# Bag of n-gram driven decoding for LVCSR system harnessing

Fethi Bougares [#1], Yannick Estève [#2], Paul Deléglise [#3],Georges Linarès [*4]

[#] *LIUM, University of le Mans, France*
[1,2,3] firstname.lastname@lium.univ-lemans.fr

[*] *LIA, University of Avignon, France*
[4] firstname.lastname@univ-avignon.fr

***Abstract*—This paper focuses on automatic speech recognition systems combination based on driven decoding paradigms. The driven decoding algorithm (DDA) involves the use of a 1-best hypothesis provided by an auxiliary system as another knowledge source in the search algorithm of a primary system. In previous studies, it was shown that DDA outperforms ROVER when the primary system is guided by a more accurate system. In this paper we propose a new method to manage auxiliary transcriptions which are presented as a bag-of-n-grams (*BONG*) without temporal matching. These modifications allow to make easier the combination of several hypotheses given by different auxiliary systems. Using *BONG* combination with hypotheses provided by two auxiliary systems, each of which obtained more than 23% of WER on the same data, our experiments show that a CMU Sphinx based ASR system can reduce its WER from 19.85% to 18.66% which is better than the results reached with DDA or classical ROVER combination.**

***Index Terms*—speech recognition, system combination, bag of n-gram driven decoding.**

## INTRODUCTION

One of the main challenges in automatic speech recognition (ASR) researches are to get accurate system working in real-life situations and with different kind of speaking styles. To achieve this goal, studies have taken many directions to look for better models or more sophisticated algorithms, meanwhile many works propose different combination schemes to benefit from systems complementarity. In previous studies, a variety of combination approaches were proposed. These combination schemes are distinguishable depending on the method used to share information and the application levels. Cross-adaptation techniques [1] and feature concatenation [2] are two examples of combination before the decoding process, while ROVER [3], lattice combination [4] and CNC [5] operate after. The DDA[6] framework is more than a combination method: applied during the decoding process, this method modify search space exploration and brought out new hypothesis not proposed by initial system.

In order to keep the search space at a manageable size, the recognition process prunes many hypotheses according to its knowledge base and its internal heuristics. The pruning process is generally local and local information is used to reject some word hypotheses. But these rejected words can be the words uttered by the speaker, and could be retained in a more global pruning process: a better pruning method could give more accurate search. Motivated by these considerations, we have chosen to explore the use of the DDA. This algorithm takes into account the output given by an auxiliary ASR system to evaluate a partial hypothesis during the decoding process of a primary system. DDA helps improving the internal pruning decision made by the primary ASR system using the output of another recognizer.

In a previous work [6], it was shown that the DDA approach gives good results in system combination. It significantly improves the output of the primary ASR system when the auxiliary system is initially better.

In this paper, we introduce the bag-of-n-gram driven decoding approach as modified DDA combination. Experimental results show that we can improve a primary ASR system and outperform DDA when using less efficient single auxiliary ASR system. Additionally, an efficient method is proposed to deal with multiple auxiliary ASR system. The first section presents the principle of DDA. Experimental framework is then presented in the section two. In the third section we investigate the DDA algorithm when primary system is more accurate than auxiliary. Before concluding along with future work, section four introduces the *BONG* method, obtained results, and their analysis.

## I. DRIVEN DECODING ALGORITHM

DDA is presented in [6] as a speech recognition system combination method. Initially DDA was proposed in [7] to help ASR systems process audio documents associated to imperfect manual transcripts (for example subtitles). This method is based on linguistic score reevaluation during the decoding process in a primary system using a recognition hypothesis computed by an auxiliary system. During the decoding process, each evaluated hypothesis is aligned to the auxiliary hypothesis using the edit distance. After finding a synchronized point, a matching score $\alpha$ is estimated depending of the number of words correctly aligned. Then the linguistic score L is computed using the following rule:

$$L(w_i/w_{i-2}, w_{i-1}) = P(w_i/w_{i-2}, w_{i-1})^{1-\alpha(w_i)}$$

where $P(w_i/w_{i-2}, w_{i-1})$ is the initial probability of the trigram and $\alpha(w_i)$ is the DDA matching score depending on

the similarity between the current and the auxiliary hypothesis in a finite-size window and depending to the confidence measures associated to the word contained in this window in the auxiliary hypothesis. $\alpha(w_i)$ is computed according to the following rules :

$$\alpha(w) = \begin{cases} \frac{\phi(w1)+\phi(w2)+\phi(w3)}{3} & \text{if } (w_1..w_3) = (hw_1..hw_3) \\ \frac{\phi(w1)+\phi(w2)}{2} & \text{if } (w_1, w_2) = (hw_1, hw_2) \\ \phi(w1) - \gamma & \text{if } (w_1) = (hw_1) \text{ and } \phi(w_1) \geq \gamma \\ 0 & \text{if } \phi(w_1) < \gamma \end{cases}$$

where $\phi(w_i)$ is the word confidence measure of $w_i$ and $\gamma$ is a confidence threshold which is a priori fixed.

