Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News

S. Galliano[1], E. Geoffrois[1], G. Gravier[2], J.-F. Bonastre[2], D. Mostefa[3], K. Choukri[3]

(1) DGA/Centre d'Expertise Parisien
16 bis av Prieur de la Côte d'Or
94114 Arcueil cedex

(2) Association Francophone
de la Communication Parlée

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

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

Abstract

This paper presents the audio corpus developed in the framework of the ESTER evaluation campaign of French broadcast news transcription systems. This corpus includes 100 hours of manually annotated recordings and 1,677 hours of non transcribed data. The manual annotations include the detailed verbatim orthographic transcription, the speaker turns and identities, information about acoustic conditions, and name entities. Additional resources generated by automatic speech processing systems, such as phonetic alignments and word graphs, are also described.

1. Introduction

The aim of the ESTER[1] evaluation campaign is to evaluate automatic broadcast news transcription systems for the French language (Gravier et al., 2004). Most of these systems are generally based on statistical methods and require a large amount of quality development data. The production of such crucial data is an investment that not every laboratory can afford. One of the goal of evaluation campaigns like ESTER is to develop and provide access to high quality resources, at a low cost for the participants.

As the corpus is both a key element in the development of broadcast news transcription systems and a useful ressource for other speech science related fields, this paper focuses on the description of the audio corpus developed in the framework of the ESTER evaluation campaign. We also describe additional resources that can be generated by automatic speech processing systems, such as phonetic alignments and word graphs. Text resources provided for the development of language models are not detailed in this paper. They consisted of articles from the French newspaper "Le Monde", from 1987 to 2003, with approximately 400M words.

The ESTER campaign implemented several tasks 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). The corpus is therefore designed to make possible the implementation of all those tasks on the same data. In designing the corpus, we benefited from the experience of the corresponding DARPA/NIST evaluation campaigns for the English language (and to some extent the Chinese and Arabic languages) in the domains of transcription (HUB, 1999; RT, 2004), speaker recognition (Martin and Przybocki, 2001) and information extraction (ACE, 2001).

In the process of creating a corpus, the first step is to analyse the tasks to implement and to list all the required in-

formation to extract from the audio signal (speech, noise, speakers...). The second step is to define the way this information will be represented and normalized. In ESTER specific annotation guidelines were defined for speaker related information, for the orthographic transcription and for named entities tagging.

The paper first presents the content of the corpus and give an overview of the annotation guidelines that were used. We then give some statistics of occurence for the various events considered, before listing resources automatically produced by transcription system and shared via a public repository. Information concerning the results of the campaign can be found elsewhere (Galliano et al., 2005) and will not be presented in this paper.

2. Corpus description

2.1. Audio recordings

The acoustic resources come from six different sources, namely: France Inter, France Info, Radio France International (RFI), Radio Télévision Marocaine (RTM), France Culture and Radio Classique. The first part of this audio corpus is made up of a total of 100 hours of radio broadcast news shows recorded in 1998, 2000, 2003 and 2004, which were manually transcribed. The second part is made up of 1,677 hours of non transcribed shows recorded between October 2003 and September 2004. The amount of data per broadcaster is listed in Table 1. Audio files were recorded in wave format at 16kHz 16-bits from a standard audio card on a PC without any compression. 90 hours of the transcribed data were devoted to training and development while the 10 remaining hours composed the test set. The non transcribed data was also made available for training, in order to investigate the improvements that can be achieved by using large amount of such data.

The two broadcasters *France Culture* and *Radio Classique* were included in the test set in order to study the performance of the transcription systems on broadcasters for which no training data were available. In the case of *France Culture*, only non transcribed data were available for train-

---

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

ing purposes while no data at all were available for training for *Radio Classique*.

Table 1: Amount of transcribed and non transcribed by broadcaster (in hours).

	transcribed	non transcribed
France Info	12	643
France Inter	37	337
RFI	27	445
RTM	22	-
France Culture	1	252
Radio Classique	1	-
Total	100	1,677

Figure 1: Three layers transcription scheme

2.2. Orthographic transcriptions

All speech in the recorded broadcast news shows was orthographically transcribed, except non-news passages such as commercials, songs, etc.

