The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News

Sylvain Galliano (1) *, Edouard Geoffrois* (1)*, Djamel Mostefa* (2)
Khalid Choukri (2)*, Jean-François Bonastre* (3) *, Guillaume Gravier* (3)

(1) Délégation Générale pour l'Armement / Centre d'Expertise Parisien
(2) Evaluations and Language resources Distribution Agency
(3) Association Francophone de la Communication Parlée

http://www.afcp-parole.org/ester

Abstract

This paper gives the final results of the ESTER evaluation campaign which started in 2003 and ended in January 2005. The aim of this campaign was to evaluate automatic broadcast news rich transcription systems for the French language. The evaluation tasks were divided into three main categories: orthographic transcription, event detection and tracking (*e.g.* speech vs. music, speaker tracking), and information extraction. The last one, limited to named entity detection in this evaluation, was a preliminary test. The paper reports on protocols and gives the results obtained in the campaign.

1. Introduction

Objective evaluation of performance in the fields of speech and natural language processing is a major issue in scientific research and technology development.

As far as the French language is concerned, a first wave of evaluation campaigns was initiated in the nineties. In particular, this effort resulted in a first evaluation campaign on automatic transcription of read speech [2]. The ESTER[1] campaign is part of this ongoing effort for developing evaluation campaigns, corpora and evaluation paradigms for the French language [3]. This campaign, organized jointly by the Francophone Speech Communication Association (AFCP), the French Defense expertise and test center for speech and language processing (DGA/CEP), and the Evaluation and Language resources Distribution Agency (ELDA), is part of the EVALDA project dedicated to the evaluation of language technologies for the French language[2]. The ESTER campaign started in 2003, with a Phase I dry run in January 2004, and a Phase II official test in January 2005.

ESTER focuses on the evaluation of rich transcription and indexing of radio broadcast news in French. This task, largely inspired by the NIST Rich Transcription evaluations [1], was chosen for several reasons. First, dealing with broadcast news is a logical progression with respect to the previous campaign on read speech transcription. Second, the tasks considered offer a strong application potential. And third, it complements the above-mentioned NIST evaluations on the English, Arabic and Chinese languages. Compared to these evaluations, though, ESTER does not include information intended to help human readability, such as punctuation or disfluencies. It does include, however, information about thematic content, and could thus also be related to other NIST evaluations such as Spoken Document Retrieval.

This paper describes the results of the recently completed phase II of the campaign.

[1] ESTER is the French acronym for "Evaluation de Systemes de Transcription enrichie d'Emissions Radiophoniques" (Evaluation of Radio Broadcast Rich Transcription Systems).

[2] The EVALDA project is sponsored by the Technolangue program, initiated by the French Ministry of Research.

2. Tasks and corpus

2.1. Overview of the evaluation tasks

The ESTER evaluation implements three categories of tasks, namely transcription (T), segmentation (S), and information extraction (E). The first two constitute the core of the campaign while the information extraction category is prospective and was limited to named entity detection. Two T tasks, real time (TTR) and unconstrained transcription (TRS), were implemented, while in category S the implemented tasks were sound event tracking (SES), speaker tracking (SVL) and speaker diarization (SRL). A brief description of each task is given as an introduction to the results in the next sections. Though not independent in practice, each task is evaluated separately with the appropriate paradigm, in order to best characterize the various components of a radio broadcast indexing system.

2.2. Corpora

A corpus of about 90 hours of manually transcribed radio broadcast news shows was given to the participants for training purposes, 8 hours of which were identified as a development set. This acoustic corpus contained shows from four different sources, namely France Inter (Inter), France Info (Info), Radio France International (RFI) and Radio Télévision Marocaine (RTM). In addition, 1,600 hours of non transcribed radio broadcast news shows, from the same four sources plus France Culture (Cult.), were provided for non supervised training purposes. Transcribed data were recorded in 1998, 2000 and 2003. Non transcribed data were recorded between October 2003 and September 2004. These two resources were complemented by a corpus of articles from the newspaper Le Monde, taken over the period 1987–2003 and containing approximately 400M words.

The test set was recorded from October to December 2004 and consists of 10 hours of radio broadcast news shows taken from the five stations of the training corpus (transcribed or not) plus Radio Classique (Class.) for which no specific training data was available (see table 1 for some statistics).

Table 1: Statistics on the 10 hours test

Length	10 h 07 min
Number of words	103,203
Number of speakers	343
Non scored passages for T task	5.99 %
Non scored passages for S task	2.47 %
Simultaneous speech	0.43 %
Non speaker passages (music, jingle, etc.)	4.95 %

For all the implemented tasks, participants were allowed to use any data recorded prior to May 2004, whether distributed specifically for the campaign or not. The use of the provided non transcribed data, partially recorded after May 2004, was also allowed.

