The ETAPE speech processing evaluation

Olivier Galibert[1], Jeremy Leixa[2], Gilles Adda[3], Khalid Choukri[4], Guillaume Gravier[5]
[1] olivier.galibert@lne.fr, LNE, France
[2] leixa@elda.org, ELDA, France
[3] gilles.adda@limsi.fr, LIMSI, CNRS, France
[4] choukri@elda.org, ELDA, France
[5] guillaume.gravier@irisa.fr, IRISA, France

Abstract

The ETAPE evaluation is the third evaluation in automatic speech recognition and associated technologies in a series which started with ESTER. This evaluation proposed some new challenges, by proposing TV and radio shows with prepared and spontaneous speech, annotation and evaluation of overlapping speech, a cross-show condition in speaker diarization, and new, complex but very informative named entities in the information extraction task. This paper presents the whole campaign, including the data annotated, the metrics used and the anonymized system results. All the data created in the evaluation, hopefully including system outputs, will be distributed through the ELRA catalogue in the future.

Keywords: ETAPE; evaluation campaign; overlapping speech; speaker diarization; ASR; named entities detection

1. Introduction

Starting in 2003, a number of evaluations have been conducted to assess the state-of-the-art of speech processing in French. The first two evaluations, called ESTER (Gravier et al., 2004; Galliano et al., 2009) put the emphasis on the handling of broadcast news, with an added twist of accents in the second one. This paper presents the third evaluation, ETAPE, which proposed some new challenges:

- TV and radio shows with prepared and spontaneous speech
- Handling of overlapping speech (detection, speaker segmentation and transcription)
- Cross-show condition in speaker diarization
- Complex named entities detection

2. The ETAPE evaluation

2.1. Overlapping speech detection

The first proposed task was to ask the systems to detect overlapping speech. They simply had to give time intervals in which overlapping speech happened.
Two metrics were proposed. The first one is a simple error metric, compute the false alarm and miss time amount and divide by the reference speech time:

$$OSDER = \frac{\text{miss} + \text{false alarm}}{\text{total reference time}}$$

An alternative metric, built to compensate for the difficulty for human to position precise frontiers for overlapping speech, is an event detection quality metric. For every overlapping speech interval of the hypothesis, compute the temporal position of its middle and check whether that time is noted as overlapping in the reference. That gives a precision. The same calculation the other way around gives a recall. The final metric is the F-measure of the two.

2.2. Cross-show Speech diarization

Speaker Diarization, also called "who spoke when", is the process of identifying, for each speaker of an input audio recording, all the regions where he/she is talking. Each temporal region containing speech should be labeled with at least one speaker-tag and segments from the same speaker shall be labeled with the same tag. Speaker tags are not identities but abstract labels. Systems were also required to collapse together interventions of the same speaker in multiple shows, hence the name of the task.
The main metric for diarization performance measurement is the *Diarization Error Rate*. It has been introduced by the NIST in 2000 within the Speaker Recognition evaluation (NIST, 2000) for their then-new speaker segmentation task. The metric is computed in two steps: the first step is to establish a mapping between the speaker tags provided by the system and the speaker identities found in the reference. The second step then computes the error rate using that mapping. Computing an error rate requires defining what the errors can be. Three error types are defined in the diarization context:

- The *confusion* error, when the system-provided speaker tag and the reference do not match through the mapping.
- The *miss* error, when speech is present in the reference but no speaker is present in the hypothesis.
- The *false alarm* error when speech has incorrectly been detected by the system.

These errors happen on segments of speech, of which the durations are summed together. Adding these durations gives us a time in error. This time in error is finally divided by the total reference speech time for normalization purposes. That gives us the final DER definition as:

$$DER = \frac{\text{confusion} + \text{miss} + \text{false alarm}}{\text{total reference speech time}}$$

The mapping establishment methodology and other subtleties due to the overlapping speech are described in (Galibert, 2013).

