CROSS-SITE AND INTRA-SITE ASR SYSTEM COMBINATION: COMPARISONS ON LATTICE AND 1-BEST METHODS

B. Hoffmeister†, D. Hillard‡, S. Hahn†, Ralf Schlüter†, M. Ostendorf‡, H. Ney†

†Informatik 6, Computer Science Dept., RWTH Aachen University, Aachen, Germany
{hoffmeister,hahn,schlueter,ney}@cs.rwth-aachen.de
‡SSLI, Electrical Engineering Dept., University of Washington,Seattle, WA
{hillard,mo}@ee.washington.edu

ABSTRACT

We evaluate system combination techniques for automatic speech recognition using systems from multiple sites who participated in the TC-STAR 2006 Evaluation. Both lattice and 1-best combination techniques are tested for cross-site and intra-site tasks. For pairwise combinations the lattice based approaches can outperform 1-best ROVER with confidence scores, but 1-best ROVER results are equal (or even better) when combining three or four systems.

Index Terms— speech recognition, system combination

1. INTRODUCTION

Most state-of-the-art automatic speech recognition (ASR) systems today include multiple contrasting systems, which are ultimately combined to produce the final hypothesis. There is consensus that improvements from combination are usually best when systems are sufficiently different, but there is uncertainty about which system combination method performs the best. This paper investigates various system combination techniques and compares the results over configurations including two, three, and four, component systems.

The component ASR systems are the final evaluation systems from the English part of the TC-STAR 2006 Evaluation campaign. Our partners kindly provided us their best lattice sets for both the development and evaluation sets. In addition, from one partner we received four lattice sets used in their internal system combination.

The aim of system combination for ASR is to minimize the expected WER given multiple system outputs. *Bayes* decision rule with a *Levenshtein* cost function $\mathcal{L}$ provides the general framework for a minimum WER decoder:

$$\{w_1^N\}_{\text{opt}} = \underset{w_1^N}{\operatorname{argmin}} \left\{ \sum_{v_1^M} \mathcal{L}(w_1^N, v_1^M) p(v_1^M | x_1^T) \right\} \qquad (1)$$

with a word sequence w_1^N and the posterior probability $p(v_1^M|x_1^T)$ for word sequence v_1^M given the acoustic observation sequence x_1^T. The exponential size of the search and summation space forbids a direct application of this decision rule for LVCSR systems [1]. Word lattices are an efficient way to narrow the search space, but they still represent a huge number of hypotheses and a direct application of Eq. (1) still is prohibitive. The confusion network (CN) and minimum Time Frame Word Error (fWER) decoder are two approaches using different approximations to realize minimum WER decoding on word lattices [2, 3]. For both approaches, relative improvements of up to 5% in WER are reported.

The main contribution of this work is to compare multiple system combination approaches in settings with different numbers of component systems. We examine three types of system combination methods, each in its standard form and in a variant using system priors. The latter is of particular interest for the cross-site task because the provided lattice sets differ strongly in their individual performance. System dependent weights can balance the impact of each system and thus improve the combination performance.

The most widely used system combination approach to date, ROVER [4], is a simple voting mechanism over the top hypothesis from each component system. In this paper we use a variant utilizing confidence scores and system dependent weights [5]. Extensions of CN and minimum fWER decoding can be employed in system combination as well, and have the advantage of combining multiple hypotheses from each component system (rather than just the top hypothesis as in the standard ROVER approach).

Next, Section 2 describes the three system combination techniques that we explore in this work. Section 3 provides experimental setup, sketches the problems we encountered when dealing with lattices from five different sites, and presents the results. Finally, we summarize our conclusions in Section 4.

2. SYSTEM COMBINATION METHODS

2.1. ROVER

ROVER [4] is a two step procedure comprised of alignment and voting, where the alignment depends on the system permutation. Exhaustive experiments have shown that best results are obtained when systems are ordered by increasing WER.

We use a slightly modified version of the original average confidence score voting function, where the confidence scores provided by each system are weighted with additional system dependent weights $\lambda_1, \ldots, \lambda_L$:

$$\text{score}(w,i) = \sum_{l=1}^{L} \lambda_l \, \left[\alpha \, \delta(w, w_{l,i}) + (1-\alpha)\text{conf}_l(w,i)\right], \qquad (2)$$

The δ is the Kronecker-δ, i denotes the position in the alignment, and L is the number of systems. Majority vote and averaged confidence score are smoothly interpolated via α. Basic ROVER is derived by setting $\lambda_1 = \cdots = \lambda_L = 1/L$.

Confidence scores are calculated directly from the lattices using the approach described in [6].