3. Transcription tasks

The transcription task is the classical task which consists in producing the (normalized) orthographic transcription from the waveform, the performance measure being the word error rate (WER). Systems operating in real-time or less (TTR task) were evaluated beside unconstrained ones (TRS). In the TTR task, participants were asked to run a system able to process the 8 hours of the development set in a time less than or equal to 8 hours. Such systems could run in slightly more than 1xRT on other data sets, depending on their difficulty.

Table 2 analyzes some of the differences between systems and report performance for the TRS task. The best results were obtained by LIMSI, which had previous experience on the task. For all participants which took part in the dry run, a significant improvement of performance was observed in about a year. This improvement is partly due to the increase of the amount of transcribed broadcast news data provided for training and development (90h vs. 35h in the dry run). Indeed, such resources are crucial for building efficient transcription systems, both for acoustic and language modeling. However, the amount of data used to build a system does not account for the performance gap observed between participants. Other components of the system, such as estimation and adaptation techniques, dictionaries and tuning also make a crucial difference.

Performance for the TTR task are reported in table 3. Most systems being variants of the corresponding TRS systems, it is interesting to compare those results with the related ones in the TRS task. In most cases, real time systems implement fast gaussian computation, use a tighter beamwidth and no speaker adaptation. These limitations result in a WER increase between 6% and 10% absolute (35% and 42% relative for the top-3 sites).

Table 3: TTR task overall performance for each participating site.

Sites	IRIT	LIA	LORIA	VR*
WER	70.4	36.3	37.4	**16.8**
Real time factor	0.63	1.23	0.93	1.09

*The Vecsys Research (VR) system is based on LIMSI's technology.

Table 4 shows the WER obtained by the four best systems on the different radio stations, for the TRS task. For all systems, results vary greatly from one broadcaster to another with a WER difference of more than 40% relative between the Radio Classique and RTM. Unseen sources, such as Radio Classique (which was the "surprise" radio) and France Culture (since most participants did not use the non transcribed data for training) did not result in a loss of performance.

Table 4: Detail of the TRS task performances for each broadcaster for the top-4 sites.

Site	LIA	LIMSI	LIUM	LORIA
Info	23.8	10.3	20.2	23.8
Inter	26.8	12.2	24.3	27.7
RFI	25.0	11.8	23.4	27.3
RTM	32.1	14.4	29.3	34.1
Cult.	30.2	14.3	25.2	29.2
Class.	19.2	7.9	16.8	21.8

A more detailed analysis of the results outlined that, unsurprisingly, systems are sensitive to degraded speech quality and to background noise. The best system obtained around 10% of WER for clean speech (studio or telephone), to be compared with 17.9% in the presence of background music or noise. It was also observed that systems perform significantly better for female speakers than for male, with a relative WER increase of more than 20%.

4. Segmentation tasks

The segmentation tasks aim at detecting, tracking and grouping together audio "events", known a priori or not. Three tasks were implemented in the 2005 evaluation, namely sound event tracking (SES), speaker diarization (SRL) and speaker tracking (SVL).

In the two tracking tasks, possible errors are miss detection and insertion of an event, and the system performance is a tradeoff between the two errors. Errors are computed based in time marks, in seconds, with a tolerance of 0.25s at reference segment boundaries. Systems were evaluated in terms of F-measure, defined as $2RP/(R+P)$ where

$$R = \frac{\sum_i t(c_i;c_i)}{\sum_i t(c_i;c_i) + t(\overline{c}_i;c_i)} \text{ and } P = \frac{\sum_i t(c_i;c_i)}{\sum_i t(c_i;c_i) + t(c_i;\overline{c}_i)},$$

where $t(c_i;c_i)$ is the amount of correctly detected target event, $t(c_i;\overline{c}_i)$ the amount of inserted target event and $t(\overline{c}_i;c_i)$ the amount of target event missed, for an event i. As the interpretation of the F-measure is not intuitive, errors are also analyzed in terms of miss and false alarm rates.

For the diarization task, a specific performance measure [1] is considered in order to take into account deletions and insertions of speech in addition to speaker substitutions, after optimal matching between true and arbitrary speaker names.

4.1. Sound event tracking

The sound event tracking task consists in identifying, on the one hand, parts of the document containing music, whether in the foreground or in the background, and, on the other hand, parts of the document containing speech, possibly with background music. Results are reported in table 5.

The results are good for speech detection where very low miss detection rates are achieved. This is due to the fact that most systems where tuned to detect speech accurately as a front

Table 2: TRS task overall performance for each participating site and comparison of some system parameters accross sites.