In [8], a generalized driven decoding is presented. This method investigates the benefit of using more information from auxiliary system presented as confusion network (CN).

## II. Experimental framework

The ASR systems used in this study are ASR systems which participated at the ESTER [9] evaluation campaign. These systems were developed by 3 different French laboratories (IRISA, LIA, LIUM).

### A. Experimental data

Experiments were carried out using three different shows made by three different French radio stations : *France Inter*, *France info* and *RFI*. Each show contains one hour of broadcast news extracted from the official ESTER development set [10]. Manual transcriptions of these 3 hours audio recordings contain 36K words. A *leave-one-out* method was used to make the experiments.

### B. ASR systems

The LIUM ASR system was built by integrating and modifying tools coming from the CMU Sphinx project [11]. This ASR system was the best open source system during the ESTER 1 [10] and ESTER 2 [9] French evaluation campaigns on processing French radiophonic broadcast news. The LIUM ASR system used in this paper is a 2-passes system using 39 dimensional features (PLP) with energy, delta and double-delta. The first pass proposes a one-best recognition hypothesis computed with a 3-gram LM and acoustic models adapted on gender and acoustic band (studio *vs.* telephone). Gender and band information are provided by the LIUM speaker diarization system [12]. The one-best output of the first pass is used to compute constrained MLLR (CMLLR) transformation for each speaker [13]. The second pass uses this CMLLR speaker adaptation to refine the speech recognition process. The two first passes are processed by a modified Sphinx3 decoder. A full description of the LIUM ASR system can be found in [11].

The LIA ASR system [14] relies on the Speeral decoder and the Alize segmenter. A particularity of the Speeral decoder is that is based on an A* search algorithm applied to phoneme lattice. 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.

The IRISA ASR system [15] is based on word-synchronous beam-search algorithm with HMM acoustic modeling and n-gram linguistic models with a vocabulary of 64k words. This 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 4-gram LM and context-dependent models.

The three hours of audio data were initially transcribed by each ASR system separately. Table I summarizes the word error rate (WERs) for each system.

TABLE I
Word error rates of the baseline primary system (LIUM) and the auxiliary systems (LIA and IRISA)

| System | F.Inter | F.Info | RFI |
|---|---|---|---|
| LIUM-base | **19.34** % | **17.92** % | **22.59** % |
| LIA | 22.52 % | 21.97 % | 24.95 % |
| IRISA | 21.96 % | 21.61 % | 26.03 % |

In our experiments, the LIUM ASR system is considered as the primary system. The LIA and IRISA are the auxiliary systems.

## III. One-best hypothesis driven decoding

So far, DDA was used to combine a primary ASR system using outputs of one or several more accurate systems. Furthermore, the generalized DDA uses Confusion Network as structure of auxiliary hypotheses [8] in order to represent hypotheses coming from several auxiliary systems. Such structure increases significantly computational cost for DDA alignment without WER reduction in comparison to one-best driven decoding. In this section we propose to investigate this combination when the primary system has better initial performance. Then we propose an efficient new generalized driven decoding without significant additional computational cost.

Firstly, the DDA combination is implemented within the LIUM ASR system as presented in section I. Each one-best auxiliary hypothesis is aligned to the speech segmentation proposed by the primary system. At this stage we use temporal information given by auxiliary system instead of using DTW during the alignment process. The DTW algorithm was really useful for the initial objectives of the DDA which consisted of helping ASR system to process audio documents associated to imperfect transcripts (for example subtitle). For an ASR system combination task, exact alignment could be used since each word coming from auxiliary hypotheses is provided with

a temporal information which are directly useful. In addition DTW, contrarily to exact alignment, computes an optimal alignment to minimize the global cost defined by similarity measure which could have some wrong local alignments.

Secondly, a decoding process is performed for each speech segment and auxiliary hypotheses are used to compute linguistic score re-evaluation.

Table II report results of DDA combination on LIUM ASR system by using IRISA as an auxiliary system. Significant improvements are obtained on F.Info and F.Inter with marginal gain on RFI show.

