The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts

Sylvain Galliano (1) *, Guillaume Gravier* (2) *, Laura Chaubard* (1)

(1) Délégation Générale pour l'Armement / Centre d'Expertise Parisien

(2) Association Francophone de la Communication Parlée

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

Abstract

This paper reports on the final results of the ESTER 2 evaluation campaign held from 2007 to April 2009. The aim of this campaign was to evaluate automatic radio broadcasts rich transcription systems for the French language. The evaluation tasks were divided into three main categories: audio event detection and tracking (*e.g.*, speech vs. music, speaker tracking), orthographic transcription, and information extraction. The paper describes the data provided for the campaign, the task definitions and evaluation protocols as well as the results.

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 emphasized by the numerous evaluation campaigns in these areas over the last decades, among which the annual NIST RT evaluations[1] since 2002.

However, as far as evaluations in the area of spoken document processing in the French language are concerned, few campaigns have been implemented. In the nineties, a first wave of evaluation campaigns targeted, among other tasks, transcription of read speech [1]. From 2003 to 2005, the ESTER evaluation campaign was dedicated to transcription and indexing of radio broadcast news [2]. This first ESTER evaluation campaign mostly implemented tasks related to the segmentation (speech detection, speaker tracking and diarization) and the transcription of broadcast news shows. Pilot studies on tasks related to natural language processing issues on transcripts were also carried out, such as the experimental named entity detection task. The ESTER campaign resulted in significant progress in all the area covered by the evaluation, mostly due to the availability of a large amount of resources—annotated corpora as well as annotation guidelines and evaluation protocols—widely distributed among the scientific community.

In this paper, we describe the data, protocols and results of the recently completed ESTER 2 evaluation campaign held in 2008 and 2009. The campaign was jointly organized by the French-speaking Speech Communication Association (AFCP, French-speaking ISCA Regional Branch) and the French Defense expertise and test center for speech and language processing (DGA/CEP), with the collaboration of the Evaluation and Language resources Distribution Agency (ELDA). This campaign is an extension of the initial 2003–2005 ESTER campaign, targeting a wider variety of speaking styles and accents. In particular, broadcast shows other than news, such as entertainment shows and debates, were considered as well as news shows from a French-speaking African radio channel exhibiting strong accents. On top of the standard segmentation and transcription tasks, emphasis was put on the detection of named entities on transcripts, now implemented as a regular task. New pilot tasks, such as the detection of multiple speaker segments or sentence segmentation, were proposed but could not be achieved due to the difficulty of defining an acceptable evaluation protocol.

[1]National Institute of Standard and Technology, Rich Transcription evaluation (http://www.itl.nist.gov/iad/mig/tests/rt/)

2. Tasks and corpus

We first briefly describe the tasks involved in the benchmark before describing the data provided to the participating sites.

2.1. Overview of the evaluation tasks

The ESTER 2 evaluation implemented three categories of tasks, namely segmentation (S), transcription (T), and information extraction (E). The core tasks of the campaign were sound event tracking (SES), speaker tracking (SVL), speaker diarization (SRL), transcription (TRS) and named entity detection (NE).

In addition, three prospective tasks were tentatively defined: overlapping speech detection (SES-2), transcription with contemporary data (TDC), and segmentation into syntactic units (SP). Unfortunately, none of the prospective tasks were finally implemented either due to the lack of interest or manpower from the participants (SVL, TDC) or to the difficulty of defining a valid evaluation protocole (SES-2 and SP).

A brief description of each task is given as an introduction to the results in the following 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

The audio training data given to the participants consisted of three sets of data. Firstly, the ESTER 2 corpus, specifically released for the campaign, consists of 150 hours of manually transcribed radio broadcast news recorded between 1999 and 2003 (100 hours of rich transcription and 50 hours of rapid transcription). Secondly, 45 hours of semi manually transcribed radio broadcast news data were provided, courtesy of the EPAC project[2]. The data of the EPAC corpus is a subset of the

[2]The EPAC project, sponsored by the French National Agency for Research (ANR), is dedicated to the study of conversational speech. For more information, see http://epac.univ-lemans.fr/.

Table 1: Statistics on the 7 hours test

Duration	7h12
Number of words	72,534
Number of speakers	255
Non scored portions for T task	7.86 %
Non scored portions for S task	0.37 %
Simultaneous speech	2.27 %
Non speaker passages (music, jingle, etc.)	8.33 %

Table 2: SES task overall performance for each participating site (100 * error rate, miss rate, false alarm rate).

