Frame Based System Combination and a Comparison with Weighted ROVER and CNC

Björn Hoffmeister, Tobias Klein, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Aachen, Germany
{hoffmeister,schlueter,ney}@cs.rwth-aachen.de

Abstract

In this paper we present a novel ASR system combination technique able to combine systems producing word graphs of different structure and with different segmentations. The new method is based on the definition of a time frame-wise word error cost function in a minimum *Bayes* risk framework. In contrast to confusion network combination (CNC), it preserves both the word graph structure and the word boundaries.
First experimental results are presented on the European Parliament Plenary Sessions (EPPS) task for European Spanish and British English. The new approach to system combination is compared to both ROVER and CNC. In addition, we also apply data-driven weighting schemes for all system combination approaches addressed in this work. For the experiments presented, a variety of internal systems as well as an additional external system were combined.
Index Terms: speech recognition, system combination, word posteriors.

1. Introduction

System Combination is a promising way to obtain a significant reduction in word error rate (WER). For example, the five English systems participating in the Second TC-STAR ASR Evaluation campaign 2006 gave word error rates ranging from 8.3% to 11.0% WER. System combination of these systems via ROVER lead to a WER of 6.9%.

Usually, system combination gives largest improvements, if the individual systems to be combined lead to similar performance and are complementary w.r.t. the errors they produce. Nevertheless, parallel development of complementary systems with comparable performance can be time consuming. On the other hand, the development cycle of a state-of-the-art ASR system involves subsequent creation of suboptimal systems due to techniques like adaption and discriminative training. Therefore, here we investigate the use of such suboptimal systems by applying system combination methods. Due to the corresponding performance range of the systems to be combined, we also investigate the use of system priors estimated on a development set.

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

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

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 graphs 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 Error (fWER) decoder are two approaches using different approximations to realize minimum WER decoding on word graphs [2, 3]. For both approaches, relative improvements of up to 5% in WER are reported.

Confusion network combination (CNC) is a system combination approach based on the alignment of CNs [4]. Therefore, CNC is based on the same approximations to the word graph structure as CNs, i.e. word boundary information is relaxated during CN construction and then discarded. In contrast to this, the presented minimum fWER combination scheme does not affect the word graph structure. This is achieved by replacing the *Levenshtein* cost function with a frame-wise word error cost function. The resulting algorithm is of low computational complexity.

ROVER is employed to obtain a baseline for the system combination experiments. We use an extended voting rule which incorporates system weights.

The rest of the paper is structured as follows. In the next two sections we give a short review of ROVER and CNC. For ROVER we introduce the used weighting scheme, which is also applied to the other combination schemes discussed. In Sec. 4 we review minimum fWER decoding and extend it for system combination. Experimental results including a comparison of the system combination methods are presented in Sec. 5. The final Sec. 6 gives conclusions and an outlook.

2. ROVER

ROVER [5] is a two step procedure comprised of alignment and voting. The alignment depends on the system permutation. Exhaustive experiments have shown that best results are obtained when systems are ordered by increasing WER.

We modified the voting function by weighting the confidence scores provided by each system with additional system dependent weights $\lambda_1, \ldots, \lambda_L$:

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

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 $\lambda_1 = \cdots = \lambda_L = 1$. Besides the linear weights we tested system dependent exponents, but the linear weights gave better results for all corpora.

3. CNC

A CN is a directed graph with the following property: all outgoing arcs of a given node have the same target node. For this structure, Eq. (1) has a simple solution. In [2] an iterative algorithm is presented that transforms a word graph into a CN by successive arc alignments.

A generalized ROVER algorithm is used to align the CNs derived from several systems [4]. The result is a new CN. The word posterior probabilities for the ith confusion set in the super-CN can easily be calculated as the joint probability of the system specific posteriors:

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

4. Frame Based System Combination

4.1. Minimum fWER Decoding

In Sec. 3 we pointed out how Eq. (1) can be simplified by changing the structure of a word graph. Alternatively, an approach to instead simplify the decision rule Eq. (1) is introduced in [3]. The idea is to replace the Levenshtein distance $\mathcal{L}$ by a computationally cheap cost function C: the time frame word error (fWER). The fWER takes the word boundary times of a word into account and calculates the cost based on the time frames covered:

$$C([w;t]_1^N, [v;\tau]_1^M) = \sum_{n=1}^{N} \frac{\sum_{\hat{t}=t_{n-1}+1}^{t_n} 1 - \delta(w_n, v_{\hat{t}})}{1 + \alpha(t_n - t_{n-1} - 1)} \tag{4}$$

$[w;t]_1^N$ denotes a sequence of words together with their ending times, where $t_0 = 0$ and $t_N = T$, and $v_{\hat{t}}$ is the word in $[v;\tau]_1^M$ which intersects time frame $\hat{t}$.

