Dynamic Combination of Automatic Speech Recognition Systems by Driven Decoding

Benjamin Lecouteux, Georges Linares, Yannick Estève, Guillaume Gravier

Dynamic Combination of Automatic Speech Recognition Systems by Driven Decoding

Benjamin Lecouteux[1], Georges Linares[2], Yannick Estève[3], Guillaume Gravier[4]
[1] benjamin.lecouteux@imag.fr (Laboratoire Informatique de Grenoble, LIG - France),
[2] georges.linares@univ-avignon.fr (Laboratoire Informatique d'Avignon, LIA- France),
[3] yannick.esteve@univ-lemans.fr (Laboratoire d'Informatique de l'Université du Maine, LIUM - France),
[4] guig@irisa.fr (Institut de recherche en informatique et systèmes aléatoires, IRISA - France)

***Abstract*—Combining automatic speech recognition (ASR) systems generally relies on the posterior merging of the outputs or on acoustic cross-adaptation. In this paper, we propose an integrated approach where outputs of secondary systems are integrated in the search algorithm of a primary one. In this driven decoding algorithm (DDA), the secondary systems are viewed as observation sources that should be evaluated and combined to others by a primary search algorithm. DDA is evaluated on a subset of the ESTER I corpus consisting of 4 hours of French radio broadcast news. Results demonstrate DDA significantly outperforms vote-based approaches: we obtain an improvement of 14.5% relative word error rate over the best single-systems, as opposed to the the 6.7% with a ROVER combination. An in-depth analysis of the DDA shows its ability to improve robustness (gains are greater in adverse conditions) and a relatively low dependency on the search algorithm. The application of DDA to both A* and beam-search-based decoder yields similar performances.**

***Index Terms*—Speech processing, automatic speech recognition, system combination**

I. Introduction

The general ASR system combination principle consists in using complementary ASR systems that exchange information at different levels of the decoding process. Numerous approaches rely on the combination of acoustic information. In [1], the authors aggregate multiple feature streams into a single one. Other approaches combine acoustic scores at the frame or at the model level [2], [3], [4]. One of the most popular techniques being cross-adaptation which consists in adapting acoustic models on transcriptions produced by other systems [5], [6]. Considering that the combination efficiency strongly depends on the system (or model) complementarity, several papers propose techniques that aim at augmenting system diversity while preserving accuracy [7], [8], [9].

Another family of combination techniques relies on the posterior re-estimation of the hypotheses generated by various systems: in this case, re-estimation is done through posterior merging of recognition hypotheses. In this combination scheme, each ASR system computes, independently from the others, the best hypothesis; as a final step, all text outputs are then aligned and merged. In the ROVER algorithm [10], this merging is performed by a simple weighted vote that is applied to the N single-system hypotheses by taking into account ASR system posteriors. ROVER is usually limited to the 1-best hypotheses; [11] proposes to apply it to the N-best list from each ASR system. In [12], [13], improvements of ROVER are evaluated based on machine learning algorithms against voting. In [14], the authors refine ROVER by using a language model concurrently with the voting algorithm. Another approach, proposed in [15], uses Bayesian decision theory for an optimal weighting of ASR system hypotheses which take into account confidence measures. A ROVER generalization was proposed in [16], where the voting mechanism relies on the posterior combination of confusion networks (CNC). However, experiments assessed in [3] exhibit only a slight gain with CNC compared to ROVER. These two methods, ROVER and CNC, yield significant performance improvement, especially when the sub-systems have a good level of complementarity and relatively close performance. Nevertheless, they have some serious limitations.

The main issue with posterior combination is that merging the system outputs leads to discarding some crucial information related to the decoding process, especially word boundaries, which are omitted in confusion networks or word-utterance synchronization and linguistic stream continuity. Furthermore, during decoding, each system prunes hypotheses according to its current knowledge and its specific decoding strategy, though sharing earlier cross-system information could avoid the pruning of correct hypotheses. Globally, we can expect better precision for the combination by integrating downstream the information from the multiple auxiliary sources [17].

In this paper, we propose an integrated approach that relies on the basic idea of combining integrated systems. We present a framework based on the driven decoding algorithm (DDA) that consists in joining an ASR system with outputs provided by auxiliary systems [18], [19] or manual transcripts [20]. This method relies on a new decoding strategy where the search algorithm is dynamically driven by transcripts from other systems, thus dynamically modifying the search space.

The paper is structured as follows: DDA is formalized in Section II while Section III details implementation of DDA with two types of decoder, namely A^* and beam search. Section IV presents the experimental framework. Section V reports the evaluation of the DDA approach applied on both a A^* decoder and a beam-search decoder; results are discussed and compared to a classical combination based on a ROVER tech-

This research was supported by the ANR (Agence Nationale de la Recherche) under contract number ANR-09-BLAN-0161 (ASH).

nique. Various cross-adaptation strategies applied on DDA-based decoding are evaluated in Section VI. In Section VII, we present a generalization of this algorithm to confusion network driven decoding and to N system combination with $N \geq 2$. Finally, we conclude in the last section.

II. The Driven Decoding Algorithm

The general idea of DDA is to give credit to partial search hypotheses of the primary system which are consistent with the hypotheses of the auxiliary systems. The algorithm works by integrating on-the-fly assumptions derived from the auxiliary systems. The principle of the algorithm consists in aligning the auxiliary transcripts with partial hypotheses of the primary system during decoding. The quality of the fit between a partial hypothesis and auxiliary transcripts is then used to reevaluate the language model component of the primary system, where a good fit increases the language model score.

We first describe the synchronization step to locate the best portion of the auxiliary transcript corresponding to the current partial hypothesis. We then propose an alignment measure to modify the language model score.