2.2. Confusion Network Combination

A Confusion Network (CN) is a directed graph where all outgoing arcs of a given node have the same target node. CNs consist of

a series of words slots, where each slot contains the hypothesized words for that position in the segment. For this structure, Eq. (1) has a simple solution that selects the maximum posterior word in each slot. In [2], an iterative algorithm is presented that transforms a word lattice into a CN by successive arc alignments.

Confusion network combination (CNC) is a generalized ROVER algorithm that aligns CNs derived from several systems [7]. The result is a new CN. The word posterior probabilities for the ith slot in the combined CN can easily be calculated as the joint probability of the system specific posteriors:

$$p(w|i, x_1^T) = \sum_{l=1}^{L} p(S_l|i, x_1^T) p(w|S_l, i, x_1^T) \quad (3)$$

In this work the system priors $p(S_l|i, x_1^T)$ are approximated by a system dependent constant λ_l.

N-Best ROVER [8] is another common approach to system combination. It is a special case of CNC where the system CNs are constructed from a multiple alignment of each component system N-Best list, rather than from lattices.

2.3. Minimum fWER Combination

An alternative approach to simplify the decision rule is to replace the Levenshtein distance by a computationally cheap cost function C: the fWER [3]. The important property of the fWER is that its calculation does not require an expensive word alignment. Replacing the Levenshtein distance in Eq. (1) by the definition of the fWER gives the Minimum fWER decision rule:

$$\{[w;t]_1^N\}_{\text{opt}} = \underset{[w;t]_1^N}{\operatorname{argmin}} \sum_{n=1}^{N} \frac{\sum_{\hat{t}=t_{n-1}+1}^{t_n} \left[1 - p(w_n|\hat{t}, x_1^T)\right]}{1 + \alpha(t_n - t_{n-1} - 1)} \quad (4)$$

The term $p(\cdot|t, x_1^T)$ is the frame-wise word posterior distribution, which can be efficiently calculated by a modified forward/backward algorithm.

As opposed to CN decoding, where word times and boundaries are only used to align the words (and posterior probabilities then depend solely on the resulting word positions), the fWER decoding approach preserves the lattice structure and thus the output is produced with correct word boundary times. The minimum fWER decoding approach for a single lattice can easily be extended to minimize the WER over multiple lattices, as in [5]. According to Eq. (4) we have to change the calculation of the word posteriors and to define the search space.

From each lattice G_l of each system S_l we derive a sequence of frame-wise word posterior distributions $p(\cdot|S_l, 1, x_1^T), \ldots, p(\cdot|S_l, T, x_1^T)$. In our experiments we use the joint probability over the system dependent posteriors to calculate a multiple system frame-wise word posterior probability:

$$p(w|t, x_1^T) = \sum_{l=1}^{L} p(S_l|t, x_1^T) p(w|S_l, t, x_1^T) \quad (5)$$

Similar to CNC, the system priors $p(S_l|t, x_1^T)$ are approximated by a system dependent constant λ_l.

The search space is simply the union of all lattices $G_1, \ldots, G_L$. The number of hypotheses can be increased by building the time conditioned lattice from the union.

3. EXPERIMENTS

3.1. Corpora

We present results on the EPPS 2006 English corpus. The corpus contains parliamentary speeches from the European Parliament and was collected within the TC-STAR project. All audio files are monaural with 16-bit resolution at a sampling rate of 16kHz.

The 2006 TC-STAR Evaluation campaign took place in February 2006. Besides RWTH Aachen [9] the following project partners participated in the English task: LIMSI [10], IBM [11], UKA [12], and IRST [13]. Afterward all project partners kindly provided us their best performing lattice set. In addition, one site provided four lattice sets that they used for internal system combination.

3.2. Experimental Setup

The original lattice sets from the five sites are pairwise completely different. They use different formats, different segmentations, have different density, and are not normalized, e.g. they still include compound words like "it_has" or different forms of abbreviations like "EU" as a single word or as two words: "E. U.". Unfortunately, we had to exclude one lattice set, because we were not able to prepare it such that we could reproduce the Viterbi decoding result.

For the remaining four sets, the first task was to unify the segmentation. Our basic approach was to concatenate lattices until all sites had a common pause of at least one second. The number of segments for the original lattice sets ranged from 183 to 1607. Our final unified segmentation consists of 77 segments for the development corpus and 66 for evaluation.

The next step was to normalize the lattices. We mapped all filler words, noises and silence to a single "non-word" label. Then we applied a filter on each lattice in order to remove "non-word" clouds, see [5]. This filtering makes the posterior scores calculated from a lattice more reliable. For compound words and abbreviations we split them into the largest possible sequence of chunks, e.g. "EU" to "E. U.". The time frames of an arc were distributed over the chunks according to the number of characters per chunk.