2.3. Speech transcription

Speech transcription is a task were the systems are required to tell what words were spoken and when. The Word Error Rate (WER) is the usual primary evaluation metric for this evaluation. That metric basically counts the number of word deletions, insertions and substitutions in the output of the automatic transcription system compared to a reference transcription produced by humans.
More precisely, the word error rate can be computed as shown in Equation 2.3.:

$$WER = \frac{S + D + I}{N}$$

where:

- S is the number of substitutions,
- D is the number of the deletions,
- I is the number of the insertions,
- N is the number of words in the reference transcription.

Due to the presence of overlapping speech we tried multiple ways of evaluating the system outputs:

- Dispatch the words in overlapping zones optimally to the different speakers (Optimally speaker-attributed word error rate)
- Ask the systems to add a speaker identity to the words, use that as a hard constraint (Speaker-attributed)
- Ask the systems to add a speaker identity to the words, add a confusion error type (Speaker-attributed with confusion)

As for the previous metric a complete description is available in (Galibert, 2013).
In addition the Normalized Cross-Entropy (NCE) metric has been used to evaluate the quality of the confidence values when provided. The cross-entropy between the confidence values and a perfect prediction (e.g. 0 for incorrect words and 1 for correct words) is computed and normalized with the maximum possible value, yielding a result in the interval $]-\infty, 1]$. Noting for each word w of the hypothesis its associated confidence $c(w)$, and noting n the number of correct words out of the N words of the hypothesis:

$$NCE = \frac{H_{max} + \sum_w \begin{cases} \log_2 c(w) & \text{if } w \text{ correct} \\ \log_2(1 - c(w)) & \text{otherwise} \end{cases}}{H_{max}}$$

$$H_{max} = -n \log_2 \frac{n}{N} - (N - n) \log_2 \frac{N - n}{N}$$

A well tuned system usually produces a result of over 0.2.

2.4. Named entities detection

The Named Entities detection task aims at detecting and classifying multi-word expressions useful for building fact databases from news data. The annotations followed a complex but powerful schema introduced within the Quaero project (Grouin et al., 2011). In that annotation schema the entities are both hierarchical, e.g. their types are structured in a tree-like fashion, and compositional, dividing them in typed sub-spans called components. That structuration allows for a better coverage of the interesting entities from an information retrieval point of view while avoiding an explosion in the number of different possible entities classes.
The evaluation follows the method described in (Galibert et al., 2011). It is done through a variant of the Slot Error Rate (SER) done in two steps:

- Associate annotations of the hypothesis and the reference
- Compute an error rate using these associations

All annotations are considered independently, ignoring the structure. The error rate is computed by counting the errors in classification, boundaries or both, plus insertions and deletions, and giving a weight for each error type. And, as usual, dividing by the number of annotations to find:

$$SER = \frac{I + D + 0.5E_c + 0.5E_b + E_{bc}}{R}$$

where:

- I is the number of inserted annotations
- D is the number of missed annotations
- E_c is the number of classification only errors
- E_b is the number of boundaries only errors
- E_bc is the number of both classification and boundaries errors
- R is the number of annotations in the reference

The evaluation is also done on the output of the ASR systems. The references on the manual transcription are projected on the automatic one with a tolerance on the annotation frontiers. That allows to use the same metric and hence have comparable results for the detection in manual and automatic transcriptions.

3. The Data

A pre-version of the corpus was presented in (Gravier et al., 2012), we will present here the final state.
The data proposed in the evaluation consisted of almost 22 hours of training speech, 7 hours of development speech and 7 hours for evaluation. The sources come from one radio, France Inter, and three TV channels, BFM TV, LCP and TV8 Mont Blanc. The raw amounts of transcribed audio are presented Table 1 and some statistics in Table 2.

