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

G. Gravier (1), J-F. Bonastre (1), E. Geoffrois (2), S. Galliano (2), K. Mc Tait (3), K. Choukri (3)

(1) Association Francophone
de la Communication Parlée
http://www.afcp-parole.org

(2) DGA/Centre technique d'Arcueil
16 bis av Prieur de la Côte d'Or
94114 Arcueil cedex

(3) ELDA
55-57 rue Brillat Savarin
75013 Paris

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

Abstract

This paper gives an overview of the ESTER evaluation campaign. The aim of this campaign is to evaluate automatic broadcast news transcription systems for the French language. The evaluation tasks are divided into three main categories: orthographic transcription, event detection and tracking (*e.g.* speech vs. music, speaker tracking), and information extraction (*e.g.* named entity detection, topic tracking). Each category is evaluated separately. This paper gives details about the tasks to be performed and the corpus, with particular emphasis on the manually transcribed reference transcription.

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. However, it is a difficult task as it requires crucial resources, usually manually validated, the production of which is not usually accessible to a single laboratory. Moreover, comparing performance can only be carried out on a well defined task, *i.e.* using standard databases and evaluation metrics.

In the United States, a long tradition of evaluation campaigns on speech and natural language technologies permitted the development of large annotated corpora with well defined evaluation paradigms. For example, evaluation campaigns organized by NIST and DARPA on automatic transcription (HUB 4, 1999; RT, 2003), topic retrieval (Wayne, 2000), named entity detection (ACE, 2001), and speaker recognition (Martin and Przybocki, 2001), to name a few, strongly contributed to fostering research in those fields.

As far as the French language is concerned, a first wave of evaluation campaigns had been initiated by AUPELF in the 1990s. In particular, this effort resulted in a first evaluation campaign on automatic transcription of read speech (Dolmazon et al., 1997). The ESTER campaign[1] is part of this ongoing effort for developing evaluation campaigns, corpora and evaluation paradigms for the French language. This campaign, organized jointly by the Francophone Speech Communication Association (AFCP), the French Defense expertise and test center for speech and language processing (DGA/CTA), and the Evaluations and Language resources Distribution Agency (ELDA), is part of the EVALDA project dedicated to the evaluation of language technologies for the French language[2], which started in 2003 and is due to finish in 2005.

ESTER focuses on the evaluation of rich transcription and indexing of radio broadcasts news in French. The rather recent notion of Rich Transcription (RT), introduced in NIST evaluations in 2002, consists in enriching the orthographic transcription with additional information. This task was chosen for three main reasons. First, dealing with broadcast news is a logical progression with respect to the previous AUPELF campaign on read speech transcription. Second, the tasks considered offer a strong application potential. And third, it complements the NIST Rich Transcription campaign on the English, Arabic and Chinese languages (RT, 2003). Compared to this campaign, though, ESTER does not includes 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 goals and organization of the campaign, the tasks considered and the corpora used in the evaluation.

2. About the campaign

This section first describes the scientific goals of the campaign before giving details on its implementation.

2.1. Objectives

The ESTER campaign has several objectives. The first goal is the promotion of an evaluation environment for speech processing in French by setting up a widely accepted evaluation framework. The second is to develop resources for evaluation on broadcast news material. These resources and related information are meant to be made available to as many laboratories as possible. In addition, beyond the pure evaluation of system performances, we also hope to federate research efforts by encouraging laboratories to share information and to collaborate. Workshops are organized throughout the campaign to meet this goal. The expected consequence of all this is a global improvement of transcription performance and new indexing approaches for broadcast news in the French language.

[1] ESTER is the French acronym for "Évaluation de Systèmes de Transcription enrichie d'Émissions Radiophoniques" (Evaluation of Radio Broadcast Rich Transcription System).

[2] The EVALDA project is sponsored by the French national Technolangue program.

