# Évaluations en Traitement Automatique de la Parole

# (ETAPE)

## Evaluation Plan

# ETAPE 2011

## Version 2.0


Contact Persons

Guillaume Gravier guillaume.gravier@irisa.fr
Gilles Adda gilles.adda@limsi.fr



ETAPE is partially financed by the French National Research Agency (ANR) under grant agreement ANR-09-CORD-009


http://www.afcp-parole.org/eval/etape.html

# Contents

# 1 Preamble

ETAPE is a project targeting the organisation of evaluation campaigns in the field of automatic speech processing. Partially funded by the French National Research Agency (ANR), the project brings together national experts in the organisation of such campaigns under the scientific leadership of the AFCP, the French-speaking Speech Communication Association, a regional branch of ISCA.

Partners of the ETAPE projects are, in alphabetical order: Association Francophone de la Communication Parlée, Direction Générale de l'Armement, ELDA S.A., Laboratoire National d'Essais, Laboratoire de Linguistique Formelle (Univ. Paris A), Laboratoire de Phonétique et Phonologie (Univ. Paris B).

The ETAPE 2011 campaign follows the series of ESTER campaigns [1, 2] organized in 2003, 2005 and 2009, targeting a wider variety of speech quality and the more difficult challenge of spontaneous speech. While the initial ESTER campaigns targeted radio broadcast news, the 2009 edition introduced accented speech and non news shows with spontaneous speech. The ETAPE 2011 evaluation will focus on TV material with various level of spontaneous speech and multiple speaker speech. Apart from spontaneous speech, one of the originality of the ETAPE 2011 campaign is that it does not target any particular type of shows such as news, thus fostering the development of general purpose transcription systems for professional quality multimedia material.

Though the ESTER and ETAPE campaigns targets, among other goals, the comparison of speech technologies, we wish these campaigns to have a spirit of collaboration rather than competition. We expect sites to share ideas and resources and to jointly improve speech technology.

The evaluation calendar is scheduled as follows:

| | |
|---|---|
| Mar. 2011 | release of the evaluation plan |
| Jun. 2011 | distribution of the training and development data |
| Nov.-Dec. 2011 | benchmark on test data |
| | *detailed scheduled to be announced* |
| Feb. 2012 | final workshop |

# 2 Tasks definition

As in the past, several tasks are evaluated independently on the same dataset. Four tasks are considered in the ETAPE 2011 benchmark. For historical reasons, tasks belong to one of the three following categories: segmentation (S), transcription (T) and information extraction (E). Table 1 summarizes the tasks considered, whose detailed descriptions are provided in the respective section below.

The multiple speaker detection task (SES-2) is implemented as an exploratory task given the lack of background.

## 2.1 Multiple speaker detection

**Task definition.** Multiple speaker detection is the task of finding all regions within a show containing speech uttered simultaneously by several speakers. The input to the system is a waveform file: speech detection and transcripts will not be provided. Participants are expected to return for each test file the start and end times of segments

| category | task | description |
|---|---|---|
| S | SES-2 | multiple speaker detection |
| | SRL | speaker turn segmentation |
| | SRL-X | cross-file speaker turn segmentation |
| T | TRS | lexical transcription |
| E | EN-ref | named entity detection on reference transcripts |
| | EN-asr | named entity detection on automatic transcripts |

Table 1: List of tasks for the ETAPE 2011 evaluation campaign.

containing speech from multiple speakers, as defined by the annotation guidelines below. Optionally, the type of each segment of superimposed speech can be provided though it will be ignored for scoring. Identities of the speakers involved are not required.

**Evaluation metrics.** Several performance indicators—such as recall, precision, detection error rate, accuracy, BAC, etc.—will be calculated for diagnostic purposes. The official metric for comparing systems will be the F1-measure defined from the following recall and precision definitions

$$\text{recall} = \frac{\text{duration of multiple speech correctly detected}}{\text{total duration of multiple speech in the reference}}$$

$$\text{precision} = \frac{\text{duration of multiple speech correctly detected}}{\text{total duration of multiple speech detected}}$$

Regions not containing speech will not be considered for scoring.

