System Combination by Driven Decoding

Benjamin Lecouteux, Georges Linarès, Yannick Estève, Julie Mauclair

SYSTEM COMBINATION BY DRIVEN DECODING

B. Lecouteux, G. Linarès

Laboratoire Informatique d'Avignon (LIA)
University of Avignon, France

Y. Estève, J. Mauclair

LIUM - Université du Maine
Le Mans, France

ABSTRACT

Combination of Automatic Speech Recognizer (ASR) generally relies on an a-posteriori merge of the different system outputs or on a cross-adaptation process. In this paper, we propose an integrated approach where the search of a primary system is driven by the outputs of a secondary one. This method allows to drive the primary system search by using the one-best hypotheses and the word posteriors gathered from the secondary system. Experiments are carried out within the experimental framework of ESTER evaluation campaign ([1]). Results show that the driven decoding algorithm significantly outperforms the two single ASR systems (-8% of relative WER, -1.7% absolute). Finally, we investigate the interactions between driven decoding and cross-adaptations. The best cross-adaptation strategy in combination with the driven decoding process brings to a final absolute gain of about 1.9% WER.

Index Terms— system combination, decoding algorithms, broadcast news transcription, confidence measures

1. INTRODUCTION

Combination of ASR systems have been largely investigated and used the last few years. Some authors propose cross-processing methods, where the sub-systems share intermediate outputs at each decoding pass. Moreover, final transcripts are generally combined by ROVER method [2] or by Confusion Networks Combination (CNC) [3]. These methods allow significant performance improvement, especially when the sub-systems have a good level of complementarity and relatively close performance. Nevertheless, the resulting hypotheses are built by merging the single-system outputs and some critical information may be lost, such as word-utterance synchronization or linguistic stream continuity. Recently, few works deal with more integrated approaches where the search-graph and/or the evaluation function are combined [4].

In this paper, we propose a Driven Decoding Algorithm (DDA) which consists in driving the search algorithm by the transcripts supplied by an auxiliary system. This follows some works we have presented in [5].

In the second section, we present the driven-decoding principle. The search algorithm of the main system is detailled and we show how the primary decoding process can be guided by the auxiliary system.
The third section presents the two broadcast news systems and the experimental framework.
The section 4 reports the evaluation of the DDA method; results are discussed and compared to a classical combination based on a ROVER technique.
Various cross-adaptation strategies applied on DDA-based decoding are evaluated in section 5.
Finaly, we conclude and suggest some perspectives.

2. THE DRIVEN DECODING ALGORITHM

2.1. Principle

The proposed combination technique consists in performing a first recognition pass using an auxiliary ASR system which provides a one-best hypothesis $h_{aux} = \{w_i\}$. For each word $w_i$from h_{aux}, a local confidence score $\phi_{aux}(w_i)$ is evaluated. Then, these informations are integrated into the search of the primary system, which is able to dynamically rescore the linguistic probabilities according to both h_{aux} and the corresponding confidence scores $\phi_{aux}(w)$. Next sections present the components involved in the DDA method.

2.2. Anatomy of the Speeral decoder

LIA has developed a large vocabulary continuous speech recognition system named Speeral [6]. This decoder is derived from a A* search algorithm operating on a phone lattice. The exploration of the graph is supervised by the estimate function $F(h_n)$ which evaluates the probability of the hypothesis h_n crossing the node n:

$$F(h_n) = g(h_n) + p(h_n) \quad (1)$$

where $g(h_n)$ is the probability of the current hypothesis which results from the partial exploration of the search graph (from the starting point to the current node n); $p(h_n)$ is the probe which estimates the probability of the best hypothesis from the current node n to the ending node.

In Speeral, the probe p combines an acoustic probability and a linguistic look-ahead score. The acoustic term is com-

puted by an acoustic decoding carried out by the Viterbi-back algorithm operating on a phone lattice.

The graph exploration is based on the function of estimate $F()$. Indeed, the stack of hypotheses is ordered on each node according to $F()$. The best paths are then explored firstly. This deep search refines the evaluation of the current hypothesis. Low-probability paths are cutted-off, leading to search backtrack. In such situations, the search is desynchronized from the audio stream.

