Generalized Driven Decoding for Speech Recognition System Combination

Benjamin Lecouteux, Georges Linarès, Yannick Estève, Guillaume Gravier

GENERALIZED DRIVEN DECODING FOR SPEECH RECOGNITION SYSTEM COMBINATION

Benjamin Lecouteux [1], Georges Linarès [1], Yannick Estève [2], Guillaume Gravier [3]

LIA, Avignon (France) [1], LIUM, Le Mans (France) [2], IRISA, Rennes (France) [3]

ABSTRACT

Driven Decoding Algorithm (DDA) is initially an integrated approach for the combination of 2 speech recognition (ASR) systems. It consists in guiding the search algorithm of a primary ASR system by the one-best hypothesis of an auxiliary system. In this paper, we generalize DDA to confusion-network driven decoding and we propose new combination schemes for multiple system combination. Since previous experiments involved 2 ASR systems on broadcast news data, the proposed extended DDA is evaluated using 3 ASR systems from different labs. Results show that generalized-DDA outperforms significantly ROVER method: we obtain a 15.7% relative word error rate improvement with respect to the best single system, as opposed to 8.5% with the ROVER combination.

Index Terms— Speech recognition, system combination

1. INTRODUCTION

Substantial efforts have been made by the ASR community for the combination of multiple speech recognition systems. Various collaboration schemes have been proposed, depending on the methods used for information sharing and on the level where combination operates. Several papers propose low level approaches at the acoustic level [1, 2, 3]. However, most combination techniques rely on the *a posteriori* re-estimation of the hypotheses generated by various systems. Such combination techniques can be implemented as a vote [4] or by merging confusion networks [5].

However, merging system outputs leads to discard some crucial information related to the decoding process, especially word boundaries, which are omitted in confusion networks. Furthermore, during decoding, each system prunes hypotheses according to its current knowledge and specific decoding strategy, though the information from the other systems could avoid pruning good hypotheses. Globally, we can expect better precision of the scoring and pruning processes by integrating earlier the information from the multiple sources and some recent papers investigate more integrated approaches [6].

In a previous work [7], we proposed an algorithm which consists in integrating the one-best hypothesis of an auxiliary ASR system into the search algorithm of a primary system. In this paper, we present a generalization of this algorithm to confusion network driven decoding and to multiple system combination. The first section presents the general principle of the Driven Decoding Algorithm (DDA). In the second section, we investigate the extension of DDA based on confusion networks rather than single best hypotheses. The third section presents integration schemes where several systems are combined with DDA. Results are compared to 2-system combination and to ROVER-based combination. Finally, we conclude and suggest some potential improvements.

This research is supported by the ANR (Agence Nationale de la Recherche), EPAC project ANR-06-MDCA-006.

2. ONE-BEST HYPOTHESIS DRIVEN DECODING

Driven decoding consists in integrating the outputs of an auxiliary system in the search algorithm of a primary system. This integration relies on two steps. Firstly, the current hypothesis of the primary system and the auxiliary transcript are aligned by minimizing the edit distance. Then, linguistic probabilities are combined according to posteriors and to an hypothesis-to-transcript matching score. The next two sections provide details on the driven decoding algorithm.

2.1. A* search algorithm in the Speeral system

The LIA speech recognizer is used as primary system. It is based on the A* search algorithm operating on a phoneme lattice. Decoding relies on the estimate function $F(h_n)$ which evaluates the probability of the hypothesis h_n crossing the node n according to

$$F(h_n) = g(h_n) + p(h_n), \tag{1}$$

where $g(h_n)$ is the probability of the current partial hypothesis up to node n, which results from the partial exploration of the search graph, and $p(h_n)$ is the probe which estimates the remaining probability from the current node n to the last node.

In order to be able to take into account information resulting from the output of an auxiliary system, the linguistic part of g in (1) is modified according to the auxiliary hypothesis as described below.

2.2. Driven decoding algorithm

Speeral ASR system generates word hypotheses as the phoneme-lattice is explored. The best hypotheses at time t are extended according to the current hypothesis probability and the probe results. In order to combine the information provided by the auxiliary transcript H_{aux} and the main search process, a synchronization point has to be found for each word node the engine evaluates. These points are found by dynamically mapping the provided transcripts to the current hypothesis minimizing the edit distance. This process allows to identify, in the auxiliary transcript H_{aux}, the best sub-sequence matching the incomplete theory h_{cur}. This sub-sequence, noted h_{aux}, is used for a new estimate of linguistic score, according to both a *matching score* $\theta(w_i)$ and h_{aux} posteriors $\phi(w_i)$.