TABLE II
WER BY RADIO FOR DDA COMBINATION IN LIUM ASR SYSTEM USING IRISA AS AN AUXILIARY SYSTEM.

| System | F.Inter | F.Info | RFI |
|---|---|---|---|
| LIUM-base | 19.34% | 17.92% | 22.59% |
| IRISA | 21.96% | 21.61% | 26.03% |
| LIUM DDA-P2-IRISA | **18.94%** | **17.59%** | **22.54%** |

In order to understand the behavior of DDA combination as implemented in LIUM ASR system, we analyze combination output. This analysis show that temporal alignment introduce error because of difference in words boundaries between combined ASR system. An experiment is performed in order to measure the impact of the constraint relaxation on the temporal alignment. When the constraint is strictly applied (DDA), the current word $w_i(pri)$ evaluated in the primary system during the decoding process is only compared to the word $w_j(aux)$ observed in the auxiliary hypotheses if they overlap. When this constraint is relaxed, the word $w_i(pri)$ can be compared to some words around $w_j(aux)$. For example, when the relaxation margin is equal to 1, $w_i(pri)$ is compared to $w_j(aux)$, but also to $w_{j-1}(aux)$ and $w_{j+1}(aux)$ respectively the predecessor and the successor of $w_j(aux)$ in the auxiliary hypothesis. A correspondence is complete if the $w_i(pri)$ is equal to $w_k(aux)$ and if their history are the same. In our experiments, we fixed the history to the two previous words: we looked for 3-gram correspondences.

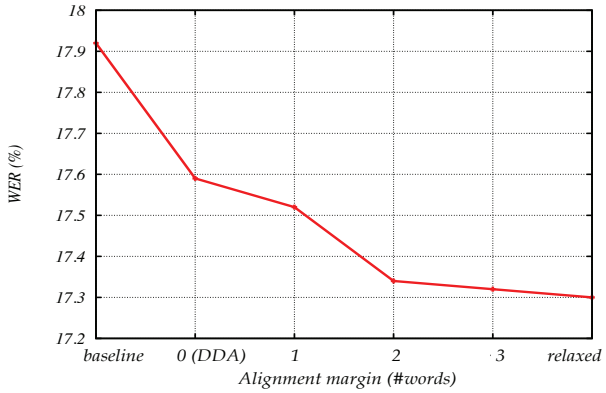


Fig. 1. *WER variation depending of alignment constraint relaxation on the France Info show, using the IRISA ASR system as the auxiliary system. DDA correspond to nil alignment margin value.*

Figure 1 presents the word error rate as a function of alignment constraint relaxation on the *France Info* show when the IRISA ASR system is used as the auxiliary system. With DDA combination (alignment margin = 0) we obtained an absolute gain of 0.33% WER . The lowest word error rate is obtained when the relaxation is maximal (0.62% of absolute WER). The maximal relaxation of temporal alignment permits to overcome the problem of differences at word boundaries and internal signal representation of each combined system.

## IV. BAG-OF-NGRAMS DRIVEN DECODING

Following the result obtained in previous section, no temporal constraint will be applied in the next experiments: when the 3-gram $(w_{i-2}w_{i-1}w_i)(pri)$ is evaluated during the decoding process, its matching score will depend only on the existence of this 3-gram in the auxiliary hypothesis, whatever its position. This seems to be reasonable due to the size of speech segments which contains rarely more than about twenty words.

Compared to original DDA, in our combination schema there is no alignment process which makes combination faster especially when we have to deal with many auxiliary systems. In addition all auxiliary hypotheses are presented as a *BONG* with $n = 3$ at segment level which allows faster search on auxiliary transcription.

### A. single auxiliary system

TABLE III
WER BY RADIO FOR *BONG* COMBINATION IN LIUM ASR SYSTEM USING LIA AND IRISA SEPARATELY AS AN AUXILIARY SYSTEM.

| System | F.Inter | F.Info | RFI |
|---|---|---|---|
| LIUM-base P1 | 20.93% | 20.33% | 25.21% |
| LIUM-base P2 | 19.34% | 17.92% | 22.59% |
| LIA | 22.52% | 21.97% | 24.95% |
| LIUM BONG-P1-LIA | 20.17% | 19.85% | 23.53% |
| LIUM BONG-P2-LIA | **19.04%** | **18.01%** | **21.17%** |
| IRISA | 21.96% | 21.61% | 26.03% |
| LIUM BONG-P1-IRISA | 19.74% | 19,39% | 23.16% |
| LIUM BONG-P2-IRISA | **18.47%** | **17.30%** | **21.38%** |

