The ETAPE corpus for the evaluation of speech-based TV content processing in the French language

Guillaume Gravier[1], Gilles Adda[1], Niklas Paulsson[2], Matthieu Carré[2], Aude Giraudel[3], Olivier Galibert[4]

Association Francophone de la Communication Parlée (AFCP)[1]
Evaluations and Language Resources Distribution Agency (ELDA)[2]
Direction Générale de l'Armement (DGA)[3]
Laboratoire National d'Essais (LNE)[4]
Contacts: guillaume.gravier@irisa.fr, gilles.adda@limsi.fr
http://www.afcp-parole.org/etape.html

Abstract

The paper presents a comprehensive overview of existing data for the evaluation of spoken content processing in a multimedia framework for the French language. We focus on the ETAPE corpus which will be made publicly available by ELDA at the end of 2012, after completion of the evaluation, and recall existing resources resulting from previous evaluation campaigns. The ETAPE corpus consists of 30 hours of TV and radio broadcasts, selected to cover a wide variety of topics and speaking styles, emphasizing spontaneous speech and multiple speaker areas.

Keywords: evaluation, speech recognition, named entity

1. Introduction

Evaluation of speech and language technology have long been a critical issue. From the early 70s, many efforts were made to design comparative evaluation campaigns, mostly in the USA under the guidance of NIST and DARPA. Thanks to clear evaluation protocols and data availability, impressive progress were made since the early evaluations on Resource Management, now making it possible to tackle large vocabulary spontaneous speech as done in the serie of rich transcription campaigns implemented since 2002. Being driven by American agencies, these evaluations mainly focus on English, Arabic and Chinese, all languages deemed vital from the security and commercial point of view.

In contrast, European languages other than English have received far less attention with most evaluation focusing on broadcast news processing (D'Halleweyn et al., 2006). The French language has been no exception up to now with the serie of ESTER evaluation campaign targeting the transcription of radio broadcast data, with particular emphasis on news (Galliano et al., 2005; Galliano et al., 2009).

This paper presents data sets and tasks of the ETAPE evaluation campaign, a follow-up of the ESTER campaigns targeting a wider variety of data coming both from radio and TV sources and emphasizing spontaneous speech and overlaping speakers. We first present in Section 2. the context of the evaluation campaign and describe in Section 3. the targeted scientific objectives and the task implemented. We then present in Section 4. the data that was made available specifically for the ETAPE evaluation. We also take the opportunity of this paper to review publicly available language resources in the French language for access to multimedia data, as an outcome of 10 years of evaluation campaigns and of national projects exploiting similar data.

2. The ETAPE project

ETAPE is a project targeting the organization of evaluation campaigns in the field of automatic speech processing for the French language. Partially funded by the French National Research Agency (ANR), the project brings together national experts in the organization of such campaigns under the scientific leadership of the Association Francophone de la Communication Parlée (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 (AFCP), Direction Générale de l'Armement (DGA), Evaluations and Language Resources Distribution Agency (ELDA), Laboratoire National d'Essais (LNE), Laboratoire de Linguistique Formelle de l'Univ. Paris 7 (LLF), Laboratoire de Phonétique et Phonologie de l'Univ. Paris 3 (LPP). AFCP is responsible for the organization of evaluation campaigns. DGA and ELDA are in charge of data acquisition and reference transcription, ELDA being in charge of making the data publicly available after competion of the benchmark. LNE implements the practical aspects of the benchmark, providing scoring tools and computing the official results.

The ETAPE campaign follows the series of ESTER campaigns (Galliano et al., 2005; Galliano et al., 2009) organized in 2003, 2005 and 2009. The initial ESTER campaigns, in 2003 and 2005, targeted radio broadcast news and the 2009 edition introduced accented speech and non news shows with spontaneous speech, while ETAPE targets a wider variety of speech quality and the more difficult challenge of spontaneous speech. To do so, The ETAPE evaluation focuses on TV material with various level of spontaneous speech and overlapping speech from multiple speakers. Indeed, spontaneous speech, sometimes with overlapping speech and with various background noises (studio, indoor, outdoor), is one of the characteristics of real-life TV

contents that still challenge current speech processing technology. Apart from spontaneous speech, one of the originality of the ETAPE 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.

3. Scope of the evaluation

As in the past, several tasks will be evaluated independently on the same data set. 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.
Most tasks are similar to previous evaluations, the main change coming from the material under study rather than from the task definitions. Moreover, evaluation metrics will take into account regions of superimposed speech that were discarded in previous evaluations for the French language. In particular, the multiple speaker detection task (SES-2), the task of finding all regions within a show containing speech uttered simultaneously by several speakers, is implemented as an exploratory task given the lack of experience with this task. Note also that, while the name entity detection task has remained unchanged, annotation conventions, rapidly discussed below, were modified from previous evaluations.