$\theta(w_i)$ is a simple count of matching words between h_{cur} and h_{aux}. This score is combined to posteriors for linguistic probabilities weighting, according to the following rule :

$$L(w_i|w_{i-2}, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})^{1-\beta}.\alpha(w_i)^{\beta} \tag{2}$$

where $L(w_i|w_{i-2}, w_{i-1})$ is the resulting linguistic score, $P(w_i|w_{i-2}, w_{i-1})$ the initial probability of the trigram, β an empirical fudge factor and $\alpha(w_i)$ is the confidence score of w_i given by :

$$\text{if } \theta(w_i) > 0 \text{ then } \alpha(w_i) = \phi(w_i) \cdot \frac{\theta(w_i)}{\gamma} \text{ and } \beta = 0.6$$
$$\text{else } \beta = 0 \quad .$$

where γ is the analysis window size reported by the edit distance ($\gamma = 4$) and $\phi(w_i)$ posteriors from word w_i of the auxiliary system.

2.3. Experimental framework

2.3.1. Evaluation corpus

Experiments are carried out in the framework of the French ESTER evaluation campaign ([8]). The ESTER corpus contains French radio broadcast news, including some ad-hoc interviews, non-native speakers, on-the-fly translations... Results are reported on a test of 3 hours from three broadcasters (F.inter, F.info and RFI), extracted from the official ESTER development set.

Three ASR systems were used for testing DDA, namely the LIA speech recognition system *Speeral*, the LIUM speech transcription system and the IRISA speech transcription system *Irene*. The LIA system is used as the primary system while the LIUM and IRISA systems are used as auxiliary ones. Those three systems are briefly described in the following sections. All systems rely on the same official ESTER resources for the training of their respective acoustic and language models. The training data consists of 80 hours of manually transcribed audio data, corresponding to 1M words, and about 200M words from the newspaper "Le Monde".

2.3.2. The LIA broadcast news system

The LIA Broadcast News system relies on the Speeral decoder and the Alize-based segmenter. Cross-word context-dependent acoustic models with 230k Gaussians are used. State tying is achieved by decision trees. The language models are classical trigrams with a vocabulary of 65K words. The system runs two passes. The first one provides intermediate transcripts which are used for MLLR adaptation.

2.3.3. The LIUM speech recognition system

The LIUM speech transcription system is based on the CMU Sphinx 3.3 (fast) decoder [9]. This decoder uses fully continuous acoustic models with 3 or 5-state left-to-right HMM topologies. The LIUM Speech Project has added a Speaker Adaptive Training module, a 4-gram word-lattice rescoring process, and a segmentation toolkit ([10]). The decoding process can be decomposed into two passes in addition to the segmentation process: a first pass using band- and gender- specialized acoustic models and a trigram language model with a vocabulary of 65K words; a second pass using adapted acoustic models and a word-lattice rescoring process with a 4-gram language model.

2.3.4. IRISA transcription system

Irene is the recognition system developed at IRISA. It is based on word-synchronous beam-search algorithm with HMM acoustic modeling and n-gram linguistic models with a vocabulary of 64k words.

	F. Inter	F. Info	RFI
LIA	21.1	22.2	24.6
LIUM	18.5	18.9	25.6
IRISA	21.4	21.8	25.6
DDA-IRISA-P1	19.6	19.3	23.5
DDA-IRISA-P2	18.7	18.7	22.2
DDA-LIUM-P1	17.8	18.1	22.4
DDA-LIUM-P2	17.2	17.8	21.5

Table 1. Word error rates for DDA combination of Speeral with an LIUM system (*DDA-LIUM*) and IRISA system (*DDA-IRISA*) with (*P1* and without (*P2*) unsupervised speaker adaptation. Experiments performed of 3 hours of French broadcast news from the ESTER corpus.

The system operates in three steps plus a linguistic post-processing step. The first step uses context-independent acoustic models with a trigram LM to generate a large word graph which is then rescored with a 4gram LM and context-dependent models. A final word graph is generated in a third pass after MLLR speaker adaptation. Finally, consensus decoding is applied to the 1000-best sentence hypotheses list based on a combined acoustic, linguistic and morpho-syntactic score [11].

2.4. Results

Results are reported in table 1 for each auxiliary system combined with Speeral before (P1) and after (P2) speaker adaptation. The 2-pass strategy is assessed after speaker adaptation based on the transcription from the first driven decoding combination. We also report word error rates for each individual system. Results show a significant improvement with system combination compared to single systems. Performance achieved by DDA with the LIUM system remains better than the ones obtained with Irene (about 1% absolute WER), the latter exhibiting a higher error rate. Nevertheless, the combination with the IRISA systems still improves significantly the initial Speeral performance.