Sites	speech			music		
	err	%m	%fa	err	%m	%fa
IRISA	1.5	0.37	16.42	—	—	—
IRIT	1.3	0.72	9.28	5.5	43.13	0.77
LIMSI	**1.1**	0.80	4.91	—	—	—
TPT/RTL	1.2	0.50	11.01	**5.2**	12.56	4.33

1,600 hours of non transcribed data from the ESTER 1 campaign, recorded between 2003 and 2004, selected to mostly include non planned speech (*e.g.*, interviews). Finally, the ESTER 1 corpus, consisting of 100 hours of manually transcribed radio broadcast news shows recorded between 1998 and 2004, was also provided as training data. Named entities were (re)annotated in the ESTER 1 corpus according to the new guidelines defined for ESTER 2. Apart from the training data, a development set containing 6 hours of radio broadcast news recorded in 2007 was provided.

Overall, the three corpora contain shows from different sources, namely France Inter (Inter), Radio France International (RFI), France Culture, Radio Classique, Africa number one (Africa 1), Radio Congo and TVME (previously known as Radio Télévision du Maroc). The last three sources are French-speaking radios, where the two African radios (Africa 1 and Congo) contain strong accents while TVME shows exhibit a large amount of instantaneous translation implying speech from two speakers.

The audio resources were complemented by a corpus of articles from the French newspaper Le Monde, taken over the period 1987–2006 containing approximately 450M words. For all tasks, participants were allowed to use any data recorded prior to January 2008, whether distributed specifically for the campaign or not.

The test set, recorded from January to February 2008, consisted of 7 hours of radio broadcast shows taken from RFI, Inter, TVME and Africa 1 (see Table 1 for some statistics). Most shows were broadcast news though a small amount of data from Inter consisted in other types of programs with more spontaneous speech material, including a debate with questions from listeners over the telephone (Le Téléphone sonne).

2.3. Participants

Most French labs working in the areas of speech processing or natural language processins participated in the ESTER 2 campaign, along with some industrial sites. In alphabetical order, the participating sites were: IRISA, IRIT, LIA, LIG, LIMSI, LINA, LI Tours, LIUM, LORIA, LSIS, Vecsys Research (VR), Synapse Development (Synapse), Télécom ParisTech/RTL (TPT/RTL) and Xerox.

3. Segmentation tasks

Segmentation tasks aim at detecting, tracking and grouping together audio "events", known *a priori* or not. Three tasks were implemented in the ESTER 2 evaluation: sound event tracking (SES), speaker diarization (SRL) and speaker tracking (SVL), where the last one received no submission.

In tracking tasks, possible errors are miss detection and insertion of an event, the system performance being a trade-off between the two errors. Errors are computed based on time marks, in seconds, with a tolerance of 0.25s at reference segment boundaries. For a particular sound class, systems were evaluated in terms of error rate, defined as the total error time (amount of time due to insertions and miss detections) divided by the total duration of signal. The error rate can be interpreted as the mean error rate by unit of time. Errors were also computed in terms of miss alarm rate, false alarm rates and mean F-measure (mean of the F-measure of each event) in order to have a more precise analysis.

For the diarization task, the specific diarization error rate measure was 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.

3.1. Sound event tracking

The sound event tracking task consists in two main sub-tasks: music tracking, whether the music is in the foreground or in the background (SES-M), and speech tracking, possibly with background music (SES-P). A prospective overlapping speech tracking task (SES-2) was submitted but could not be achieved due to difficulties to define overlapping speech precisely and, as a consequence, to produce enough training data for participants.

Results reported in Table 2 are good for speech detection where very low miss detection rates are achieved. This is due to the fact that most systems were tuned to detect speech accurately as a front-end for transcription. On the other hand, the music detection task is particularly difficult since purely music portions are very limited in broadcast news shows and the signal to noise ratio of the music is quite low in the case of background music, *e.g.*, during headlines. Moreover, few participating sites developped a music specific tracking system.

3.2. Speaker segmentation

Speaker diarization (SRL) aims at segmenting documents into speaker turns and to group 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. Contrary to ESTER 1, portions of signal with several speakers were considered for scoring.