**Annotation guidelines.** Four types of multiple speaker situations are considered:

- back-channel: minimal speech showing that the listener is following (hmm, oui, ok, ...)
- approbation/opposition: complementary speech with actual content but without trying to take the turn
- early start: the next speaker anticipates the end of previous speaker's turn, leading to a (short) period of overlapping
- voluntary jamming: active "hostile" turn taking attempt, successful or not

All these situations were annotated. Note however that idenitfying the situation is not part of the task. No strict annotation guidelines were given for this exploratory task and annotators identified such regions with high tolerance. In particular, a region where two persons speak simultaneously often contains in reality a limited amount of truly multiple speech (i.e., multipitch signal) because of pauses and of communication strategies. To avoid oversegmentation, the entire region is marked as containing multiple speakers. See examples in the training data for a better idea of the annotations.

## 2.2 Speaker turn segmentation

**Task definition.** Speaker turn detection, aka speaker diarization, is the task of partitioning a document into speakers, grouping into the same anonymous class all

segments from the same speaker. The input to the system is a waveform file: speech detection and transcripts will not be provided. Two variants will be considered, depending on whether speaker turn detection is to be performed independently on each input file (SRL) or simulataneously on an input collection (SRL-X). In this last case, one has to group together all segments from the same speaker across all files in the collection while in the first case, attributing segments from two distinct input files to the same speaker is not required. In both cases, the output of the system is a list of arbitrary speaker IDs with, for each speaker found, a list of segments, each identified by a start and an end time in the corresponding file, in which this speaker is present.

**Evaluation metrics.** The evaluation metric will be the standard diarization error rate (DER). Speakers of the hypothesis are mapped with speakers of the reference in a one to one fashion (some may be left out). The time in error is then enumerated:

- Confusion: Time of a hypothesis speaker overlapping a reference speaker with which he is not associated
- False alarm: Time of a hypothesis speaker with no reference speaker in front of it
- Miss: Time of a reference speaker with no hypothesis speaker in front of it

$$DER = \frac{Confusion + FalseAlarm + Miss}{Referencespeakertime}$$

Of all possible mappings the one chosen (using the Hungarian algorithm, for the curious) is the one giving the lowest DER.

**Annotation guidelines.** Speaker turns were annotated as defined in the transcription guidelines accompanying Transcriber (Transcription guidelines version X). Known speakers are identified by their names throughout all files. Unknown speakers are local to the file were they appear. Multiple speaker regions were annotated as described in the previous section.

## 2.3 Lexical transcription

**Task definition.** Lexical transcription aims at providing a normalized orthographic transcription of an input document. The input to the system is a waveform file: speech detection and speaker turn segmentation are not provided. Output is a set of time- and speaker- stamped words, where the term *word* is used for simpliciy but rather refers to a (normalized) lexical token (see evaluation metrics).

**Evaluation metrics.** Word error rate (WER), obtained by alignment between a normalized reference transcript and the hypothesized transcripts, will be used as the primary measure for comparing systems. For diagnostic purposes, various performance measures will be implemented such as lemma error rate (LER). Note that submissions should be normalized by participants according to the guidelines described in Appendix B. However submissions will be renormalized to some extent before scoring. Multiple-speakers segments will be scored using the NIST RT09 methodology.

The main secondary metric, designed to better handle multiple-speaker regions, will be the Speaker-Attributed Word Error Rate. The words are tagged with the

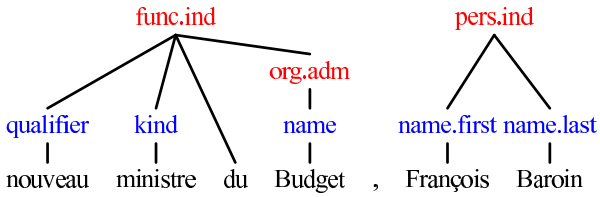


Figure 1: Samples of annotation in types (red tags) and components (blue tags) with different levels of annotation: *new minister of budget , François Baroin.*

diarized speaker label, the speaker mapping is done a per the DER method, and the word alignements are done per-speaker.

**Annotation guidelines.** Annotations are performed according to the standard transcription guidelines (see link to the transcription guide on the web site).

## 2.4 Named entity detection

**Task definition.** The named entity task consists in detecting all direct mentions of named entities and in categorizing the entity type. The taxonomy follows the LIMSI Quaero definition as per the version 1.22 of the guide. Two conditions will be evaluated, detection on manual transcriptions and detection on ASR. At least one of the ASRs will be a rover.