The last step was to reduce the density in order to make the posterior decoding methods more stable and to generate lattice sets of similar density. For the cross-site task the maximum density was given by the least dense lattice set, see Table 1.

Table 1. Densities for the EPPS 2006 English lattices. For the cross-site task all densities were calculated on the unified segmentation.

Lattice Set	CROSS-SITE avg. density				INTRA-SITE avg. density			
	Unpruned		Pruned		Unpruned		Pruned	
	dev	eval	dev	eval	dev	eval	dev	eval
1	24	31	24	31	347	284	89	80
2	412	337	24	22	356	298	90	83
3	333	268	33	29	342	292	88	82
4	37	29	37	29	347	296	90	82

The combination experiments for ROVER used software provided by NIST, CNC experiments used the SRILM toolkit [14], and Minimum fWER experiments used software from RWTH.

Finally, after CNC combination the word hypothesis has no word time boundaries, but NIST tools require reasonable boundaries

for WER scoring. The standard approach is to run a forced alignment on the CNC output, but we did not have access to the systems used for creating the source lattices. Therefore, we developed an algorithm that solved the problem by extracting word time and posteriors from the source lattices. First, we reduce the source lattice to its time-conditioned form. Then, if it contains a path matching the CNC output, we take the boundaries from that path and stop. If it does not contain a matching path, we build the time frame-wise word posterior distributions $p(w|t, x_1^T)$ from the lattice and align the CNC output against it and then use the boundaries from the best alignment as word boundaries.

3.3. Results

We evaluate the system combination approaches from Section 2 for various configurations of the component systems we have available. Table 2 presents WER results for combination among systems from four different institutions (cross-site), as well as for combination of a set of four intra-site lattices.

Single system performance for the cross-site systems is presented in the first major row of numbers. The first major column provides the Viterbi decoding WER for each individual lattice, while the second and third major columns show the CN and fWER decoding WER. While CN and fWER are sometimes better, there is no consistent trend for Viterbi versus consensus decoding.

The second major row contains the six pair-wise system combinations. Individual system weights were tuned on the development set and then applied on the evaluation set for the weighted system combination results. For the case of two component systems, weighted CNC and fWER consistently outperform ROVER combination.

The third major row provides all three-way system combinations for the cross-site systems. Here, ROVER achieves equal or lower WER than both the CNC and fWER combination. Finally, the same trend continues when all four systems are combined. Again, ROVER achieves equal or lower WER.

The lower half of Table 2 shows WER for four within-site lattices. The first major row compares Viterbi, CN, and fWER decoding, where there are again no clear trends. The final row of the table reinforces the conclusions from the cross-site system combination: all three combination methods are close, but ROVER is always best.

The ROVER result for the intra-site combination has the same as the WER published by that site, which validates the preprocessing in our experiments. The best result published to date on the eval set is 6.9% WER, which is ROVER over the best output from the five participating partners. We can also achieve 6.9% with just the four sites available in this work, by applying ROVER across four results: the weighted cross-site ROVER, CNC, fWER, and intra-site ROVER.

4. CONCLUSIONS

We found that when more than two complementary systems are available for system combination, ROVER most consistently achieves the best results. CNC and and fWER combination only out-perform ROVER when just two systems are available for combination. With only two systems, combination approaches that include multiple hypotheses from each system can obtain better results. But, with increasing numbers of systems the result converges to the ROVER result (or can actually be slightly worse).

One benefit of fWER combination that is not present in ROVER or CNC is that the hypothesized output preserves the word context (as well as word times) from the lattice. When the hypothesis space is restricted to the union of the lattices, the final result is a valid path from one of the original lattices. For this work we did not observe a degradation in WER when using this union, compared to a time-conditioned form of the union (which is not restricted to paths in the lattice). The improved fluency in the ASR output that results from the preserved context might benefit downstream tasks such as machine translation.

Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (FP6-506738).

5. REFERENCES

[1] V. Goel and W.J. Byrne, "Minimum bayes-risk automatic speech recognition," *Computer Speech and Language*, vol. 14, pp. 115–136, 2000.

[2] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. European Conf. on Speech Communication and Technology*, 1999, vol. 1, pp. 495 – 498.

[3] F. Wessel, R. Schlüter, and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities," in *Proc. ICASSP*, 2001, vol. 1.

[4] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, 1997.

[5] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *Proc. ICSLP*, 2006.

[6] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288 – 298, 2001.

[7] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, 2000.

[8] A. Stolcke et al., "The SRI March 2000 Hub-5 conversational speech transcription system," in *In Proc. Speech Transcription Workshop*, 2000.

