INTERSPEECH 2009 BRIGHTON

Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?

Paul Deléglise, Yannick Estève, Sylvain Meignier, Teva Merlin

LIUM, University of Le Mans, France

firstname.lastname@lium.univ-lemans.fr

Abstract

This paper describes the new ASR system developed by the LIUM and analyzes the various origins of the significant drop of the word error rate observed in comparison to the previous LIUM ASR system. This study was made on the test data of the latest evaluation campaign of ASR systems on French broadcast news, called ESTER 2 and organized in December 2008.

For the same computation time, the new system yields a word error rate about 38 % lower than what the previous system (which reached the second position during the ESTER 1 evaluation campaign) did. This paper evaluates the gain provided by various changes to the system: implementation of new search and training algorithms, new training data, vocabulary size, etc. The LIUM ASR system was the best open-source ASR system of the ESTER 2 campaign.

Index Terms: automatic speech recognition system, acoustic model, language modeling, evaluation

10.21437/Interspeech.2009-607

1. Introduction

The LIUM automatic speech recognition system is based on the CMU Sphinx system. The tools distributed in the CMU Sphinx open-source package, although already reaching a high level of quality, can be supplemented or improved to integrate some state-of-art technologies. It is the solution LIUM has adopted to develop its own ASR system, by building on this base and gradually extending it to bring it to new performance levels.

The present article focuses on the evolution and technical improvements between two major versions of the LIUM ASR system, with an analysis of how these improvements translated into performance gains, either individually or combined.

The two versions of the system were evaluated at a 3 year interval in the same series of evaluation campaigns: the ESTER campaigns, which aim to evaluate ASR systems for French broadcast news, in a similar way to what the NIST RT campaigns offer in English. The first campaign of the series (ESTER 1) started in 2003 and ended in January 2005 [1], and the second edition (ESTER 2) took place from 2007 to 2008. Though not identical, the data sets and conditions of the two editions are similar enough to facilitate comparison between the results, highlighting the performance gain achieved over this period.

The version of the LIUM system used for ESTER 2 (which we will designate as LIUM'08, after the date of the campaign) yields a word error rate (WER) about 38 % lower than what the version used for ESTER 1 (LIUM'05, again after the date of the campaign) does, without impairing transcription speed.

This research was supported by the ANR (Agence Nationale de la Recherche) under contract number ANR-06-MDCA-006.

The rest of this paper will present the features of LIUM'05, then the changes brought to it as part of LIUM'08, as well as how the latter relates to the original system, CMU Sphinx. We will then highlight how the various changes between the two versions of the system contribute to the performance gain.

2. The ESTER evaluation campaigns

2.1. ESTER 1

The ESTER 1 evaluation campaign was organized within the framework of the TECHNOLANGUE project funded by the French government under the scientific supervision of the AFCP[1] with the DGA[2] and ELDA.

About 100 hours of transcribed data make up the corpus, recorded between 1998 and 2004 from six radio stations: *France Inter, France Info, RFI, RTM, France Culture* and *Radio Classique*. Shows last from 10 minutes up to 60 minutes. They consist mostly in prepared speech such as news reports, and a little conversational speech (such as interviews).

The corpus of articles from the French newspaper "Le Monde" from 1987 to 2003 can be used in addition to the transcription of the broadcast news to train the language model.

2.2. ESTER 2

The ESTER 2 campaign falls under the continuity of ESTER 1. It was organized by the DGA and the AFCP during 2007 to 2008. The new campaign builds on the previous edition by reusing its corpus and extending it to cover new types of data. In particular, it includes more programs with foreign accents, as well as more programs spontaneous speech: in addition to French national broadcast news, ESTER 2 includes talk-shows and African programs (from the station *Radio Africa No 1*).

ESTER 2 supplements the ESTER 1 corpus with about 100 hours of transcribed broadcast news recorded from 1998 to 2004, with additional 6 hours for development and 6 hours for test from 2007-2008. Fast transcriptions of 40 hours of African broadcast news are also available. Textual resources are extended by articles from the newspaper "Le Monde" from 2004 to 2006.

3. System LIUM'05 (for ESTER 1)

The system described below (and in [2]) ranked second in the ESTER 1 evaluation campaign.

3.1. Diarization

The diarization system developed for the ESTER 1 campaign is composed of an acoustic BIC-based segmentation followed by

[1] AFCP: Association Francophone de la Communication Parlée

[2] DGA: Délégation Générale de l'Armement

a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian. Viterbi decoding is used to adjust the segment boundaries using GMMs for each cluster.

Music and jingle regions are removed using Viterbi decoding with 8 GMMs, for music, jingle, silence, and speech (with wide/narrow band variants for the latter two, and clean/noised/musical background variants for wide-band speech). Gender and bandwidth are then detected using 4 gender- and bandwidth-dependent GMMs.