3. CONFUSION NETWORK DRIVEN DECODING

The information used by driven decoding based on the output transcription of an auxiliary system remains relatively poor. We investigate in this section the benefit of a richer information about the previous run of the auxiliary system. We apply the idea by integrating not only the one-best hypothesis but the word confusion network (WCN) generated by the auxiliary system.

3.1. Principle

As with the single best output, the combination method operates at the search level, by dynamically mapping the current word utterance to the confusion network. This is achieved by minimizing the edit-distance between the hypothesis and the WCN. In order to reduce the computational cost of this alignement, partial paths are saved and retreived on-demand, according to the paths explored in the search graph. In spite of a slight cpu-ressource increase, the time required by this step remains negligible compared to the time required for the full decoding process. The alignment step allows to extract the best

	F. Inter	F. Info	RFI
LIUM	18.5	18.9	25.6
DDA-LIUM-P1	17.8	18.1	22.4
DDA-LIUM-P2	17.2	17.8	21.5
DDA-WCN-LIUM-P1	17.7	18.1	22.3
DDA-WCN-LIUM-P2	17.2	17.8	21.5

Table 2. Word error rates for confusion network driven decoding (*DDA-WCN*), according to the decoding pass. Results are compared to the ones of the best single system (LIUM) and to the best one-best DDA system (*DDA-LIUM*)

.

projection of the hypothesis in the network; at this point, the rescoring problem is similar to the one-best driven decoding case : linguistics probabilities are rescored according to WCN-to-hypothesis matching-score and word posteriors (cf. equation 2).

3.2. Results

We tested confusion network driven decoding using confusion networks from the LIUM system. Results are reported in the table 2. We observed a significant improvement compared to the single systems (+1.5% absolute WER). Nevertheless, the gain with respect to the one-best driven decoder remains marginal (about -0.15% WER) for the first pass, and no gain is observed after speaker adaptation.

Two reasons could explain this disappointing gain provided by WCN:

- driven decoding based on the one best hypothesis uses both the confidence measures and the final decision taken by the auxiliary search; the latter guides the main search algorithm toward good hypotheses; this is probably a low-risk strategy which leads to a more robust combination;
- word confidence measures used in the one-best output are more reliable than the posteriors used in WCN scoring, especially due to a better support of linguistic information. As the confidence score is crucial for linguistic rescoring, the difference of confidence measure relevance could impact significantly the final results.

4. MULTI-SYSTEM COMBINATION

So far, DDA was limited to a single auxiliary system. In this section, we propose several extensions to generalize the DDA to several auxiliary systems.

4.1. Principle

Following the general DDA combination paradigm, combining several auxiliary systems can be done in one of two different ways.

The first approach consists in merging the set of auxiliary one-best hypotheses using a vote-based method such as ROVER. The resulting hypothesis drives the decoding performed at the second level using DDA.

The second approach consists in considering all information sources as independent word-streams, which can be integrated at the search level. In this full DDA-combination scheme, the current hypothesis is synchronized with each of the auxiliary transcripts, and independent matching scores are computed. Final linguistic rescoring integrates posteriors in order to estimate new linguistic scores according to each information sources and to the primary language model.

We tested and compared the two approaches on a 3 system configuration. Finally, we test a last scheme where all single systems and the DDA system outputs are merged by ROVER.

	F. Inter	F. Info	RFI
LIUM	18.5	18.9	25.6
ROVER-3	17.1	18.2	22.5
2-Level DDA-ROVER	16.8	17.3	21.3
DDA-3	16.7	17.0	20.6
DDA-3+ROVER	16.0	16.4	20.7

Table 3. Word error rates of multiple-system combination according to the combination schemes : the baseline ROVER combination of the 3 single systems (*ROVER-3*), the 2-level method (*2-Level DDA-ROVER*), the full DDA-integration (DDA-3) of auxiliary systems, and the ROVER combination of all systems including DDA-3 (DDA-3+ROVER). This last one obtains the best results with a WER decrease of about 15.7% relative with respect to the best single system (LIUM).

4.2. Two-Level ROVER-DDA combination

The principle of 2-level scheme relies on a first merging step where all auxiliary transcripts are merged. In our experiment, we use ROVER for merging LIUM and IRISA system outputs. The word confidence scores of the output are computed by averaging the confidence scores of words in each single system output. The resulting transcript is then used as an auxiliary hypothesis, following the classical scheme of a 2-system DDA combination with Speeral as primary system.

