The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news

Yannick Estève[1], Thierry Bazillon[1], Jean-Yves Antoine[2], Frédéric Béchet[3], Jérôme Farinas[4]

[1]LIUM, University of Le Mans, France
[2] LI, Tours, France
[3] LIA, Avignon, France
[4] IRIT, Toulouse, France
yannick.esteve@lium.univ-lemans.fr

Abstract

This paper presents the EPAC corpus which is composed by a set of 100 hours of conversational speech manually transcribed and by the outputs of automatic tools (automatic segmentation, transcription, POS tagging, etc.) applied on the entire French ESTER 1 audio corpus: this concerns about 1700 hours of audio recordings from radiophonic shows. This corpus was built during the EPAC project funded by the French Research Agency (ANR) from 2007 to 2010. This corpus increases significantly the amount of French manually transcribed audio recordings easily available and it is now included as a part of the ESTER 1 corpus in the ELRA catalog without additional cost. By providing a large set of automatic outputs of speech processing tools, the EPAC corpus should be useful to researchers who want to work on such data without having to develop and deal with such tools. These automatic annotations are various: segmentation and speaker diarization, one-best hypotheses from the LIUM automatic speech recognition system with confidence measures, but also word-lattices and confusion networks, named entities, part-of-speech tags, chunks, etc. The 100 hours of speech manually transcribed were split into three data sets in order to get an official training corpus, an official development corpus and an official test corpus. These data sets were used to develop and to evaluate some automatic tools which have been used to process the 1700 hours of audio recording. For example, on the EPAC test data set our ASR system yields a word error rate equals to 17.25%.

1. Introduction

The EPAC project is related to process non-structured audio data. It involves four French academic laboratories: IRIT, LI, LIA, and LIUM. The EPAC project started on March 2007 and will end on August 2010.

The goal of the EPAC project is to propose methods for information extraction and document structuring which would be specific to audio data, taking into account all the information channels present in those data: signal segmentation (speech/music/jingle/etc.), speaker diarization, speaker named identification, speech transcription, topic detection and tracking, opinion speaker detection, speech analysis, conversational interactions, etc.

The audio data processed during the EPAC project come from the ESTER 1 campaign (Galliano et al., 2005), the evaluation campaign of automatic speech recognition systems on French broadcast news. The ESTER 1 corpus contains about 1700 hours of radio programs recorded in French, included 100 hours manually annotated: a large part of these data was not transcribed, only the audio data being distributed.

The EPAC project focuses on conversational speech processing. Among the various radiophonic shows, the part of conversational speech is often minimal and overlooked: a method for detecting spontaneous speech from large audio database has been proposed and evaluated (Dufour et al., 2009) in order to localize spontaneous speech in the ESTER 1 untranscribed audio recordings.

Among the main objectives of the EPAC project, two ones concern the building of a corpus composed by rich transcriptions of the ESTER 1 untranscribed data:

1. a set of 100 hours of conversational speech manually transcribed;
2. the outputs of automatic tools (automatic segmentation, transcription, POS tagging, etc.) applied on the entire ESTER 1 audio corpus and concerning about 1700 hours of audio data.

This paper describes these two different parts of this new corpus which will be easily accessible to the scientific community.

This research was supported by the ANR (Agence Nationale de la Recherche) under contract number ANR-06-MDCA-006.

2. The EPAC corpus

As seen above, the EPAC corpus was built from the audio data distributed during the ESTER 1 evaluation campaign. The EPAC corpus will be available as a part of the ESTER 1 corpus distributed by ELRA without additional cost: the ESTER 1 corpus, composed by 100 hours of manually annotated audio data is yet distributed by ELRA for 300 euros to academic labs which are ELRA members. Those which have already the ESTER 1 corpus will be able to acquire the EPAC corpus freely.

The EPAC corpus is composed by two different parts: manual transcriptions of conversational speech on French broadcast news and automatic multi-level annotations provided by the different work-packages from the EPAC projects.

2.1. Manual transcriptions

The manual transcriptions contained in the EPAC corpus have been realized between March 2007 and March 2009. It includes transcription and annotation for approximately 100 hours of speech, mostly spontaneous. These data have been taken from the untranscribed ESTER 1 corpus, and deal with three french radio stations: France Inter (30

hours), France Culture (40 hours) and RFI (25 hours). It is mainly shows, such as *Le Téléphone sonne*, *Quartiers d'Été*, *Sous les étoiles exactement*, *Culture vive* or *Les Matins*, with one animator/interviewer and one or more guests. Transcriptions and annotations have been made with *Transcriber*, an open source transcription tool (Barras et al., 1998).