Sites	CLIPS	ENST (ENST/TSI)	IRENE (IRISA-ENST/INF)	IRIT	LIA	LIMSI	LIUM	LORIA
WER	40.7	45.4	35.4	61.9	26.7	**11.9**	23.6	27.6
Audio corpus	90h	90h	90h	31h	90	90h+100h*	90h+90h nt†	90h
#states	1,500	114	6,000	117	3,600	12,000	7,000	6,000
#gaussians	24k	14k	200k	3,7k	230k	370k	154k	90k
#words	21k	65k	65k	61k	65k	200k	65k	60k
#pronunciations	38k	118k	118k	119k	130k	276k	107k	112k
Broadcast news	1M	1M	1M	1M	1M	92M	1M	1M
News paper	400M	400M	400M	400M	400M	500M	400M	400M
Web	75M	–	–	–	–	14M	–	–
2-gram	7M	4M	4M	16M	16.9	23M	18M	7M
3-gram	9M	4M	4M	87M	19.9	40M	26M	25M
4-gram	–	–	–	–	–	37M	20M	–
#pass	1	1	1	2	2	3	3	2
Real time factor	40	8	15	4	100	7.5	15	20

*The additional audio data are from both TV and radio sources and date from 1994 to 1999.
†The additional audio data are from the 1,600 hours non transcribed (nt) audio data.

end of the transcription system. The music detection task is particularly difficult when the SNR of the music is low.

Table 5: SES task overall performance for each participating site (F-measure * 100, %fa and %fr).

	speech & music			speech			music		
Sites	F	%fa	%fr	F	%fa	%fr	F	%fa	%fr
FTR&D	——	——	——	99.1	25.5	1.1	——	——	——
IRISA	93.1	1.3	12.1	98.9	9.7	1.9	33.7	1.0	78.5
IRIT	**94.2**	2.1	9.5	98.8	30.1	1.5	52.7	1.2	61.7
LIA	92.7	11.6	5.7	**99.2**	36.6	0.7	**54.8**	10.9	38.7
LIUM	90.7	1.3	16.2	97.4	8.0	4.9	17.8	1.1	89.6
LORIA	——	——	——	97.5	34.2	4.0	——	——	——
SIS*	83.7	11.5	20.9	93.4	82.2	10.4	12.7	10.4	89.2
UOB	88.2	3.9	18.6	95.1	20.1	8.9	26.2	3.4	82.0

*Late submission.

4.2. Speaker tracking

Speaker tracking is somewhat similar to sound event tracking with speakers being the events to track. The task consists in detecting portions of the document that have been uttered by a given speaker known beforehand and for which training data are available before the test stage. A list of 279 speakers with at least 2 minutes of speech in the training set was provided to the participants. The task consisted in tracking each speaker in the 10h of the test set. Hard decision results are given in table 6 for each participating sites while detection error tradeoff curves are given in figure 1.

Table 6: SVL overall performance for each participating sites.

site	ENST/TSI*	IRISA	LIA
%fa	9.8	0.3	2.8
%fr	25.3	23.6	30.6
F-measure * 100	46.9	**84.3**	66.0

*Late submission.

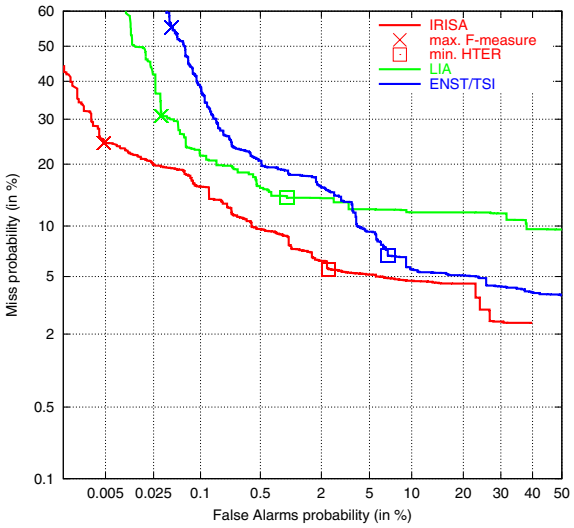

Figure 1: Detection error tradeoff curves of all participating sites for the SVL task.

The speaker tracking performance remains very variable, show per show, partially because of the mismatch between the training and test conditions. Due to the evaluation criterion, the F-measure, actual operating points correspond to very low false alarm rates. This is a major drawback of the F-measure criterion which only depends on the amount of target events and not on the total amount of data.

4.3. Speaker segmentation

Speaker diarization (SRL) aims at segmenting documents into speaker turns and at grouping together portions of the document uttered by the same speaker. Speakers are not known beforehand and identification is not required. Systems must return a segmentation of the document with a possible arbitrary speaker identifier for each segment.

Results are given in table 7, where performance are detailed in terms of miss speech, false alarm speech and speaker confusions. Performance comparable to the ones obtained in the NIST RT04 evaluations on the English broadcast news data [4] were achieved.

