Automatic audio and manual transcripts alignment, time-code transfer and selection of exact transcripts

C. Barras, G. Adda, M. Adda-Decker, B. Habert, P. Boula de Mareüil, P. Paroubek

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{barras,gadda,madda,habert,mareuil,pap}@limsi.fr

Abstract

The present study focuses on automatic processing of sibling resources of audio and written documents, such as available in audio archives or for parliament debates: written texts are close but not exact audio transcripts. Such resources deserve attention for several reasons: they represent an interesting testbed for studying differences between written and spoken material and they yield low cost resources for acoustic model training. When automatically transcribing the audio data, regions of agreement between automatic transcripts and written sources allow to transfer time-codes to the written documents: this may be helpful in an audio archive or audio information retrieval environment. Regions of disagreement can be automatically selected for further correction by human transcribers. This study makes use of 10 hours of French radio interview archives with corresponding press-oriented transcripts. The audio corpus has then been transcribed using the LIMSI speech recognizer resulting in automatic transcripts, exhibiting an average word error rate of 12%. 80% of the text corpus (with word chunks of at least five words) can be exactly aligned with the automatic transcripts of the audio data. The residual word error rate on these 80% is less than 1%.

1. Introduction

In the present study we focus on automatic processing of sibling resources of audio and written documents, such as available in audio archives or for parliament debates: written texts are close but not exact audio transcripts. They are instances of bona fide or trustworthy transcriptions, which are being used for quotations, and even for legal purposes. Such resources become more and more within reach and deserve attention for several reasons: manual bona fide transcripts may provide a "draft version" of exact ones. The manual transcripts are carefully edited, they can be readily tagged and parsed: the resulting annotations, which cannot be produced as reliably from an automatic transcript, may then be transfered to a speech recognizer's output. Differences between press-oriented and exact manual transcripts focus on spontaneous speech specificities.

Automatic transcription systems on sibling audio and written documents allow to produce enriched resources. Time-alignment between the written text and the related audio source via automatic transcripts, allows for precise audio retrieval when scanning a written document. This question is addressed in section 4. on time-code transfer.

Exact audio transcripts are particularly useful both for acoustic and language model training and for studies on spontaneous speech specific phenomena. However exact transcriptions are very expensive to produce from scratch. Draft transcriptions can be generated automatically for any audio document via an automatic speech recognition system (Lamel et al., 2000): the quality of the automatic transcript depends on the recognizer. If audio-related texts are available they first allow to tune the recognition system for improved automatic transcription. They can further be used to select in the latter those regions which agree with the written texts. The comparison between written documents and automatic transcripts provides an automatic partitioning of the audio corpus in very probably exact and very probably erroneous transcribed data. Fast exact transcripts can then be produced manually, by focusing on those regions which are labeled as very probably erroneous. These aspects are developed in sections 5. and 6.

2. Corpus and Transcripts

The current experiment makes use of 10 hours of French radio archives, recorded about 10 years ago. In each one hour show a major personality from either political or civil society (e.g. nonprofit humanitarian organizations) undergoes a detailed questioning by three or four journalists. For each show we have both the audio data and press-oriented transcripts. These press-oriented transcripts (TPress henceforth) are intended to be rather close to the audio (as quotations are being extracted from them for other media) while lying somewhere in between written text and exact transcript: they stick to implicit conventions for speech rendering. As a matter of fact most disfluencies and linguistic errors have been discarded or edited. We produced exact audio transcripts (TExact) for 10% of the data: all audible phenomena, in particular disfluencies and overlapping speech have been manually transcribed.

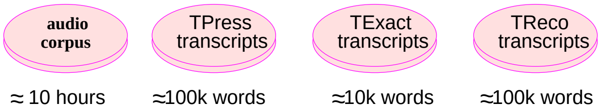

Figure 1: Audio corpus and transcripts. TPress corresponds to the archive transcripts, TExact has been produced manually for 10% of the data and transcribes all audible speech phenomena, including overlapping speech segments. TReco is obtained automatically via speech recognition.

Even though edited, the press-oriented transcripts remain fairly close to the audio. To get an idea of the differences between both TPress and TExact versions, we used the NIST `sclite` tool (`http://www.nist.gov/speech/tools/`), with, as a reference, the TExact version where all disfluencies have been filtered out. The average word difference rate amounts to 9%. This is mainly due to deletions (omitted parentheticals, asides and overlapping speech) entailing locally high difference rates.