A. ASR synchronization to the auxiliary transcript

In order to locate an anchor point in the auxiliary hypothesis T of size m (being the number of words), each word from the current hypothesis H is aligned to T using an edit distance. We use the method presented in [20]: the partial hypothesis H of size n is built by collecting the current word and its history from the best path found during the search process. To align sequences we construct an n-by-m matrix where the element (i, j) of the matrix contains the distance between the two words T_j and H_i. We use a basic distance function, where $d(T_j, H_i) = 0$ when $T_j = H_i$ and $d(T_j, H_i) = 3$ for deletion, $d(T_j, H_i) = 4$ for insertion and $d(T_j, H_i) = 6$ for substitution. Costs were computed via the estimated probability of each event in the auxiliary systems (i.e., number of insertion, deletions and substitutions). In practice these costs are the same for all auxiliary systems and the impact of distance values in a $\pm$ 1 range is limited. The cumulative distance $\gamma(i, j)$ between H_i and T_j is computed as follows:

$$\gamma(i,j) = d(T_j, H_i) + min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (1)$$

This alignment is performed each time a word end is hypothesized. To efficiently compute the alignment, a cache of the previous alignment can be used in order to quickly compute the cumulative distances $\gamma(i, j)$.

The dynamic synchronization of the search algorithm driven by the alignment on an auxiliary transcript is illustrated Figure 1.

The best hypothesis-to-reference matching provides a synchronization point $\Gamma(i)$ defined as follows:

$$\Gamma(i) = \arg\min_j \gamma(i, j) \quad (2)$$

$\Gamma(i)$ allows to find the best synchronization point s_i:

$$s_i = (T_{\Gamma(i)}, H_i) \quad (3)$$

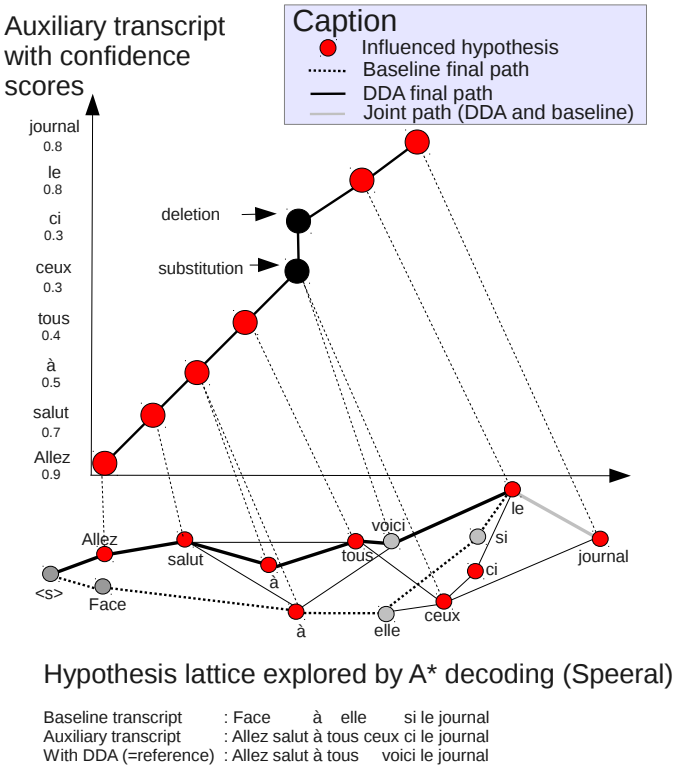

Fig. 1. Synchronization of the search beams with the auxiliary transcript by the DTW algorithm during an asynchronous decoding. Linguistic probabilities are biased according to the similarity of aligned transcripts.

Given the synchronization point between the hypothesized and auxiliary transcripts, one has to evaluate a transcript-to-hypothesis score reflecting how well the two sequences match.

B. Transcript-to-hypothesis matching score

In order to use information of the auxiliary transcript, the linguistic part of the primary ASR cost function is reevaluated according to a transcript-to-hypothesis matching score $\theta_\delta(s_i)$ combined with a confidence score $\phi(s_i)$ of the word $T_{\Gamma(i)}$ in s_i. This mechanism drives the search by dynamically rescoring the language model value, according to the alignment and word confidence scores. The better the alignment and confidence score, the higher the impact on the linguistic score.

The matching score denoted $\theta_\delta(s_i)$ is based on the number $\epsilon_\delta(\Gamma(i))$ of words in the language model short-term history of size δ that are correctly aligned with the auxiliary transcript. $\theta_\delta(s_i)$ is greater when the trigram is aligned and linearly decreases with the misalignments of the history:

$$\theta_\delta(s_i) = \frac{\epsilon_\delta(\Gamma(i))}{\delta} \quad (4)$$

where δ is the size of the history used to compute the matching score

Finally, the matching score $\alpha_\delta(H_i)$ of the word H_i is computed by using the word confidence score $\phi(s_i)$ of the word T_j as

$$\alpha_\delta(H_i) = \phi(s_i)\theta_\delta(s_i) \quad (5)$$

C. On-the-fly linguistic rescoring

Once the matching score is fully estimated according to the synchronization process and the confidence measures, the n-gram probabilities are reevaluated. We evaluate two types of strategies: the first one consists of a dynamic language model scaling factor and the second one of a dynamic word penalty.

1) Dynamic language model: The first proposed combination dynamically evaluates the language model scaling factor according to the auxiliary transcript T. The linguistic probabilities are modified according to the following rescoring rule:

$$\tilde{P}(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\alpha_\delta(w_i)} \quad (6)$$

where $\tilde{P}(w_i|w_{i-2},w_{i-1})$ is the resulting modified linguistic score. $P(w_i|w_{i-2},w_{i-1})$ is the initial trigram probability and $1-\alpha_\delta(w_i)$ is the word LM scale factor for w_i.

2) Log-linear combination: [21] conducted experiments which demonstrated that the product combination rule gives good results when all classifiers are quite accurate. On the other hand, if one (or more) of the classifiers significantly fails, then the arithmetic mean rule yields better results. In this case, the combination consists of a dynamic word penalty with a rescaling of the initial linguistic probability. Following this principle, the linguistic probabilities are modified using the following rescoring rule:

$$\tilde{P}(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\beta}\alpha_\delta(w_i)^{\beta} \quad (7)$$

where β is an empirical scaling factor.

III. DDA IMPLEMENTATION

The DDA principle of modifying the linguistic score of an ASR system based on auxiliary transcripts is generic and can be implemented in various ASR systems. We provide details on the implementation with two widely used algorithms, namely A^* and beam search.

A. A^* implementation

The LIA laboratory has developed a large vocabulary continuous speech recognition system named Speeral [22]. This decoder is derived from the A* search algorithm dedicated to the search of the best path in a graph. It has been used in several speech recognition engines, generally for word-graph decoding. In Speeral, the A* algorithm operates on a phoneme lattice, which is estimated using cross-word and context-dependent HMM.