In order to be able to take into account information resulting from the auxiliary transcript, the linguistic part of the $F()$ function is rescored according to a transcript-to-hypothesis matching score ($\alpha(w)$). Matching score and linguistic rescoring are described into the next two sections.

2.3. On-the-fly linguistic rescoring

Speeral speech recognition system generates hypotheses as the phone-lattice is explored. The best hypotheses at time t are extended according to the current hypothesis probability and the probe results. In order to locate anchorage points in an auxiliary transcript h_{aux}, each evaluated word from the current hypothesis h_{cur} is aligned to h_{aux} by using a Dynamic Time Warping (DTW) algorithm. Once h_{cur} is synchronized with h_{aux}, the algorithm estimates the matching transcript-to-hypothesis score $\phi_{aux}(w_i)$. This score is based both on the local confidence score and on the number of words in the short-term history which are correctly aligned with the transcript. Then, the linguistic probabilities are modified using the following rescoring rule:

$$\tilde{P}(w_i|w_{i-2}, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})^{1-\alpha(w_i)} \quad (2)$$

where $\tilde{P}(w_i|w_{i-2}, w_{i-1})$ is the updated trigram probability of the trigram (w_{i-2}, w_{i-1}, w_i) and $P(w_i|w_{i-2}, w_{i-1})$ is the initial probability of the trigram. $\alpha(w_i)$ is the confidence score of w_i

2.4. Transcript-to-hypothesis matching score

$\alpha()$ is a similarity measure between the current hypothesis h_{cur} and h_{aux}. This score is evaluated during the word graph exploration, by combining the confidence scores $\phi(w_i)$ and the number of words from h_{aux} which match to the current hypothesis. The computation of $\alpha(w)$ is achieved according to the following rules :

$$\alpha(w) = \begin{cases} \frac{\phi(w_1)+\phi(w_2)+\phi(w_3)}{3} & \text{if } (w_1..w_3) = (hw_1..hw_3) \\ \frac{\phi(w_1)+\phi(w_2)}{2} & \text{if } (w_1, w_2) = (hw_1, hw_2) \\ \phi(w_1) - \gamma & \text{if } w_1 = hw_1 \text{ and } \phi(w_1) \geq \gamma \\ 0 & \text{if } w_1 \neq hw_1 \text{ or } \phi(w_1) < \gamma \end{cases}$$

where γ is a confidence threshold which is *a priori* fixed. This filtering value allows to cut-off segments from h_{aux} where the auxiliary system probably fails.

2.5. Local confidence measure

Each word w_i of the h_{aux} hypothesis is associated to a local confidence measure $\phi(w_i)$. In this paper, the ASR which provides this one-best hypothesis is the LIUM speech recognition system [7]. This ASR is described in the section 3.2.

The confidence measure used by this system is described in [8], and is called WP/LMBB.

This measure is a combination of classical word posteriors (WP) with a measure based on the language model back-off behaviour (LMBB). Using the normalized cross entropy (NCE) as an evalution metric of confidence measures (this is the one used during the NIST campaigns), the WP/LMBB measure obtains 0.266 on the data used for the experiment presented below. This is an interesting score which shows that the WP/LMBB provides a reliable information on the correctness of the recognized words.

2.6. Hypothesis completion

Segmentation errors lead to the miss of speech segments or to non-speech decoding, increasing significantly the WER. We take advantage of the dual decoding on the segmentation level. When the main system misses some speech segment which have been recognized by the auxiliary one with a confidence score greater than a fixed threshold, the corresponding transcript is integrated to the final hypothesis.

3. EXPERIMENTAL FRAMEWORK

In our experiments, the main system is based on Speeral decoder which has been developped at the LIA, and the auxiliary hypotheses (and associated confidence measure) is supplied by LIUM laboratory. These 2 systems are described in the next sections.

3.1. The LIA broadcast news system

The LIA broadcast system relies on Speeral decoder and Alize-based segmenter ([9]). Here, we use the system involved in the ESTER evaluation campaign [1]. Context-dependent acoustic models are used. Tying is achieved by decision trees. We train the acoustic models on ESTER materials (about 80 hours of anotated speech). The language models are classical trigrams estimated on about 200M of words from the French newspaper *Le Monde* and the broadcast news manual transcripts provided during the ESTER campaign. The system runs two passes. The first one provides intermediate transcripts which are used for MLLR adaptation. The first pass takes about 3xRT and the second one about 5xRT on a standard desktop computer.