As mentioned previously, one of the objective of this first broadcast news evaluation campaign for the French language is to make available a large annotated corpus for the tasks considered. This corpus, described in more detail in section 4, is the main element of the evaluation package which will be made available to the scientific community at the end of the campaign for a very low cost, in order to promote research activities in this field.

2.2. Implementation

The ESTER campaign is divided into two phases. Phase 1 is a pilot evaluation on a subset of the final corpus while phase 2 corresponds to the evaluation campaign itself. Each phase is followed by a workshop.

The phase 1 pilot evaluation, which started in June 2003 and was completed in January 2004, was aimed at validating and improving the evaluation paradigms and metrics using feedback from the participating sites. About ten sites, academic and industrial laboratories, participated with various levels of involvement. Only transcription and segmentation tasks were implemented in phase 1 (see section 3 for a detailed description of the tasks). For the transcription task, five sites returned results with word error rates ranging from about 20 to 50 percent. For most sites, the pilot study consisted in getting acquainted with the broadcast news transcription task and developing a system. Indeed, even if some sites had previous experience in read speech transcription, very few had experience with planned and spontaneous speech, and only one with broadcast news material.

The evaluation in phase 2 will be conducted on a larger corpus, for training as well as for testing (cf. section 4). It officially starts with the release of the additional data, scheduled for April 2004.

Participation in the ESTER campaign is opened to all interested participants on a voluntary basis. Participation is free and remains possible until the official start of the test phase, scheduled for late 2004. During the campaign, participating sites have access to the entire evaluation resource set. Sites actually participating in the final test stage, *i.e.* sites submitting results, will be allowed to keep all the data at no additional cost for research purposes. The evaluation data, in the form of an ”evaluation package”, will be made available by ELRA/ELDA to non participating sites shortly after the end of the test phase to enable reproduction of the test conditions of the campaign. Different licenses will be proposed, ranging from the right to use the data solely for evaluation puposes to the unlimited use of the data. A low cost package will be proposed to enable academic laboratories to work on the data.

3. Evaluated tasks

The ESTER evaluation implements three categories of tasks, namely transcription (T), segmentation (S), and information extraction (E). The first two constitutes the core of the campaign while the “information extraction” tasks are more prospective. The tasks are listed in table 1.

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.

Table 1: Evaluated tasks

abbrev.	description
T / TRS	orthographic transcription
T / TTR	real time transcription
S / SES	sound event tracking
S / SRL	speaker diarization
S / SVL	speaker tracking
S / SIL	interactive speaker tracking
E / EN	named entity detection
E / SD	document segmentation
E / ST	topic detection and tracking
E / QR	information retrieval (question answering)

3.1. Transcription

The transcription task is the classic task which consists in producing the orthographic transcription from the recordings, and is evaluated in terms of word error rates. In addition to the unconstrained transcription task (TRS), the TTR task will evaluate systems operating in real-time or less.

The use of resources other than the distributed ones is authorized, provided the additional resources are prior to the test corpus (prior to April 2004). Participants using additional data are encouraged to submit contrastive results solely based on the official training data.

3.2. Segmentation

The segmentation tasks aim at detecting, tracking and grouping together audio “events”, priorly known or not. Four tasks are considered, namely sound event tracking, speaker diarization, speaker tracking and interactive speaker tracking.

Sound event tracking (SES) consists in detecting portions of the document containing a particular event known beforehand. In this evaluation, sound events considered are speech and music. The task is therefore to identify, 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.

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.

Speaker tracking (SVL) 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.

Finally, interactive speaker tracking (SIL) is a variant of the speaker diarization category where a system may ask questions to an oracle to disambiguate decisions, thus simulating interaction with a human operator. For example, systems may ask whether two segments were uttered by the same speaker or not. Results will be evaluated as a function of the number of questions asked.