Results are given in Table 3, where performance are detailed in terms of miss speech, false alarm speech, speaker confusion rate and speaker error rate. A detailed analysis of the results outlines results highly dependent on the show, with error rates ranging from 0.22% to 23.86% for the best system. Despite the good average level of performance reached, the systems are very dependent on the nature of the show. This point is clearly linked both to the performance measure criterion and to the variable nature of the shows (length, number of speakers, accent).

Table 3: SRL task overall performance of each site primary system.

site	IRIT	LIA	LIG	LIMSI	LIUM
%miss	1.4	1.4	1.6	1.5	1.0
%ins	0.6	1.3	1.7	0.4	1.6
%sub	12.0	12.4	7.6	10.5	8.2
%err	14.0	15.1	10.9	12.4	**10.8**

Table 5: Detail of the TRS task performances for each broadcaster for the top-4 sites.

Site	IRISA	LIMSI	LIUM	VR
Africa 1	29.7	12.1	18.3	15.0
Inter	28.2	14.0	20.2	17.5
RFI	16.9	7.0	10.8	9.3
TVME	23.9	11.2	16.7	13.3

4. 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). The proposed pilot TDC task, whose goal was to evaluate unsupervised adaptation of ASR systems to new topics, received no submission.

Table 4 analyzes some of the differences between systems and report performance for the TRS task. A comparison with the ESTER 1 campaign results show that most systems have improved, even though the ESTER 2 test data was much more difficult than the ESTER 1 test set, due to more spontaneous speech, a larger proportion of telephone speech, the presence of strong accent and of background noises. This improvement is partly due to the increase in the amount of training data available and partly to the expertise gained by the participants.

Table 5 shows the WER obtained by the four best systems on the different radio channels. Results vary greatly depending on the accent and on the speaking style. Shows from the RFI channel, containing non accented broadcast news, exhibit fairly low error rates. As dialectal accents become stronger, *e.g.*, on TVME or Africa 1, word error rates increase as the training data are mostly from non accented speakers. Finally, shows containing more spontaneous speech (with non accented speakers), as in Inter where the selected shows correspond to debates (with a large proportion of telephone speech) and entertainments, were obviously more difficult than non accented broadcast news.

All ASR systems rely on the same hidden Markov model paradigm either with proprietary softwares or with publicly available toolkits such as HTK, Sphinx or Julius. It is interesting to note that most participating sites developed specific systems both for Africa 1 and TVME to tackle the issues of specific vocabulary (in particular named entities) and accented speech.

5. Information extraction tasks

The information extraction tasks aim at extracting higher level information useful for indexing or document retrieval purposes. The main sub-task was named entity extraction and recognition (NE). A prospective task of transcription segmentation into syntactic units (SP) was proposed but could not be achieved due to difficulties to define a precise evaluation protocol.

5.1. Named entity detection

The named entity (NE) detection task on French was first implemented in ESTER 1 as a prospective task in order to define the first annotation guideline, corpus and scoring tools. In ESTER 2, the task was proposed as a standard one. Two subtasks were defined: detection on the reference transcriptions and detection on automatic transcriptions. For the automatic transcript subtask, in order to precisely measure the impact of the WER on named entity detection, three automatic transcripts with different WER were given to the participants.

The NE tag set consists of 7 main categories—persons, locations, organizations, human products, amounts, time and functions—and 38 sub-categories. The tag set considered is therefore much more complex than the one used in the NE extraction tasks of the MUC 7 and DARPA HUB 5 programs [6, 7], where only 3 categories were considered. The official error measure used was the slot error rate (SER) [8] but precision, recall and F-measure were also given for further analysis. For automatic transcripts, the slot error rate is computed after optimal alignment between the (normalized) reference and automatic transcripts.

As far as data are concerned, the entire ESTER 1 corpus was updated to fit the new annotation guidelines. As the annotation guideline had to be improved during the evaluation, only the dev and test set could be updated that implied a light lag with the training set.

More than twice the number of ESTER 1 participants have been involved in this task. This clearly shows an increase of the interest for the information extraction subject. Among the systems, three were based only on rules (LIMSI, LINA, LIT), two based on syntactic analysis in addition to rules (Synapse, Xerox) and two followed a machine learning approach based on conditional random fields (LIA, LSIS). Results are reported in Table 6. Good average performance can be achieved on reference transcripts while transcription error drastically impact named entity recognition. Moreover, transcripts from LIA and IRISA lacked capital letters which proved detrimental to most named entity detection systems. Apart from the LIA and LIMSI systems, most systems were originally designed for standard texts rather than transcripts. Though adaptations were made by the participants, systems specifically designed to deal with automatic transcripts remain more robust on such material.

