A New Framework for System Combination Based on Integrated Hypothesis Space

I-Fan Chen and Lin-Shan Lee

Graduate Institute of Communication Engineering
National Taiwan University, Taipei, Taiwan, Republic of China
ifanchen@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

In this paper, a new concept of integrated hypothesis space for large vocabulary continuous speech recognition (LVCSR) system combination is proposed. Unlike the conventional systems combination approaches such as ROVER, the hypothesis spaces are directly integrated here without string alignment. In this way the timing information for all word hypotheses is well preserved and the new framework is more flexible on rescoring approaches used. Four rescoring criteria on the integrated hypothesis space were further explored and experiments on Chinese broadcast news corpus indicated improved performance.

Index Terms: system combination, word graph, integrated hypothesis space

1. Introduction

Substantial efforts have been made in various areas towards the goal of improving the performance of large vocabulary continuous speech recognition (LVCSR) technologies. Two important areas towards this goal, among many others, are rescoring over the word graphs as well as combination of multiple systems.

In the first area of rescoring over the word graphs, a graph of limited number of word hypothesis is generated for each input utterance, referred to as word graph or hypothesis space, and thus more complicated acoustic/linguistic models or search algorithms can be applied with low computational requirements [1]. The language model rescoring approach with utterance level MAP criterion is a typical example in this area. Alternatively, in the word level MAP approach the word graphs are first reduced to confusion networks, and then the words with the highest posterior probabilities in each segment are selected as the recognized word sequence [2]. Minimum Bayes-Risk (MBR) rescoring, on the other hand, used the expected Levenshtein distance as the object cost function [3]. A time frame error cost function was also proposed to replace the Levenshtein distance in MBR rescoring criterion [4]. Also, an optimal Bayes classification (OBC) was proposed recently, in which a smoothed Bayesian factor is used as the risk function in the classification process, as versus the Levenshtein distance in MBR search [5].

In a similar but different area, a very useful approach to improve the recognition performance is to combine the outputs of several different systems to produce a more reliable output. ROVER [6] is the most widely used technique in this area. The output of each component system here can be 1-Best word sequence [6], N-Best word sequence [7], or confusion networks [2][8]. With multiple string alignment, a sequence of confusion word slots is first constructed. A voting process is then performed on each word slot based on word frequencies, confidence scores and so on to produce the best word sequence.

It is certainly highly desirable to integrate the above two areas of approaches together. However, the multiple system combination is usually performed on the output word sequences or confusion networks, while the rescoring approaches were designed for word graphs. Furthermore, the multiple system combination very often ignores the endpoints of the word hypothesis, in which some valuable information may be lost during decoding, and the multiple string alignment may thus inevitably distort the original hypothesis spaces, and introduce errors during voting. This is why in this paper a new framework for system combination is proposed, in which the hypothesis spaces of different systems can be efficiently integrated and rescoring processes can be effectively performed. Reasonable performance improvements have been observed in preliminary experiments.

This paper is organized as follows. Section 2 presents the proposed approach. Section 3 describes the speech corpus used in the experiments and the baseline systems, and Section 4 gives the experimental results. Section 5 finally makes the conclusion.

2. Proposed Approach

Conventionally the system combination is usually performed on N-best lists or confusion networks. But certainly this can be accomplished one stage earlier to directly integrate the hypothesis space, on which the various rescoring process can be directly applied. This is the way the word graph rescoring and system combination approaches can be integrated as proposed here, as illustrated in figure 1. As will be clear below, no string alignment or confusion network construction preprocess is needed in this framework. Thus the implementation is easy and efficient. This new frame work includes two steps: the hypothesis space integration and rescoring over the integrated hypothesis space.

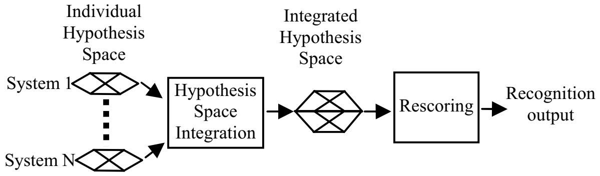
Individual Hypothesis Space
Integrated Hypothesis Space
System 1
System N
Hypothesis Space Integration
Rescoring
Recognition output

Figure 1. The proposed approach of system combination with integrated hypothesis space.