Speech segments are then limited to 20 s by splitting overlong segments using a GMM-based silence detector.

3.2. Speech recognition system

3.2.1. Features

The transcription decoding process is based on multi-pass decoding using 39 dimensional features (PLP with energy, delta, and double-delta). Two sets of features are computed for each show, corresponding to broadband and narrowband analysis.

3.2.2. Decoding

After speaker diarization, a first decoding pass permits to compute a CMLLR transformation for each speaker [3]. A second decoding pass is then performed using Speaker Adaptive Training (SAT) based on CMLLR acoustic models. The word-graph generated during the second decoding is re-evaluated with a quadrigram language model. The word-graph is pruned to avoid combinatorial explosion before proceeding with exploratory search.

3.2.3. Acoustic models

Acoustic models for 35 phonemes and 5 kinds of fillers are trained using a set of 80 hours from the ESTER 1 training corpus. These models are composed of 5500 tied states, each state being modeled by a mixture of 22 diagonal Gaussians. Both decoding passes employ tied-state word-position 3-phone acoustic models which are made gender- and bandwidth-dependent through MAP adaptation of means, covariances and weights.

The CMLLR technique for SAT in the second decoding pass generates a full 39x39 matrix for each speaker.

3.2.4. Vocaburary and Language models

The textual data provided for training in the ESTER 1 evaluation consist partly of manual transcriptions of 100 hours of broadcast news, representing 1.35M occurrences of 34k distinct words. The major part, however, came from the French newspaper "Le Monde".

The articles from "Le Monde" of the year 2003 (19M occurrences of 220k distinct words) are used to filter the 34k words of the manual transcriptions: of these words, only those occurring more than ten times in the 2003 articles are kept for inclusion in the vocabulary of the ASR system. The most frequent words in the rest of the articles from "Le Monde" (from 1987 to 2002 – 300M word occurrences) are used to complete the vocabulary, up to 64k words.

Phonetic transcriptions for the vocabulary are taken from the BDLEX database, or generated by the rule-based, grapheme-to-phoneme tool LIA_PHON[4] for words not in the database.

Using this vocabulary, all the textual data of the training corpus is used to train trigram and quadrigram language models. To estimate and interpolate these models, the SRILM is employed using the modified Kneser-Ney discounting method. Unigrams and bigrams are all kept, but trigrams or quadrigrams occurring only once are discarded.

4. System LIUM'08 (for ESTER 2)

This system was the best open-source ASR system participating in the ESTER 2 evaluation campaign, with 24.2 % of WER on the development corpus and 19.2 % on the test corpus. The development of the system used the official ESTER 2 development corpus, consisting of 6 hours recorded in June-July 2007.

4.1. Diarization

The diarization system is very close to the one in LIUM'05. The only 2 differences are a lower threshold for BIC clustering and the use of the Sphinx toolkit to compute the feature vectors whereas LIUM'05 used the SPRO toolkit. This system, completed by a CLR-based clustering phase, obtained the best diarization error rate during the ESTER 2 campaign.

4.2. Speech recognition system

The system is an evolution of system LIUM'05, to which new data and new skills are added, as described below.

4.2.1. Decoding

The new decoding strategy involves 5 passes. Pass # 1 is the same as in LIUM'05. The other passes are as follows:

- # 2 The best hypotheses generated by pass # 1 permit to compute a CMLLR transformation for each speaker. Decoding # 2, using SAT and Minimum Phone Error (MPE) acoustic models and CMLLR transformations, generates word-graphs.
- # 3 In the third pass, the word-graphs are used to drive a graph-decoding with full 3-phone context with a better acoustic precision, particularly in inter-word areas. This pass generates new word-graphs.
- # 4 The fourth pass consists in recomputing with a quadrigram language model the linguistic scores of the updated word-graphs of the third pass.
- # 5 The last pass generates a confusion network from the word-graphs and applies the consensus method to extract the final one-best hypothesis [5].

4.2.2. Acoustic models

The acoustic models are quite similar to the ones in LIUM'05, but trained on a different corpus, composed of 240 hours from ESTER 1 & 2 plus 40 hours of transcribed French broadcast news provided by the EPAC project. Models for pass # 1 are now composed of 6500 tied states. Models for passes # 2 to # 5 are composed of 7500 and are now trained in a MPE [6, 7] framework applied over the SAT-CMLLR models.

4.2.3. Vocaburary and Language models

Data used to build the linguistic models are of three kinds:

1. Manual transcriptions of broadcast news. They correspond to the transcription of the data used to train the acoustic models. We have also used manual transcriptions of conversations from the PFC corpus[8];
2. Newspaper articles: in addition to 19 years of "Le Monde" newspaper corpus, we also use articles from an-

other French newspaper, "L'Humanité", from 1990 to 2007, and the French Giga Word Corpus;

3. Web resources drawn from "L'Internaute", "Libération", "Rue89", and "Afrik.com".

To build the vocabulary, we generate a unigram model as a linear interpolation of unigram models trained on the various training data sources listed above. The linear interpolation was optimized on the ESTER 2 development corpus in order to minimize the perplexity of the interpolated unigram model. Then, we extract the 122k most probable words from this language model.