6. Discussion

Few French speaking laboratories had experience in broadcast news transcription when the ESTER project started in 2003. Probably a major result of the two evaluation campaigns 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 two ESTER campaigns is the availability of a significant amount of specifically created resources. These resources include annotation conventions, protocols, scoring tools, about 300 hours of transcribed radio broadcast speech and around 1,600 hours of non transcribed speech data. 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.

Table 4: TRS task overall performance for each site primary system and comparison of some system parameters across sites.

Sites	IRISA	LIA	LIMSI	LIUM	LORIA	VR
WER	26.1	26.8	**12.1**	17.8	26.3	15.1
Audio corpus	150h	250h	na	300h	170h	500h
#states	8,000	5,200	12,000	7,500	10,900	11,600
#gaussians	200k	272k	370k	165k	272k	370k
#words	65k	88k	200k	120k	63k	100k
#pron	132k	na	268k	300k	126k	na
Broadcast news	2.4M	6M	100M	3.3M	36M	100M
News paper	600M	600M	1.4G	1.0G	570M	1.4G
Web	–	na	na	88M	80M	na
#pass	4	4	6	5	4	2
Real time factor	15	10	na	10	na	na

Table 6: NE task overall performance for each participating site (Slot Error Rate, Precision, Recall and F-measure)

	ref. transcript			LIMSI transcript			LIA transcript			IRISA transcript		
%WER				12.11			17.83			26.09		
Sites	%S	%P	%R	%S	%P	%R	%S	%P	%R	%S	%P	%R
LIA	23.9	86.46	71.85	**43.4**	79.52	59.45	**51.6**	76.51	55.02	**56.8**	72.26	49.02
LIMSI	30.9	81.15	70.94	45.3	75.13	62.33	55.5	70.50	57.52	61.2	66.13	50.67
LINA	37.1	80.75	55.48	54.0	71.98	44.01	60.4	68.76	40.84	65.2	63.66	35.66
LI Tours	33.7	79.39	65.82	50.7	71.36	54.16	80.8	56.59	46.46	82.9	51.28	42.38
LSIS	35.0	82.65	73.07	55.3	70.23	58.39	86.5	70.36	28.66	88.6	67.03	25.22
Synapse	9.9	93.02	89.37	44.9	76.39	67.16	60.7	70.26	59.21	66.2	65.95	52.71
Xerox	**9.8**	93.61	91.50	44.6	58.91	70.06	—	—	—	—	—	—

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 error rate which 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 occurrence 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 named entity detection task in ESTER 2 gathered much more participants, both labs and company, than in ESTER 1, which clearly shows the increasing interest in information extraction tasks and the maturity achieved by transcription tools. Thanks to this dynamics, the guideline, corpus and scoring tools for the named entity task have been significantly improved during the campaign and a specific working group has been created to think about the next named entity evaluation which will probably also evaluate sub-task such as information retrieval or named entity tracking.

Finally, we are currently working on a continuation of the ESTER 2 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 new campaign focusing on challenging issues such as processing large TV streams containing reports, talshows, debates and games in addition to news-like shows.

7. References

[1] J.-M. Dolmazon, F. Bimbot, G. Adda, M. El-Bèze, J. C. Caërou, J. Zeilinger and M. Adda-Decker. Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. In Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF, 3–18, 1997.

[2] 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 European Conference on Speech Communication and Technology, 2005.

[3] Jonathan G. Fiscus, Jerome Ajot, John S. Garofolo. The Rich Transcription 2007 Meeting Recognition Evaluation. In Multimodal Technologies for Perception of Humans: Inter. Eval. Workshops CLEAR 2007 and RT 2007.

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

[6] E. Marsh and D. Perzanowski. MUC-7 Evaluation of IE Technology: Overview of Results, Proc. of the Message Understanding Conference, 1998.

[7] The 1998 Hub-5 Evaluation Plan for Recognition of Conversational Speech over the Telephone, in English. Version 3.0, July 1998.

[8] J. Makhoul, F. Kubala, R. Schwartz and R. Weischedel. Performance measures for information extraction, Proc. of the DARPA Broadcast News Workshop, pp. 249-252, Virginia, USA, 1999