2.1. Hypothesis Space Integration

Suppose there are N systems. For an input utterance, $W_1, W_2, \cdots, W_N$ denote the hypothesis spaces or word graphs produced by systems 1 to N. First consider two word graphs produced by systems 1 and 2, W_1 and W_2, and let q_1, q_2 represent respectively a word arc in W_1 and W_2, which can be expressed as $q_1 = [w_i; t_{start}, t_{end}]$ for word w_i from time t_{start} to t_{end} produced by system 1 and $q_2 = [w_j; \tau_{start}, \tau_{end}]$ for word w_j from time τ_{start} to τ_{end} produced by system 2. So we have $W_1 = \{q_1\}$ and $W_2 = \{q_2\}$, or W_1 and W_2 are the collections of all q_1 and all q_2 respectively. Also each word arc has a certain score denoted as score(q). Here we define the condition for two word arcs from different systems to be equal as:

$$q_1 = q_2 \text{ iff } w_i = w_j, t_{start} = \tau_{start}, t_{end} = \tau_{end}. \tag{1}$$

If two arcs q_1 and q_2 are equal, we can merge them together into $q=q_1+q_2$ with the scores combined.

$$score(q = q_1 + q_2) = combine(score(q_1), score(q_2)) \text{ if } q_1 = q_2. \tag{2}$$

Now we can define the composition of two word graphs as

$$W_1 + W_2 \equiv \{q = q_1 + q_2 \mid q_1 = q_2\} \cup \{q_1 \mid q_1 \notin W_2\} \cup \{q_2 \mid q_2 \notin W_1\}, \tag{3}$$

or W_1+W_2 includes all word arcs in either W_1 or W_2, with equal ones merged. From eq(3), the hypothesis space integration for all systems 1 to N can be expressed as

$$W = W_1 + W_2 + \cdots + W_N = \sum_{i=1}^{N} W_i, \tag{4}$$

where the integrated hypothesis space W is still a word graph, but including all word arcs of the word graph from all component systems with equal ones merged.

2.2. Rescoring on the Integrated Hypothesis Space

In equations (3)(4), the hypothesis space integration is accomplished without any alignment or any loss of timing information for word arcs. Also, the combine function for scores in equation (2) can be flexibly defined, therefore any reasonable score combination operations can be applied. With different acoustic/language models and different scales of the different systems, rescoring over the integrated hypothesis space across different systems is challenging. Here we propose four possible approaches for rescoring across different systems as given below. We will also show that the concept of discriminative decoding can also be applied here.

2.2.1. Consensus Score (CONS)

In this approach the score for a word arc q in an original word graph, score(q) as defined above, is simply the posterior probability of the word arc q, $P_i(q)$, evaluated using the forward-backward algorithm within the original word graph for the original system i [9][10], and the score combination in equation (2) is simply a summation,

$$score(q_1 + q_2) = score(q_1) + score(q_2). \tag{5}$$

Now, in the integrated hypothesis space, the score of each word arc q, referred to as the associated posterior probability regardless of whether it was merged or not:

$$score(q) = score_{CONS}(q) = P(q) \equiv P_i(q) \tag{6}$$

if q was generated by the system i alone, and

$$score(q) = score_{CONS}(q) = P(q) \equiv \sum_i P_i(q) \tag{7}$$

if q has been merged from several word arcs q in several component word graphs, where the summation is over those merged word arcs and equation (7) is directly from equation (5).

The decoding procedure is then as usual:

$$w^* = (q^1 \cdot q^2 \cdots q^M) = \underset{w \in W,\ q^k \in w}{\arg\max} \prod_{k=1}^{M} score(q^k), \tag{8}$$

where score(q) is as define in equations (6)(7) and w^* is the word sequence output. With this consensus score, the output word sequence w^* tends to include most probable word arcs and be the most probable path with consensus of all component systems.