Source	Show	Train		Dev		Test	
		Total	Overlap	Total	Overlap	Total	Overlap
BFM TV	BFM Story	4:01:47	0:17:28	0:44:30	0:05:55	0:44:38	0:05:19
LCP	Ca Vous Regarde	2:18:35	0:09:42	0:53:53	0:07:46	0:55:52	0:10:44
	Entre les Lignes	2:39:38	0:14:36	0:52:52	0:10:26	0:53:31	0:08:02
	Pile et Face	3:30:59	0:46:23	0:26:44	0:01:38	0:26:34	0:05:37
	Top Questions	1:19:50	0:00:30	0:28:38	0:00:19	0:28:23	0:00:19
TV8 Mont Blanc	La place du village	-	-	0:47:21	0:01:53	0:50:19	0:06:26
France Inter	Le Fou du Roi	1:17:57	0:06:02	-	-	-	-
	Un Temps de Pauchon	1:00:08	0:01:32	0:20:44	0:00:39	0:23:10	0:00:16
	Comme on nous parle	1:29:03	0:01:49	0:31:23	0:00:44	0:25:35	0:00:24
	Le Masque et la Plume	3:05:41	0:13:57	0:54:14	0:04:32	0:54:06	0:07:10
	La Tête au Carré	0:55:11	0:02:49	-	-	-	-
	Service Public	-	-	0:52:58	0:02:05	0:55:04	0:03:01
Total		21:38:49	1:54:48	6:53:17	0:35:57	6:57:12	0:47:18

Table 1: Amount of speech present in the ETAPE corpus

	Train	Dev	Test
Speech segments	23,017	7,189	8,581
Words	335,387	109,646	115,803
Entities	19,270	5,913	5,933
Components	27,656	8,410	8,609

Table 2: Some statistics for the ETAPE corpus

Lab.	Run	Time error	F-measure
A	all64m p0	46.3%	13.8%
	no64ms128m p0	40.4%	10.9%
	no64ms128m p10	40.0%	13.5%
	no64ms256m p0	39.5%	7.4%
B	primary	36.9%	15.2%
C	primary	37.7%	31.6%
	cms	34.8%	33.5%

Table 3: Overlapping speech detection results

4. Evaluation results

4.1. Overlapping speech detection

The overlapping speech detection results are available in table 3. We can notice that the time error is similar for all systems while the F-measure, which tries to evaluate the presence detection, varies a lot more. Looking at the data we can't help but notice that annotating overlapping speech is not an easy task for humans. Segements annotated as overlapping in the transcription are often overestimated, in particular in the case of short backchannels. Further experiments in that area will require the use of systems to refine the boundaries, but what systems is unclear in the first place. In particular the usual method of forced alignement on the reference on the acoustic signal may not be reliable on overlapping speech.

4.2. Speaker diarization

The speech diarizarion results are available in table 4. We can see that numerous experiments were done by the participants. We can notice that efficient methods were used for the cross-show condition, since the loss is really low. That validates that condition as what should be the default one in future evaluations.

Lab.	Run	DER ind.	cross.
A	primary	23.53	-
	2	22.10	-
	X primary	23.53	27.99
	X 2	24.54	28.61
B	primary	30.27	-
	no_map	31.05	-
	purif_mapetape_cms	28.70	-
C	primary	26.65	-
	secondary	27.65	-
	X primary	26.65	36.93
D	primary	22.84	-
	X primary	20.54	21.95
E	primary bic_ilp_flt2_jfa	18.46	-
	bic_ilp	20.83	-
	bic_ilp_flt2	19.19	-
	clr_sr	21.23	-
	clr_sr_flt2	19.82	-
	clr_sr_flt2_jfa	18.89	-
	clr_sr_flt2_jfa_gender	18.89	-
	primary bic_ilp_flt2_jfa_clr	18.96	19.71
	bic_sr_clr_ftl2_clr	20.58	23.10
	bic_sr_clr_flt2_jfa_clr	19.50	21.63
F	primary	21.96	-
	2	23.12	-
	3	25.54	-
	4	21.63	-
	5	22.79	-
	6	25.62	-
G	primary	15.61	-