The exploration of the lattice is supervised by an estimation function $F(h_n)$, which evaluates the probability of the hypothesis h_n crossing the node n:

$$F(h_n) = g(h_n) + p(h_n) \quad (8)$$

where $p(h_n)$ is the probe that estimates the probability of the best hypothesis from the current node n to the ending node. In Speeral, the probe p combines the acoustic probability and a language model look-ahead (LMLA) score [23]. The acoustic term is determined via an acoustic decoding process, carried out as a Viterbi algorithm on the phone-lattice. The LMLA used in Speeral enables the comparison of competing hypotheses before reaching a word boundary. The probability of a partial word corresponds to the best probability in the list of words sharing the same prefix:

$$P(W^*|h) = max_i P(W_i|h) \quad (9)$$

where W^* is the best possible continuation word and h the word history (partially present in $g(h_n)$). $g(h_n)$ is the probability of the current hypothesis that results from the partial exploration of the search graph (from the starting point to the current node n):

$$g(h_n) = \max P(W)^{\Theta}\Delta^{|W|}P(X|W) \quad (10)$$

where $P(W)$ is the linguistic probability of the current word sequence, $P(X|W)$ is the acoustic probability according to the word sequence W, Θ is the language model scale factor, Δ is the linguistic penalty and $|W|$ the number of words in W.

The Speeral speech recognition system generates hypotheses as the phone-lattice is being explored: the best hypotheses at time t are extended according to the current hypothesis probability and the probe results: best paths are then explored in a depth-first manner. In practice, a partial hypothesis H of size n is built by collecting the current word (i.e., the best word end according to the LMLA for the node currently considered) and its history from the path found during the search process. The sequence alignment is achieved as presented in II-A at each newly encountered word in the phone-lattice.

With the DDA (in log-linear version), Eq. 10 becomes:

$$\tilde{g}(h_n) = \max \prod_{k=1}^{k=|W|} P(W^k)^{(\Theta-\beta+1)}\alpha_\delta(W^k)^{\beta}\Delta^{|W|}P(X|W) \quad (11)$$

where $\alpha(W^k)$ is dynamically evaluated according to the auxiliary transcript alignment with the currently explored hypothesis (i.e., using the δ last words of the sequence W).

B. Viterbi implementation

The DDA can be transposed in a Viterbi decoding with the requirement to know the best history alignment of the currently explored word in the beam search in order to scale the linguistic score. The framework proposed by [24] aligns subtitles in a similar manner. In their approach the rescoring trigram rule is based on the conditional distribution $P(w_i|w_{w_1\ldots i-1}c_1\ldots c_m)$ where w_i is the following word in a transcript whose subtitle is $c_1\ldots c_m$.

C. DDA and decoding time

We observe that the DDA algorithm improves decoding speed slightly on the A^* algorithm, in spite of the additional computational cost due to search synchronization. This gain in terms of execution time is due to the earlier exploration of the best paths provided by auxiliary systems.

IV. Experimental framework

We implemented DDA both in an A^* algorithm and in a Viterbi-based algorithm. The A^* version is assessed by the *Speeral* decoder. The Viterbi version is tested on the Speeral word-graphs (not obtained with DDA) by using *Fastnc*, a LIA Viterbi-based graph decoder. The auxiliary hypotheses (and associated confidence measures) are supplied by a Sphinx-derived system developed by LIUM [25].

A. Development and test corpus

Development (1 hour) and test (3 hours) data constitutes a subset of the development corpus provided for the ESTER evaluation campaign [26]. The ESTER corpus consists of radio broadcast news from four channels, namely *France Inter*, *France Info*, *Radio France International* (RFI) and *Radio Télévision Marocaine* (RTM). The corpus, designed to evaluate broadcast news transcription systems, contains mostly high quality speech. Nevertheless, some segments are recorded in more difficult acoustic conditions, including *ad-hoc* interviews, speech from non-native speakers, and on-the-fly translations.

We used RTM (1h) to tune the scale factor β of DDA. The combination values were determined empirically by using a grid-search method in order to reduce the WER. We have observed a relation between parameter values and the quality of auxiliary ASR systems. The quality of confidence measures has a strong impact.

The test is performed on 3 shows from 3 radio broadcasts: one hour from *France Inter*, one hour from *France Info* and one hour from *RFI*.

B. Segmentation issue

Each ASR system uses its own segmentation algorithm. Segmentation is used in order to extract speech segments from audio data. Each segmentation system gives different results in terms of timestamps and speech detection. Segmentation errors lead to the miss of speech segments or to non-speech decoding, increasing significantly the WER. We take advantage of the multiple ASR segmentations. When the main system misses some speech segments 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.

C. The LIA broadcast news system

Experiments were carried out with the LIA system used in the ESTER evaluation campaign, this system was based on the Speeral decoder and the Alize-based segmenter [27]. Segments are automatically segmented by speaker (in order to adapt the acoustic models using Maximum Likelihood Linear Regression), and their size is limited to 30 seconds.

Context-dependent acoustic models are used. We train the acoustic models on ESTER materials (about 80 hours of annotated speech). The language models are classical trigrams estimated from 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 takes about 5xRT (exploration cut-offs are decreased) on a standard desktop computer.

We use the *Fastnc* decoder in order to test DDA with a Viterbi search. Initially the Speeral graphs do not contain linguistic informations. As a first step the *Fastnc* decoder performs a static graph expansion by using the 3-gram language model. For each new expanded node, DDA is applied to bias the language model probabilities. When the graph is expanded a classic Viterbi algorithm is applied to find the best path in the graph.

The LIA confidence measures (used for ROVER experiments) are computed in two stages. The first one extracts low-level features related to the acoustic and search graph topology, and high-level features related to linguistic information. Each word from the hypothesis is represented by a feature vector composed of 23 features that are grouped into 3 classes. Then, a first bad word detection hypothesis is produced by a classifier based on the boosting algorithm. More details can be found in [28].

D. The LIUM broadcast news transcription system

The LIUM speech transcription system is based on the CMU Sphinx 3.3 (fast) decoder [29]. The Sphinx 3.3 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 Adaptive 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 4-gram language model.