2.2.2. Expected Phone Accuracy Score (EPA)

The consensus score mentioned above focuses on the word level optimization; but sometimes focusing on the phone level may be helpful too, as evidenced by the well known MPE training [11]. We therefore borrowed the concept in MPE training and define the expected phone accuracy for rescoring purposes as follows. Given a hypothesis phone p with its start and end time, we consider all possible paths w in the word graph W as the reference sequence, and p' is a phone on a reference sequence w which overlaps in time with p. If the proportion of the length of p' which is overlapped with p is e(p, p'), then

$$A(p) = \sum_{w \in W} P(w \mid O) \max_{p' \in w} \begin{cases} -1 + 2e(p, p') & \text{if p' and p are the same phone} \\ -1 + e(p, p') & \text{if p' and p are different phones} \end{cases} \tag{9}$$

For a word arc q including phones $\{p_1, p_2, \ldots, p_k\}$, the phone accuracy for the word arc q is then

$$A(q) = \sum_{\substack{i=1 \\ p_i \in q}}^{K} A(p_i) \tag{10}$$

The expected phone accuracy (EPA) score of the word arc q is then

$$score(q) = score_{\text{EPA}}(q) = E[A(q)] = A(q) \cdot P(q), \tag{11}$$

where P(q) is the merged posterior probability of word arc q in equations (6) and (7). The decoding procedure will be as usual as in equation (8), except score(q) in equation (11) is used. With this approach, the search is to find the best path with highest phone accuracy.

2.2.3. Combination of consensus score and expected phone accuracy score (CONS+EPA)

Because the consensus score optimizes on the word level while the expected phone accuracy score optimizes on the phone level, it is thus reasonable to combine them together,

$$score(q) = score_{CONS+EPA}(q) = score_{CONS}(q)[score_{EPA}(q)]^{\beta} \tag{12}$$

where β is a weighting parameter. The same decoding procedure of equation (8) equally applies.

2.2.4. Minimum Time Frame Error (TFE)

A nice property of the framework proposed here is that the integrated hypothesis space is still a word graph. Thus all relevant approaches designed for word graphs can be applied. Minimum time frame error decoding is a modified version of minimum Bayes risk decoding proved very useful [4], in which the traditional Levenshtein distance loss function is replaced by a frame level loss function. We can thus have

$$score(q) = score_{TFE}(q) = \frac{(t_{end} - t_{start} + 1) - \sum\limits_{q'=[w_i;t'_{start},t'_{end}]} overlap(q,q') \cdot P(q')}{1 + \alpha \cdot (t_{end} - t_{start})}, \tag{13}$$

where $q = [w_i;t_{start},t_{end}]$, α is a normalization parameter, q' is any other word arc in the word graph which is also for the word w_i, *overlap*(q,q') is the number of frames overlapped by the word arcs q and q', and P(q') is the posterior probability of q' which can be calculated according to equations (6) and (7).

Since the time frame error score here is a risk measurement, so the decoding procedure is slightly different, i.e., to find the path giving the minimum score,

$$w^* = (q^1 \cdot q^2 \cdots q^M) = \mathop{\arg\min}_{w \in W,\, q^k \in w} \sum_{k=1}^{M} score_{TFE}(q^k) \tag{14}$$

3. Experimental Setup

Two large vocabulary Mandarin speech recognition systems were tested here as system 1 and 2. The primary difference between them was in the acoustic feature parameters. System 1 used the conventional 39-dimensional MFCC feature vectors, which consisted of 12 MFCC and log energy, and their first and second derivatives. Utterance-based cepstral mean subtraction (CMS) was applied to all the training and testing materials. System 2 used acoustic features directly derived from the Mel Scale Filter Bank. We applied Heteroscedastic Linear Discriminant Analysis (HLDA) on the Mel Scale Filter Bank outputs to construct 39-dimensional feature vectors as well. Maximum Likelihood Linear Transform (MLLT) and Cepstral Normalization (CN) were then applied on these feature vectors.