**Evaluation metrics.** Performance will be measured using the slot error rate (SER) metric [3]. It uses an *error enumeration* approach: collect the individual errors, associate a cost to each one, sum the costs and divide the total by the number of elements in the reference (the slots). We will use a simple weighting scheme where insertions ($I$), deletions ($D$) and elements with errors both in span and in type ($S_{ST}$) cost 1, while elements with errors only in either span ($S_S$) or type ($S_T$) cost 0.5. Span or type errors are substitutions ($S_S$, $S_T$, and $S_{ST}$). The final score is then given by

$$\text{Slot Error Rate} = \frac{D + I + S_{ST} + 0.5 \times (S_S + S_T)}{Ref}$$

**Annotation guidelines.** Entity *types* are organized in a hierarchical way (7 types and 32 sub-types):

1. Person: *pers.ind* (invidual person), *pers.coll* (collectivity of persons);
2. Location: administrative (*loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup*), physical (*loc.phys.geo, loc.phys.hydro, loc.phys.astro*);
3. Organization: *org.ent* (services), *org.adm* (administration);
4. Amount: quantity (with unit or general object), duration;
5. Time: date *time.date.abs* (absolute date), *time.date.rel* (date relative to the discourse), hour *time.hour.abs, time.hour.rel*;

6. Production: *prod.object* (manufactury object), *prod.art, prod.media, prod.fin* (financial products), *prod.soft* (software), *prod.award, prod.serv* (transportation route), *prod.doctr* (doctrine), *prod.rule* (law);

7. Functions: *func.ind* (individual function), *func.coll* (collectivity of functions).

An entity is composed of at least one *component* and can include unannotated spans (e.g, determiners, prepositions). We distinguish components that are specific to an ENE type from transverse components that can be used in multiple ENE types. Transverse components are: *name* (the entity name), *kind* (hyperonym of the entity), *qualifier* (a qualifying adjective), *demonym* (inhabitant or ethnic group names), *val* (a number), *unit* (a unit), *object* (an object), *range-mark* (a range between two values). Specific components are: *name.last, name.first, name.middle, title* for pers.ind, *address.number, po-box, zip-code, other-address-component* for loc.add.phys, and *week, day, month, year, century, millenium, reference-era, time-modifier* for time.date.

# 3 Resources

Resources in adequation with the targeted type of speech will be made available to participating sites for training and development purposes, providing the signature of an agreement engaging the site to participate in at least one of the tasks[1].In addition, participants are free to use all resources available. We are aware that (partial) transcripts of the test data might accidentally appear on the web, thus possibly introducing a potential bias in transcription and named entity detection systems. Though the probability of such a fact happening is low, it is not null. We will therefore provide diagnostic tools for a post-evaluation analysis of language models and lexicons by participants so as to flag potential inclusion of test data transcripts in the ressources[2]. Participants will be required to report such figures in their system description.

We review below the data from the ESTER 2009 campaign which are very relevant to the ETAPE organization and which can be acquired at low cost from ELDA and DGA.

All scoring tools will be made available to participants for development purposes and for validation of the results.

## 3.1 ETAPE 2011 data set

Training and development data will be made available to participants for the sole purpose of benchmarking. The ETAPE 2011 data consists of 30 hours of radio and TV data, selected to include mostly non planned speech and a reasonable proportion of multiple speaker data. Table 2 below summarizes the data available and the sources. Note that the number of hours are reported in terms of recordings, not speech. It was measured that, in the ETAPE TV data, about 80 % of the recording contains speech. A more detailed table is provided in appendix. Samples can be obtained from the ETAPE 2011 web site.

All data were carefully transcribed, including named entity annotation, according to the respective guidelines mentioned in Section 2.

In the scope of the ETAPE ANR project, phonetic alignments and syntactic trees will enrich part of the ETAPE data set. Such data will not be made available to

[1]The agreement will be provided by the organizers upon sign up.

[2]e.g., OOV rate, named entity OOV rate, number of 3 and 4 gram hits per document, etc.

| genre | train | dev | test | sources |
|---|---|---|---|---|
| TV news | 7h40 | 1h40 | 1h40 | BFM Story, Top Questions (LCP) |
| TV debates | 10h30 | 5h10 | 5h10 | Pile et Face, Ça vous regarde, Entre les lignes (LCP) |
| TV amusements | – | 1h05 | 1h05 | La place du village (TV8) |
| Radio shows | 7h50 | 3h | 3h | Un temps de Pauchon, Service Public, Le masque et la plume, Comme on nous parle, Le fou du roi |
| Total | | | | |

Table 2: ETAPE 2011 data summary

participants for the evaluation campaign. However, participants will be provided this supplementary material for research purposes after completion of the project.