The LIUM system has earned second position in the transcription task (TRS) in the ESTER evaluation campaign. More details about this system are presented in [25].

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

Each word H_i in hypothesis H is assigned to a local confidence measure $\phi(H_i)$. In the next experiments, the LIUM speech recognition system [25] provides the one-best hypothesis. The confidence measure used by this system is described in [30], and is called WP/LMBB.

This measure is a combination of classical word posteriors (WP) with a measure based on the language model backoff behavior (LMBB). Using the normalized cross entropy (NCE) as an evaluation 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 reliable information on the correctness of the recognized words.

V. Evaluation of combination by driven decoding

The auxiliary system (the LIUM one) runs a full decoding process as described in Section IV-D, thus providing a transcript with confidence measures. These results are integrated into the primary system using the driven decoding algorithm (DDA) in the Speeral search process as detailed in section II.

The baseline results are the recognition outputs from the LIA, LIUM and a ROVER based combination: *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) and *ROVER* is a ROVER using LIUM and LIA confidence scores.

A. DDA experiments

The DDA is used here during the second pass of the LIA system: unsupervised acoustic adaptation is applied using the first pass transcript of the LIA system; the second pass using the adapted acoustic models. The first pass is the same as the one used in the LIA baseline system. Results of this DDA process are called *LIA-P1 DDA-P2* in the case of Speeral and DDAf in the case of Fastnc (Fastnc experiments use graph computed by the second pass of the LIA baseline system, without DDA).

Table I shows that the DDA process obtains a significant reduction of the word error rate (WER) in comparison with the best baseline system for a given show (up to 2.4% absolute WER reduction). The best global reduction is 2% absolute in comparison with the best baseline system (21.2% WER for the LIUM system, 19.2% WER for the DDA system).

Moreover, these experiments show that DDA is able to outperform significantly the ROVER baseline. Especially on the F.Info show where ROVER does not allow for improvement compared to the ASR LIUM system. Finally DDA presents a global WER reduction of 1.4% absolute in comparison with the ROVER baseline system (about 7% relative).

Better results are obtained by using the log-linear combination which is more efficient on all shows.

The Viterbi decoder (DDAf) shows a slight improvement: the word-graphs generated by the LIA system are not fully explored by the A^* algorithm because of cut-off, while the Viterbi decoder performs an exhaustive search. However, these experiments highlights that DDA works similarly in a Viterbi framework and in an A^* framework.

	F. Inter	F. Info	RFI	Average
LIA-P2 (base. LIA)	21.1	22.2	24.6	22.6
LIUM (base. LIUM)	19.5	18.8	25.4	21.2
ROVER	**19.0**$_{(-0.5)}$	**18.9**$_{(+0.1)}$	**24.0**$_{(-0.6)}$	**20.6**$_{(-0.6)}$
LIA-P1 DDA-lin-P2	**18.1**$_{(-1.4)}$	**18.4**$_{(-0.4)}$	**22.7**$_{(-1.9)}$	**19.7**$_{(-1.5)}$
LIA-P1 DDA-log-P2	**17.8**$_{(-1.7)}$	**18.1**$_{(-0.7)}$	**22.4**$_{(-2.2)}$	**19.4**$_{(-1.8)}$
LIA-P1 DDAf-lin-P2	**17.9**$_{(-1.6)}$	**18.3**$_{(-0.5)}$	**22.4**$_{(-2.2)}$	**19.5**$_{(-1.7)}$
LIA-P1 DDAf-log-P2	**17.6**$_{(-1.9)}$	**17.9**$_{(-0.9)}$	**22.2**$_{(-2.4)}$	**19.2**$_{(-2.0)}$

TABLE I
Evaluation of DDA (*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*). DDA is tested with an A* framework (DDA) and a Beam search framework (DDAf). This test is achieved on 3 shows of French broadcast news from the official ESTER development corpus.

	F. Inter	F. Info	RFI	Average
LIA-P2⊕LIUM	14.9	13.8	19.5	16.0
LIA-P2⊕LIUM⊕DDA	13.0$_{(-1.9)}$	12.1$_{(-1.7)}$	18.8$_{(-0.7)}$	14.6$_{(-0.4)}$
LIA-P2⊕LIUM⊕DDAf	12.9 $_{(-2.0)}$	12.1$_{(-1.7)}$	18.7$_{(-0.8)}$	14.6$_{(-0.4)}$

TABLE II
Word error rates for *oracle* ROVER combination of system outputs: combination of the baseline systems only *vs.* combination of the baseline systems with DDA (and then DDA*f*).

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 transcription of the utterance is computed. This allows to determine the *oracle* WER using a ROVER method [10] to merge the results of these two systems. Moreover, the *oracle* WER using a ROVER among the 3 systems (*LIA-P2* baseline, *LIUM* baseline and *DDA* system) is also computed.

Results reported in Table II show that the optimal best potential gain obtained using the DDA system is very significant. Mainly, these results underline an interesting feature of the DDA in comparison with combining the two systems through a single-weighted ROVER: by using an *oracle* between LIA, LIUM and DDA, we observe a gain compared to the LIA-LIUM *oracle*. This highlights that the DDA approach allows to propose new word hypotheses which were not present in the initial one-best results of the baseline systems. This aspect shows that with the same information used by ROVER, DDA is able to combine information and to explore new paths in the graph (Figure 4). These new paths allows to introduce new hypotheses correlated with auxiliary transcripts, thus confirming the potential gains of an integrated approach.

The next experiments generalize the DDA approach by driving the search process using confusion networks instead of single one-best hypotheses.

B. Confusion network driven decoding

The information used by the 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 also the word confusion network (WCN) generated by the auxiliary system.