The transcribed data of the ESTER campaign is natively in Transcriber format and was produced using the Transcriber software (Barras et al., 1998)[2]. This tool enables the creation of a corpus which contains all the necessary information required for each task. Transcriptions are encoded in XML format and follow common guidelines of orthographic transcription. From the transcription in the Transcriber format, task specific information can be exported into appropriate formats (stm, etf, ...).

The transcription contains 3 layers of annotation as described in Figure 1. The first one is the **section** segmentation. A section can be a *report*, a *filler*, or a *nontrans* section. At the section level, *report* sections structure the audio file in journalistic topics like news, politics, economy, sport, etc. The *filler* sections concern part of the recordings that are not related to a specific theme but rather to transition between journalists or shows. Finally, *non-trans* sections are used when the signal is not transcribed. This is the case for example for advertisements or songs. Sections of type *report* are labeled with some topic information. However, this information is free form and not normalized throughout the corpus. Each section gathers **speaker turns** between different speakers, where a turn contains speech from a single speaker (or from the same two speakers in case of segments with multiple speakers). Each speaker turn includes information on the name and gender of the speaker, if he/she is a native speaker, the speech type (spontaneous vs. planned), the channel (telephone vs. studio). Each speaker turn is further divided into **speech segments**, usually corresponding to breath groups, for which a transcription is provided.

The transcriptions include punctuations and are case-sensitive, with the exception that no capital letter is used at the beginning of sentences. In particular, the transcription indicates filled pauses (such as the word 'euh' in French), truncated words, mispronunciations and non standard pronunciations. Non lexical phenomena are also marked with specific tags including noises, music, jingles, mouth noise,
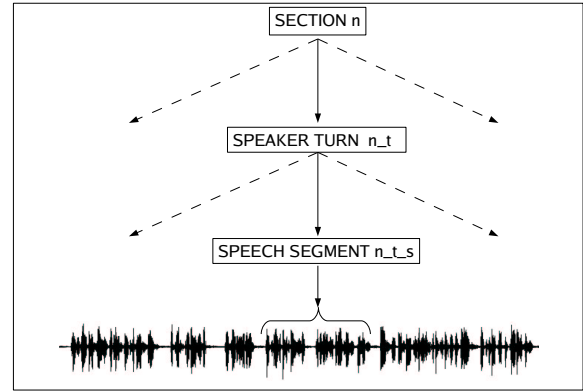
etc. In the case of truncated words, whenever possible, the word actually targeted by the user is indicated. For example, the transcription *consti(tution)* would indicate the the speaker intended to utter the word 'constitution' but only pronounced the first two syllables.

2.3. Named entities annotation

Named entities (NE) extraction is a well-established part of the field of information extraction with significant practical application. However there are very few speech corpora annotated in named entities, and, in particular, none publicly available for the French language. Direct mentions of named entities were tagged in the 100h transcribed corpus. We have distinguished 8 classes of entities:

- **Amount**. This category includes quantifiable data (age, duration, temperature, high, weight, width, distance, area, volume, speed, currencies).

- **Facility**. Facilities include buildings such as hospitals, factories, houses, museums, stadiums, . . .

- **GPE**. Geo Political Entities refer to politically geographical regions. These entities don't distinguish between a geographical region, its people or its government.

- **Localisation**. This category includes geographical areas, circulation axes, postal and electronic addresses and telephone numbers.

- **Organization**. Expressions, names, acronyms that refer to an organisation that can be of political, religious, cultural nature are annotated as organisation entities.

- **Person**. Real persons as well as imaginary persons are considered in this category.

- **Product**. This category includes awards, vehicles, artistic word, and printed work.

- **Time**. Both date and time expressions are annotated as Time entities.

- **Unknown**. This category is used for terms which are supposed to be Named entities, but are difficult to classify in one of the above categories.

---

[2]Transcriber is distributed as free software and is available at http://trans.sourceforge.net.

2.4. Statistics

The transcribed database comprises 100 hours of recordings including 2,172 different speakers and more than 1.2M running words with a vocabulary of 45k words and 15k different named entities. Non transcribed parts, mainly corresponding to advertisements and music areas, correspond to 1.76% of the data duration. In the transcribed part of the corpus, speech accounts for 97% of the signal and music for 2.3%, the rest being pauses.