## 3.2 Data sets from ESTER 2 (2009)

In addition to the ETAPE 2011 specific data, participants are allowed to use any data, audio or textual, provided they were collected prior to May 1, 2011. In particular, participants are invited to make extensive use of the ESTER data sets distributed by ELDA and by the DGA[3].

**ESTER 2.** The data set resulting from the ESTER 2 evaluation campaign comprises about 250 hours of radio broadcast transcribed by human listeners, as well as the newspaper corpus Le Monde from 1987 to 2003. For the most part, the transcribed audio corpus contains radio broadcast news, from French, Moroccan and African radio stations. More spontaneous data can be found in limited amount in the test set of the ESTER 2 final evaluation campaign. See [2] for more details.

| type | amount | comments |
|---|---|---|
| French news | 185h | radio broadcast news from national French channels (mostly planned speech) |
| Moroccan news | 35h | radio broadcast news from the Moroccan channel RTM (mostly planned speech, light accent, with Arabic pronunciations of proper names) |
| African news | 15h | radio broadcast news from the African channels (mostly planned speech but strong accents and sometimes heavily degraded acoustic conditions) |
| Radio debates | 4h | Debates and interactive programs from the French national radio channel France Inter (Le Téléphone sonne ; etc.) |

Table 3: ESTER 2 data set description

During the ESTER 2009 campaign, DGA also made available fast transcriptions of 37h of African news for training purposes.

Note that part of the ESTER data include named entity annotations. However, the original annotations distributed with the package were made with conventions which

[3]The ESTER evaluation data can be partly purchased from ELDA under the reference ELDA-E0021 at a very limited price for evaluation purposes. An additionnal part of the ESTER data set is made available directly by DGA upon signature of an bilateral license agreement. Contact the ETAPE organizers should you need these data.

substantially differ from the ones used in the ETAPE campaign. A reannotation of this data set will be made available by partners of the OSEO funded French Quaero project.

Please contact us should you be interested in ESTER 2 data.

**EPAC.** As a complement to the ESTER 2 data set, a large amount of untranscribed data ($\simeq$ 1,700 hours) was made available during the ESTER 2 evaluation campaign in 2009. Part of this data was transcribed by the Laboratoire d'Informatique de l'Université du Maine (LIUM) in the framework of the project "Exploration de masses de documents audios pour l'extraction et le traitement de la parole conversationnelle[4]" (EPAC). About 100 hours of mostly conversational speech from the ESTER 2 untranscribed audio corpus were transcribed by human listeners [4]. In addition, automatic transcripts of the entire untranscribed data are made available. The EPAC corpus can be obtained from ELDA under the reference ELDA-S0305.

# 4 Evaluation rules

The participants shall provide a description of the system or systems including the list of used resources, algorithms and methods chosen, and the memory usage and CPU time that have been necessary. They should be ready to present their work during the Etape workshop.

For every task, the following rules apply:

- The document origin (radio/tv channel) and the recording times will be available and are allowed to be used. Some of all of the test data may be recorded at a time of the day for which no training data is provided.
- The resources used must follow the constraints described in section 3 of this document.
- For every task, participants submitting multiple results shall identify one as *primary*, which will then be used for the official ranking. Other submissions will be considered as contrasts.
- Results submitted after the deadline will not appear in the official ranking.

In addition, the usual set of ground rules apply:

- The test audio data must not be listened to before or during the evaluation period.
- The evaluated systems must not be changed one the computation is started. A system should be run only once.
- The data handling must be entirely automatized. The results must not be modified in any case. The only allowable manual interventions are for starting computations, checking for correct behaviour and restarts in case of crashes.
- If multiple systems are evaluated for the same task, no results shall be examined until the last one is submitted. One and only one system must be identified as primary.

[4]Exploring audio data for conversational speech processing. http://projet-epac.univ-lemans.fr/doku.php

# References


[1] Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier. The ESTER Phase II evaluation campaign for the rich transcription of French broadcast news. In *European Conference on Speech Communication and Technology*, pages 1149–1152, 2005.

[2] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Conf. of the Intl. Speech Communication Association (Interspeech)*, pages 2583–2586, 2009.