The denominator in Eq. (4) allows a smooth normalization of the time frame errors. $\alpha = 1$ gives time frame-wise normalization, and $\alpha = 0$ gives word-wise normalization of the error. For all tested corpora the best results were obtained with $\alpha = 0.05$.

In contrast to the Levenshtein distance, the advantage of the fWER is that no sentence alignment is required. That makes the fWER computationally cheap. In [3], a strong relation between fWER and WER is shown empirically, which justifies the usage of the fWER as an approximation of the WER.

Inserting Eq. (4) into Eq. (1) gives the minimum fWER decision rule:

$$\{[w;t]_1^N\}_{\text{opt}} = \underset{[w;t]_1^N}{\operatorname{argmin}} \sum_{n=1}^{N} \frac{\sum_{\hat{t}=t_{n-1}+1}^{t_n} \left[1 - p(w_n|\hat{t}, x_1^T)\right]}{1 + \alpha(t_n - t_{n-1} - 1)} \tag{5}$$

The term $p(\cdot|t, x_1^T)$ is the frame-wise word posterior distribution.

The frame-wise word posteriors are calculated by a modified forward/backward (FB) algorithm. Figure 1 illustrates the algorithm. The rectangles in b) represent the word-wise accumulated FB-scores of the arcs in a). For time frame t the posterior probability $p(\text{“}have''|t, x_1^T)$ is the normalized sum of the FB-scores of all arcs labeled with “have” and intersecting time frame t.

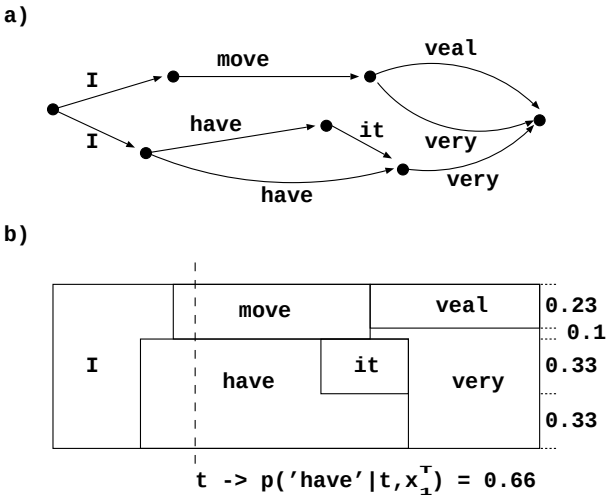

Figure 1: *Illustration of the calculation of the posterior distribution $p(\cdot|t, x_1^T)$ from a word graph. The rectangles in b) represent the word-wise accumulated fwd./bwd.-scores of the arcs in a).*

The evidence space is the set of all hypotheses considered in the decoding step. In the original paper, the set of hypotheses in the word graph is used as evidence space. The graphs produced by our Viterbi decoder are word conditioned. We can enlarge the evidence space by transforming the word graphs into time-conditioned ones. This resulted in a little but insignificant decrease in WER for all corpora.

There is a substantial difference between CN and fWER decoding. In CN decoding word boundaries are only used to align words. Once the CN is built, all time information is lost and the calculation of the word posterior probabilities depends only on the resulting word positions. Time boundaries for the output have to be produced in a post-processing step.

The fWER decoding approach preserves the word graph structure and thus the output is produced with correct word boundary times.

4.2. Minimum fWER over Multiple Word Graphs

The minimum fWER decoding approach for a single word graph can easily be extended to minimize the WER over multiple word graphs. According to Eq. (5) we have to change the calculation of the word posteriors and to redefine the evidence space.

From each word graph G_l of each systems 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:

$$p(w|t, x_1^T) = \sum_{l=1}^{L} p(S_l|t, x_1^T) p(w|S_l, t, x_1^T) \tag{6}$$

The system priors $p(S_l|t, x_1^T)$ are approximated by a system dependent constant λ_l. We also tried a log-linear combination model, but the joint probability model turned out to be superior for all tested corpora.

The new evidence space is simply the time conditioned word graph derived from the union of all word graphs $G_1, \ldots, G_L$.