3.1. Segmentation tasks

Two subtasks are considered, namely multiple speaker detection and speaker turn segmentation.

Multiple speaker detection. 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 containing speech from multiple speakers. Several performance indicators—such as recall, precision, detection error rate, accuracy, BAC, etc.—will be calculated for diagnostic purposes, either time based or event based.

Speaker turn segmentation. 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. The evaluation metric will be the standard diarization error rate (DER).

3.2. Transcription task

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 (enventually) speaker- stamped words, where the term *word* is used for simpliciy but rather refers to a (normalized) lexical token. 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. To account for overlapping speech, a bag of word representation is adopted in overlapping regions, a word being counted as correct if present in the bag of word. Alternately, speaker attributed word error rate will be used.

3.3. Entity detection task

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 defined in (Rosset et al., 2011). Performance will be measured using the slot error rate (SER) metric (Makhoul et al., 1999). Two conditions will be evaluated, detection on manual transcriptions and detection on ASR where at least one of the ASRs will be a rover from all the submissions of the transcription task. The ROVER as well as some of the participants' submissions will be made available by ELDA with the evaluation package after completion of the campaign.

4. The Data

We describe briefly the amount of data collected and annotated specifically for the ETAPE evaluation campaign. Details on annotation guidelines are provided with the tasks descriptions in the previous sections. In this section, we focus on describing the sources and the amount of data made available. As ETAPE builds upon previous benchmarking initiatives targeting the rich transcription of radio broadcast news, we briefly review in Section 5. the data made available as a result of the ESTER campaigns. Indeed, the ESTER and ETAPE data jointly constitute a large corpus for the development of spoken processing technology for the French language for access to multimedia data.

4.1. Data description

The ETAPE data consists of 13.5 hours of radio data and 29 hours of 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, 77 % of the recording contains speech. From the about 22 hours of speech material present in the data, about 1.5 hours correspond to multiple speaker areas, which corresponds to about 7 % of the time over all shows. This amount of overlapping speech is significantly higher to what can be observed in traditional broadcast news data and therefore validates the research directions promoted by ETAPE, focusing on the processing of overlapping speech with the pilot task on overlapping speech detection and the diarization and transcription metrics taking into account such regions which were traditionally discarded from previous evaluations.

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.

genre	train	dev	test	sources
TV news	7h30	1h35	1h35	BFM Story, Top Questions (LCP)
TV debates	10h30	2h40	2h40	Pile et Face, a vous regarde, Entre les lignes (LCP)
TV amusements	–	1h05	1h05	La place du village (TV8)
Radio shows	7h50	3h00	3h00	Un temps de Pauchon, Service Public, Le masque et la plume, Comme on nous parle, Le fou du roi
Total	25h30	8h20	8h20	42h10

Table 2: ETAPE 2011 data summary

News shows follow the standard pattern of such shows with stories introduced by one or several anchor. Though lot of similar material already exist, the idea is to include well-known types to measure progress from previous evaluations. Debates, often with several guests, include more interactive material and therefore contain more spontaneous speech and overlapping regions. The amount of overlapping speech in such shows is of 9 % as opposed to 2 % in news shows. The radio shows of the ETAPE corpus are mostly debates, sometimes with difficult acoustic conditions because of outdoor interviews (e.g., Un temps de Pauchon). Finally, the show "La PLace du Village" comes from a regional channel and exhibit difficult acoustic conditions and strongly accented speech.

4.2. Annotation guidelines

All data were carefully transcribed, including named entity annotation.

Lexical transcription of the data follows the classical guidelines for the French language (D; 2008). Disfluencies, repairs and discourse markers were also annotated. An example of a transcription is given in Fig. 1.

Overlapping speech regions were indicated in the data where four types of multiple speaker situations were 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

Apart from the above definitions, 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.

Direct mentions of named entities were also annotated throughout the entire data set. We provide here a brief overview of the entity annotation guidelines. For details, the reader is refered to either the full document (in French) or to (Grouin et al., 2011).