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 (Figure 2). At the decoder level the changes are minor: we only modify the distance function of the dynamic alignment. The distance function becomes :

$$\begin{array}{l} d(WCN_j, T_i) = 0 \text{ when } T_j = WCN_i^n \\ \text{else } d(WCN_j, T_i) = 4 \text{ for insertion} \\ \text{else } d(WCN_j, T_i) = 3 \text{ for deletion} \\ \text{else } d(WCN_j, T_i) = 6 \text{ for substitution} \end{array} \quad (12)$$

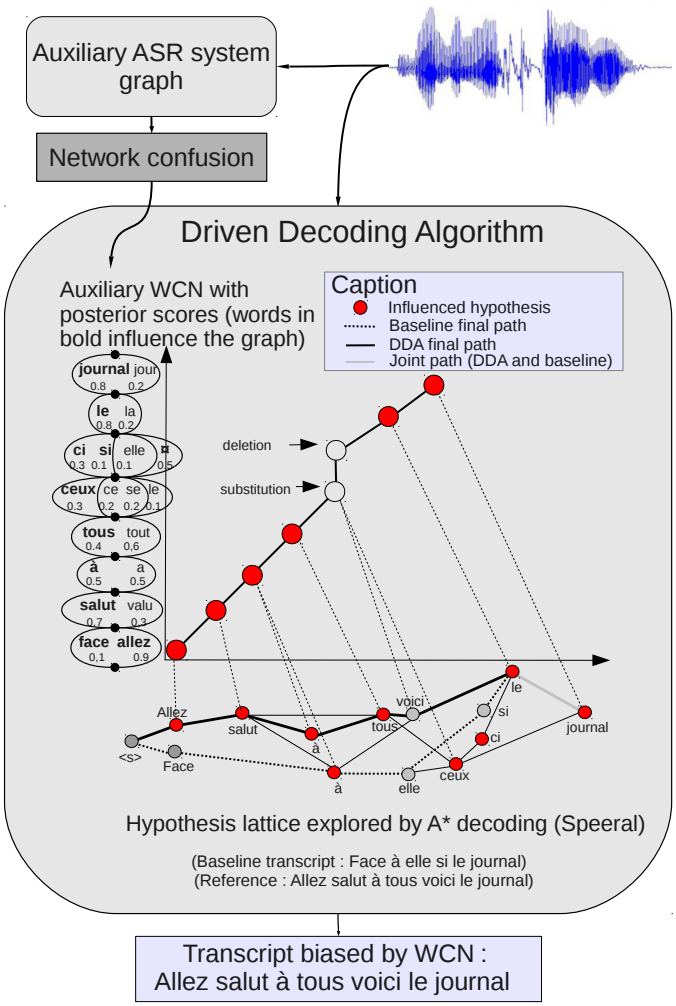

Fig. 2. DDA using WCN.

where WCN_i is the i^{th} node of the WCN and WCN_i^n is the n^{th} word of the i^{th} node. The alignment step allows to extract the best projection of the hypothesis in the network; at this point, the rescoring problem is similar to the one-best driven decoding case: linguistic probabilities are reevaluated according to WCN-to-hypothesis matching-score and word posteriors.

We tested confusion network driven decoding using confusion networks from the LIUM system. Results are reported in Table III. We observed a significant improvement compared to the single systems (1.6% absolute WER). Nevertheless, the gain with respect to the one-best driven decoder remains marginal (about -0.1% WER) for the first pass, and no gain is observed after speaker adaptation.

Three reasons might 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;
- 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. Since the confidence score is crucial for linguistic rescoring, the difference of confidence measure relevance could impact significantly the final results;
- WCN introduces too much noise: all hypotheses are credited.

In the next experiments, we use DDA in order to adapt acoustic models for the second pass.

	F. Inter	F. Info	RFI	Average
LIUM	18.5	18.9	25.6	21
DDA-LIUM-P1	17.8	18.1	22.4	19.4$_{(-1.6)}$
DDA-LIUM-P2	17.2	17.8	21.5	18.8$_{(-2.2)}$
DDA-WCN-LIUM-P1	17.7	18.1	22.3	19.3$_{(-1.7)}$
DDA-WCN-LIUM-P2	17.2	17.8	21.5	18.8$_{(-2.2)}$

TABLE III
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*)

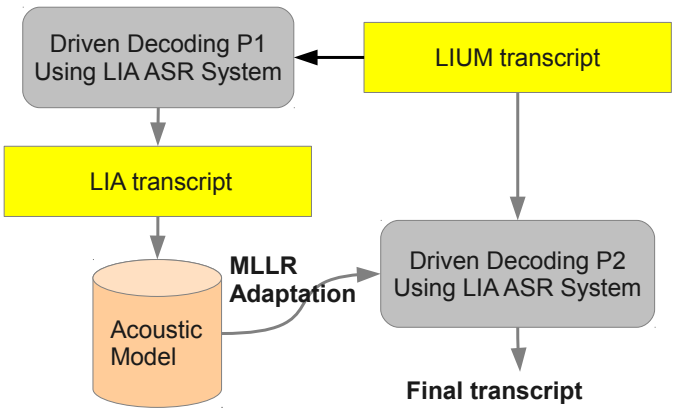

Fig. 3. Cross adaptation (DDA-P1 DDA-P2)

VI. CROSS ADAPTATION AND DRIVEN DECODING

Cross adaptation has shown to be an efficient and relatively simple method for system combination [5]. 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 the acoustic modeling complementarity among systems. 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 LIA decoding without any unsupervised adaptation (*LIA-P1*), a DDA decoding without adaptation, and a cross adaptation to transcripts followed by a LIA decoding (*LIUM-P1 LIA-P2*). Finally, we evaluate 3 acoustic adaptation strategies for the DDA system: acoustic model mapping to the transcript (*LIUM-P1 DDA-P2*), adaptation using the first pass of LIA decoding (*LIA-P1 DDA-P2*), adaptation using the DDA first pass decoding (*DDA-P1-DDA-P2*, Figure 3). Results are reported in table IV.

	F. Inter	F. Info	RFI	Average
LIUM	18.5	18.9	25.6	21
LIA-P1	22.5	23.3	26.3	24
LIA-P2	21.1	22.2	24.6	22.6
LIUM-P1 LIA-P2	20.4	21.8	24.1	22.1
DDA-lin-P1	18.1	18.7	23.6	20.1$_{(-2.0)}$
DDA-log-P1	17.8	18.1	22.4	19.4$_{(-2.7)}$
LIA-P1 DDA-lin-P2	18.1	18.4	23.1	19.9$_{(-2.2)}$
LIUM-P1 DDA-lin-P2	17.9	18.1	22.7	19.6$_{(-2.5)}$
DDA-P1 DDA-lin-P2	17.9	18.1	22.7	19.6$_{(-2.5)}$
LIA-P1 DDA-log-P2	17.8	18.0	22.5	19.4$_{(-2.7)}$
LIUM-P1 DDA-log-P2	17.5	17.9	22.0	19.1$_{(-3.0)}$
DDA-log-P1 DDA-log-P2	17.2	17.8	21.5	18.8$_{(-3.3)}$

TABLE IV
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 LIA DECODING (*LIA-P1*), DDA FIRST PASS DECODING (*DDA-P1*), AND LIA DECODING BY ADAPTING TO TRANSCRIPTS (*LIUM-P1 LIA P2*).