Table 4: Speech diarization results

		No overlap		Optimal		Speaker conf.		Speaker att.	
		WER	NCE	WER	NCE	WER	NCE	WER	NCE
A	p1_rasr	32.66	-	39.11	-	-	-	-	-
	p1_speer	38.86	-	44.60	-	-	-	-	-
	bong	32.13	-	38.66	-	-	-	-	-
	rov_bong	29.95	-	36.72	-	-	-	-	-
B	primary	26.40	-0.527	33.06	-0.527	39.11	-0.717	57.62	-0.705
	2	27.24	0.230	33.89	0.226	39.99	0.031	56.70	0.024
	3	27.00	0.184	33.73	0.175	39.73	-0.016	56.77	-0.024
	4	25.27	0.240	32.03	0.236	38.07	0.039	55.77	0.032
C	primary	37.50	-	43.25	-	50.35	-	70.66	-
	2	36.28	-	42.22	-	57.52	-	93.89	-
	3	38.43	-	44.19	-	59.03	-	94.64	-
D	primary	21.83	$-\infty$	28.84	$-\infty$	33.80	$-\infty$	50.33	$-\infty$
	contraste	23.16	$-\infty$	30.06	$-\infty$	35.12	$-\infty$	51.98	$-\infty$
E	primary	26.23	-	32.71	-	43.17	-	74.06	-
	contraste1	32.62	-	38.84	-	49.44	-	80.95	-
	contraste2	31.10	-	37.12	-	57.25	-	107.33	-
All	rover	28.68	-	35.63	-	-	-	-	-
	oracle	10.95	-	-	-	-	-	-	-

Table 5: Speech transcription results

4.3. Speech transcription

The speech transcription results are available in table 5. The transcriptions were corrected through checking the results of an oracle rover, where all the system outputs are merged together and the best solution is chosen given the reference. The error rates of the oracle rover go from 1.63% (Top Questions) up to 39.55% (Un Temps de Pauchon). The standard rover results are not very good, due to a poor choice of system order when adding them together.

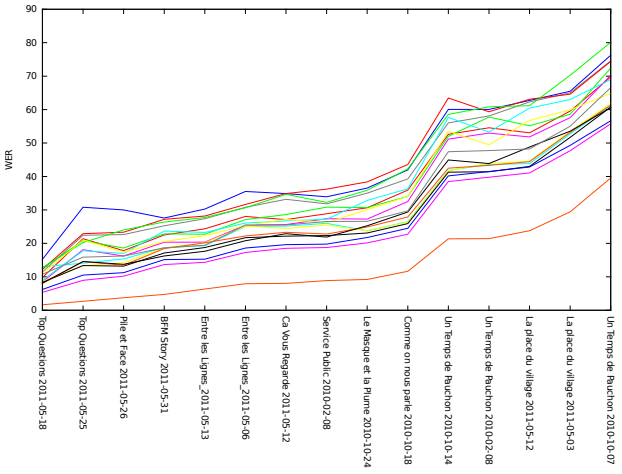

Figure 1: Per-system WER depending on the file. The lowest line is the oracle rover.

Figure 1 shows the results obtained by every system on each show after sorting them by oracle rover score. We can see a clear difference between *Un temps de Pauchon* and *La place du village* together and all the other shows. Those two shows are built around field reporting, interviewing run-of-the-mill people in often difficult sound conditions, on subjects such as a culinary metting or naturalization oath proceedings for *Un temps de Pauchon*, or the history of a remote village for *La place du village*. The accents are heavy, the vocabulary and topics unusual, and the speakers not trained. That accumultes difficulties for the systems.
The difference between the no-overlap and the optimal mapping scores, which is within the range of the overlapping speech ratio, shows that the systems have been extremely bad on the overlapping speech. Hopefully having both transcribed data and the evaluation tools available will make deeper work in that area possible in the future. It is, in any case, definitively not a solved problem.

