approach in the "Rescoring Algorithm" box, adjust the structure of recognition candidates, and tune thresholds in any particular tasks.

According to Fig.1, the performance of a rescoring system is decided by two key factors: the reliability of the detectors and the effectiveness of the rescoring algorithms under different scenarios. In this paper, a systematic investigation upon these two key issues has been organized. First, to enhance the reliability of the detectors, some effective discriminative training criteria are employed. The minimum verification error (MVE) training [9, 10] is a well-suited approach that aims at minimizing the empirical estimate of the total detection error. We have seen a number of solid manifestations of the effectiveness of the MVE modeling method for the detector design in detection-based ASR literatures [9, 11]. However, the original MVE training is designed for using isolated speech segments hence not consistent to be combined with most of the rescoring algorithms for continuous speech recognition tasks. To alleviate the mismatch between the detector training and rescoring scenario, we propose two modified versions of the MVE training criteria. Second, we examine various rescoring algorithms under multiple rescoring scenarios. We investigate two types of rescoring algorithms in terms of their information fusion strategy – the scoring fusion rescoring and the

decision fusion rescoring. For the first type of algorithm, we combined the scores from different information sources and make the final rescoring decision based on those scores. For the second one, the independent decisions are made before being transformed into the final result. The study of the rescoring algorithms are conducted on both the intra-semantic level rescoring (e.g., the rescoring is made on the phone level and the objective of rescoring is to improve phone recognition accuracy) and the inter-semantic level rescoring (e.g., the rescoring is made on the phone level and the objective of rescoring is to improve word recognition accuracy) to find out the appropriate rescoring configurations under different scenarios. Note that we are not claiming any "optimal" system. The objective of this paper is to justify suitable combinations of the detector design and the rescoring algorithms for future tasks. Further more, it is a helpful intermediate step towards the pure detection-based ASR applications.

This paper is organized as follows. Before any detailed discussion of the detector design and the rescoring algorithms, we give an overview of different rescoring scenarios. The introduction of the intra-semantic level and the inter-semantic level rescoring scenarios is presented separately in the next section. In Section 3, we briefly review the theory of MVE and its modification. Rescoring algorithms are discussed in Section 4. Experiments and results in regard to various combinations of training and rescoring methods are presented in Section 5. Finally, we provide the conclusion and the future work in Section 6.

2. RESCORING SCENARIOS

The function of a rescoring system is to improve the task performance in terms of the ultimate design objective. With various situations of different speech recognition tasks, the detector design and the information fusion algorithm, need to be tuned in order to optimize the performance of the entire rescoring framework. Briefly speaking, there are two prevalent ASR rescoring scenarios in continuous speech recognition tasks.

2.1. Intra-Semantic Level Rescoring

The first rescoring scenario is the *intra-semantic level rescoring*. In this case, the accuracy of the decoding decisions determines the system performance directly. For example, the decoders are constructed by phone models and the system performance metric is the phone recognition accuracy. If we built detectors on the same level with the decoder, the scores or decisions generated by detectors can be directly fused into the results of the decoder to affect the system performance.

2.2. Inter-Semantic Level Rescoring

The second situation is the *inter-semantic level rescoring*. In this case, the system performance is not decided directly by the accuracy of the decoding decisions. We need to organize the decoding results to form the output on the level of the performance metric. For example, the decoders are constructed by phone models and the system performance metric is the word recognition accuracy. Therefore, to optimize the rescoring performance in terms of the system objective, the scores or decisions of detectors have to be manipulated with the decoding results to conduct a cross-level rescoring.

3. DETECTOR TRAINING METHODS

3.1. Original Definition for MVE Training

The MVE method can be viewed as a special version of the MCE method [12] for detection and verification problems. Analogous to the MCE criterion, the essence of the MVE training [9] is to directly minimize the total detection errors. In detection problems, there are two different kinds of errors: type I error (miss) 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 and g_a are the normalized log likelihood of the target model and anti-model for the ith class, respectively. Θ_t and Θ_a are 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 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

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

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

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.

$$\begin{aligned} l_{total}^i(\mathbf{O}_k|\Theta^i) &= PE_I l_I^i(\mathbf{O}_k|\Theta^i) \\ &+ PE_{II} \sum_{j=1, j\neq i}^{M} l_{II}^j(\mathbf{O}_k|\Theta^j) \end{aligned} \tag{5}$$

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 [12] w.r.t. all parameters.