Each task in this category results in a segmentation of the document in terms of presence or absence of a particular event, hence the category name. The performance measure for such tasks is the classification error rate, computed with respect to time marks. For the tracking categories (SES and SVL), the considered measure is a weighted sum of the false acceptance and false rejection rates, relative to the classification rate of a "dummy" system that does not detect anything. A specific performance measure is considered for the diarization tasks 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. This is the same measure as used in the NIST evaluations (RT, 2003).

3.3. Information extraction

The information extraction tasks are aimed at extracting higher level information useful for indexing or document retrieval purposes, with an application oriented question answering task. The tasks in this category are named entity detection, document segmentation, topic tracking and question answering.

Named entity detection (EN) is the task of detecting, in the audio document, occurrences of an identified entity. In the ESTER framework, we limit ourselves to the detection of direct mentions of person, location, organization and event (historical, social, etc) names as well as dates and physical measures (*i.e.* followed by a unit). Indirect mentions are not considered in the scope of the current evaluation campaign. Performance will be evaluated based on the (automatic) transcription by counting the number of (correct) words correctly tagged as named entities after alignment of the automatic transcription with the reference. Alternate performance measures based on time rather than words will be explored.

Broadcast material is structured in terms of shows, reports and topics, possibly with advertisements between and within shows. The document segmentation (SD) task aims at retrieving this structure from the audio stream. This task is limited to document structure analysis and systems are not required to give any information on the (thematic) content of the different sections. However, we plan to also evaluate in this task systems that cluster together reports on the same topic (without topic identification).

In a similar way to speaker tracking, topic tracking (ST) aims at detecting portions of the document that match a given topic. We will limit ourselves to broad and general topics in the scope of the current evaluation. Examples of broad topics are 'sport' or 'politics' with corresponding general topics 'volley-ball' and 'Gulf war'.

The final information extraction task (QR) is dedicated to the evaluation of complete question answering systems. The goal is to answer a question formulated in natural language. As this is a prospective task, only a few questions will be considered and performances will be evaluated by human experts.

4. Corpora

Three main resources are distributed to participating sites, two of them being released or created in the framework of the ESTER project. The principal resource is the broadcast news corpus which consists of manually transcribed radio broadcast news shows. Text resources are also given for language modeling purposes. A less traditional resource consists of large amount (about 2000h) of non transcribed broadcast news material intended to explore research issues in unsupervised training and adaptation.

4.1. Transcribed audio resources

The main resource for the ESTER evaluation is a corpus containing 100h of manually transcribed radio broadcast news shows from various French speaking radio stations. The layout of the corpus is summarized in table 2. The availability of a 40h subset of the corpus, provided by DGA/CTA, made it possible to start the pilot evaluation early on in the project.

The training and development portions of the corpus contain material from four radio stations, namely Radio France International (RFI), France Inter, France Info and Radio Télévision Marocaine (RTM). The three first stations are French national radio stations while the last one is a Moroccan radio station. Only news shows were recorded, including advertisements.

The corpus is divided into three separate parts: a training, development and test corpus. The training corpus contains 82 hours of shows and the development corpus 8 hours. The test corpus contains 10 hours, 2 hours from each of the above mentioned sources plus 2 hours from a different source, unknown to the participating sites and thus labeled "surprise" in table 2. The unseen source in the test data is meant to evaluate the impact of prior knowledge of the document source on performance.

The corpus was carefully transcribed and annotated. Recordings are divided into sections that roughly correspond to the development of one of the news headlines, with separate (non-transcribed) sections for advertisements. Topic indices are associated to sections. Note that the section structure corresponds to the news broadcast structure considered in the document segmentation task. Sections are divided into speaker turns. For each speaker turn, the speaker identity (or possibly the speaker identities for multiple speaker turns), the channel and bandwidth and the detailed orthographic transcription are provided. The orthographic transcription is synchronized with the audio at regular intervals roughly corresponding to breath groups. Additional information is provided as necessary about pronunciation, non-linguistic events (such as lip noises or laughters), and named entity tags. Independently of sections or speaker turns, background events such as the presence of music are also indicated.