4.4. Named Entities Detection

The named entities detection results are available in table 6. The complexity of the Quaero named entities require a large amount of work in the adjudication phase of the evaluation, which, in addition to individual remarks, ended up defining a dozen rules to detect possible problems and have humans judge and fix them. Still, some problems were not fixable due to a lack of foresight: defining entities boundaries in the presence of overlapping speech requires taking the speaker into account. Extracting the text and annotating it doesn't work well due to intermixed speech turns. Retrying such an exercise will require some experiments to define the best annotation approach.
Two systems were essentially linguistically based, while the others relied more on statistical methods. The non-statistical systems got similar results to the statistical ones, instead of the much better results they already obtain. This was probably due to the structuration on the Quaero entities, requiring new rule design methodologies which were not fully ready by the time of the evaluation. It is interesting to note that they did not do worse on the asr outputs than the stochastic ones, giving them an unexpected resilience.

Lab.	Run	manual	rover	s23	s24	s25	s30	s35
A	1	85.6%	98.1%	100.7%	94.2%	98.9%	98.4%	100.9%
	2	156.6%	147.4%	178.8%	160.4%	168.0%	163.9%	168.2%
B	1	36.6%	57.2%	59.3%	64.7%	62.0%	61.7%	71.8%
C	1	50.5%	88.0%	98.8%	76.8%	92.8%	94.9%	99.6%
D	1	44.8%	69.7%	73.8%	72.1%	73.7%	74.8%	86.0%
E	e-a-p		79.2%	79.5%	66.8%	80.8%	80.0%	87.0%
	e-a-pt+r		67.8%	68.4%	67.6%	70.9%	69.9%	85.2%
	m	37.5%						
F	1	62.5%	75.8%	79.2%	76.9%	79.8%	80.5%	90.5%
G	1	39.3%	65.0%	69.9%	66.3%	70.5%	69.9%	87.0%
H	1				68.4%			
	2	38.4%	63.7%	67.5%	64.1%	69.1%	68.6%	80.4%
	3	51.6%			72.7%			

Table 6: Named entities detection results

We were not fully satisfied with the SER metric and are developing a new one for future evaluations which gives better insights on the system qualities, see (Ben Jannet et al., 2014) for details.

5. Conclusion

This paper presented the ETAPE evaluation, including data, metrics and anonymized results. Some encountered difficulties were described and solutions will have to be devised if overlapping speech-handling systems are to be evaluated further.

The overlapping speech detection are promising, especially given the difficulty in establishing the references. Speaker diarization worked rather well, and the cross-show condition was well handled, validating from a feasability standpoint this very useful condition from an applicative point of view.

Speech transcription worked reasonably well on news and debates data, and rather badly on very spontaneous field speech, as expected. In addition, overlapped speech was very badly handled. Hopefully the availability of the data will make progress possible.

Finally the named entitities detection task was the first time the new Quaero named entities were open to the scientific community. They are complex and their new structure requires new approaches to handle them, giving not so impressive results which can only get better in the future. The community answer to these entities was rather positive though, which gives good hope for the creation of powerful systems handling them.

6. References

Ben Jannet, M. A., Adda-Decker, M., Galibert, O., Kahn, J., and Rosset, S. (2014). Eter : a new metric for the evaluation of hierarchical named entity recognition. In *Proc of LREC*, Reykjavik, Iceland. ELRA.

Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., and Quintard, L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. In *Proc of IJCNLP*, Chiang Mai, Thailand.

Galibert, O. (2013). Methodologies for the evaluation of Speaker Diarization and Automatic Speech Recognition in the presence of overlapping speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proc. of InterSpeech*.

Gravier, G., Bonastre, J., Geoffrois, E., Galliano, S., McTait, K., and Choukri, K. (2004). Ester, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Proceedings of JEP'04*, Fèz, Maroc.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In Chair), N. C. C., Choukri, K., Declerck, T., ur Do an, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc. of the Fifth Linguistic Annotation Workshop (LAW-V)*, Portland, OR, june. Association for Computational Linguistics.

NIST. (2000). 2000 Speaker Recognition Evaluation - Evaluation Plan.