2.4.1. Speakers

In the ESTER transcriptions, a speaker is described by his name, which is considered either as a global identifier for the whole corpus if it is his/her real full name, or as an identifier local to the transcribed file if the real name of the speaker cannot be retrieved. In addition to the name, we also describe if the speaker is male or female, if he/she is non native, and his/her accent if any.

The original transcriptions contains a total of 2,172 different speakers. About one third (744) are female speakers while 1,398 are male speakers. For about 20 young speakers (children) it was not possible to determine unambiguously the sex of the speaker. The majority of the speakers are native French speakers with a total of 1,778 speakers. The others, 394, are non-native speakers and come mainly from the RTM radio, which is a radio broadcasted in Morocco. However, professional speakers from RTM, though strictly speaking non native speakers, do not have a strong accent.

The transcribed data include 14,872 speaker turns with an average turn duration of 24 seconds. Overlapping speech or multiple speaker segments occur when two or more persons speak at the same time. This kind of data is challenging for automatic speech recognition systems. 840 turns out of the 14,872 turns are multiple speaker segments, representing 0.68% of the data duration. Not surprisingly, overlapping speech segments are shorter than other single speaker turns with an average duration of 2.8 seconds.

2.4.2. Transcriptions

The transcriptions include about 1,2M words for a lexicon of 45k words before normalization, and 1.1M words for a lexicon of 37k words after normalization. The proportion of mispronounced words is 0.16% while the one of truncated words is of 0.34%. Words for which the orthography remains unsure represent 0.14% of the corpus.

Narrow band telephone speech represents about 18% of the corpus, while speech with background music represents 8.5%. Most of the time, music is related to the news headlines with a high signal to noise ratio.

2.4.3. Named entities

The 100 hours of transcription contain 74,082 occurrences of named entities from 15,152 different named entities. The more important category is by far person names which correspond to nearly one third of the entities. Figure 2 gives an overview of the classification. Table 2 gives the 10 most frequent named entities of the transcribed data. If we except the two words "today" (*aujourd'hui*) and "yesterday" (*hier*) which are very frequent, we can see that the actuality

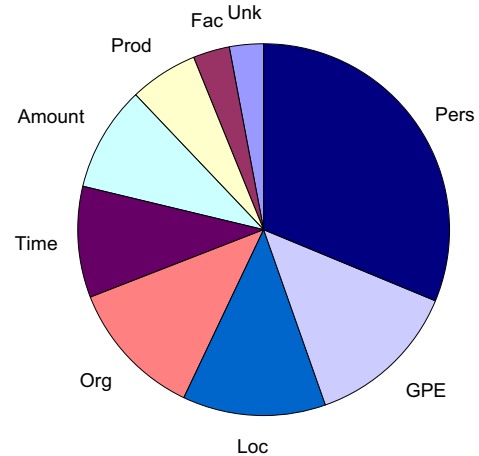has been domined by the Irak crisis with entities like Irak, Bagdad or Saddam Hussein.

Figure 2: Named entities repartition

Table 2: Top ten named entities

Named entity	#occurences
aujourd'hui	1735
hier	1456
France	1251
Irak	1236
Paris	834
États-Unis	763
France Inter	629
Bagdad	612
demain	538
Saddam Hussein	470

2.5. Quality assessment

The quality of an automatic system directly depends on the quality of the data used for its development. During the campaign, modifications have been made on the data as well as on the guidelines in order to improve their quality. Automatic alignments were used for finding errors in the transcriptions.

For the transcription task, an equivalence dictionary was created to take into account the alternative forms of certain words or expressions. For example, some common words can be written in two different forms (*e.g.*, *cdrom* or *cédérom*) or be commonly split (*e.g.*, *antiterroriste* or *anti terroriste*). This dictionary also encodes alternative spellings of proper names (*e.g.*, *baat* and *baath*) and expressions for which the number agreement is undetermined (*années de souffrances* vs. *années de souffrance*).