2.1.1. Assisted transcription

In order to accelerate the manual transcription process and to reduce the financial cost, we have used a CAT (Computer Assisted Transcription) approach. Instead of transcribing from scratch, the human annotator was working from outputs of the 2005 LIUM automatic speech recognition system (Deléglise et al., 2005). The human annotator had to correct segmentation and transcription errors, to add punctuation and to mark some specific phenomena like noise, applause, *etc.*.
The gain of time provided by this CAT approach was measured and published in (Bazillon et al., 2008).

2.1.2. Additional annotations

In addition to orthographic transcription, a lot of metadata have been annotated. First, informations about speakers : full name, type (male or female), non-native speakers, kind of speech, used channel (studio or telephone), signal quality, etc. *Transcriber* software interface is optimized to process these data, as the annotator just has to choose between predefined choices for each of them. Then, several phenomena were accentuated with anchors:

1. every kind of noises: breathes, laughs, background noises, applauses...
2. specific pronunciations: erroneous or unclear ones, foreign words, some proper nouns, spelled or read acronyms...
3. grammatical incorrections: gender or number agreement, incorrect conjugations...
4. lexical problems: unknown words, uncertain spellings, neologisms...
5. (on a part of 10 hours of corpus) each "spoken" pronunciation: *"j'pense"* for *"je pense"*, *"pas d'problèmes"* for *"pas de problèmes"*, etc.

More, some in-text annotations have also been added, following the ESTER conventions. It concerns mainly truncations, elided words and syntactic breaks which were indicated with some brackets, as seen below:
"(il) y avait vrai() vraiment un () enfin bon, c'était curieux."
We have also experimented a new kind of annotations, which deals with broadcast types and speakers roles. Taking each file of the EPAC corpus, we have categorized transcribed data under tags such as "debate", "interview", "news", etc. When possible (especially for debate), complementary informations such as topics have been added. Concerning the speakers, some roles have been definished (animator, guest, presenter, interviewer, interviewee, chronicler, expert...) and often extended with speaker's profession. What's more, in case of contradictory debate, favourable or unfavourable opinions were mentionned. Lastly, a special tag was added when a speaker was "external": special correspondent, guest speaking by phone, etc. All these annotations have also been realized with *Transcriber*, thanks to section and speaker windows.

2.1.3. Features and difficulties of the transcription task

The first difficulties we had to take into account were about acoustic perception: prepared speech does not require specific attention to be transcribed, as it is usually pronounced by professional speakers (journalists in most case) who have a fluent and coherent diction. But as far as spontaneous speech is concerned, things are radically different: speakers may be very confused in their elocution, making repetitions, false starts, truncations...
Overlapping speech was undoubtedly the toughest difficulty we had to deal with: in spontaneous speech, this kind of phenomenon is quite frequent and poses several problems. First of all, the transcriber must be able to distinguish each speaker and each speech stream ; in case of more than two speakers, it is far from being an easy task. Second, the transcription software is particularly uneasy for spontaneous speech: it only allows the alignment of two simultaneous speech streams. It means that if a third (or more) speaker is concerned, the human transcriber won't succeed in temporally bringing his speech into line with the others two.

3. Automatic processing

The other part of the EPAC corpus concerns the 1700 hours of untranscribed audio data provided during the ESTER 1 evaluation campaign (including the 100 hours manually annotated as presented above). These 1700h have been automatically processed by using tools developed by the EPAC partners.
Several kinds of automatic treatments were applied:

1. audio segmentation and speaker diarization using the LIUM (reaching the 1^{st} position during the ESTER 2 campaign) system (Meignier and Merlin, 2010) and the IRIT system (El Khoury et al., 2009);
2. transcription using the LIUM ASR (Deléglise et al., 2009) which participated to the ESTER 2 evaluation campaign (Galliano et al., 2009): 1-best hypotheses are provided with confidence measures in addition to network confusions and word-graph generated during the transcription process;
3. POS tagging, chunk segmentation (Antoine et al., 2008);
4. Named entity detection, by using both the LI and LIA named entity detection systems which participated to the ESTER 2 evaluation campaign (Galliano et al., 2009). The LIA named entity detection system reach the 1^{st} position during this campaign.

