# The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news

Paul Deléglise, Yannick Estève, Sylvain Meignier, Teva Merlin

# The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news

*Paul Deléglise, Yannick Estève, Sylvain Meignier, Teva Merlin*

LIUM/CNRS
Université du Maine, Le Mans, France
{paul.deleglise,yannick.esteve,sylvain.meignier,teva.merlin}@lium.univ-lemans.fr

## Abstract

This paper presents the system used by the LIUM to participate in ESTER, the french broadcast news evaluation campaign. This system is based on the CMU Sphinx 3.3 (fast) decoder. Some tools are presented which have been added on different steps of the Sphinx recognition process: segmentation, acoustic model adaptation, word-lattice rescoring.
Several experiments have been conducted on studying the effects of the signal segmentation on the recognition process, on injecting automatically transcribed data into training corpora, or on testing different approaches for acoustic model adaptation. The results are presented in this paper.
With very few modifications and a simple MAP acoustic model estimation, Sphinx3.3 decoder reached a word error rate of 28.2%. The entire system developed by LIUM obtained 23.6% as official word error rate for the ESTER evaluation, and 23.4% as result of an unsubmited system.

## 1. Introduction

The ESTER evaluation campaign of french radiophonic broadcast news [1] has allowed to stimulate research on speech recognition in French. This evaluation campaign is similar to the Rich Transcription evaluation organized by the NIST, in terms of the tasks offered and of evaluation rules. It offers a significant amount of train and test data, giving the participants the opportunity to develop robust systems.

The Laboratoire d'Informatique de l'Université du Maine (LIUM) participated in the tasks of automatic transcription and speaker diarization, as well as the prospective task of named entity detection. For the first two tasks, LIUM ranked second. This paper presents the automatic transcription system developed by the LIUM for this campaign. The development was based on the CMU Sphinx Project which provides high quality tools for speech recognition. We also present the experiments which validated our technical choices.

## 2. CMU Sphinx III

CMU Sphinx Project has been funded for many years by the DARPA to develop a robust speaker-independent large vocabulary continuous speech recognizer. Since 2000, first with the CMU Sphinx II decoder and then with SphinxTrain and CMU Sphinx III decoders, a large part of the CMU Sphinx Project has been made available as open-source packages by Carnegie Mellon University. The fast decoder [2, 3] called s3.3 from the CMU Sphinx III family was used in the work reported in this paper.

### 2.1. The fast decoder

The fast decoder s3.3 is a branch from CMU Sphinx III project. This branch has been developed to include some speed improvements such as sub-vector clustered acoustic models [2] or use of few static lexical trees. This decoder uses only fully continuous acoustic models with 3 or 5-state left-to-right HMM topologies. Only bigram or trigram language models can be used. At last, the vocabulary size is limited to about 65K words.

### 2.2. Added features

Although the tools distributed in the CMU Sphinx open-source package reach a high level of quality, they can be supplemented or improved to integrate some state-of-art technologies. Until now, we have focused on adaptation of acoustic models and on word-lattice rescoring.

#### *2.2.1. SAT*

We completed the SphinxTrain and decoder modules with a Speaker Adaptive Training (SAT) procedure based on CMLLR [4].

The CMLLR transformation is a block-diagonal matrix composed of 3 blocks of $13 \times 13$ coefficients. Classicaly, CMLLR can be computed either on a sentence-by-sentence basis or on a speaker-by-speaker basis.

In the first case, diagonal initialization is very important because of the lack of data: we based the initialization on a solution[1] proposed by [4].

#### *2.2.2. 4-gram lattices rescoring*

The last release of the fast decoder, CMU Sphinx 3.5, is distributed with a tool to rescore word-lattices (generated by the decoder) with trigram language models. In fact, it is more interesting to rescore this lattice with an higher order language model: we have modified the provided tool to make it able to use quadrigram language models. This modification implies first a lattice pruning to avoid a combinative explosion: to suppress some transitions, the *a posteriori* probability of each transition is computed and only the most probable transitions are preserved for a given frame[2]. Then a exploratory search is processed using a quadrigram language model.

[1] More precisely this initialization relies on equation 64 of [4] but differs from equations 66 to 76.

[2] On average, 20 transitions are kept per frame.

## 3. Signal processing

Cepstral features are classical: 13 Mel frequency cepstra are computed for each window of 25ms with an overlap of 10ms, completed with delta and delta-delta. Two sets of features are computed for each show, corresponding to broadband (130Hz - 6800Hz) and narrowband (440Hz - 3500Hz) analysis.