Performance by the DDA without speaker adaptation (*DDA-P1*) is greater than the one obtained by the initial LIA decoding (-2.7% WER) and relatively close to those obtained with the best configuration (-0.6% WER). Moreover, the cross adaptation of LIA models using the LIUM transcripts (*LIUM-P1 LIA-P2*) outperforms dramatically the classical scheme where the system is adapted using its own transcripts (*LIA-P2*, reported in the Table IV). Results show that DDA outperforms the baseline ROVER. Nevertheless, it seems clear that the gains are not cumulative: we obtain a maximum of absolute additional gain of 0.3% compared to the driven decoding with models adapted to transcripts from LIA first pass (*LIA-P1 DDA-P2*).Finally, by combining DDA and cross adaptation, we reach an absolute WER gain of 3.8% compared to the initial Speeral decoding and about 2.2% compared to the LIUM system.

In the next section we investigate DDA method to n-system combination (with $n > 2$). According to the previous results the best configuration will be used: the one-best with DDA log-linear combination.

VII. MULTI-SYSTEM COMBINATION

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.

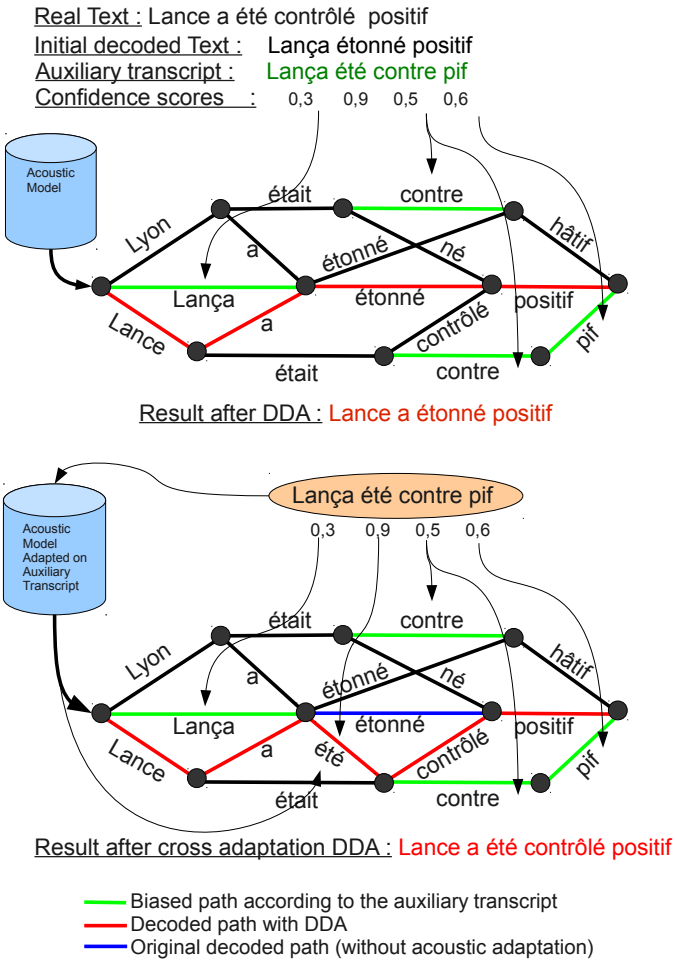

Fig. 4. Cross adaptation and DDA examples at the graph level.

In this section, the two approaches are compared, using the IRISA ASR system as the second auxiliary source. Finally, we test a last scheme where all single systems and the DDA system outputs are merged by ROVER.

A. 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. 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 reevaluated 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 [31]. IRISA confidence measures are assessed by the *posteriors*.

B. Generalized DDA

1) Integrated DDA-based combination: In this approach, the outputs of all auxiliary systems are submitted to the primary search. A matching score is computed for each transcript

	F. Inter	F. Info	RFI	Average
LIA	21.1	22.2	24.6	22.6
LIUM	18.5	18.9	25.6	21.0
IRISA	21.4	21.8	25.6	22.9
DDA-IRISA-P1	19.6	19.3	23.5	20.8*(-0.2)*
DDA-IRISA-P2	18.7	18.7	22.2	19.9*(-1.1)*
DDA-LIUM-P1	17.8	18.1	22.4	19.4*(-1.6)*
DDA-LIUM-P2	17.2	17.8	21.5	18.8*(-2.2)*

TABLE V
WORD ERROR RATES FOR DDA COMBINATION OF LIA SYSTEM WITH AN LIUM SYSTEM (*DDA-LIUM*) AND IRISA SYSTEM (*DDA-IRISA*) WITH (*P1* AND WITHOUT (*P2*) UNSUPERVISED SPEAKER ADAPTATION. EXPERIMENTS PERFORMED ON 3 HOURS OF FRENCH BROADCAST NEWS FROM THE ESTER CORPUS.

	F. Inter	F. Info	RFI	Average
ROVER-3	17.1	18.2	22.5	19.3
DDA-LIUM-P1	17.8	18.1	22.4	19.4*(+0.1)*
DDA-LIUM-P2	17.2	17.8	21.5	18.8*(-0.5)*
2-Level DDA-ROVER	16.8	17.3	21.3	18.5*(-0.8)*
DDA-3	16.5	16.9	20.6	18.0*(-1.3)*
DDA-3+ROVER	15.9	16.4	20.7	17.7*(-1.6)*

TABLE VI
WER 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).

according to its synchronization to the hypothesis. Finally, all linguistic scores are merged by the log-linear combination extended to n systems:

$$\tilde{P}(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\beta} \prod_{k=1}^{N} \alpha_{\delta_k}(w_i)^{\beta_k} \quad (13)$$

where β_k is the weight of the system k: $\sum_{k=1}^{N} \beta_k = 1$, $\alpha_{\delta_k}(w_i)$ are the posteriors of the words w_i provided by the system k and N the number of auxiliary systems.