3.2. MVE Modifications for Continuous Speech Recognition

In most of ASR tasks, the MVE training routine is applied to the phone level using isolated speech tokens which are usually not segmented as consistent and accurate as expected. Further more, the decoded candidates are generated in a fashion of continuous recognition that may cause mismatch between the detector construction and rescoring conditions. Therefore, in this section, we propose two modifications of the MVE method that are more suitable in the context of continuous ASR but still inherit the merits from the original MVE criterion.

3.2.1. Substring MVE Training

The first modification of the MVE training is named *substring MVE training (S-MVE)*. This method concatenates the target and anti-target models of contiguous phones respectively to form a set of substring detectors which may contain arbitrary number of phones. The MVE training is conducted from the start and shifts all along the utterance. For example, if the utterance is "sil sh iy hh aa s", the first substring training could be applied to the detector model "sil+sh+iy" and the second one could be applied to the model "sh+iy+hh", etc. This method inherits the discriminative ability of the MVE criterion and avoid setting the phone boundaries inside the substring explicitly. The other advantage of this modification is that it exploits some context dependency.

3.2.2. Relaxed-Boundary MVE Training

Though the substring MVE method could alleviate the effects of the unreliable phone boundaries, the start and end time of each substring are still subject to errors. Thus, we develop the second modification of the MVE criterion, the *relaxed-boundary MVE training (RB-MVE)*. This is a more advanced modification of the original MVE criterion based on the S-MVE method. The essence of the RB-MVE method is that the phone boundaries are re-defined by the detector models. We form the utterance target model and anti-model as the concatenation of all phone models in the utterance. For each state j in the sequence of the phone models, the forward and backward likelihood ratio vector α_t^j and β_t^j can be computed for each frame t. Assume there are N states in each phone model, we then can determine the "best" segment $[t_s, t_e]$ for each phone in terms of the highest forward likelihood ratio represented by $\alpha_{t_e}^N - \alpha_{t_s}^1$. The S-MVE method is then carried out based on the adjusted segments. It is a data-driven procedure to set up better phone boundaries based on the best detector models we have. The training and boundary determination can be repeated iteratively until satisfaction.

4. RESCORING ALGORITHMS

In this section, we investigate two types of rescoring algorithms: the score-fusion algorithm and the decision-fusion algorithm. Three algorithms are proposed respectively for each type.

4.1. Score-Fusion Algorithms

Score fusion is a technique that combines the detectors scores and the decoder scores. The decoding candidates are re-ranked based on the new scores. In this section, we review three score-fusion methods proposed in [4]. Suppose there are M corresponding detectors that each of them consists of a target model and an anti-model. For a speech 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 $S_{new}^{(i)}$.

4.1.1. Naive-Adding

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

$$S_{new}^{(i)} = S_{decode}^{(i)} - S_{anti}^{(i)} + ratio^{(i)} \quad (6)$$

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.

4.1.2. Competitive-Rescoring

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

$$S_c^{(i)} = ratio^{(i)} - \log\{\frac{1}{M-1}\sum_{j\neq i}^{M}\exp(\eta \cdot ratio^{(j)})\}^{1/\eta} \quad (7)$$

and

$$S_{new}^{(i)} = S_{decode}^{(i)} - S_{anti}^{(i)} + S_c^{(i)} \quad (8)$$

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 [12] but using the corresponding detectors' likelihood ratio and there is a sign difference.

4.1.3. Remodeled Posterior Probability

The third method is called *Remodeled Posterior Probability (RPP)*. Borrowing from the idea of the recognition phone graph, we formed a pseudo-graph for each phoneme segment using detector arrays. 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,

$$S_{new}^{(i)} = \frac{\exp(S_{decode}^{(i)} - S_{anti}^{(i)}) + \exp(ratio^{(i)})}{\exp(S_{decode}^{(i)} - S_{anti}^{(i)}) + \sum_{j=1}^{M}\exp(ratio^{(j)})} \quad (9)$$

4.2. Decision-Fusion Algorithms

In many rescoring tasks, the detector design is on the different semantic level compared to the recognized candidates thus the inter-semantic level rescoring is required. The score-fusion mechanisms such as the RPP method may only gain incremental impact in the inter-semantic level rescoring because they do not affect the rescoring decisions directly. One alternative rescoring method is to fuse the independent decisions from both the decoder and the detectors to prune the recognized candidates. In this paper, three decision-fusion methods are proposed and compared under the cross-semantic level rescoring scenario in which the phone-level detectors are employed to improve the word accuracy.

To apply the phone-level detectors upon the word graph, the decision-fusion methods prune the candidates in word graphs based on the reliability of the phone sequence in each word. In other words, each phone in every decoded word is examined by the corresponding detectors. We define a "miss" error in this situation if a recognized phone belongs to the ith class but the likelihood ratio of the corresponding ith detector is less than the threshold. Similarly, a "false alarm" error occurs when a recognized phone belongs to the ith class but the likelihood ratio of any detectors other than the ith detector is larger than the threshold.