## 4. Segmentation process

The segmentation process splits the signal into homogeneous parts in terms of speaker, gender, and bandwidth. For transcription, accuracy of segment boundaries is as important as correct gender and bandwidth labels. Errors in terms of speaker label have less impact in this task.

Figure 1 shows the two systems that were investigated: a word-based segmentation system and a phone recognizer. In both cases, the segmentation process relies a speaker segmentation system based upon the Bayesian Information Criterion (BIC [5, 6]).

### 4.1. Speaker segmentation system

The acoustic speaker segmentation is based upon a BIC framework composed of three modules:

- The signal is first split in small homogeneous segments.
- Then, the segments are clustered by speaker without changing the boundaries.
- Finally, the boundaries are adjusted.

The initial segment boundaries are determined according to a Generalized Likelihood Ratio (GLR) computed over two consecutive windows of 2.5s sliding over the features (12MFCC+E). No threshold is employed, except for the minimal segment length which is set to 2.5s. The signal is oversegmented in order to minimize misdetection of boundaries but the minimum segment length is set long enough to allow for correct estimation of a speaker model.

The clustering is based upon a bottom-up hierarchical clustering. In the initial set of clusters, each segment is a cluster. The two closest clusters are then merged at each iteration until the BIC stop criterion is met. The speaker, *ie* the cluster, is modeled by a full covariance Gaussian as in the segmentation process. The BIC penalty factor is computed over the length of the two candidate clusters instead of the standard penalty computed over the length of the whole signal [7]. To minimized the clustering time, a first pass of clustering is performed only over adjacent clusters.

Viterbi decoding is performed to adjust segment boundaries. A speaker is modeled by a one-state HMM containing a diagonal covariance GMM of 8 components learned by EM-ML over the set of speaker segments. The log-penalty of switching between two speakers is fixed experimentally to 250.

### 4.2. Phone recognizer based system

A segmentation is obtained using a context-independent phone recognizer with a null language model over narrowband features. The output phones are labeled as speech or filler[3]. Small filler segments are re-labeled as speech according to a set of heuristics based upon the segment length. Finally, gender and bandwidth of speech segments are detected using a set of GMMs.

[3] *ie* silence, music, breath, hesitation

### 4.3. Phone- vs Word-based segmentation

In the phone-based segmentation, the boundaries of the segments from the speaker segmentation system are adjusted to those of the speech segments. The bandwidth and gender for each speaker segment are set to the most represented (in terms of duration) corresponding labels in the underlying speech segments.

In the word-based segmentation system, bandwidth, then gender, are detected (through GMMs) directly on each speaker segment. Segment boundaries finally get adjusted according to sentence boundaries provided by the first pass transcription decoding (see section 7).

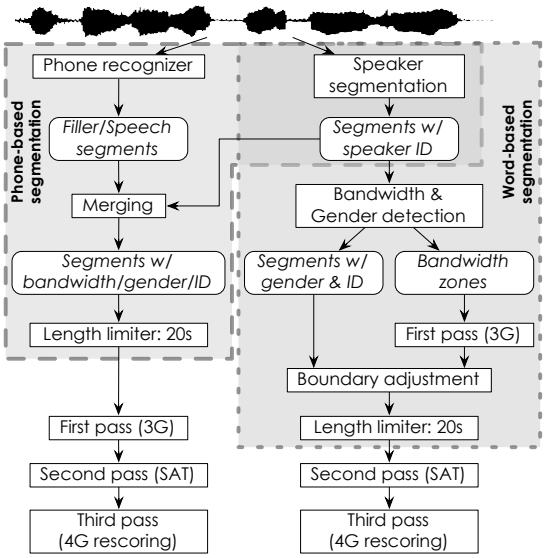


Figure 1: Overview of the speech transcription processing.

## 5. Acoustic modeling

Acoustic models for 35 phonemes and 5 kinds of fillers were trained using a set of 72 hours of broadband data and 8 hours of narrowband data with manual transcription. For training, the acoustic trainer toolkit SphinxTrain, associated to CMU Sphinx decoders, was used. The final models are composed of 5500 tied states (up to 6000 when training data were injected, see section 8.3), each state being modeled by a mixture of 22 Gaussians. The broadband (BB) model was built from the broadband data only, whereas the narrowband (NB) model relied on the whole 80 hours (narrowband analysis) and was then adapted by MAP method [8] to the 8 hours of narrowband data.

For transcription alignment (selection of phonetic variant and insertion of fillers), two different methods were used:

- reliance on the manual transcription; only silence labels are added, detected with a context-independent, bandwidth-dependent, mono-gaussian model;
- automatically detection of all fillers using a context-dependent, bandwidth-independent model with 22 Gaussians per state.