2) Two-Level ROVER-DDA combination: The principle of the 2-level scheme relies on a first merging step where all auxiliary transcripts are merged. In our experiment, we use ROVER for merging auxiliary system outputs (using the *avg-conf* method). The resulting transcript is then used following the classical scheme of a 2-system DDA combination with LIA as primary system.

C. Results

Baseline results are reported in Table V for each auxiliary system combined with LIA system 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 the IRISA system (about 1% absolute WER), the latter exhibiting a higher error rate. Nevertheless, the combination with the IRISA system still improves significantly the initial LIA system performance.

Table VI compares results obtained by the different fusion strategies. First, we observe that the addition of the third system improves systematically the system accuracy. Nevertheless, the ROVER of the 3 single-systems (ROVER-3) obtains results that are close to the best 2-system combination (DDA-LIUM P2). The 2-level method (2-Level DDA-ROVER) provides a more significant WER decrease (0.9% better than DDA-LIUM P1 and 0.3% better than DDA-LIUM P2), but this configuration remains significantly worse than the full DDA approach (DDA-3) in a 3-system configuration (additional gain of -0.5%WER). The last combination method consists in merging all available system outputs including DDA-3 (DDA-3+ROVER). This hybrid method further improves the system accuracy by about 0.3% absolute WER. Globally, the optimal DDA-based combination outperforms both the best single system LIUM (of about 3.3% absolute) and the classical ROVER combination of the 3 single-systems (-1.6% absolute WER).

The results obtained confirm the interest of fusing all information sources into the primary search algorithm. This enables the evaluation of competing hypotheses while taking into account all the constraints and knowledge available. Moreover, accuracy benefits substantially from the addition of the Irene system. Indeed, compared to LIA+LIUM DDA configuration, the relative WER gain is about 7%. The final WER gain is about 4.6 points compared to the single primary system, and to 3.0 points compared to performance of the LIUM system, which is the best single system. The best gain is obtained on RFI show, which is the more difficult session (5.0 points less than to the LIUM-RT, and 4.0 points less than the LIA WER).

D. Driven decoding analysis

1) Oracle DDA: The analysis of multiple system driven decoding is completed, by conducting experiments similar to those proposed in Section V, in order to study the behavior of DDA in this configuration. 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 outputs (*ORACLE DDA+ROVER*) we show in Table VII that linguistic rescoring allows to guide the search toward alternative paths: the *oracle* WER improvement means that DDA outputs introduce new correct words; this point confirms that DDA may not be considered as an on-line voting method. However, it is really an integrated approach where additional information is integrated to the global cost function, allowing a new exploration of the search graph.

2) DDA+ROVER: Moreover, we note that ROVER combination of DDA-3 and all single-systems outperforms the pure DDA approach. This result demonstrates that while DDA finds new correct hypotheses, it also deletes some correct ones compared to single-systems. This demonstrates that DDA may benefit from more efficient tuning in order to systematically select more good hypotheses from the auxiliary transcripts.

3) WER at segment level: In this section, we analyze the driven decoding algorithm behavior at the segment level.

	F. Inter	F. Info	RFI	Average
ORACLE-3	10.3	10.5	14.5	11.8
ORACLE DDA-3+ROVER	9.8	10.0	13.6	11.1

TABLE VII
COMPARISON OF 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-3+ROVER*). THE *oracle* WER IMPROVEMENT MEANS THAT DDA OUTPUTS INTRODUCE NEW CORRECT WORDS.

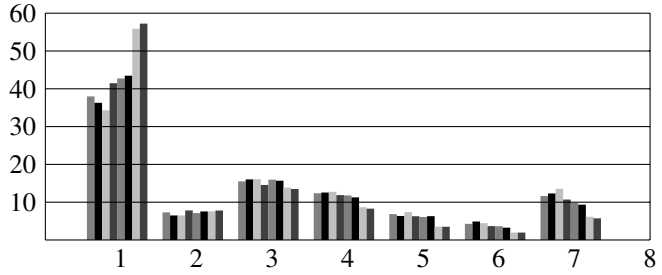

Fig. 5. Segment distribution for each WER range. Results are presented in 7 WER ranges: 0-5, 5-10, 10-20, 20-30, 30-40, 40-50 and 50-100. Systems are presented in the following order: baseline system (i.e. the best single ASR system: LIUM), the LIA system, the IRISA system, the ROVER 3-systems, DDA LIA+LIUM+IRISA, the ROVER 3-systems+DDA, the *oracle* 3-systems, the *oracle* 3-systems-DDA.

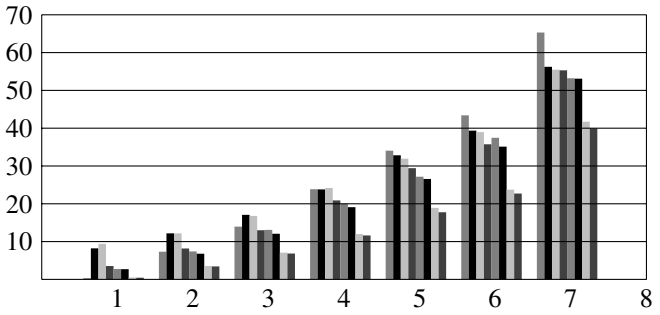

Fig. 6. WER compared to the baseline: segment classification is based on baseline. Results are presented in 7 WER ranges: 0-5, 5-10, 10-20, 20-30, 30-40, 40-50 and 50-100. Systems are presented in the following order: baseline system (i.e. the best single ASR system: LIUM), the LIA system, the IRISA system, the ROVER 3-systems, DDA LIA+LIUM+IRISA, the ROVER 3-systems+DDA, the *oracle* 3-systems, the *oracle* 3-systems-DDA.

According to the reference segments, we sort each segment according to its WER. We present two analyses which compare the systems at different levels. The first analysis introduces the WER distribution into the same ranges. The second one shows the relative WER for each range based on the WER of the baseline system.

In order to compute WER ranges we use the reference segmentation. Then the WER is computed for each segment. Results are presented in the WER ranges 0-5, 5-10, 10-20, 20-30, 30-40, 40-50 and 50-100. In the distribution based analysis, each segment is assigned into a range according to its score. In the analysis based on the relative WER, the segment classification is according to the baseline (i.e., the best single system LIUM): this allows one to compare the WER evolution inside the segments.