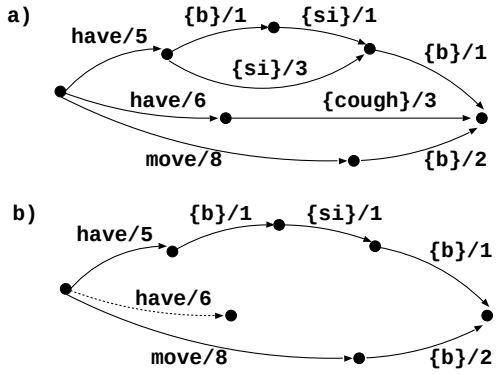

Figure 2: *Illustration of the word graph pre-processing step. Alternative arcs labeled with non-speech events $\{\cdot\}$ are removed.*

It should be noted that the frame based combination works, even if the individual systems use different segmentations to produce their word graphs: the calculation of the frame-wise word posterior distributions is independent of the segmentation, since it is applied on a frame-by-frame basis. Again, the evidence space is obtained by the union of all systems word graphs. The corresponding search in this case still is efficient, since the corresponding decision rule does not use context (the acoustic and language model context is considered on the level of word posterior computation already).

4.3. Word graph pre-processing

The RWTH LVCSR system uses different acoustic models to represent non-speech events like silence, hesitation, articulatory and non-articulatory noise. These models tend to be very similar. As a consequence, all non-speech models are hypothesized in parallel having similar scores, and if they survive the pruning steps they occur as "non-speech event clouds" in the word graphs as illustrated in Figure 2 a). These clouds bias the word posteriors calculated from the word graph. The posteriors of words and non-speech events lying on a path through a "non-speech cloud" are over-estimated.

The basic idea to get rid of the bias is to discard alternative non-speech events. Figure 2 illustrates the function of the filter. In 2 a) two arcs labeled with "have" start from the leftmost node. Both arcs are followed by non-speech events. From all the alternative paths starting with one of the "have"-arcs and ending in the rightmost node, we only want to keep a single one. For all the nodes in the "non-speech event cloud", all incoming arcs but the best scoring one are discarded. The result is the graph 2 b). The dotted arc is removed by a subsequent trimming step.

5. Experiments

5.1. Corpora

We present results on two different corpora: the EPPS 2005 Spanish corpus and the EPPS 2006 English corpus. Both corpora contain parliamentary speeches from the European Parliament and were collected within the TC-STAR project. All audio files are monaural with 16-bit resolution at a sampling rate of 16kHz. The training material contained 30h for the EPPS 2005 Spanish task and 100h for the EPPS 2006 English task.

Table 1: *Baseline for the EPPS Spanish Evaluation 2005 task. The baseline system is a standard MFCC system.*

	WER[%]		CN WER[%]		fWER [%]	
	dev.	eval.	dev.	eval.	dev.	eval.
3-gram LM						
w/o LDA	13.6	14.9	13.6	14.8	13.4	14.9
	12.2	13.1	12.2	13.0	12.1	13.0
with VTN	11.8	12.6	11.9	12.5	11.7	12.5
4-gram LM						
w/o LDA	13.2	14.6	13.2	14.5	13.2	14.6
	11.9	12.8	11.9	12.8	11.9	12.9
with VTN	11.7	12.1	11.7	12.1	11.5	12.2
LIMSI	11.2	12.3	11.2	12.2	-	-

Table 2: *Baseline for the EPPS English Evaluation 2006 task. The baseline system is a standard MFCC system with an additional voicedness feature and VTN.*

	WER[%]		CN WER[%]		fWER [%]	
	dev.	eval.	dev.	eval.	dev.	eval.
+ (C)MLLR	14.1	11.8	14.1	11.8	13.9	11.8
+ MMI	13.7	11.7	13.7	11.7	13.5	11.5
+ SAT	13.3	10.8	13.4	10.7	13.1	10.8
new lex/LM	12.9	10.3	13.0	10.4	12.7	10.3

5.2. Systems and Experimental Setup

For the Spanish task of the first TC-STAR Evaluation campaign 2005 we trained three acoustic models. A baseline model similar to the one described in [6], a model using vocal tract length normalization (VTN) and a model without linear discriminant analysis (LDA). Two different language models were used, a trigram and a fourgram. In addition, word graphs were kindly provided by J.-L. Gauvain (LIMSI) for this corpus. Initial experiments indicated that best performance can be expected from a combination of three systems including the best systems, i.e. VTN+fourgram and LIMSI, and one system using the trigram LM.

For the EPPS 2006 Evaluation English task we did system combination on the set of suboptimal acoustic models that evolved during the training of the final evaluation system [6].