4.2.1. Strict-Pruning

The first method, the *strict-pruning (SP)* prunes the whole word if any phone in the word is detected as an "miss" error. This method maps the phone errors and the word errors directly in a strict one-by-one manner.

4.2.2. Relaxed-Pruning I

The second method prunes the word only if at least two or over half of the phones are detected as "miss" errors. This method relaxes some constraints compared to the first method (SP) but still concentrates on the "miss" errors. We name it *Relaxed-Pruning I (RP-I)*.

4.2.3. Relaxed-Pruning II

The third method is similar to the second one except an additional provision in which the pruning shall not occur unless there exist "false alarm" errors at the same time. We call this method *Relaxed-Pruning II (RP-II)*.

5. EXPERIMENTAL RESULTS

The experiments are 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 41 CI phones that are folded from the 48 monophone set defined in [13]. The phones "vcl cl epi" are folded into "sil". The phones "ix el em en" are folded into "ih l m n" respectively and there is no phone labeled as"dx". Each phone is modeled by a 3-state HMM. The decoder uses 32 mixtures for each state in the intra-semantic level rescoring and 70 mixtures in the inter-semantic level rescoring. The model parameters are trained by embedded Baum-Welch algorithm [14] using 39 dimensional feature vectors with 12MFCC,12Δ, 12Δ^2 and 3 log energy values. The recognition candidates are organized using phone/word graphs rather than N-best lists because phone/word graphs represent more information in a much compact topology than N-best lists. The phone/word graphs are generated using HVite in the HTK toolbox (http://htk.eng.cam.ac.uk/) in the way that the pruning criterion is set that only 3 recognition candidates can survive simultaneously. In one graph, each node represents a time instance and each arc represents a phone/word.

Three taxonomical phonetic category detectors are defined and trained first by the Baum-Welch algorithm [14] then adjusted by the variations of the MVE method. These categories include 6 classes [14], 14 classes [15], and 41 classes phonemes respectively. Table 1 and 2 show the mapping rules from the 41-class phone set to the 6-class and 14-class set, respectively. The target models and anti-models in detectors are constructed using 3-state HMM with 32 Gaussian mixtures in each state. In our experiments, we employ these three detectors separately to conduct the cross-category rescoring for the 41-class phone/word graphs in each scenario.

Based on the decoder and the detectors described above, we conduct extensive investigations on rescoring performance under different detector building and information fusion scenarios. We examine two prevalent rescoring situations – the intra-semantic level rescoring and the inter-semantic level rescoring. The experiments of the intra-semantic level rescoring are organized using the phone-graph rescoring to enhance the phone recognition accuracy. Comparative results are organized to show the performance difference between various combinations of different detector training methods and score-fusion algorithms. On the other hand, the inter-semantic level rescoring experiments are presented using the phone-level information integration on word-graphs in order to boost the word recognition accuracy. We concentrate on the result comparison between the decision-fusion algorithms in this case.

Table 1. Mapping rule from the 41-class to the 6-class category.

6-class	monophones
fricatives	ch dh f jh s sh th v z zh
vowels	aa ae ah ao aw ax ay eh er ey ih iy ow oy uh uw
nasals	m n ng
stops	b d g k p t
others	hh l r w y
silence	sil

Table 2. Mapping rule from the 41-class to the 14-class category.

14-class	abbreviation	monophones
front vowels	fv	ae eh ey ih iy
mid vowels	mv	ah ax er
back vowels	bv	aa ao ow uh uw
voiced fricatives	vf	dh v z
unvoiced fricatives	uf	f th s sh zh
affricatives	aff	ch jh
voiced consonants	vc	b d g
unvoiced consonants	uc	k p t
nasals	na	m n ng
diphthongs	di	aw ay oy
liquids	li	el l r
glides	gli	w y
whispers	wh	hh
silence	sil	sil

5.1. Intra-Semantic Level Rescoring Using Phone Graphs

The system performance reaches its upper bound when selecting the candidate from the phone graph which best matches the reference phone transcription. 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. In this section, we only focus on the score-fusion methods. Table 3 shows phone recognition accuracy of the baseline decoder and the upper bound of the phone graph using 0-gram and bigram, respectively.

Table 3. Baseline phone accuracy and upper bounds.

Acc(%)	0-gram	bigram
Baseline	56.78	63.93
Upper bound	63.27	70.75

We first compare the results between the different rescoring algorithms using the conventional MVE training. Then, three variations of MVE training methods are compared using the best rescoring method.