Figure 5 presents the segment distribution according to the WER range: We observe that DDA have a significant impact on the 4 last ranges and the first range. Its behavior is similar to that of the ROVER. The increase on the first range show that DDA is able to improve significantly the ASR output quality, by decreasing WER in very degraded areas.

Figure 6 presents the WER improvement compared to the baseline. This graph allows one to observe the systems complementarity. According to the best system, similar segments in other systems show better WER. Only the range 0-5 presents some degradations, but in this range segments are often very small: an error have a very big WER impact at the segment level. In all other ranges DDA produces better results than the best single ASR system. The 4 last ranges exhibit about 20% WER improvement, while the ranges 5-10 and 10-20 present respectively 7% and 13% relative WER improvement. We observe also that the combination ROVER-DDA yields higher improvements in the 3 first ranges. While the DDA allows one to correct degraded segments, rover allows one to improve well decoded segments. These results explain why DDA has a big impact on the most difficult show RFI and why a slight degradation was observed on the same show when DDA was combined with ROVER.

VIII. CONCLUSION

We have proposed an algorithm for driven-by-transcript decoding (DDA). This method allows an efficient combination of ASR systems, by rescoring linguistic probabilities according to auxiliary transcripts and word posteriors from auxiliary systems. Once the auxiliary transcripts synchronized to the ASR search algorithm, the current decoding hypothesis is reassessed by exploiting the confidence scores of the auxiliary transcripts. Different configurations were evaluated on the ESTER evaluation corpus. Different strategies have been considered: a simple combination using one pass, an *a priori* combination by ROVER, a combination using two passes with DDA and two DDA-pass followed by ROVER between all systems. We also experimented several cross-adaptation strategies between systems. The results show that the DDA allows a significant reduction in word error rate. The best configuration is based on two DDA passes followed by ROVER. Moreover, our experiments show that the combination with DDA using the one-best auxiliary hypothesis performs better than the DDA WNC-based system. Finally, the integration of several auxiliary systems (instead of one) brings an additional substancial reduction and significantly exceeds the combination of 3 ROVER system. Finally, using the DDA with a final pass in ROVER we get an overall reduction of 3 points of WER (14.5%) compared to the best selected baseline system. The reduction observed over a ROVER therefore shows that DDA is not always an optimal combination. Nevertheless, this gain is higher than those cited in the literature using either ROVER or CNC.

REFERENCES

[1] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, March 18–23, 2005, pp. 457–460.

[2] L. Barrault, C. Servan, D. Matrouf, G. Linarès, and R. De Mori, "Frame-based acoustic feature integration for speech understanding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 4997–5000.

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

[4] H. Nock and S. Young, "Loosely coupled HMMs for ASR," in *ICSLP*, 2000.

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

[6] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, "Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end," in *Conference on Speech Communication and Technology, Interspeech, Pitsburg, USA*, 2006.

[7] L. Burget, "Complementarity of speech recognition systems and system combination," Ph.D. dissertation, FIT VUT, Brno,CZ, 2004.

[8] O. Siohan, O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, B. Ramabhadran, Ed., vol. 1, 2005, pp. 197–200.

[9] C. Breslin and M. Gales, "Complementary system generation using directed decision trees," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, vol. 4, 15–20 April 2007, pp. IV–337–IV–340.

[10] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 347–354.

[11] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum bayes-risk ASR voting strategies," in *ICSLP*, 2000.

[12] D. Hillard, B. Hoffmeister, M. Ostendorf, R. Schlüter, and H. Ney, "iROVER: Improving system combination with classification," in *HLT*, 2007.

[13] R. Zhang and A. Rudnicky, "Investigations of issues for using multiple acoustic models to improve continuous speech recognition," in *ICSLP*, 2006.

[14] H. Schwenk and J.-L. Gauvain, "Combining multiple speech recognizers using voting and language model information," in *ICSLP*, 2000.

[15] A. Sankar, "Bayesian model combination (baycom) for improved recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, March 18–23, 2005, pp. 845–848.

[16] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation, and system combination," in *In Proceedings NIST Speech Transcription Workshop,College Park P. (2000a).*, 2000.

[17] I.-F. Chen and L.-S. Lee, "A new framework for system combination based on integrated hypothesis space," in *Interspeech'06-ICSLP*, Pittsburgh, Pensylvania, USA, 2006.

[18] B. Lecouteux, G. Linarès, Y. Estève, and J. Mauclair, "System combination by driven decoding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, G. Linares, Ed., vol. 4, 2007, pp. IV–341–IV–344.

[19] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 1549–1552.

[20] B. Lecouteux, G. Linarès, and S. Oger, "Integrating imperfect transcripts into speech recognition systems for building high-quality corpora," *Computer Speech & Language*, vol. 26, pp. 67–89, 2012.

[21] D. Tax, R. Duin, and M. Breukelen, "Comparison between product and mean classifier combination rules," in *Workshop on Statistical Techniques in Pattern Recognition*, 1997.

[22] P. Nocéra, C. Fredouille, G. Linarès, 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.

[23] D. Massonié, P. Nocéra, and G. Linarès, "Scalable language model look-ahead for LVCSR," in *InterSpeech'05, Lisboa, Portugal*, 2005.

[24] P. Placeway and J. Lafferty, "Cheating with imperfect transcripts," in *Proc. Fourth International Conference on Spoken Language ICSLP 96*, J. Lafferty, Ed., vol. 4, 1996, pp. 2115–2118 vol.4.

[25] 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.

[26] 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 *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

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

[28] B. Lecouteux, G. Linarès, and B. Favre, "Combined low-level and high-level features for out-of-vocabulary word detection," in *International conference of the Speech Communication Association, ISCA, Inter-Speech'09, Brighton, UK*, 2009.

[29] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, "The 1996 Hub-4 Sphinx-3 system," in *Proceedings of the 1997 ARPA Speech Recognition Workshop, pp. 85-89*, 1997.

[30] 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.

[31] S. Huet, G. Gravier, and P. Sébillot, "Morpho-syntactic post-processing of n-best lists for improved French automatic speech recognition," *Computer Speech and Language*, 2010.