Table III shows the word error rates obtained for each radio show according to the auxiliary hypothesis used with *BONG*. Results are given for pass 1 (P1) and pass 2 (P2) of the primary ASR system. For each one, the *BONG* outperforms the baseline performance. The global gain obtained with *BONG* is presented in Table IV. The best improvement with *BONG* combination is reached when the IRISA system is used as the auxiliary system with 18.96% of global WER in pass 2 instead of 19.85%. Although LIA and IRISA systems have same initial WER, the combination is better with IRISA system. This can be due to the IRISA post-processing linguistic step.

These results shows that it is possible to improve system performance by using **BONG** driven decoding even with a less accurate auxiliary system.

### B. Multiple auxiliary systems

In the previous experiments, auxiliary hypotheses can be considered as bags of 3-grams. For each speech segment, we

TABLE IV
GLOBAL WER FOR *BONG* COMBINATION IN LIUM SYSTEM USING LIA AND IRISA SYSTEMS SEPARATELY AS AUXILIARY SYSTEMS. (BONG-P1 WHEN COMBINATION IS MADE IN THE FIRST PASS AND BONG-P2 WHEN COMBINATION IS MADE IN PASS 1 AND 2)

| | Global | |
|---|---|---|
| System | LIA-aux | IRISA-aux |
| LIA | 23.06% | – |
| IRISA | – | 23.07% |
| LIUM-base P1 | *22.03%* | *22.03%* |
| LIUM BONG-P1 | 21.09% (-0.97) | 20.66% (-1,37) |
| LIUM-base P2 | *19.85%* | *19.85 %* |
| LIUM BONG-P2 | **19.34%** (-0.51) | **18.96%** (-0.89) |

propose to take into consideration the hypotheses coming from the two auxiliary systems by merging all the 3-grams observed in each auxiliary hypotheses into the same bag of 3-grams: in the next experiment, this bag of 3-grams will be used as the auxiliary hypothesis during the decoding process of the primary system in the same way as the one used to deal with single auxiliary systems.

TABLE V
GLOBAL WER ACCORDING TO COMBINATION SCHEMA: THE BASELINE ROVER COMBINATION OF THE 3 SINGLE SYSTEMS (ROVER-3), THE *BONG* COMBINATION WITH MULTIPLE AUXILIARY SYSTEM (LIUM-BONG-IRISA-LIA-P1-P2) AND THE ROVER COMBINATION FOR ALL SYSTEMS (BONG+ROVER)

| System | Global |
|---|---|
| LIUM-base P1 | 22.03% |
| LIUM-BONG-IRISA-LIA-P1 | **20.48**% (-1.55) |
| LIUM-base P2 | 19.85 % |
| ROVER-3 | 18.91% |
| LIUM-BONG-IRISA-LIA-P1-P2 | **18.77** % (-1.08) |
| BONG+ROVER | **18,66** % (-1.19) |

Results presented in the table V show that the word error rate reached with the use of multiple auxiliary systems is lower than the lowest word error rate obtained with the use of a single auxiliary system. The *BONG* combination reaches an accuracy improvement of 0.74 % relative points compared to ROVER. Even if the gain is limited, an inter-system comparison (BONG-IRISA-LIA and Rover-3) was computed with the NIST *sc_stat* tool, by doing a Matched Pairs Sentence-Segment Word Error (MAPSSWE) test. It indicates that the improvement with *BONG* combination is statistically significant at the level of $p < 0.001$.
The *BONG* combination generates two additional outputs which can be added to the initial ROVER combination. This combination (BONG+ROVER) provides more accurate output with 18.66 % WER. This amelioration is due to the fact that *BONG* combination generates new word hypotheses not present neither in the baseline primary system nor in the auxiliary systems.

### C. BONG combination analysis

Since our combination is performed at the segment level, our method efficiency is evaluated segment by segment: we divide segments in 11 different classes according to their baseline WER: for instance 0-10 class contains segments with baseline WER belongs to the interval $]0, 10]$. Figure 2 presents the impact of this combination in each WER class. The x-axis represents the WER classes while y-axis represents the *BONG* combination impact and is calculated using this formula :

$$\frac{(WER_{baseline} - WER_{combination})\ \#word_c}{\#word_t}$$

where $WER_{baseline}$ and $WER_{combination}$ represent the WER respectively before and after the combination; and $\#word_c and\ \#word_t$ represent the number of words in the class and the total number of words respectively.

