Manual *vs* assisted transcription of prepared and spontaneous speech

Thierry Bazillon, Yannick Estève, Daniel Luzzati

LIUM (Laboratoire d'Informatique de l'Université du Maine)
Avenue Laënnec - 72085 Le Mans, FRANCE
first.lastname@lium.univ-lemans.fr

Abstract

Our paper focuses on the gain which can be achieved on human transcription of spontaneous and prepared speech, by using the assistance of an ASR system. This experiment has shown interesting results, first about the duration of the transcription task itself: even with the combination of prepared speech + ASR, an experimented annotator needs approximately 4 hours to transcribe 1 hours of audio data. Then, using an ASR system is mostly time-saving, although this gain is much more significant on prepared speech: assisted transcriptions are up to 4 times faster than manual ones. This ratio falls to 2 with spontaneous speech, because of ASR limits for these data. Detailed results reveal interesting correlations between the transcription task and phenomena such as Word Error Rate, telephonic or non-native speech turns, the number of fillers or propers nouns. The latter make spelling correction very time-consuming with prepared speech because of their frequency. As a consequence, watching for low averages of proper nouns may be a way to detect spontaneous speech.

1. Introduction

Transcription of speech is a complex task: fully manual transcription is very expensive because of the time it requires, while ASR systems are not accurate enough when perfect transcriptions are needed. Assisted transcription appears as a good compromise in order to obtain such transcriptions in a reasonable amount of time. The aim of this paper is to quantize the time gain which can be achieved, for prepared and spontaneous speech, between manual and assisted transcriptions. The latter have been made with LIUM's ASR system based on CMU Sphinx (Deléglise et al., 2005). LIUM RT training has been mainly based on articles from French newspaper *Le Monde*: as a consequence, it is not optimized to process spontaneous speech, such as debates or interviews. This work is linked with project EPAC (http://epac.univ-lemans.fr), selected by the ANR in Call for projects 2006 *Masse de Données - Connaissances Ambiantes*. In addition to LIUM, which is the coordinator, Avignon (LIA), Tours (LI) and Toulouse (IRIT) laboratories have been working on project EPAC since March 2007. EPAC's main purpose is exploring large sets of audio documents for extraction and processing of spontaneous speech. More specifically, acoustic markers, named identification, relations between Automatic Speech Recognition systems and spontaneous speech, or transcription and annotation are some of EPAC's tasks. One of EPAC's purposes is to provide, by March 2009, annotated transcriptions of one hundred hours of radio data from the ESTER campaign (Galliano et al., 2005). The obtained corpus will be essentially made of spontaneous speech, because currently the French speech community does not have enough of this kind of data. Beyond the proper scientific interest, it is thus in order to optimize this task that the present work has been done.

2. Spontaneous speech *vs* prepared speech

The obvious difference between these two kinds of speech is the number of disfluencies (Adda-Decker et al., 2003), generally much more present in spontaneous speech. As a consequence, spontaneous speech fails to be well recognized by ASR systems. However, a previous work on spontaneous and prepared speech (Jousse et al., 2008) showed that this distinction may be ambiguous, because some spontaneous speeches (for instance when speakers are politicians) are very similar to prepared ones. Besides, a prepared speech with several false starts or repetitions may sound spontaneous.

So a study was carried out, considering speech quality instead of spontaneous / prepared distinction. On a corpus of about 10 hours, two annotators were in charge of tagging each speech segment, according to the following categories:

- high quality speech: no or few disfluencies;
- average quality speech: some false starts, truncated words, repetitions..., but speech is still coherent and understandable;
- low quality speech: speech with a lot of false starts, etc., making it incoherent and/or non-intelligible.

Part of these 10 hours was marked jointly by the two annotators, to be sure they use the same criteria. Then the KAPPA coefficient (Cohen, 1960) was calculated to validate the process. A score of 0.852 was obtained: scores over 0.81 are usually regarded as excellent. This proves that despite the subjectivity of the concept of speech quality, annotators agreed on what was high quality speech and what was not. Nevertheless, for the current study, a simpler way to select the test files has been chosen, that is to say the conventional distinction spontaneous / prepared (what is considered here as prepared is broadcast news; spontaneous means debates or interviews). Indeed to provide relevant results, files of about 10 minutes were needed, and it was impossible to find in such a file only segments with same speech quality tags. Dissociating spontaneous speech and all the phenomena it implies (false starts, truncated words, repetitions, overlapping speech, French filler words like *euh*, *ben*...) from prepared speech was enough to allow our experiment to get interesting results and prospects.