Annotations were carried out using the Transcriber software and more details on the annotation guidelines can be found in the software documentation[3].

[3] http://www.etca.fr/CTA/gip/Projets/Transcriber

Table 2: Content of the training (train), development (dev) and test (test) sets for the two phases of the campaign. All of phase 1 data becomes training data for phase 2.

source	phase 1		phase 2		
	train/dev	test	train/dev	non-trans	test
France Inter	19h40/2h40	2h40	8h/2h	∼ 200h	2h
France Info	–	–	8h/2h	∼ 800h	2h
RFI	11h/2h	2h	8h/2h	∼ 900h	2h
RTM	–	–	18h/2h	∼ 100h	2h
“surprise”	–	–	–	–	2h
total	40h		42h/8h	∼ 2000h	10h
period	1998–2000		2003	2004	2004

4.2. Untranscribed audio resources

In order to encourage work on the use of raw, non-transcribed, audio data for unsupervised training and adaptation, the audio corpus contains an additional part of approximately 2000 hours of non transcribed broadcast news shows. Participating sites willing to use this corpus are required to make available all transcriptions, whether manual or automatic, they may produce, thus enabling the production of a non-controlled transcribed corpus at very low cost. The resulting corpus will be distributed by ELRA/ELDA at the end of the evaluation campaign.

4.3. Textual resources

Two text corpora intended for language modeling are provided. The first consists of articles from the French newspaper “Le Monde”. Articles cover the period from 1987 to 2003 and contain approximately 300 million words plus topic tags for each article. The second corpus consists of transcriptions of debates of the European Council. This corpus, known as MLCC, contains 5.5 million words. Note that these are edited debates, that is elaborated transcriptions which reflect the content of the debates, rather than exact transcriptions. The manual transcriptions of the audio corpus described above provide a third textual resource.

Text resources are intended to be used for language modeling in transcription tasks but also as training material for the topic characterization related tasks.

4.4. Other resources

Other, unofficial, resources such as grapheme to phoneme conversion software or silence detectors are made available by participating sites to other participating sites for the sake of convenience. Such resources are listed on the campaign web site and most of them are freely accessible to non participating sites.

Furthermore, most current participants agreed to make available resources derived from their development work such as word graphs and automatic transcriptions of the development and test part of the audio corpus, or phonetic alignments on the training part of the corpus. These resources will be distributed with the audio corpus at the end of the campaign. They will also be made available on the campaign web site.

5. Conclusion

We have described the organization of the ESTER evaluation campaign for the rich transcription of French radio broadcast news. The recently completed pilot evaluation was very succesful in gathering most of the French speech recognition community, with many sites who have been participating enthusiastically and actively even though they are not specifically funded for their effort, and let us expect many interesting results for the official evaluation in early 2005.

In the future, we hope that this logic of ongoing evaluations will help create a strong and dynamic community in the field of spoken document transcription and indexing in the French language and that new techniques will emerge from these evaluations. One of our additional goals is the enlargement of this community to other speech related fields such as phonetics and linguistics. Making derived resources such as phonetic alignments, word graphs or automatic transcriptions available is a first step toward this goal and the organizing committee welcomes any request or suggestion in this direction.

6. References

ACE, 2001. *ACE 6-Month Meeting*. http://www.nist.gov/speech/tests/ace/phase2/doc/nyu-meeting.htm.

Dolmazon, Jean-Marc et al., 1997. 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*.

HUB 4, 1999. *Broadcast News Workshop*. http://www.nist.gov/speech/publications/darpa99.

Martin, Alvin and Mark Przybocki, 2001. The NIST Speaker Recognition Evaluations: 1996-2001. In *Speaker Odyssey*.

RT, 2003. *Spring 2003 Rich Transcription Workshop*.

Wayne, C., 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference*.