All experiments were done on word graphs. The word graphs were pruned to a density of approximately forty. The graph error rates (GER) for the Spanish systems are around 5%. For the English development set the GER is approx. 3% and $\sim 1\%$ for the evaluation set.

For CN and CNC decoding the SRILM toolkit was used and for ROVER experiments the ROVER tool provided by NIST. The fWER experiments were done with our own software based on the RWTH FSA toolkit [7]. The confidence scores for the ROVER experiments were calculated as described in [8]. System priors and the ROVER parameters were optimized on the development sets. Oracle error rates were calculated on the best hypothesis of each system using the ROVER tool.

5.3. Results

In Tables 1 and 2 the results of the single systems to be combined are summarized and compared to CN and minimum fWER decoding. Possibly due to the low initial word error rates, the EPPS 2006 English development set was the only condition for which

Table 3: *Results on the EPPS Spanish Evaluation 2005 task for the combination of RWTH internal systems.*

combination method	systems	WER[%] dev.	WER[%] eval.
	best single system	11.7	12.1
Oracle	lm4+VTN, lm3, lm4 w/o LDA	8.1	8.7
ROVER	lm4+VTN, lm3, lm4 w/o LDA	11.3	12.2
	+ conf. scores	11.2	12.0
	+ weighted conf. scores	11.2	11.9
CNC	lm4+VTN, lm3, lm4 w/o LDA	11.3	12.2
	+ weights	11.3	12.1
Frame Based	lm4+VTN, lm3, lm4 w/o LDA	11.2	12.2
	+ weights	11.1	12.1

Table 4: *Results on the EPPS Spanish Evaluation 2005 task for the combination of RWTH internal systems and the LIMSI system.*

combination method	systems	WER[%] dev.	WER[%] eval.
	best single system	11.2	12.1
Oracle	Limsi, lm4+VTN, lm3	6.6	7.3
ROVER	Limsi, lm4+VTN, lm3	10.4	11.4
	+ conf. scores	10.3	11.2
	+ weighted conf. scores	10.0	10.8
CNC	Limsi, lm4+VTN, lm3	10.6	11.3
	+ weights	10.3	11.2

we observed (small) decreases in WER. In all other cases of fWER and CN no improvements were observed.

Table 3 shows the result for the internal system combination experiments on EPPS 2005 Spanish. Here, we used only RWTH systems and did not include the word graphs from LIMSI. Although the oracle WER indicates a potential for system combination, the final gain is small. The inclusion of the LIMSI lattices lowered the oracle WER by 1.4%. Weighted ROVER was able to decrease the WER by almost the same amount, cf. Table 4. Also CNC benefited from the LIMSI lattices, but less than ROVER.

Table 5 summarizes the results on the EPPS 2006 English corpus. For the development set, system combination seemed to capitalize on the suboptimal systems. But on the evaluation set none of the combination methods considered achieved a significant improvement.

6. Conclusions And Outlook

In this paper, a new system combination method based on minimum fWER decoding was presented. The new approach preserves the structure of the word graphs and the corresponding word boundaries, and can even be applied on word graphs of different segmentation.

For the first experiments presented we could not observe a significant difference in the performance of the new frame based system combination approach, CNC, and ROVER. However, the comparison of all combination approaches was done for internal system combination only, for which none of the methods gives significant improvements on the corpora considered.

Future work will therefore concentrate on combination experiments on word graphs both from different sites and based on fundamentally different models of comparable performance.

Table 5: *Results on the EPPS English Evaluation 2006 task for the combination of RWTH internal systems.*

combination method	systems	WER[%] dev.	WER[%] eval.
	best single system	12.9	10.3
Oracle	all systems	10.8	8.6
ROVER	all systems	13.0	10.5
	+ conf. scores	12.6	10.5
	+ weighted conf. scores	12.5	10.4
CNC	all systems	13.1	10.6
	+ weights	12.9	10.2
Frame Based	all systems	12.8	10.7
	+ weights	12.5	10.3

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).

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA.

7. 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*, Budapest, Hungary, September 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. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, May 2001, vol. 1, pp. 33 – 36.

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

[5] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, CA, USA, December 1997, pp. 347 – 354.

[6] Ch. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, Ch. Plahl, M. Bisani, R. Schlüter, and H. Ney, "The 2006 rwth parliamentary speeches transcription system," in *Proc. Int. Conf. on Spoken Language Processing*, Submitted, 2006.

[7] S. Kanthak and H. Ney, "Fsa: An efficient and flexible c++ toolkit for finite state automata using on-demand computation," in *Proc. 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, July 2004, pp. 510–517.

[8] 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, March 2001.