3. Protocol

The data include 24 files of about 10 minutes, each selected from the untranscribed ESTER corpus. 4 radio stations are concerned: France Inter, France Info, France Culture and RFI. 12 of these files are considered as prepared speech (3 for each station), and the other 12 are considered as spontaneous speech. France Info was not been considered for spontaneous speech because it did not contain relevant data, so 4 files of the 3 other stations were chosen. Also, when selected files contained unrelevant data such as music, or few spontaneous speech in prepared speech file (or the other way), they were not taken into account.

Using the TRANSCRIBER software (Barras et al., 1998), each file was transcribed twice: first manually, then with the help of an automatic output generated by LIUM RT. These two tasks were realized several days apart to avoid border effects. For assisted transcription, the annotator got output files generated by LIUM RT with automatic (and potentially erroneous) segmentation for speech turns and speakers, but without named identification. The transcription work was divided into three levels:

- transcription of text and segmentation in speech turns;
- assignation of speakers to each speech turn;
- spelling correction, especially of proper nouns.

4. Main results

	Prepared speech	Spontaneous speech
Manual	17h36'	19h33'
Assisted	8h31'	15h44'

Table 1: Total transcription time (respective corpora durations: 2H08 and 2H10)

The first gain, which is rather interesting, is the global one (Table 1): assisted transcription means undoubtedly faster work, but mainly with prepared speech. Indeed for this kind of speech, total transcription time is approximately half for assisted transcription of what it is for fully manual transcription. Nevertheless, this gain decreases dramatically with spontaneous speech.

For example, considering a prepared speech file of 10 minutes, the transcriber needs approximately 40 minutes to transcribe it, assign speakers and correct spelling, starting from an automatic transcription output file. If the same tasks are done on the same file in a fully manual way, about 83 minutes are necessary.

Now, considering a spontaneous speech file of 10 minutes, assisted tasks represent a total work time of 73 minutes, which is quite different from the results for prepared speech. On the other hand, manual transcription needs about 90 minutes, i.e. only a little more than prepared speech scores. So assisted transcription represents a gain in every case, but a much more important one with prepared speech.

	Prepared speech	Spontaneous speech
Manual	13h36'	16h15'
Assisted	5h06'	12h41'

Table 2: Transcription of text and segmentation

Text transcription (Table 2) represents the most interesting gain: with prepared speech, manual transcription needs about 2.67 more time than assisted transcription (5h06' vs 13h36'). More precisely, the most signicant file obtained a score of 3.75: with a duration of 08'55", assisted transcription needs 14'49", and manual transcription, 55'34". This element is very significant, especially if it is compared with spontaneous speech. With approximately the same duration, the ratio is only about 1.28. For the file with the highest gain, it is only 1.95: for 11'18" of speech, assisted transcription needs 47'03", and manual one, 31'52". It means that many more corrections are needed, highlighting the fact that LIUM's ASR has difficulties processing spontaneous speech.

	Prepared speech	Spontaneous speech
Manual	1h17'	2h13'
Assisted	1h17'	2h13'

Table 3: Assignation of speakers

As far as the speakers are concerned (Table 3), the strictly identical results confirm the obvious fact that an ASR system does not help at all for speaker assignation, since it does not identify them by their names. One much more important fact is that speaker assignation requires twice as long for spontaneous speech than prepared speech. This can be explained quite easily: on the one hand spontaneous speech contains a lot of speech turns, so the transcriber often has to assign a speaker to a turn, although there may be only two speakers in a file. On the other hand, prepared speech contains generally more speakers but less speech turns as they are much longer than in spontaneous speech. Furthermore, spontaneous speech sometimes contains overlapping speech, and in the case of three speakers or more, it may be long and hard to define who is talking.

	Prepared speech	Spontaneous speech
Manual	2h43'	1h05'
Assisted	2h08'	0h51'

Table 4: Spelling correction

A surprising result of this study deals with spelling correction (Table 4): the specific difference between assisted and manual transcription regarding this task is not very significant, but the difference between prepared and spontaneous speech is much more. The reason for this gap is quite simple: prepared speech files are essentially broadcast news, and this kind of data contains lots of proper nouns (reporters, people, towns...), that can not all be known by the annotator. So looking for the right spelling form may be a tough task, especially with foreign names. On the contrary, spontaneous speech files are interviews or debates, in which there are a few proper nouns because their topics generally do not require many person named entities.

	Prepared speech	Spontaneous speech
Manual	16.95	35.21
Assisted	15.83	34.33

Table 5: Word Error Rate (%)

The last main results deal with the Word Error Rate (Table 5). The output files generated by LIUM RT were compared with the manual transcription, and then with the assisted transcription, using the classical formula of the Word Error Rate.