The format file is based on XML and is inspired by the format used in the LUNA European project. Figure 1 shows an example of the header of such file containing automatic

```
<tools>
  <tool type="speaker diarization" name="LIUM_SpkDiarization" version="3.1" date="October 2009"/>
  <tool type="word transcription" name="LIUM_RichTranscription" version="EPAC-ESTER2" date="October 2009"/>
</tools>
<audiofile name="20040423_0000_1159_RFI_ELDA_part7" >
  <speakers>
    <speaker name="S474" identity="" type="generic label" gender="M" generator="auto"/>
    <speaker name="S117" identity="" type="generic label" gender="M" generator="auto"/>
    <speaker name="S0" identity="" type="generic label" gender="F" generator="auto"/>
    <speaker name="S217" identity="" type="generic label" gender="F" generator="auto"/>
    <speaker name="S123" identity="" type="generic label" gender="M" generator="auto"/>
    <speaker name="S85" identity="" type="generic label" gender="M" generator="auto"/>
  </speakers>
```

Figure 1: Example of an header in a file containing automatic transcriptions under the EPAC file format

```
  <segment start="819.570000" end="826.380000" bandwidth="S" speaker="S123" generator="auto">
    <text generator="auto"><sil/> un petit peu <sil/> pas toujours qui ne dure pas <sil/> <i/>
     <sil/> autant que <carillon/> non <sil/> <i/> <sil/> mais <sil/> eh <end/> </text>
    <graph id="0" type="1-best" generator="auto">
      <link id="0" start="0" end="1" type="filler" posterior="0.984"><sil/></link>
      <link id="1" start="1" end="2" type="wtoken" posterior="0.984">un</link>
      <link id="2" start="2" end="3" type="wtoken" posterior="0.984">petit</link>
      <link id="3" start="3" end="4" type="wtoken" posterior="0.984">peu</link>
      <link id="4" start="4" end="5" type="filler" posterior="0.984"><sil/></link>
      <link id="5" start="5" end="6" type="wtoken" posterior="0.984">pas</link>
      <link id="6" start="6" end="7" type="wtoken" posterior="0.984">toujours</link>
      <link id="7" start="7" end="8" type="wtoken" posterior="0.775">qui</link>
      <link id="8" start="8" end="9" type="wtoken" posterior="0.984">ne</link>
      <link id="9" start="9" end="10" type="wtoken" posterior="0.820">dure</link>
      <link id="10" start="10" end="11" type="wtoken" posterior="0.984">pas</link>
      <link id="11" start="11" end="12" type="filler" posterior="0.984"><sil/></link>
      <link id="12" start="12" end="13" type="filler" posterior="0.984"><i/></link>
      <link id="13" start="13" end="14" type="filler" posterior="0.984"><sil/></link>
      <link id="14" start="14" end="15" type="wtoken" posterior="0.961">autant</link>
      <link id="15" start="15" end="16" type="wtoken" posterior="0.961">que</link>
      <link id="16" start="16" end="17" type="filler" posterior="0.885"><carillon/></link>
      <link id="17" start="17" end="18" type="wtoken" posterior="0.493">non</link>
      <link id="18" start="18" end="19" type="filler" posterior="0.885"><sil/></link>
      <link id="19" start="19" end="20" type="filler" posterior="0.885"><i/></link>
      <link id="20" start="20" end="21" type="filler" posterior="0.796"><sil/></link>
      <link id="21" start="21" end="22" type="wtoken" posterior="0.615">mais</link>
      <link id="22" start="22" end="23" type="filler" posterior="0.709"><sil/></link>
      <link id="23" start="23" end="24" type="wtoken" posterior="0.481">eh</link>
      <link id="24" start="24" end="25" type="filler" posterior="0.984"><end/></link>
    </graph>
  </segment>
```

Figure 2: Example of an automatic transcription under the EPAC file format

segmentations and transcriptions. This header presents the tools used to build these outputs, gives the processed audio file name (only one audio file by XML file) and shows the list of speakers detected by the speaker diarization system. Information about speaker gender is also provided.

Figure 2 shows an example of an automatic transcription of a speech segment. First, the temporal position of this segment is given with the detected acoustic condition (studio or telephone) and the anonymous label associated to this speaker. Then, in a raw text format, the automatic transcription is provided, including *fillers* like silence ($< sil/ >$), *inspiration* ($< i/ >$) and other ones. Last, this automatic transcription (an 1-best hypothesis) is presented as a word-graph: in other files entire word-graph and network confusion are given in the same format. The nature of each item is provided (*filler* or word token) with its word posterior value: the posterior value permits to have an indication about the reliability of this word hypothesis.

In order to have an indication about the performances of the automatic tools used to build this corpus, we have used the manual EPAC corpus containing 100 hours of audio recordings manually annotated to make some experiments. This corpus was split into three data sets: a training set (80h), a development set (10h), and a test set (10h, corresponding to 120,785 words).