3.2. The LIUM speech recognition system

The LIUM speech transcription system is based on the CMU Sphinx 3.3 (fast) decoder [10]. The s3.3 decoder is a branch of the CMU Sphinx III project which has been developed to include some speed improvements. 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 Adaptative Training module, a 4-gram word-lattice rescoring process, and a segmentation toolkit. The decoding process can be decomposed into two passes (plus the segmentation process): a first pass using band- and gender- specialized acoustic models and a trigram language model; a second pass using adapted acoustic models and a word-lattice rescoring process with a quadrigram language model. The entire process runs under 12xRT on a standard Intel Pentium IV computer.

The LIUM system has reached the second position in the transcription task (TRS) on the ESTER evaluation campaign [1]. More details about this system are presented in [7].

For the experiment presented in this paper, the acoustic and linguistic models were trained on the ESTER training corpus.

4. EVALUATION OF COMBINATION BY DRIVEN DECODING

The two ASR systems are assessed on 3 shows (3h) of radio broadcast (one hour from *France Inter*, one hour from *France Info* and one hour from *Radio France International*) extracted from the official ESTER development corpus.

The auxilary system (the LIUM one) runs a full decoding process as described in section 3.2. Then, confidence scores are estimated as presented in section 2.5. These results are integrated using the Data Driven Algorithm (DDA) in the Speeral search process as detailed in section 2.

The baseline results are the recognition outputs from the two ASR systems: *LIA-P2* is the result of the entire decoding process of Speeral (performing two passes), *LIUM* is the result of the entire decoding process of the LIUM system (two passes).

The DDA is used here during the second pass of the Speeral system (unsupervised acoustic adaptation is applied on the first pass of Speeral decoding). The first pass is the same as the one used in the Speeral baseline system. Results of this DDA process are called *LIA-P1 DDA-P2*.

Table 1 shows that the DDA process allows to obtain a significative reduction of the word error rate (WER) in comparison with the best baseline system for a given show (up to 1.9% absolute WER reduction). The global reduction is 1.7% absolute in comparison with the best baseline system (21.1% WER for the LIUM system, 19.4% WER for the DDA system).

	F. Inter	F. Info	RFI
LIA-P2 (base. LIA)	21.1	22.2	24.6
LIUM (base. LIUM)	19.5	18.8	25.4
LIA-P1 DDA-P2	**18.1** *(-1.4)*	**18.4** *(-0.4)*	**22.7** *(-1.9)*

Table 1. Evaluation of Driven Decoding Algorithm (*LIA-P1 DDA-P2*) performance in terms of Word Error Rate (WER). Results are compared to those obtained by the LIA system (*LIA-P2*) and by the LIUM system (*LIUM*). This test is achieved on 3 shows of French broadcast news from the official ESTER development corpus.

In order to evaluate the optimal combination of the one-best hypotheses of the two baseline systems, the best combination of the two hypotheses knowing the correct word utterance is computed. This allows to determinate the *oracle* WER using a ROVER method [2] to merge the results of these two systems. Moreover, the *oracle* WER using a ROVER between the 3 systems (*LIA-P2* baseline, *LIUM* baseline and *DDA* system) is also computed.

The results reported in table 2 show that the optimal potential gain obtained in using the DDA system is very significative. Mainly, these results underline an interesting feature of the DDA in comparison with a simple ROVER to combine two systems: the DDA approch allows to propose new word-hypotheses which were not present in the initial results of the baseline systems.

	F. Inter	F. Info	RFI
LIA-P2⊕LIUM	*14.9*	*13.8*	*19.5*
LIA-P2⊕LIUM⊕DDA	*13.0 (-1.9)*	*12.1 (-1.7)*	*18.8* (-0.7)

Table 2. Word error rates obtained according to the *oracle* ROVER combination of the outputs of the baseline systems and the oracle ROVER combination of these outputs and the output of the DDA system.

5. CROSS ADAPTATION AND DRIVEN DECODING

Cross adaptation has shown to be an efficient and relatively simple method for system combination ([11]). It consists in adapting acoustic models of a system by mapping them to transcripts provided by another system. This method leads to significant improvements by taking advantage of sub-systems complementarity at the level of acoustic modeling. We investigate various cross-adaptation schemes by using intermediate transcripts provided by the auxiliary system or by the DDA system.