The averages shown in table 5 validate what was said before: LIUM's ASR system is not as competitive on spontaneous speech as on prepared speech. The differences between manual and assisted results can be explained by the fact that the annotator did not necessarily output the same text for both transcriptions of the same files: phenomena such as repetitions, false starts or overlapping speech are sometimes hard to perceive, and as a consequence their transcription may be inconsistent, even between two sessions by the same transcriber.

Taken as a whole, WER for assisted transcription is in keeping with the ratio between total transcription time and total files duration: a professional annotator needs approximately twice as long to correct a spontaneous file than a prepared one, and WER is approximately twice as high for spontaneous speech than for prepared speech.

5. Some detailed results

Table 6 presents detailed results of transcription of text and segmentation for prepared speech, coming with several data: the ratio between assisted transcription task duration and file duration; the ratio between manual transcription task and file duration; the Word Error Rate; the number of proper nouns; the part of segments with telephonic or non-native speakers.

This last criterion is particularly significant with prepared speech, as it is essentially broadcast news, where many reports are made through the telephonic channel, or sometimes by foreign reporters who speak French: in these two cases, ASR accuracy is generally reduced, either because signal quality is average, or because pronunciation is unclear.

The number of proper nouns is an important element in prepared speech, as shown before with the results of spelling correction. These results are validated by the percentage of proper nouns, which varies from 6 % to 10.5 % in table 6; in comparison, the average percentage is 1.9% for prepared speech. It is difficult to extract reliable informations from these results, as proper nouns can be French or foreign ; the LIUM RT system is obviously trained with French speech, and is then more accurate with French proper nouns. This can explain the WER of files 2 or 4, although their percentages of proper nouns is high. On the other hand, this element may be an explanation for scores concerning file 9: it contains very little telephonic or non-native data, but obtains a WER of 17.5, whereas file 1 gets a WER of 9.7 with 48 seconds of telephonic speech. But file 9 contains more than 10% of proper nouns (the highest one in the corpus), and especially some foreign ones, which may explain the value of the WER.

More generally, the results reveal that assisted transcription is always faster than manual transcription. Besides, there is an obvious distinction between the first 5 files and the others: from file 1 to 5, the ratio for assisted transcription is under or around 2, and the WER is under 15%. Over file 5, the ratio is between 2.5 and 3, and the WER is from 17 to 22%. As we supposed, part of telephonic and non-native speech segments is often linked to these two attributes: files 6, 8, 11 and 12 have some of the highest WER, and contain at least three minutes of these particular data.

All ratios for assisted transcription and WER are very tight, except for file 1; it may be due to its length, as this file is shorter than the others (less than 9 minutes).

The manual transcription ratio is very homogeneous, since scores are set between 5.96 and 6.95. It means phenomena such as telephonic speech and non-native speakers delay much less the ASR systems than a professional annotator.

The same kind of results for spontaneous speech are presented in Table 7, however the number of proper nouns has been replaced by the number of French fillers *euh*, much more significant for spontaneous speech: as they are more frequent, speech is less fluent and as a consequence harder to detect and transcribe. Results here are not necessarily correlated to assisted transcription ratio, but some isolated facts are interesting: first of all, file 16 contains less than 2% of fillers, but is quite particular, as we explain later. If we consider the other 11 files, file 13 comes in first position, which is rather logical. But then, the number of fillers and WER or assisted ratio are not especially linked, even if files 17, 21 and 24 (with WER of about 40%) contain approximatively 5% of fillers, while files 14, 18 and 19 (WER between 20 and 30%) contain between 2 and 3%. On the other hand, the percentage of fillers for specific cases like file 15 (low assisted ratio, WER of 30% but more than 5% of fillers) or 23 (quite high assisted ratio, WER of about 40% and only 3% of fillers) seems to be less efficient.

Then, it is obvious that telephonic speech and non-native speakers are less important in spontaneous speech, as these files are essentially interviews or debates in studio. Nevertheless, file 23 gets an important part of this kind of data because of listeners telephonic interventions, which partially explains the WER of 40% and the high ratio for assisted transcription. File 14 contains the same programme, named *Le téléphone sonne*, but with a little telephonic interventions; it may be a part of the explanation of better ratio and WER for this file.