The trigram and quadrigram models are trained in the same way as the LIUM'05 language models but no cut-off is applied on trigrams and quadrigrams. The models are composed of 121k unigrams (respectively 65k for LIUM'05 LMs), 29M bigrams (resp. 18M), 162M trigrams (resp. 25M), and 376M quadrigrams (resp. 20M).

5. System LIUM'08 and CMU Sphinx tools

We have added large extensions to the SphinxTrain toolkit since 2005: MAP adaptation of means, but also weights and covariances of the models, as well as SAT based on CMLLR and MPE, are the most remarkable.

Passes # 1 and 2 of system LIUM'08 use version 3.7 of the Sphinx decoder, slightly modified to employ the CMLLR transformation applied to the features. Pass # 4 is based on *sphinx3_astar* which we extended to handle quadrigram LMs. Passes # 3 and 5 are based on Sphinx version 4, which we heavily modified to develop the acoustic graph decoder and the confusion network generation.

Other parts, such as computation of PLP features and the diarization system, do not rely on Sphinx and are entirely in-house developments.

6. Experiments

The experiments are carried out using the official test corpus of the ESTER 2 campaign. This corpus consists in 6 hours (26 shows) recorded between December 2007 and February 2008.

6.1. Global results

The WER over the test data for system LIUM'05 is 29.4 %. For system LIUM'08, the WER is 19.2 %. This 10 point (or 34 %) WER drop is very significant and, as such, motivates a thorough analysis of the contributions provided by the various modifications integrated into the new system.

System LIUM'08 is based on a multi-pass architecture: table 1 shows the WER after each pass of the decoding process.

Table 1: Word error rates for each pass of LIUM'08

Pass	Word error rate
# 1 (general acoustic models, trigram)	27.1 %
# 2 (acoustic adaptation)	22.5 %
# 3 (word-graph acoustic rescoring)	20.4 %
# 4 (word-graph quadrigram rescoring)	19.4 %
# 5 (consensus)	**19.2** %

We can observe that adaptation of the acoustic models allows a large gain in pass # 2, as does the better acoustic precision given by the full 3-phone search algorithm used to rescore a word-graph in pass # 3 (the acoustic models used theses two passes were trained using the MPE method). Rescoring this word-graph with a quadrigram model in pass # 4 allows to lower the WER by one extra point. The last pass does not have a significant impact on WER, but it allows the ASR system to provide confidence measures.

6.2. Decoder comparison

In order to precisely evaluate the contribution of the new software architecture of our system, independently from any gain brought by differences of training data, we have decoded the test data with LIUM'05 and LIUM'08 using the same vocabulary, the same language models and the same acoustic models for both systems. Table 2 shows that the changes in software architecture alone allow to decrease the WER by more than five percentage points.

Table 2: Word error rates of LIUM'05 and LIUM'08 using the same vocabulary, linguistic and acoustic models

ASR system	Word error rate
LIUM'05	29.4 %
LIUM'08 with LIUM'05 models	24.1 %

In another experiment, we have evaluated the contribution of the modifications to the diarization system: as seen in table 3, these modifications permit to gain 0.8 point.

Table 3: WER of LIUM'08 depending on diarization system

Diarization system	Word error rate
LIUM'08	19.2 %
LIUM'08 using LIUM'05 diarization	20.0 %

6.3. Contribution of training data and training process

To evaluate the contribution of the training data of acoustic models, we have decoded the ESTER 2 test data using LIUM'08 with 3 sets of acoustic models: models trained on the ESTER 2 training data and using the MPE method (the baseline of system LIUM'08), models trained on the same data but without using the MPE method, and the acoustic models of the LIUM'05. Table 4 shows two cumulative gains in WER: a first one by using MPE and another one by injecting new data into an already relatively large training data set (80 hours of data to train the LIUM'05 acoustic models *vs.* 280 hours for LIUM'08).

A similar experiment was made by substituting the languages models of LIUM'08 with those of LIUM'05 (still using the LIUM'08 acoustic models). Table 4 shows that adding new data while increasing the vocabulary size (LIUM'05 LMs are computed with a lexicon compounded by 65k words instead of 120k in the case of LIUM'08) allowed to decrease the WER

Table 4: Word error rates of LIUM'08 according to the acoustic models, the linguistic knowledge or the vocabulary size

ASR system configuration	WER
LIUM'08	19.2 %
without MPE	20.3 %
using LIUM'05 acoustic models (no MPE)	21.6 %
using LIUM'05 language models	22.6 %
using a vocabulary size of 65k words	20.4 %

by 3.4 point. The gain here is higher than the one observed when injecting new training data for acoustic models, which is 2.4.But we have to notice that the amount of data injected into the training corpus of language models is more massive that the one injected for the acoustic models. These results can be compared with the results presented in table 2 where both the language and acoustic models were replaced.