3.1. Automatic tools evaluation

Table 1 summarizes three experimental results on the EPAC test data. First, the speaker diarization error rate gives the performance of the automatic speaker diarization system. This diarization error rate obtained is really interesting and confirms the quality of the LIUM speaker diarization tool.

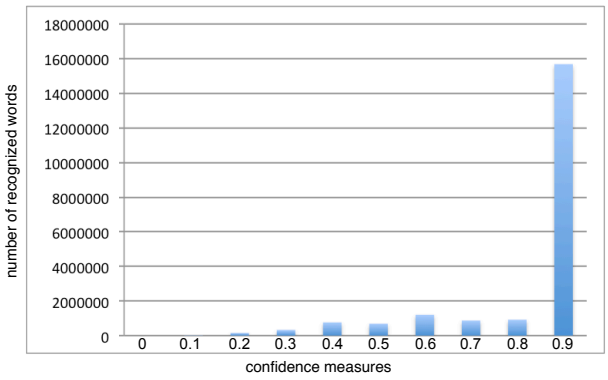

Figure 3: Number of recognized words according to their confidence measure

The word error rate gives the performance of the automatic speech recognition system on this test set and the normalized cross entropy gives an indication about the confidence measure which seems to be competitive.

Metric	Result
Speaker diarization error rate	5.71 %
Word error rate	17.25 %
Normalized cross entropy	0.306

Table 1: Evaluation of automatic tools on the test data from the EPAC corpus

While we cannot throw these results 'as is' on the entire EPAC corpus, they help us to get an idea about the accuracy which could be reached by processing these 1700h of audio with these automatic tools.

Figure 3 presents the distribution of the 20,603,957 recognized words in the audio data according to their confidence measure. If we assume, according to the experimental results on the EPAC test data, that the confidence measure is reliable, we can expect that the automatic transcriptions contain a reasonably low number of errors because the major part of recognized words is associated to very high values given by the confidence measure.

Last, table 2 shows the repartition of the 603,445 detected speech segments according to speaker gender or acoustic conditions. These 603,445 speech segments correspond to 1739 hours of audio data which are recorded in 2505 audio files.

	Male	Female	Studio	Telephone
#seg.	449,001	154,444	44,484	558,961
duration	1325h	414h	1613h	126h

Table 2: Number and duration of segments proposed by the automatic segmentation tool according to the proposed speaker gender or the acoustic condition. The total number of speech segments is 603,445 for a duration of 1739h

4. Conclusion

This paper presents the EPAC corpus built during the EPAC project funded by the French Research Agency (ANR). This corpus increases significantly the amount of French manually transcribed audio recordings easily available: a part (about 30h) of this corpus was already used during the ESTER 2 campaign in 2008 (Galliano et al., 2009). The EPAC corpus it is now included as a part of ESTER 1 corpus available in the ELRA catalog without additional cost. By providing a large set of automatic outputs of speech processing tools, the EPAC corpus should be useful to researchers who want to work on such data and who have no access to such tools: experimental results obtained on the manually annotated test data set of EPAC indicate that the automatic outputs of the tools applied to the 1700h of audio data should be precise enough in order to conduct various kinds of research work (linguistics, computational linguistics, speech recognition, multimedia indexing, etc.).

5. References

Jean-Yves Antoine, Abdenour Mokrane, and Nathalie Friburger. 2008. Automatic rich annotation of large corpus of conversational transcribed speech. In *European conference on Language Resources and Evaluation*, Marrakesh, Morocco.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Thierry Bazillon, Yannick Estève, and Daniel Luzzati. 2008. Manual vs assisted transcription of prepared and spontaneous speech. In *LREC*, Marrakech, Maroc.

Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin. 2005. The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news. In *Interspeech*, Lisbonne, Portugal.

Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin. 2009. Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate? In *Interspeech*, Brighton, Royaume-Uni.

Richard Dufour, Yannick Estève, Paul Deléglise, and Frédéric Béchet. 2009. Local and global models for spontaneous speech segment detection and characterization. In *The eleventh biannual IEEE workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, Merano, Italie.

Elie El Khoury, Christine Senac, and Julien Pinquier. 2009. Improved speaker diarization system for meetings. In *ICASSP 2009*, pages 4241–4244, Taipei, Taiwan, April.

Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *Interspeech*, Brighton, Royaume-Uni, Septembre.

Sylvain Meignier and Teva Merlin. 2010. Lium_SpkDiarization: An open source toolkit for diarization. In *CMU SPUD Workshop 2010*, Dallas, Texas, USA.