[9] J. Lööf et al., "The 2006 RWTH parliamentary speeches transcription system," in *Proc. ICSLP*, 2006.

[10] L. Lamel et al., "The LIMSI 2006 TC-STAR transcription systems," in *Proc. TC-STAR Workshop*, 2006.

[11] B. Ramabhadran et al., "The IBM 2006 speech transcription system for european parliamentary speeches," in *Proc. ICSLP*, 2006.

[12] S. Stüker et al., "The ISL TC-STAR spring 2006 ASR evaluation systems," in *Proc. TC-STAR Workshop*, 2006.

[13] F. Brugnara et al., "The ITC-irst transcription systems for the TC-STAR-06 evaluation campaign," in *Proc. TC-STAR Workshop*, 2006.

[14] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.

Table 2. Results for the EPPS 2006 English lattices. The upper part contains the results on the cross-site task where lattice sets from four different sites were combined. The lower part contains the results on a separate four lattice set from a single site. Individual lattice results are derived via Viterbi, Confusion Network (CN), and Minimum fWER decoding. Three system combination approaches are applied: ROVER with confidence scores, CN combination (CNC), and fWER combination. For the standard setup no system weights are used, for the weighted setup system dependent weights were tuned on the dev-set. The ROVER Oracle WER is the minimum possible WER when combining the single best hypothesis from each individual system. Individual systems cannot have weighted or oracle combinations, so those cells are blank.

Lattice Set 1	Lattice Set 2	Lattice Set 3	Lattice Set 4	Viterbi/ROVER WER [%]				CN/CNC WER [%]				Minimum fWER WER [%]				Oracle WER[%]	
				Standard		Weighted		Standard		Weighted		Standard		Weighted			
				dev	eval	dev	eval	dev	eval	dev	eval	dev	eval	dev	eval	dev	eval
CROSS-SITE LATTICES																	
Individual Cross-Site Systems																	
X				10.5	9.0			10.4	8.8			10.2	8.8				
	X			11.4	9.0			11.5	9.2			11.4	9.2				
		X		12.8	10.4			12.8	10.3			12.6	10.4				
			X	13.9	11.9			14.0	12.2			13.8	11.7				
Pairwise Cross-Site System Combinations																	
X	X			10.2	9.2	10.2	8.8	9.6	8.0	9.6	8.0	9.4	7.9	9.4	7.9	6.4	5.2
X		X		10.4	9.2	10.2	8.8	12.2	7.9	12.2	7.9	9.8	8.4	9.8	8.4	6.7	5.5
X			X	10.8	9.9	10.2	8.8	12.2	10.5	10.1	8.7	11.0	9.9	10.1	8.7	7.2	6.1
	X	X		11.0	9.0	11.0	9.0	10.3	8.3	10.3	8.3	10.4	8.5	10.4	8.5	7.4	5.4
	X		X	11.4	9.5	11.4	9.5	11.9	9.7	10.9	8.8	11.8	10.3	10.8	9.0	7.5	5.9
		X	X	12.7	10.8	12.7	10.8	12.9	10.6	12.9	10.6	12.2	10.5	11.1	9.4	8.0	6.2
Three-way Cross-Site System Combinations																	
X	X	X		9.1	7.2	9.1	7.2	9.5	7.6	9.5	7.6	9.0	7.6	9.0	7.6	5.3	4.2
X	X		X	9.2	7.5	9.2	7.5	10.0	8.0	9.6	7.7	9.5	8.2	9.2	7.8	5.5	4.6
X		X	X	9.5	7.8	9.5	7.8	10.2	8.3	9.6	7.8	9.7	8.3	9.4	8.0	5.7	4.6
	X	X	X	9.8	7.8	9.8	7.8	10.8	8.3	10.7	8.2	10.1	8.5	9.9	8.1	6.0	4.5
Four-way Cross-Site System Combinations																	
X	X	X	X	8.9	7.3	8.9	7.3	9.6	7.5	9.4	7.4	9.1	7.7	8.9	7.3	4.8	3.9
INTRA-SITE LATTICES																	
Individual Intra-Site Systems																	
X				11.4	9.0			11.6	9.3			11.3	9.0				
	X			11.6	9.4			11.7	9.7			11.5	9.5				
		X		11.8	9.5			11.9	9.7			11.7	9.4				
			X	11.7	9.4			11.8	9.6			11.5	9.3				
Four-way Intra-Site System Combinations																	
X	X	X	X	10.7	8.6	10.7	8.6	11.1	8.8	11.0	8.9	10.7	8.7	10.6	8.7	8.0	6.2