3. Automatic Speech Recognition

The audio corpus has been transcribed using the standard LIMSI speech recognition system for French (Adda-Decker et al., 1999) resulting in the TReco transcripts. The acoustic models were trained on about 100 hours of French broadcast news data; they consist in context-dependent models of 33 French phonemes, plus 3 generic models for silence, filler words and breath noises. The standard language model (LM) is an interpolation of 4-gram back off language models trained on different data sets: press-oriented transcriptions of various broadcast shows (48M words), exact transcriptions of broadcast news (BN) data, mainly radio shows (0.95 M words) and newspapers texts (311M words). The lexicon contains 65k words, chosen for optimizing the coverage of broadcast news development data (very different in date and source from the archive corpus). The pronunciations are derived from grapheme-to-phoneme rules and manually checked. The system runs at about 10 times real-time on a standard PC. Using the TPress transcripts provided with the corpus (about 580k words), an *informed* LM was designed by interpolation with the standard n-gram LM; the lexicon contains only the 26k most frequent words from the standard sources, together with all the 19k words contained in the press-oriented transcripts, resulting in a 30k words lexicon.

Recognition results using the standard and informed recognition system are shown in Table 1 (left). The standard and informed TReco transcripts have word error rates (WER) of roughly 24 and 12%. In the following TReco corresponds to the informed system output.

4. Time-code Transfer

The different transcript versions have different characteristics: TPress is a correct written rendering of the audio data, but has no audio retrieval information associated. TReco, although not fully correct, allows to precisely access the audio via time-codes of the recognized words. Comparison of TPress and TReco allows for instance to study disfluencies (Adda-Decker et al., 2003) and other spontaneous speech specific phenomena. Here we take advantage of the closeness between TPress and TReco transcripts to automatically time-align the audio data with the TPress transcripts.

As TPress and TReco are transcripts corresponding to the same audio data, they can be aligned (using the standard Unix `diff` command on appropriately normalized versions) with a minimum of errors. Using this alignment as a pivot representation, we are able to transfer the accurate time-codes of the automatic transcription to TPress. Recognized disfluency events (fillers, breath noises, large pauses, repetitions of words or phrases) can also be transfered and thus contribute to a more accurate transcript, closer to the desired TExact.

TPress is closest to a standard written form and is thus appropriate for most of the current NLP tools (which is not the case for both the TExact or TReco forms). We can transfer punctuations to the TReco, as well as any NLP tags of the TPress version: lemmatization, part-of-speech tagging, chunking... Other useful information can be gained from a TPress/TReco alignment, for instance overlapping speech generally corresponds to high density error regions.

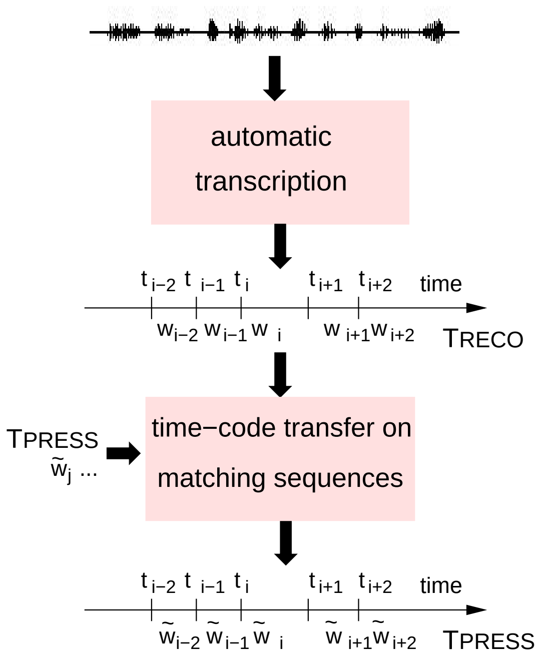
automatic transcription
t_{i-2} t_{i-1} t_i t_{i+1} t_{i+2} time
w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2} TRECO
TPRESS
$\tilde{w}_j$ ...
time−code transfer on matching sequences
t_{i-2} t_{i-1} t_i t_{i+1} t_{i+2} time
$\tilde{w}_{i-2}$ $\tilde{w}_{i-1}$ $\tilde{w}_i$ $\tilde{w}_{i+1}$ $\tilde{w}_{i+2}$ TPRESS

Figure 2: Time-code transfer: time-codes are produced automatically via automatic audio transcription. On matching sequences, i.e. sequences of N words which occur both in TPress and TReco, time-codes can be adopted by TPress.

5. Selection of Exact Transcripts

Since the errors done by the automatic recognition system are by nature different from the stylistic corrections observed in TPress, it is very likely that a perfect match between both of them corresponds to correct recognition.