A MAP procedure was used to specialize both models on gender, resulting in BB-male, BB-female, NB-male, NB-female. The procedure runs in 3 iterations and adapts means, variances, and weights.

### 5.1. SAT

The gender- and bandwidth-dependent models (BB-male, BB-female, NB-male, NB-female) were used to compute the CMLLR transformation for each sentence (or each speaker). This approach yields better results than using gender-independent models.

After applying the transformation, four gender-, bandwidth-dependent models (SAT-BB-male, SAT-BB-female, SAT-NB-male, SAT-NB-female) were estimated using the same process as described above.

## 6. Language modeling

Ideally, text corpora used to train statistical language models have to be the closest as possible to the task targeted by the speech recognition application. Because of the cost, manual transcriptions of spoken language are difficult to obtain and approximations are necessary: the training data provided for the ESTER evaluation were made partly of manual transcriptions of broadcast news from various radio stations, but the major part of the data came from articles of the french newspaper "Le Monde". In fact, spontaneous speech seldom occurs in broadcast news and, by default, we can be satisfied with the use of newspaper articles. We chose to split the training data into three homogeneous sets :

1. Manual transcriptions of 89 hours out of the 90 hours of radiophonic broadcast news provided by the ESTER organisation. The last one hour was left to test the trained language models on. These transcriptions were composed of 1.35M words, including 34K different words.
2. Articles from french newspaper "Le Monde" for the year 2003. These sentences were composed of 19M words, including 220K different words. This set comprises the most recent articles before the period of test data, 2004.
3. Articles from french newspaper "Le Monde" for the period between 1987 and 2002. These sentences were composed of approximately 300M words.

Training language models cannot be dissociated from building the vocabulary. For this reason the words of our vocabulary come from the three sets of text corpora.

### 6.1. Lexicon building

First, all the 34K words occurring in the first set are incorporated into the vocabulary. Since the sentences in this set are of comparable nature to the test sentences, it seems intuitively interesting to keep these words. Then, words occurring more than 10 times in the second training data set (about 19K words) were incorporated. Finally the most frequent words in the last training data set were used to get the vocabulary to its maximum size (65K words).

### 6.2. N-gram models estimation

Using this vocabulary, each of the three data sets defined above was used to train a trigram language model. To estimate and interpolate these trigram models, the SRILM toolkit [9] was used. Each language model was a backoff model, using the modified Kneser-Ney discounting method. All the unigrams and bigrams were kept, whereas trigrams occurring only once were not taken into account. To compute the interpolation weights, the EM implementation provided by the CMU SLM toolkit [10] was used. The resulting trigram language model was used in the first two passes of the speech recognition process. To rescore the word-lattice in the third pass, a quadrigram language model was necessary: it was estimated the same way as the trigram language model, rejecting all quadrigrams and trigrams occurring only once in training data.

Table 1 shows the number of n-grams in the resulting trigram and quadrigram models.

| Model | 1-grams | 2-grams | 3-grams | 4-grams |
|---|---|---|---|---|
| trigram | 65.5K | 18.4M | 25.4M | – |
| quadrigram | 65.5K | 18.4M | 22.2M | 19.7M |

Table 1: Number of n-grams in the trigram and quadrigram language models used in the speech recognition process

## 7. Speech transcription process

The speech transcription process is composed of three passes:

1. The first pass uses the acoustic model corresponding to the gender and the bandwidth detected by the segmentation process, and using a trigram language model.
2. The second pass applies a CMLLR transformation by speaker or by segment, and uses the same trigram language model as the first pass. A word-lattice is generated which contains words and their acoustic scores.
3. This lattice is then rescored in the last pass with a quadrigram language model.

Figure 1 summarizes the system architecture, highlighting the two segmentation schemes.

## 8. Experiment and results

The experiments described here were conduced in the strict context of the ESTER evaluation [1]: no other training data were used beside those distributed during the campaign.

Audio corpora for training and development were composed of 90 hours of audio files from four French speaking radio stations (3 from France + 1 from Morocco): France Inter, France Info, Radio France International and Radio Télévision Marocaine. The textual corpora for training and development (which include manual transcriptions of the audio corpora) were described in section 6.

The test corpus was composed of 10 hours of shows: 2 hours from each of the four radio stations included in the train and development corpora, plus 2 hours from two unknown (at evaluation time) stations (which turned out to be Radio Classique and France Culture). It comprises about 10,000 sentences amounting to about 112,000 words. The method we used to build our vocabulary (see section 6.1) induced an out-of-vocabulary word rate of 1.18% on the test data.