Table 7: SRL task overall performance per site.

site	CLIPS	LIA	LIMSI	LIUM
%miss	0.0	0.6	0.7	1.9
%ins	2.8	1.0	1.0	0.3
%sub	24.4	17.6	9.8	14.7
%err	27.2	19.2	**11.5**	16.9

As in the SVL task case, a detailed analysis of the results outline results highly dependent on the show, with error rates ranging from 1.5% to 26.1% for a particular system. So, despite the good level of performance reached, the systems are very dependent to the nature of the show. This point is clearly linked both to the performance measure criterion and to the variable nature of the shows (from 14 minutes duration with 5 speakers to 1 hour duration and 60 speakers).

5. Information extraction tasks

The information extraction tasks aim at extracting higher level information useful for indexing or document retrieval purposes. Within this framework, a prospective named entity (NE) detection task was implemented. Because of time constraints, only a dry run of this task could be done.

The NE tagset chosen is made of 8 main categories (persons, locations, organizations, socio-political groups, amounts, time, products and facilities) and over 30 sub categories. The tagset considered is therefore much more complex than the one used in the NE extraction tasks of the MUC 7 and DARPA HUB 5 programs where only 3 categories are considered. The error measure used was the slot error rate (SER) [5].

Two conditions were considered: detection on the reference transcriptions and detection on automatic speech recognition (ASR) transcriptions. The preliminary results obtained by the 3 systems participating clearly showed the impact of both the WER (between reference and ASR transcripts) and the temporal mismatch (between the training/development corpora on one hand and the test corpus on the other hand) on the SER performance.

On the reference transcriptions the SER score increased from 22% on the development set to 34% on the test set for the best system, and similarly for other systems. This shows the impact of a temporal mismatch of 6 months between the training and the test sets. In the future, NE detection systems will run on all the transcription system outputs in order to plot a WER/SER curve and measure the relation between these two error measures.

6. Discussion

Few French speaking laboratories had experience in broadcast news transcription when the ESTER project started. Probably a major result of this evaluation campaign is that many sites were eventually able to participate in the transcription task. Another outcome of the campaign is a strong community of French speaking labs willing to work together around the rich transcription topic. As none of the participants were funded to participate in this evaluation, this is a very positive outcome which illustrates the interest of evaluation campaigns and more generally of the evaluation paradigm.

Another major consequence of the ESTER campaign is the availability of significant resources, specifically created for the campaign. These resources include annotation conventions, protocols, scoring tools, about 100 hours of transcribed broadcast speech, around 1,600 hours of untranscribed speech, etc. They will be included in an evaluation package that will be made available and distributed by the organizers of the ESTER evaluation campaign via ELDA. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign.

From a scientific point of view, the transcription task is clearly better defined than the other tasks, certainly thanks to the experience and knowledge acquired from the NIST evaluations. The two tracking tasks were evaluated with the F-measure which turned out to be unsatisfactory, mostly because this measure does not take into account the total amount of data. Moreover, there is no balance between the false alarm and miss rates according to prior probabilities of occurence of an event. The speaker diarization metric is also very sensitive to the nature of the recordings. In a long duration record with few speakers, splitting a speaker has a strong impact on the performance.

The implementation of a dry run evaluation of the named entity detection task, though preliminary, resulted in an annotation manual and scoring tools for such tasks. Clearly, more work has to be done in the light of the experience gained during the campaign, in order to finalize a good evaluation protocol for the NE task and other information extraction tasks.

Finally, we are working on the continuation of the ESTER campaign in two directions. The first one consists in pursuing the broadcast news transcription effort, by consolidating the current results and by adding some information extraction tasks such as topic detection and question answering. The second direction concerns the organization of a follow-up campaign which should focus on new challenges, like dealing with long duration recordings (one week of broadcast radio or television data), with meeting data or with new tasks like automatic structuring and indexing of the information.

7. References

[1] "Spring 2003 Rich Transcription Workshop", 2003.

[2] Jean-Marc Dolmazon, Frédéric Bimbot, Gilles Adda, Marc El-Bèze, Jean-Claude Caërou, Jérôme Zeilinger, Martine Adda-Decker, "Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale", Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF, 3–18, 1997.

[3] Guillaume Gravier, Jean-François Bonastre, Edouard Geoffrois, Sylvain Galliano, Kevin McTait and Khalid Choukri, "The Ester Evaluation Campaign for the Rich Transcription of French Broadcast news", Proc. Language Evaluation and Resources Conference, 2004.

[4] D. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization", Proc. IEEE ICASSP, Philadelphia, March 2005.

[5] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel, "Performance measures for information extraction", Proc. of the DARPA Broadcast News Workshop, pages 249–252, Virginia, USA, 1999.