We test three baseline configurations: a Speeral decoding without any unsupervised adaptation (*LIA-P1*), a DDA decoding without adaptation, and a cross adaptation to h_{aux} transcripts followed by a Speeral decoding (*LIUM-P1 LIA-P2*).

Finally, we evaluate 3 acoustic adaptation strategies for the DDA system: acoustic model mapping to the h_{aux} transcript (*LIUM-P1 DDA-P2*), adaptation using the first pass of Speeral decoding (*LIA-P1 DDA-P2*), adaptation using the DDA first pass decoding (*DDA-P1-DDA-P2*). Results are reported in table 3 and compared to the one obtained by the DDA system without any adaptation (*DDA-P1*).

	F. Inter	F. Info	RFI
LIA-P1	22.5	23.3	26.3
LIUM-P1 LIA-P2	20.4	21.8	24.1
DDA-P1	18.1	18.7	23.6
LIA-P1 DDA-P2	18.1	18.4	23.1
LIUM-P1 DDA-P2	17.9	18.1	22.7
DDA-P1 DDA-P2	17.9	18.1	22.7

Table 3. Various schemes of cross adaptation combined to driven decoding : adaptation targets are provided by LIUM decoding (*LIUM-P1 DDA-P2*), LIA first pass decoding (*LIA-P1 DDA-P2*), DDA first pass decoding (*DDA-1 DDA-2*). Resulting WER are compared to single Speeral decoding (*LIA-P1*), DDA first pass decoding (*DDA-P1*), and Speeral decoding by adapting to h_{aux} transcripts (*LIUM-P1 LIA P2*).

Performance reached by the DDA decoding without speaker adaptation (*DDA-P1*) are greater than those obtained by the initial Speeral decoding (-1.0% WER) and relatively close to those obtained with the best configuration (-0.7% WER). Moreover, the cross adaptation of Speeral models using the LIUM transcripts (*LIUM-P1 LIA-P2*) outperforms dramatically the classical scheme where the system is adapted using its own trancripts (*LIA-P2*, reported in the Table 1). Nevertheless, it seems clear that the gains are not cumulative: we obtain a maximum of absolute additionnal gain of 0.27% compared to the driven decoding with models adapted to transcripts from Speeral first pass (*LIA-P1 DDA-P2*).

Finally, by combining DDA and crosss adaptation, we reach an absolute WER gain of 2.9% compared to the initial Speeral decoding and about 1.9% compared to the LIUM system.

6. CONCLUSION AND PERSPECTIVES

We have proposed an algorithm for a driven-by-transcript decoding (DDA). This method allows an efficient combination of 2 systems, by rescoring linguistic probabilities according to transcripts and word posteriors gathered from an auxiliary system.

Experimental results show that this integrated approach brings significant gains compared to classical single cross-adaptation: better results are obtained by performing a two pass driven decoding than a single cross-adaptation system. DDA leads to a WER improvment (-1.3% of absolute gain) compared to the best single cross-adaptation system. By combining this dynamic linguistic rescoring with an acoustic cross-adaptation, we observe a final absolute gain of 1.9% WER in comparaison with the best baseline ASR. Moreover, the analysis of ROVER *oracle* results shows that the DDA approach generates new correct hypotheses which were not proposed by any baseline systems.

We plan now to generalize the DDA approach by driving the search process using confusion networks instead of single one-best hypotheses. Moreover, application of DDA method to n-system combination (with $n > 2$) will be investigated.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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*, 2000.

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

[5] Benjamin Lecouteux, Georges Linares, J.F. Bonastre, and Pascal Nocera, "Imperfect transcript driven speech recognition," in *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA, 2006.

[6] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, "The LIA's French broadcast news transcription system," in *SWIM: Lectures by Masters in Speech Processing*, Maui, Hawaii, 2004.

[7] 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.

[8] J. Mauclair, Y. Estève, S. Petit-Renaud, and P. Deléglise, "Automatic detection of well recognized words in automatic speech transcription," in *LREC 2006*, Genoa, Italy, May 2006.

[9] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *ICASSP'05*, Philadelphia, USA, March 2005.

[10] 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.

[11] 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.