Table 4. Intra-semantic level rescoring performance for different rescoring algorithms (The first row is the rescored accuracy and the second row is the relative improvement).

	NA		CR		RPP	
	0gram	bigram	0gram	bigram	0gram	bigram
6	57.04	63.93	57.29	63.94	58.04	64.05
	4.01	**0.0**	**7.86**	**0.15**	**19.41**	**1.76**
14	57.08	63.93	57.40	63.96	58.01	64.20
	4.62	**0.0**	**9.55**	**0.44**	**18.95**	**3.96**
41	57.62	63.95	57.94	63.99	58.41	64.35
	12.94	**0.29**	**17.87**	**0.88**	**25.12**	**6.16**

Table 5. Intra-semantic level rescoring performance for different detection training methods(The first row is the rescored accuracy and the second row is the relative improvement).

	MVE		S-MVE		RB-MVE	
	0gram	bigram	0gram	bigram	0gram	bigram
6	58.04	64.05	58.10	64.20	58.98	64.25
	19.41	**1.76**	**20.34**	**3.96**	**33.90**	**4.69**
14	58.01	64.20	58.10	64.35	59.01	64.60
	18.95	**3.96**	**20.34**	**6.16**	**34.36**	**9.82**
41	58.41	64.35	58.90	64.40	59.21	64.78
	25.12	**6.16**	**32.67**	**6.89**	**37.44**	**12.46**

5.1.1. Rescoring Algorithms Comparison

Table 4 displays the performance of all three rescoring approaches by using all three taxonomical phonetic detectors upon phone graphs for cross-category rescoring. In Table 4, the first row of results are the rescored accuracy and the second row contains the relative improvement (%). For all three detectors, with the detectors trained using the conventional MVE method, we tried three rescoring algorithms: the Naive-adding (NA) method, the Competitive-Rescoring (CR) method and the Remodeled Posterior-Probability (RPP) method. In addition, the rescoring effect under two kinds of different language models, 0-gram and bigram, are respectively investigated in the experiments.

Since the phone graphs are generated over the 41-class phone set, we map each phone back to the 6-class and 14-class phone set and compute detection scores when conducting cross-category rescoring. Based on the experiment results, first, we can see that the Naive-adding (NA) method has the least performance boosting and the Remodeled Posterior-Probability (RPP) method obtains the most gain. It is not surprising since NA is the most naive approach among those three while the RPP method tries to find a candidate with maximum value of a remodeled posterior probability, which bears relationship to Bayes risk. Second, the 41-class detector displayed the highest performance in cross-category rescoring. Third, rescoring techniques showed much higher performance improvement when 0-gram is used. The reason of this observation might be that the use of better language model eliminates some errors due to inaccurate acoustic modeling.

5.1.2. Detector Training Methods Comparison

Upon the phone graph, the rescoring results of using all three taxonomical phonetic detectors with different detector training strategies are presented in Table 5. In Table 5, the first row of results are the rescored accuracy and the second row contains the relative improvement. For all three detectors, we tried three MVE training methods: the original one, the substring MVE (S-MVE) and the relaxed-boundary MVE (RB-MVE). The rescoring algorithm is the RPP method. In addition, the rescoring effect under two kinds of different language models, zero-gram and bigram, are respectively investigated in the experiments.

From Table 5 we can also make some conclusive observations as in the last section. First, the RB-MVE method and S-MVE method outperform the original MVE method no matter what kind of detector is employed. Second, the 41-class detector displayed the highest performance as expected. Finally, as we observed before, the improvement of rescoring using 0-gram graphs is higher than that of bigram phone graphs.

5.2. Inter-Semantic Level Rescoring Using Word Graphs

The inter-semantic level rescoring is conducted by using the phone-level detectors to rescore phones inside each recognized word candidate to improve the word recognition accuracy. In this section, we study the rescoring performance for both the score-fusion and decision-fusion methods. In the score-fusion rescoring part, as the RB-MVE method and RPP method outperformed their competitors in our previous research, we employ them as the detector training and score combination approach respectively for cross-category rescoring using three taxonomical phonetic detectors. The focus of this section is the decision-fusion rescoring, in which we compare three decision-fusion methods in terms of the final word accuracy using the best configuration obtained from the previous experiments. Table 6 shows phone recognition accuracy of the baseline decoder and the upper bound of the phone graph using bigram.