Using the alignment procedure described in figure 2, about 85% of the words from TReco were aligned with the same words in TPress. For further analysis, matching segments between TPress and TReco were extracted and filtered according to their length. By selecting matching segments of at least 5 (resp. 10, 15 or 20) words, still 78% (resp. 66%, 54% and 44%) of the corpus is kept. With a random distribution of the errors in TReco, an exponential reduction of the corpus size given the minimal segment length would be expected. On the opposite a quasi-linear decrease is observed here, suggesting that there are large contiguous segments of agreement between TPress and TReco.

Our hypothesis is that these matching segments correspond to correct recognition. We thus computed the recognition score of TReco for these segments on the TExact subset. On matching segments of at least 5 words, a word error rate (WER) lower than 1% was observed, to be compared to 12.4% WER without filtering (see Table 1). Furthermore, most of the differences in TExact consist in gen-

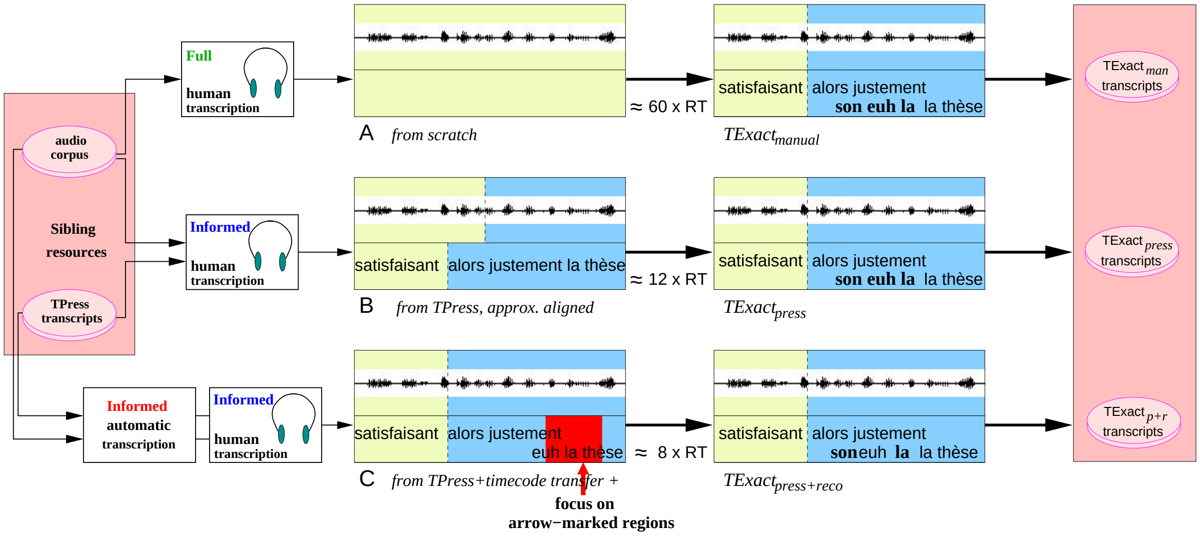

Figure 3: Exact audio transcription generation process, depending on whether TPress transcripts (B,C) and TReco (C) are available or not (A).

der or number agreement corrections, or other corrections which are acoustically ambiguous or very similar. It seems that inter-annotator agreement level is reached; our hypothesis is verified: using both TPress and TReco, it is possible to automatically select segments covering about 80% the corpus along with their exact transcription.

The selected segments with their exact transcription may then be used either for further automatic or manual processing. On the one side, they can be used for a speaker-specific supervised adaptation of the acoustic models of the recognition system, resulting in an improved automatic transcription. The process can then be iterated, providing a new alignment with TPress. On the other side, they can be provided as an auxiliary information for helping and speeding up the manual transcription of the complete document.

6. From *bona fide* to precise manual transcriptions

The manual production of a precise orthographic transcription synchronized with the audio signal is a very time-consuming process. For segmentation, speaker identification and exhaustive word transcription of news data, our experience is that about 60 times real-time is generally needed when no a priori knowledge is provided. Using an existing *bona fide* transcription and its alignment with an automatic transcription should provide a significant help. We have experienced two contrastive protocols, using Transcriber (`http://www.etca.fr/CTA/gip/Projets/Transcriber`) which is a tool for assisting the manual annotation of speech signal:

• In the first configuration, only the information found in TPress was used. An approximative alignment of the text with the audio was performed in the segment to be transcribed, by assigning to each speaker turn a length proportional to its word count.

• In the second configuration, the time-aligned version of TPress was used. Furthermore, color codes were used to bring to the fore the signal portions for which the alignment failed and which remained to be manually edited. The human annotator therefore could concentrate on this subpart of the signal.