Concerning the ratio results themselves, the three files that needed the less time to be corrected have a WER under or equal to 30%. Their speakers are either professional journalists, or people who are used to being interviewed, which explains this important gain for assisted transcription. On the other side, the annotator spent a lot of time to transcribe file 24, whose WER is one of the higher in our corpus (42.9%), because the main speaker is a shy person who does not speak very loud, and who does not seem familiar

File	Ratio assisted	Ratio manual	WER (%)	Telephonic speech/non-native speakers	Proper nouns (%)
1	1.66	6.23	9.7	0’48”	7.41
2	2.01	5.95	13.4	none	8.46
3	2.1	6.06	14.2	1’39”	6.05
4	2.14	6.29	12.6	2’04”	9.88
5	2.18	6.42	15.6	1’56”	7.85
6	2.52	5.96	20.3	3’18”	8.81
7	2.52	6.88	17.1	2’01”	8.41
8	2.53	6.32	19.1	3’05”	8.79
9	2.58	6.31	17.5	0’18”	10.35
10	2.59	6.37	20.8	1’47”	6.77
11	2.85	6.73	20.5	3’41”	7.57
12	2.94	6.95	22.6	3’22”	6.1

Table 6: Detailed results for prepared speech

File	Ratio assisted	Ratio manual	WER (%)	Telephonic speech/non-native speakers	Fillers (%)
13	4.16	8.13	20.3	none	2.2
14	4.48	8.06	21.5	1’11”	3.2
15	4.84	6.99	30	none	5.29
16	5.33	5.22	54.5	none	1.37
17	5.7	6.33	44.2	none	5.59
18	5.85	8.51	27.8	none	3.06
19	5.86	8.16	31.9	none	2.28
20	6.2	8.16	38	0’41”	3.41
21	6.59	7.48	38.8	none	5.26
22	6.62	7.75	33.1	none	4.38
23	7.04	7.55	39.5	3’35”	3.1
24	8.7	8.31	42.9	none	4.1

Table 7: Detailed results for spontaneous speech

with interviews. Concerning this file, assisted transcription needed more time than manual one, which is a very rare phenomenon. According to our table, there is no specific reason to explain that; what’s more, file 17 has a high WER and needed a reasonnable time to be transcribed. A simple explanation could be that the annotator may have been particularly efficient or unefficient when transcribing especially these files.

File 16 obtained a WER of 54.5%, but did not require a lot of time to be transcribed, because its main speaker is an old photographer who speaks with lots of blanks, in a low voice and who does not articulate at all. As a consequence, the ASR system has trouble to detect speech segments, while human annotator perception is accurate enough to transcribe without losing time. That is why assisted transcription, as for file 24, needed more time than manual one; even if the difference is lower here, it is significant. Other files are all with an assisted transcription ratio approximately between 6 and 7, and WER between 30 and 40%.

Dealing with ratio for manual transcription, as for prepared speech, things are much simpler: all files are between 7 and 8.5, except files 16 and 17. Transcription of file 16, as mentionned before, is truncated by lots of blanks. As a result, there are less words than in other files, which explains the high gain of time, even in the case of manual transcription. About file 17, this ratio is quite inexplicable and as for file 24, the explanation may be the efficiency of the annotator.

6. Conclusion and prospects

This work shows that the task of transcription, fully manual or assisted, is a long process. The best obtained results are for the combination of prepared speech with assisted transcription, but even in that case, transcribing 10 hours means 40 hours of work. With the association of spontaneous speech and manual transcription, 80 hours are necessary. Our study also proves that, except with a few files, the ASR system represents a gain even if it may change a lot from a file to another. Detailed results prove that the WER is often correlated with assisted transcription duration, but it is not the only explanation: some kind of data (telephonic speech or non-native speakers for example) are a bit of a problem for ASR systems, while they do not disturb a human annotator.

Assignation of speakers may seem rather long for spontaneous speech, but these results have to be restrained: usually transcription of text, assignation of speakers and spelling correction are done as they go along, instead of being separated. In the case of speakers, this is important because assignating them after the transcription forces the transcriber to check the whole file. In prepared speech, it does not take too much time because speech turns are generally long and well defined; conversely, spontaneous speech often contains short speech turns with many changes of speakers. As a consequence, the transcriber has to listen again to nearly every segment separately, and assignation in that case is very time-consuming.

The number of proper nouns is interesting for future works: it could be an ontological way to detect spontaneous speech. That would be very useful and time-saving, mostly concerning large corpora such as the ones in project EPAC.

7. References

M. Adda-Decker, B. Habert, C. Barras, G. Adda, Ph. Boula de Mareuil, and P. Paroubek. 2003. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. In *DISS*, pages 67–70, Göteborg, Sweden, September.

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on languages ressources and evaluation (LREC)*, pages 1373–1376, Granada, Spain, May.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 2005. The LIUM speech transcription system: A CMU sphinx III-based system for French broadcast news. In *Interspeech*, Lisbon, Portugal, September.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, September.

V. Jousse, Y. Estève, F. Béchet, and G. Linares. 2008. Caractérisation et détection de parole spontanée dans de larges collections de documents audio. In *JEP*, Avignon, France, June.