3. Additional resources

Automatic speech processing systems provide tools to automatically derive additional resources on top of the ESTER Phase II corpus. In particular, such resources include phonetic alignments of the ESTER training corpus, automatic
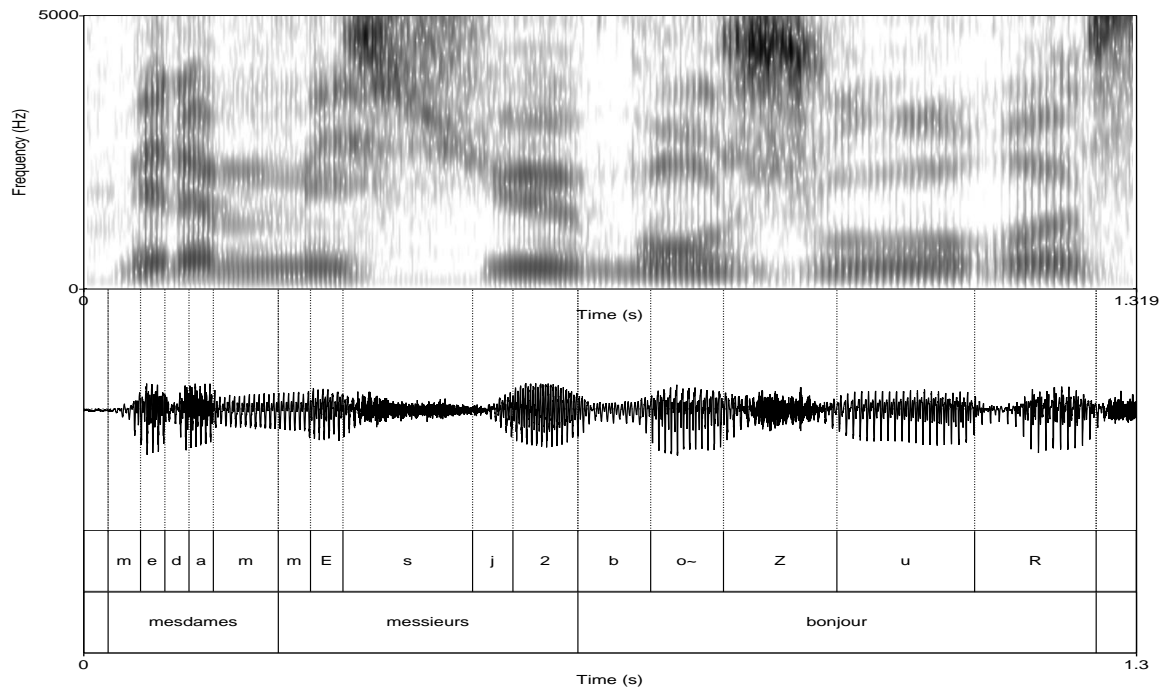
Figure 3: An example of an automatic phonetic alignment

transcripts including confidence measures, N-best hypotheses lists, as well as word graphs and sausages. As an illustration, figure 3 shows the result of an automatic phonetic alignment for one utterance.

Such resources are made available on a voluntary basis via a public repository on the ESTER web site.

4.   Conclusion

In this paper, we have presented the corpus developed for the ESTER evaluation campaign on rich transcription of broadcast news in French. Detailed information concerning the content of the corpus and the statistics of occurrences of various events that occur in speech (double talkers, fillers, etc.) was provided. We also presented additional resources that can be obtained by automatic processing.

The ESTER corpus is publicly available through ELDA's catalogue[3] for research or commercial use. In addition to the corpus itself, an evaluation package including resources which where used or produced during the ESTER campaign such as protocols, scoring tools and official results are under preparation and will be available shortly. The evaluation package will enable external players to evaluate their system and compare its results with those obtained during the official evaluation.

5.   Acknowledgement

6.   References

ACE 2001. *ACE 6-Month Meeting*, May 9–11, 2001. http://www.nist.gov/speech/tests/ace/phase2/doc/nyu-meeting.htm.

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proc. Language Evaluation and Resources Conference*, pp. 1373–1376, 1998.

S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri. The ESTER evaluation campaign for the rich transcription of French broadcast news. In *Interspeech - Eurospeech*, 2005.

G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri. The ESTER evaluation campaign of rich transcription of French broadcast news. In *Proc. Language Evaluation and Resources Conference*, 2004.

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

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

RT 2004. *Spring 2004 Rich Transcription Workshop*, 2004. http://www.nist.gov/speech/tests/rt/.

---

[3]http://www.elda.org

[4]http://www.technolangue.net