[3] Makhoul, John; Francis Kubala; Richard Schwartz; Ralph Weischedel. Performance measures for information extraction. In DARPA Broadcast News Workshop, February 1999.

[4] Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet and Jérôme Farinas. The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In Proc. European conference on Language Resources and Evaluation, May 2010.


# A ETAPE data sources

| name | source | train | dev | test | total |
|---|---|---|---|---|---|
| BFM Story | BFM | 5h20 | 1h00 | 1h00 | 7h20 |
| Pile et Face | LCP | 4h05 | 0h30 | 0h30 | 5h05 |
| Ça vous regarde | LCP | 2h50 | 1h05 | 1h05 | 5h00 |
| Entre les lignes | LCP | 3h40 | 1h05 | 1h05 | 5h50 |
| Top questions | LCP | 2h20 | 0h40 | 0h40 | 3h40 |
| La place du village | TV8 Mont Blanc | 1h05 | 1h05 | 2h10 | 4h20 |
| Un temps de Pauchon | France Inter | 1h20 | 0h30 | 0h30 | 2h20 |
| Service Public | France Inter | 1h00 | 1h00 | 1h00 | 3h00 |
| Le Masque et la plume | France Inter | 3h00 | 1h00 | 1h00 | 5h00 |
| Comme on nous parle | France Inter | 1h30 | 0h30 | 0h30 | 2h30 |
| Le fou du roi | France Inter | 1h00 | – | – | 1h00 |
| Total | | 27h10 | 8h25 | 9h30 | 45h |

# B Transcription normalization rules

The normalization rules will be described in the scoring package. In a nutshell:

- Case is not taken into account when scoring, but systems are encouraged to output truecased words.
- Punctuation is ignored.
- A space is inserted after every apostrophes present due to an elision, the apostrophe being glued to the left component. That means pretty much all of them with the exception of *aujourd'hui* and some proper nouns.
- Dash-separated compound words will be split and the dash removed.

- Numerical expressions in the reference are rewritten in textual form.
- Acronyms are left in their compact form (no spaces nor dots). Acronyms containing digits (A380, CO2...) will be accepted as an alternative form with the numerical part textualized.
- Hesitation words are replaced by *%hesitation* and are optional.
- Cut words are optional and are considered correct when the spoken part matches the prefix (or suffix, depending) of the hypothesis word.
- Badly pronounced words (but untruncated) are written in their dictionary form.
- Words with varying or uncertain spelling will have their variants added to the equivalence dictionary.
- A preliminary equivalence dictionary will be provided with the scoring package.

Spelling variants as accepted in dictionaries such as the Larousse, the Robert or the Grevisse are accepted. Very frequent variants found on the Internet may also be accepted. Participants can propose updates to the equivalence lists up to the start of the evaluation. After the results submission the organizers will create an update to the lists taking into account the test data and the systems outputs. The participants will have 48 hours to react to the proposal, and the non-rejected changes will be integrated for the official scoring.

# C Submission formats

All submissions should be encoded in iso-8859-1. Submission formats are subject to changes until the evaluation starts.

## C.1 Detection tasks

Overlapping speech detection will use the standard ETF format with "overlap" for the type, "-" for the subtype, and one of "backchannel", "complement", "early" or "jamming" as evenement. Each line in the submission file documents a segment according to

```
source 1 start duration type subtype event [score [decision]]
```

Speaker diarization will use the standard MDTM format.

## C.2 Transcription tasks

To allow experimenting with overlapping speech recognition evaluation, and in particular the speaker-attributed WER, the expected format will be a variation on the CTM format with added speaker information. The expected speaker labels are tokens as in the diarization task, and not real names.

The format will thus be a column format with an added column for speakers:

```
filename 1 start duration speaker word confidence
```

### C.3 Named entities extraction tasks

Input will be free-form text without markers of any kind. For the manual transcriptions the lines will match speaker turns, and the normalization (apostrophes, etc) will have been applied. For the automatic transcriptions ten tokens will be put per line. In addition the associated extended ctm files will be provided.

The systems will have to insert XML tags corresponding to the entities in the appropriate places. The tokenization should stay the same with the tags removed (please don't remove spaces).

Note that such transcripts will be distributed "as is" and, depending on the system, might be non capitalized. Participants are asked to make submissions on as many ASR transcripts as possible. Participants using their own ASR are encouraged to submit contrastive runs on the ASR transcripts provided.