The speech corpus for training and testing is from the Mandarin Broadcast News corpus (MATBN) collected in Taiwan [12]. Roughly 25 hours of gender-balanced data for the field reporters collected Nov 2001 to Dec 2002 were used for training, while another set of 1.5 hour data of field reporters collected within 2003 for testing. The acoustic models were trained by ML criterion first and then followed by MPE training for both systems.

The lexicon size of this task for both systems 1 and 2 is 72K words. The background language model is trained on the Chinese News Agency (CNA) 2001 and 2002 text corpus, including roughly 170 million characters. Trigram models were used. Meanwhile the reference transcriptions of the 25-hour training utterances, consisting of about 500K characters, were regarded as in-domain text corpus, and used to train an in-domain language model, to be interpolated with the background language model to be used as the final language model for the experiments.

4. Experimental Result

Table 1 shows the performance in terms of syllable error rates (SER), character error rates (CER) and word error rates (WER) for the proposed integrated hypothesis space framework as compared to those of the conventional N-Best ROVER with N varying from 1 to 20 using NIST SCTK1.3 [13]. Relevant parameters were directly tuned on the test corpus to find the upper bounds of the ROVER performance.

Comparing ROVER and the proposed approach of integrated hypothesis space, we found for the WER performance over the whole test set, the proposed approach and ROVER are quite similar, able to reduce roughly 1.7% absolute word error rate. As for CER, the results for N-Best ROVER was reduced as N goes higher, yet the lowest CER is still significantly higher than the proposed integrated hypothesis space approaches. A possible reason is that though both frameworks merged the systems on the word level, in the integrated hypothesis space approach no word sequence alignment is performed. This eliminated the possible distortion introduced by alignment and kept the output character sequence continuing. Next considering the four rescoring methods proposed here for the integrated hypothesis space. Comparing (1)CONS, (2)EPA, and (3)CONS+EPA first, we found that CONS is focused on words, thus offered the lowest word error rate; EPA is focused on phones, thus offered the lowest syllable error rate. When combining these two scores together, CONS+EPA achieved the lowest character error rate. If we further applied the (4)Minimum Time Frame Error (TFE) decoding on the integrated hypothesis space, we were able to obtain further improvements. The overall CER improvement is absolute 1.53% (19.27% vs. 20.80%) from baseline, and absolute 0.85% (19.27% vs. 20.12%) from the 20-Best ROVER upper bound.

Tested system		SER	CER	WER
Baseline	MFCC	15.89	22.19	29.93
	HLDA	14.43	20.80	28.53
ROVER upper bound	1-Best	14.90	20.39	26.92
	10-Best	14.64	20.21	26.76
	20-Best	14.49	20.12	26.79
Integrated Hypothesis Space	(1)CONS	13.67	19.62	26.88
	(2)EPA	13.41	19.73	27.70
	(3)CONS+EPA	13.55	19.54	26.97
	(4)TFE	13.35	19.27	26.71

Table 1. Syllable, character, and word error rate of different system integrating approaches

Other than the recognition error rates over the whole test set, we also wish to analyze how the combination frameworks performed on the utterance level. We therefore compared WER and CER of the utterances of the two baseline systems. Out of the all 292 testing utterances, the MFCC system gave lower word error rates than the HLDA system in 129 utterances; while the HLDA system outperformed the MFCC system in 84 utterances; and there are 79 utterances for which both systems have same word error rate. For character error rates, the MFCC system was better on 84 utterances, the HLDA system outperformed on 158 utterances; and both systems have the same CER for 50 utterances.

We can then divide the performance of the system combination results into 5 categories:

1. Better – The performance is better than both baselines.
2. As Best – The performance is equal to the best baseline.
3. Between – The performance is between the two baselines.
4. As Worst – The performance is equal to the worst baseline.
5. Worse – The performance is worse than both baselines.