4.3. Integrated DDA-based combination

In this approach, all auxiliary systems outputs are submitted independantly to the primary search. For each of them, a matching score is computed according to independent transcript-to-hypothesis synchronization. Finally, all linguistic scores are merged by the log-linear combination extended to n systems:

$$L(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\beta} \cdot \frac{1}{N} \sum_{k=0}^{N} \alpha_k(w_i)^{\beta_k} \quad (3)$$

where β is the averaged β_k as defined in equation 2, α_k are the posteriors provided by the system k and N the number of auxiliary systems.

4.4. Results

Table 3 compares results obtained by the different proposed strategies of fusion. First, we observe that the adding of the third system improves systematically the system accuracy. Nevertheless, the ROVER of the 3 single systems obtains results that are close to the best 2-system combination (-0.2% absolute WER). The 2-level

	F. Inter	F. Info	RFI
DDA-3	16.7	17.0	20.6
ORACLE-3	10.3	10.5	14.5
DDA-3+ROVER	16.0	16.4	20.7
ORACLE DDA+ROVER	9.8	10.0	13.6

Table 4. Analysis of DDA by comparison to ROVER and Oracle measures.

method provides a more significant WER decrease (1.1% better than DDA-LIUM), but this configuration remains significantly worse than full DDA approach in 3-system configuration (additional gain of -0.4%WER). The last combination method consists in merging all available system outputs (including DDA-3). This hybrid method still improve the system accuracy of about 0.3% absolute WER. Globally, our best combination scheme allows improve both the initial best single system of about 3.3% absolute while outperforming significantly the classical ROVER combination of the 3 single systems (-1.6% absolute WER).

The obtained results confirm the idea that auxiliary information sources should be integrated in the search algorithm as soon as possible, in order to evaluate competing hypothesis while taking account all the constraints and knowledge available.

4.5. Driven decoding analysis

In order to complete the analysis of multiple system driven decoding, we conducted some additional experiments aiming to learn more about the behavior of DDA in this configuration.

We try first to know if DDA allows to find hypotheses not present in any of the single transcripts. This is achieved by comparing the Oracle performance on the 3 single systems (*ORACLE-3*) with the performance obtained when combining the single system outputs with the DDA-3 system output (*ORACLE DDA+ROVER*). Results reported in table 4 show that linguistic rescoring allows to guide the search toward alternative paths; this point confirms that DDA may not be considered as an on-line vote method but is really an integrated approach where additional information is integrated to the global cost function, allowing a new exploration of the search graph.

Moreover, it is important to note that ROVER combination of DDA-3 and all single systems outperfoms the pure DDA approach. This result demonstrates that while DDA finds new correct hypotheses, it also removes some correct ones compared to single systems. Moreover, this suggests that DDA may take benefit from more efficient tuning in order to select more systematically the good hypotheses when they can be found in auxiliary transcripts.

5. CONCLUSION

In this paper, we proposed an extension of the one-best driven decoding algorithm to word confusion network driven decoding and multi-system combination. WCN-driven decoding is a more general scheme for the integration of system outputs into the search algorithm, results show that WCN-driven decoding improves the primary system. Nevertheless, the performance are very close to the one obtained with the more simple one-best driven decoding. The enrichment of auxiliary information by driving the search using mutliple system outputs is a much more efficient strategy. We compared ROVER-based combination and system fusion by the DDA approach. The latter yielded better performance, taking a substantial benefit from the diversity of auxiliary systems. Finally, by using DDA-based cross-site system combination and a final ROVER pass, we obtained a global absolute gain of about 3.3% WER (15.7% relative gain) with respect to the best single system.

6. REFERENCES

[1] B. Hoffmeister, T. Klein, R. Schluter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *International Conference on Spoken Language Processing, Interspeech*, 2006, pp. 537–540.

[2] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Philadelphia, PA, March 2005, vol. 1, pp. 197–200.

[3] R. Prasad, S. Matsoukas, C.-L. Kao, J.Z. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System," in *InterSpeech 2005*, Lisbon, 2005.

[4] J.M Fiscus, "A post processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *IEEE ASRU Workshop*, 1997, pp. 347–352.

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

[6] I-Fan Chen and Lin-Shan Lee, "A new framework for system combination based on integrated hypothesis space," in *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA, 2006.

[7] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Julie Mauclair, "System combination by driven decoding," in *ICASSP'07*, Honolulu, Hawaii, USA, 2007.

[8] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Interspeech'05-Eurospeech*, Lisbon, Portugal, 2005.

[9] K. Seymore, C. Stanley, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M.A. Siegler, R. Stern, and E. Thayer, "The 1997 CMU Sphinx-3 english broadcast news transcription system," in *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.

[10] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news," in *Interspeech'05-Eurospeech*, Lisbon, Portugal, September 2005.

[11] G. Gravier S. Huet and P. Sébillot, "Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation," in *European Conf. on Speech Communication and Technology – Interspeech*, 2007.