### 8.1. SAT: sentence vs speaker

All the experiments described below relied on a sentence-by-sentence CMLLR transformation for SAT. This proved to be more efficient than a more traditional speaker-by-speaker approach, with a gain of 0.4 point in terms of word error rate.

### 8.2. Segmentation and alignment

Table 8.2 shows comparative results for the two segmentation methods described in section 4, as well as for the two alignment

strategies described in section 5.

| Model (80h) | 3-grams | 4-grams |
|---|---|---|
| Phone-based seg - man. fillers | 24.5 | 23.6 |
| Phone-based seg - auto. fillers | 24.8 | 23.8 |
| Word-based seg - auto. fillers | 24.6 | 23.7 |

Table 2: Comparative results (w.e.r.) for segmentation methods and alignment strategies

Word-based segmentation yields slightly better results (0.1 point with quadrigrams). We believe it can be further improved because some mistakes it makes (insertion of words within music fillers) should be easily detected and corrected afterwards.

As for transcription alignment strategies, it appears that automatic filler detection does not give as good results as manual transcribers.

### 8.3. Addition of automatically transcribed data for training

We tested two iterative processes for expanding the training data set through addition of data stemmed from automatic transcription. Three data sets, of approximately 25 hours each, were used to extend the initial 80h data set (set 1: France Culture – December 2003; set 2: mixed radio stations – January-March 2004; set 3: FC – September 2004). The first process (A) consisted in adding set 3 (for a total data set of 105 hours), then set 1 (130 hours total). The second process (B) consisted in adding set 1 first, then set 2.

| Data size | Process A | Process B |
|---|---|---|
| 80h | 23.8 | 23.7 |
| 105h | 23.7 | 23.7 |
| 130h | 23.5 | 23.6 |

Table 3: Results (w.e.r.) of data addition (with 4G models)

Process A was tried only in conjunction with phone-based segmentation, while process B was tested with word-based segmentation, hence the score difference for the initial 80h data set. However, differences between various amounts of data within each process are what counts most here. From this point of view, process A is the most efficient of the two, seemingly because data from mixed stations (set 2) is more difficult to decode and this penalizes process B.

Process A was pushed further by adding set 2, for a total of 155 hours of training data. The resulting word error rate was 23.4%, improving the initial score (on the 80h data set) by 0.4 point.

## 9. Conclusion

The LIUM speech transcription system based on CMU Sphinx Project has finished at the second position during the ESTER evaluation campaign for the transcription task (TRS) with an official 23.6% word error rate. Post-evaluation scoring of an unsubmitted system showed that this system can reach 23.4% on the same test data. The difference between these results come from a bad use of the MMIE method to train acoustic models (whereas this method improved the results during the development step). In the system presented here, the discriminative MMIE training method was not used, whereas it was during the ESTER Evaluation.

Experiment results show that the features added by LIUM to the CMU Sphinx tools (segmentation process, SAT, and word-lattice rescoring with quadrigram language model) have lead to a relative reduction of about 17% of the word error rate (from 28.2% to 23.4%). Notice that the major part of the add-on tools developped by LIUM will be distributed under an open-source license.

## 10. Acknowledgements

The authors would like to thank the Canergie Mellon University for making Sphinx tools available as an open source project. In particular, many thanks to Ravi Mosur, Alex Rudnicky, Evandro Gouvea, and Arthur Chan.

## 11. References

[1] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of french broadcast news," in *Language Evaluation and Resources Conference (LREC 2004)*, Lisbon, Portugal, May 2004.

[2] M. Ravishankar, R. Bisiani, and E. Thayer, "Sub-vector clustering to improve memory and speed performance of acoustic likelihood computation," in *Proceedings of European Conference on Speech Communication and Technology (ESCA, Eurospeech 97)*, vol. 1, 1997, pp. 151–154.

[3] A. Chan, J. Sherwani, M. Ravishankar, and A. Rudnicky, "Four-level categorization scheme of fast gmm computation techniques in large vocabulary continuous speech recognition systems," in *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, 2004.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR 291, May 1997.

[5] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 91)*, vol. 2, Toronto, Canada, May 1991, pp. 873–877.

[6] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, February 1998.

[7] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *DARPA RT04 Fall*, Palisades, NY, USA, 2004.

[8] J.-L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 22, pp. 291–298, April 1994.

[9] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2002)*, vol. 2, Denver, Colorado, USA, 2002, pp. 901–904.

[10] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proceedings of European Conference on Speech Communication and Technology (ESCA, Eurospeech 97)*, vol. 1, Rhodes (Greece), 1997, pp. 2707–2710.