Table 2 lists the utterance WER comparison in terms of the numbers in the above five categories. We can find that though the ROVER approaches always have the largest numbers in the "As Best" category out of the five. They also always have significant numbers in the categories of "Between", "As worst" and "Worse". The alignment and voting scheme turned out to have less stable performance here. On the other hand, we found that the proposed CONS has very large number in the "As Best" categories and almost zero in the "Between" and "Worse" categories. So it did not actually generate better recognized word segments, but selected the best output from the component systems. The TFE approach, on the other hand, has the largest number in the "Better" category. In other words, both CONS and TFE are more stable in performance considered here. This is not necessarily true for EPA, probably because it is focused on phone accuracy instead of word errors.

(WER) MFCC better:129, HLDA better:84, Equal:79, Total:292						
	ROVER upper bound			Proposed Approaches		
	1-Best	10-Best	20-Best	CONS	EPA	TFE
Better	62	68	71	7	53	126
As Best	104	94	94	209	91	57
Between	53	58	51	2	44	47
As worst	48	34	41	74	40	37
Worse	25	38	35	0	64	25

Table 2. Utterance WER comparison in terms of the five categories. TFE offered the largest in "Better" category, which CONS offer the largest number in "As Best" category.

(CER) MFCC better:84, HLDA better:158, Equal:50, Total:292						
	ROVER upper bound			Proposed Approaches		
	1-Best	10-Best	20-Best	CONS	EPA	TFE
Better	69	81	80	97	102	145
As Best	82	70	75	81	78	54
Between	87	80	80	65	61	63
As worst	31	28	32	26	24	18
Worse	23	33	25	23	27	12

Table 3. Utterance CER comparison in terms of the five categories.

Similar results for the utterance CER is shown in table 3, from which similar observations can be made. Note that here half of the testing utterances were benefited from the TFE approach, in which the number for "As Worst" and "Worse" categories are also the smallest. This may be the reason TFE achieved the lowest CER.

5. Conclusion and Future work

In this paper, we propose a new approach of integrated hypothesis space for systems combination. Several rescoring methods over this space are also explored. Comparing with ROVER, improved recognition performance was obtained and improved utterance level performance stability was observed.

The proposed approach provides a wide flexibility on possible rescoring methods. The work to try to apply other advanced rescoring concepts onto this framework is still under progress.

6. References

[1] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43-72, Jan. 1997.

[2] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, pp. 373-400, Oct. 2000

[3] V. Goel, and W. Byrne, "Minimum Bayes-Risk Methods in Automatic Speech Recognition," *Pattern Recognition in Speech and Language Processing (CRC Press)*, chapter. 2, pp. 51-80, 2003

[4] F. Wessel, R. Schlüter, and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities," *in Proc. ICASSP*, pp. 33-36, 2001

[5] J.T. Chien, C.H. Huang, K. Shinoda, and S. Furui, "Towards Optimal Bayes Decision for Speech Recognition," *in Proc. ICASSP*, 2006

[6] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 347-352, 1997

[7] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, J. Zheng, "The SRI March 2000 HUB-5 Conversational Speech Transcription System"

[8] G. Evermann, and P. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," *in Proceedings of the NIST and NSA Speech Transcription Workshop*, College Park, MD, 2000

[9] F. Wessel, R. Schlüter, and H. Ney, "Using Posterior Word Probabilities for Improved Speech Recognition," *in Proc. ICASSP*, 2000

[10] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 288-298, March, 2001

[11] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition," PhD Thesis, Cambridge University Engineering Department, 2003

[12] H. M. Wang, B. Chen, J. W. Kuo, and S. S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, June, 2005

[13] The latest version of SCTK is available from `http://www.nist.gov/speech/tools/`

[14] A. Sankar "Bayesian Model Combination (BAYCOM) for Improved Recognition," *in Proc. ICASSP*, 2005

[15] T. J. Hazen, I. L. Hetherington, A. Park, "FST-Based Recognition Techniques for Multi-Lingual and Multi-Domain Spontaneous Speech," *in Proc. EUROSPEECH*, 2001

[16] M. Mohri, F. Pereira, M. Riley, "Weighted Finite-State Transducers in Speech Recognition," *Computer Speech and Language*, 2002