Another experiment was to evaluate the gain obtained by increasing the vocabulary size. Table 4 shows that it is not negligible: when using language models trained on the same data as the language models of LIUM'08, but with a vocabulary size of 65k words instead of 120k words, WER goes up by 1.2 points.

6.4. Latest improvements

Two improvements have been recently integrated into system LIUM'08. The first one consists in assigning a score to each pronunciation variant in the dictionary. The score is computed by observing the frequency of the variant in the LIUM'08 training corpus. Table 5 shows that this allows a gain of 0.4 point in terms of WER.

The second improvement is based on the presence of two kinds of francophone radio stations in the ESTER 2 campaign: French and African ones. We have decided to build two sets of linguistic knowledgde (lexicon and n-gram models), specific to each of these two kinds of stations. The MPE method to train acoustic models was also adapted for the African radio stations. Table 5 shows that this allows to obtain the best word error rate of all our experiments: 18.1 %.

Table 5: Gain in WER due to the use of probabilities for pronunciation variants as well as specialization according to the kind of radio station

ASR system	Word error rate
LIUM'08	19.2 %
+ pronunciation variant probability	18.8 %
+ specialization of linguistic models	**18.1** %

6.5. Confidence measures

In order to provide additional information for applications which could use it, system LIUM'08 uses the *a posteriori* probabilities computed during the generation of the confusion networks to provide confidence measures [9].

However, as seen in table 6 which presents an evaluation of these confidence measures in terms of normalized cross entropy (NCE), with no specific treatment these *a posteriori* probabilities are not very good predictors of the word error rate.

So, a mapping method is applied which consists in splitting the *a posteriori* probabilities into 15 classes of values: each confidence measure is linearly transformed using the coefficient associated with its class. These coefficients have been optimized on the ESTER 2 development corpus to maximize NCE. Such mapping approach was presented in [10]. Table 6 shows that this method makes the confidence measures provided by system LIUM'08 very competitive, with a NCE of 0.329 on the ESTER 2 test corpus.

Table 6: Contribution of the mapping method applied to the confidence measures of LIUM'08

Confidence measures	NCE
without mapping	0.064
with mapping	**0.329**

7. Conclusions

System LIUM'08 was the best open-source ASR system participating in the ESTER 2 evaluation campaign in 2008. The availability of the CMU Sphinx tools under an open-source licence was an excellent opportunity to develop a very competitive system for broadcast news. The features added by LIUM since 2005 to the CMU Sphinx tools are various: PLP features, diarization tools, CMLLR, complete MAP adaptation, SAT and MPE training, word-graph decoder with a better inter-word acoustic precision, and more. The CMU Sphinx3.x one-pass decoder employed "out of the box" (pass #1 of system LIUM'08) obtains 27.1 % of WER, while the LIUM add-ons assembled in a multi-pass decoder permit to bring the WER down to 18.1 %.

The LIUM ASR system was the only one among all the participants in the ESTER 1 and ESTER 2 campaigns to decrease significantly its WER. This paper gives some explanations about the origins of this performance.

However, while the LIUM system obtained a WER of about 18 % during ESTER 2, down from a WER of about 23.6 % during ESTER 1 (ranking second), the best system obtained about 12 % during both campaigns. We expect to continue to decrease the WER by exploring innovative approaches and integrating state-of-art methods into our ASR system.

8. Acknowledgements

The authors would like to thank Carnegie Mellon University for making the Sphinx tools available as an open-source projet.

9. References

[1] 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*, Lisbon, Portugal, 2005.

[2] 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*, Lisbon, Portugal, 2005.

[3] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, p. 7598, 1998.

[4] F. Béchet, "LIA_PHON un système complet de phonétisation de texte," in *Traitement Automatique Des Langues*. Hermès, 2001, vol. 42, pp. 47–68.

[5] H. Mangu, E. Brill, and S. A., "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[6] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Florida, USA, 2002, pp. 105–108.

[7] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. Dissertation, Department of Engineering, University of Cambridge, United Kingdom, 2004.

[8] J. Durand, B. Laks, and C. Lyche, "La phonologie du français contemporain : usages, variétés et structure," *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*, pp. 93–106, 2002.

[9] H. Jiang, "Confidence measures for speech recognition: a survey," *Speech Communication Journal*, vol. 45, pp. 455–470, 2005.

[10] G. Evermann and P. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000.