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*

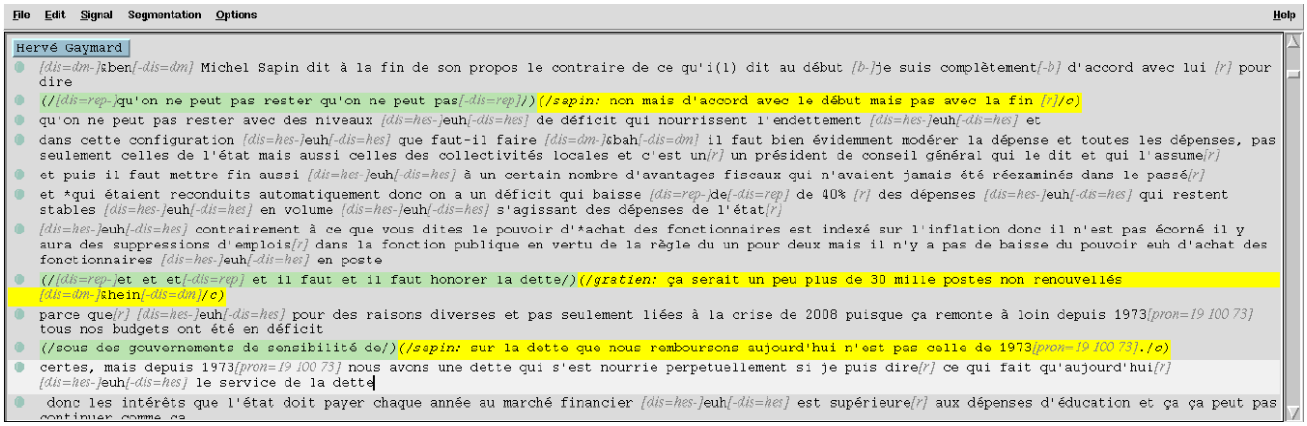

Figure 1: Example of a transcription

(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.

4.3. Phonetic and syntactic enrichment

In the scope of the ETAPE ANR project, phonetic alignments and syntactic trees will enrich part of the ETAPE data set. The idea is to have a unique rich corpus combining lexical, phonetic and syntactic information. To the best of our knowledge, this will constitutes one of the very few resource with both phonetic and syntactic annotations, thus constituting a unique opportunities for speech science. While a detailed description of these enrichment is out of the scope of this paper, we will briefly mention the methodology considered and the expected outcome.

For phonetic alignments, automatic forced alignments using hidden Markov models are first performed. Trained phoneticians are then asked to verify these alignments, adjusting word boundaries and correcting phonetic transcription wherever necessary. In order to avoid the influence of an existing phonetic alignment, words for which a correction was proposed are further processed where a different annotator is asked to provide the phonetic transcription directly from the signal.

Syntactic parses are obtained by manually correcting the output of an existing parser (Petrov et al., 2006) and functional labeller (Candito et al., 2009) which was trained for written texts and adapted to spoken content by incorporating preprocessing and postprocessing rules for handling disfluencies, repairs and overlaps.

5. Review of the ESTER data

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.

5.0.1. 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 (Galliano et al., 2009) for more details.

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 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.

5.0.2. 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'Universit 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[1]" (EPAC). About 100 hours of mostly conversational speech from the ESTER 2 untranscribed audio corpus were transcribed by human listeners (Estève et al., 2010). 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.

6. Conclusion

The paper has presented the details of the ETAPE evaluation benchmark taking place in spring 2012. This new benchmark extends existing ones with new types of spoken

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

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 Tlphone sonne ; etc.)

Table 3: ESTER 2 data set description

content, thus facing challenging conditions such as background noise, spontaneous speech and overlapping speech (up to 7 %). The paper also presents a comprehensive overview of existing data for the evaluation of spoken content processing in a multimedia framework for the French language.

7. Acknowledgments

The ETAPE project is partially supported by the Agence Nationale de la Recherche (project ANR-09-CORD-009). We wish to thank Sophie Rosset, Cyril Grouin and Pierre Zweigenbaum for their invaluable work on the definition of the guidelines for named entities annotation and for making the guidelines available.

8. References

M. Candito, B. Crabb, P. Denis, and F. Guerin. 2009. Analyse syntaxique du franais : des constituants aux dpendances. In *TALN*.

Délégation Général l'Armement, 2008. *ESTER 2 : convention d'annotation détaillée et enrichie*. http://www.afcp-parole.org/etape/docs/Conventions_Transcription_ESTER2_v01.pdf.

Elisabeth D'Halleweyn, Jan Odijk, Lisanne Teunissen, and Catia Cucchiarini. 2006. The Dutch-Flemish HLT Programme STEVIN: Essential speech and language technology resources. In *Language Resources and Evaluation Conference*.

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

Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-Franois Bonastre, and Guillaume Gravier. 2005. 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.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. 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.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karn Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *Proc. of the Fifth Linguistic Annotation Workshop*.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *DARPA Broadcast News Workshop*.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING/ACL*.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. 2011. Entités nommées structurées : guide d'annotation quaero. Technical Report 2011-04, LIMSI-CNRS, Sep.