Table 6. Baseline word accuracy and upper bounds with bigram

Acc(%)	bigram
Baseline	50.28
Upper bound	65.46

5.2.1. Score-Fusion Rescoring

We use a similar rescoring method as we did for the intra-semantic level rescoring. The detectors are trained using the RB-MVE criterion. The RPP method was applied to calculate new scores for each phone in every word in the word graphs. The final rescored word score is the summation of all rescored phone scores in the word. As we did in phone graph rescoring, we mapped each phone back to the 6-class and 14-class phone set and computed detection scores when conducting cross-category rescoring. Table 7 shows the performance of the inter-semantic level rescoring using the best score-fusion method selected from the intra-semantic level rescoring experiments. Still, among all cross-category rescoring experiments, the 41-class phonetic detectors displays the highest improvement of the word accuracy.

5.2.2. Decision-Fusion Rescoring

Table 8 displays the experimental results of using three taxonomical phonetic detectors with all three decision-fusion methods. As we did in the intra-semantic level rescoring, we mapped each phone in the 41-class category back to the 6-class and 14-class phone set when making cross-category decisions. We can see that for all types

Table 7. Inter-semantic level rescoring performance using the best score-fusion method (RPP) among our experiments

Detector		Rescored Acc(%)
6-class	Rescored	50.82
	Relative	**3.56**
14-class	Rescored	50.70
	Relative	**2.77**
41-class	Rescored	51.25
	Relative	**6.39**

Table 8. Inter-semantic level rescoring performance using decision-fusion methods.

phone class	Acc(%)	SP	RP-I	RP-II
6-class	Rescored	50.20	50.98	51.27
	Relative	**-0.53**	**4.61**	**6.52**
14-class	Rescored	50.01	50.87	51.46
	Relative	**-1.78**	**3.89**	**7.77**
41-class	Rescored	50.24	51.02	52.39
	Relative	**-0.26**	**4.87**	**13.90**

of detectors, the Strict-Pruning method (SP) overprunes and leads to a slight performance drop because of its strict constraint. The Relaxed-Pruning I (RP-I) method shows some positive gains and the Relaxed-pruning II (RP-II) method achieves the best performance in the inter-level experiments. Though there is no substantial improvement as we expected, the decision-fusion methods do show higher performance enhancement than the score-fusion approaches.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present extensive investigation for the rescoring performance on continuous speech recognition tasks. The study is based on a general framework depicted in Fig.1 and two key components of the rescoring system, the attribute detector design and the rescoring algorithms, are examined under the intra-semantic level rescoring and the inter-semantic level rescoring, respectively.

For detector design methods, two variations of the MVE training criterion, the S-MVE method and the RB-MVE method, are introduced for continuous speech recognition scenarios. We find out that the RB-MVE criterion achieves the best result in the performance comparison. We introduce two types of the rescoring algorithms, the score-fusion algorithms and the decision-fusion algorithms. The score-fusion algorithms are tested in the intra-semantic level rescoring, in which the RPP method shows the best performance. The decision-fusion algorithms is examined in the inter-semantic level rescoring and the RP-II method displays the best performance over other decision-fusion methods.

The future work includes developing more efficient training criteria for continuous speech recognition and propose more effective rescoring methods. Further more, the experiments currently are conducted on the TIMIT database, which is artificial and phoneme-balanced. We will eventually conduct experiments on conversational speech such as the switchboard database (http://www.ldc.upenn.edu/). In the future, we will migrate to construct a complete detection-based ASR system.

7. ACKNOWLEDGEMENTS

The authors would like to thank Antonio Moreno-Daniel for his help and useful discussions. This work was supported in part by a grant from Microsoft research and AT&T.

8. REFERENCES

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

[2] 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.

[3] C. Ma and C. H. Lee, "A study on detection based automatic speech recognition," in *Interspeech-2006*, Pittsburgh, PA, Sep. 2006.

[4] Q. Fu and B. H. Juang, "Investigation on rescoring using minimum verification error (mve) detectors," Pittsburgh, PA, Sep. 2006.

[5] S. M. Siniscalchi, J. Li, and C. H. Lee, "A study on lattice rescoring with knowledge scores for automatice speech recognition," in *ICSLP-06*, Pittsburgh, Pennsivania, Sep. 2006.

[6] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. Siniscalchi, Y. Tsao, and Y. Wang, "Detection-based asr in the automatic speech attribute transcription project," in *Interspeech-2007*, Antwerp, Belgium, Aug. 2007.

[7] 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.

[8] J. Hou, L. R. Rabiner, and S. Dusan, "Automatic speech attribute transcription (asat) - the front end processor," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

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

[10] A. E. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker verification using minimum verification error training," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 105–108.

[11] J. Li and C.-H. Lee, "On designing and evaluating speech event detectors," in *Interspeech-05*, Lisbon, Portugal, Sep. 2005.

[12] 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.

[13] K. Lee and H. 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.

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

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