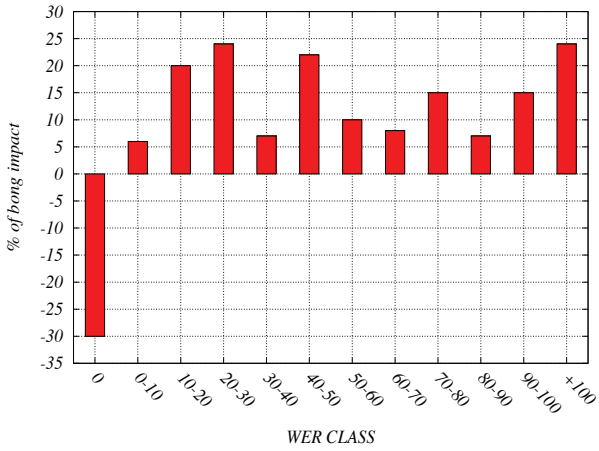


Fig. 2. *BONG combination impact by WER class.*

As shown in Figure 2, the combination effect depends of initial performance: when the primary system has perfect transcription (WER class = 0), the auxiliary system has negative impact. Following this analyze, if the *BONG* combination is applied only when primary system is not doing well, we can avoid unwanted impact and improve final WER. In future work we plan to add a decision module to our framework in order to determine for each segment if we must apply *BONG* or not. The decision may be taken using the primary system confidence measure.

## V. CONCLUSION

In this paper, we have proposed the *BONG* combination method. Inspired from DDA, this combination method computes matching score without temporal alignment. Furthermore, when dealing with utterances produced by an automatic segmentation process (about 20s per segment), no temporal alignment is necessary and this make the *BONG* framework more efficient than original DDA: using bags of 3-grams contained in 1-best hypotheses provided by an auxiliary system, is sufficient to reduce the word error rate of an ASR system. This approach have some analogies with the cache language model paradigm by boosting some n-grams already observed. In the *BONG* framework, these n-grams

are not in the history, but provided by other systems. Our approach was successfully extended to multiple auxiliary systems: experiments show that the *BONG* combination permits to reduce the word error rate of an accurate ASR system by using auxiliary hypotheses provided by really less performant systems. Moreover, the BONG+ROVER approach improves the ROVER combination by adding a new recognition hypothesis which contains complementary information.

Since the *BONG* combination method make driven decoding simpler and efficiently generalizable, we plane to integrate more information in the combination framework. This framework could be extended either by sharing more than one-best hypotheses or by including other auxiliary systems.

## REFERENCES


[1] X. Liu, M. Gales, and P. Woodland, "Language model cross adaptation for lvcsr system combination," in *Interspeech*, 2010.

[2] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, "Trapping conversational speech: Extending trap/tandem approaches to conversational telephone speech recognition," 2004.

[3] J. Fiscus, "A post-prcessing system to yield reduced word error rates : recogniser output voting error reduction (rover)," in *ASRU*, 1997, pp. 347–354.

[4] L. Xiang and R. S. R. M. Stern, "Lattice combination for improved speech recogniton," September, 2002.

[5] 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*, vol. 14, no. 4.

[6] B. Lecouteux, G. Linarès, Y. Estève, and J. Mauclair, "System combination by driven decoding," in *ICASSP*, 2007.

[7] B. Lecouteux, G. Linarès, P. Nocera, and J. Bonastre, "Imperfect transcript driven speech recognition," in *ICSLP /Interspeech*, Pittsburgh, Pennsylvania, USA, 2006.

[8] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *ICASSP*, Las Vegas, Nevada, USA, 2008.

[9] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," in *Interspeech*, Brighton, Royaume-Uni, Septembre 2009.

[10] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Interspeech 2005*, Lisbon, Portugal, September 2005.

[11] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?" in *Interspeech*, Brighton, UK, September 2009.

[12] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, Texas, USA, 2010.

[13] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Cambridge University Engineering Department, Tech. Rep., May 1997.

[14] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J. Bonastre, D. Massoni, and F. Bchet, "The lia's french broadcast news transcription system." in *SWIM: Lectures by Masters in Speech Processing*, Maui, Hawaii, 2004.

[15] S. Huet, G. Gravier, and P. Sbillot, "Morpho-syntactic post-processing with n-best lists for improved french automatic speech recognition," *Computer Speech and Language*, vol. 24, no. 4, pp. 663–684, 2010.