For each protocol, segments lasting between three and six minutes were randomly extracted from the corpus and distributed among six transcribers. Each segment had to be processed by two different transcribers in order to assess inter-annotator variability. Transcriptions had to be produced in standard orthographic French, excluding overlapping speech and noises. Globally, a large variability in time was observed between annotators due to different annotation skills and depending of the amount of overlapping speech in the segments (even if overlapping speech was excluded from word transcription, its boundaries still had to be indicated). However, the median values still show a significant trend: following the first protocol, transcriptions were performed at about 12 times real-time; using further informations from TReco as proposed in the second protocol, it decreased to about 8 times real-time. This last result is comparable with what has been obtained on aligning close captions and broadcast news in American English (Lamel & Gauvain, 2003). Both figures have to be compared with the 60 times real-time when no a priori information is provided (see Figure 3).

Time-code transfer from TReco were generally reliable and thus provided a useful help for the manual transcription. Usefulness of colored error segments was more balanced and their use depended of the transcriber strategy. We believe it is due to the ergonomy of this experimental protocol, since the segments were shown under the signal, not within the text editor, as tested by (Eickeler et al., 2002).

7. Inter-annotator agreement

Inter-annotator agreement is studied on 3 minutes excerpts from 6 shows, following the 1st configuration of section 5 (audio and approximate alignment of TPress, Fig. 3.B). We defined 9 classes of inter-annotator discrepancies:

1- hesitation insertion (*euh*);

	TReco$_{std}$	TReco	TPress	TPress ∩ TReco
WER	24%	12%	9%	<1%

Table 1: Word error rates on different transcript versions, computed with TExact as a reference. The TReco$_{std}$ (resp. TReco) rates stem from the standard (resp. informed) recognizer. The last column corresponds to automatically selected subsets of TPress (using TReco).

n^o- *class*	# occ.	n^o- *class*	# occ.
1- hesitations	20	6- homophones	41
2- connector insert.	13	7- lexical fragm.	7
3- grammat. insert.	11	8- lexical insert.	6
4- grammat. subst.	6	9- lexical substit.	1
5- oral specific	8		
Total :		113	

Table 2: Typology of inter-annotator disagreements on 18 minutes, by class

2- connector insertion (*mais vous savez* vs *vous savez*);
3- grammatical word insertion (*et de demander* vs *et demander*);
4- grammatical word substitution (*la liste* vs *ma liste*);
5- oral specific vs standard writing (expliciting the negative particle *ne* often absent in speech; writing *cela* instead of *ça*; expliciting the subject pronoun *il* in *il y a*);
6- homophonic variants: no acoustic hints for different written forms (case: *Européen* vs *européen*; singular/plural: *progressiste* vs *progressistes*; compound variants: *Mendès - France* vs *Mendès-France*; commas);
7- word fragments fully written or not (*parmi* vs *par*...);
8- lexical word insertion (*oui, ça, je crois que, je je crois* vs *je crois que*);
9- lexical word substitution (*je voudrais* vs *je vais*);

The homophonic variants represent by far the most important class of discrepancies (36%). However, this type of variation makes no difference as far as the quality of the transcription is concerned.

The "unit" for a discrepancy score is the token. For instance, the variation *toutes choses égales par ailleurs* vs *toute chose égale par ailleurs* yields 3 instances of class 6. The table 7. gives the number of disagreements per class.

In our pilot study the number of discrepancies varies widely between extracts (from 8 to 29). It is difficult to assess what is due to the show and to the interviewee's idiosyncrasies and what to the various levels of transcribing ability and to the transcriber's choices.

The Kappa coefficient is often used to measure pairwise agreement among coders making category judgments, corrected for expected chance agreement (Carletta, 1996): $K = P(A) - P(E)/1 - P(E)$, where $P(A)$ is the proportion of times that the coders agree and $P(E)$is the proportion of times that we would expect them to agree by chance.

We can consider two situations. In the first one, we use the number of tokens in the 6 extracts (3,566) as the number of category judgments and in the second the actual proportion on which at least one of the coders disagree with the initial transcription (the union of their disagreements with TPress and between them), that is 320 tokens. However, we found no way to estimate properly the proportion of chance agreement. For each word of TPress, the annotator has to choose between keeping it (the overwhelming correct choice), erase it, change it, add a new word. The probability of adding a word differs whether it is a new one or a repetition (which itself is conditioned by the grammatical category and the position in the sequence of the repeated word). We then only give the following proportions of agreement (excluding homophonic variants):
1. on the whole extract: 97%;
2. on the disagreement portion: 78%.
These figures are provisional estimates. However they hint as a rather high inter-annotator agreement.

8. Conclusions & Perspectives

We addressed the general question of information transfer between different transcript types of audio data. More particularly the problem of time-code transfer from automatic recognition transcripts to *bona fide* transcripts has been investigated. Automatic and *bona fide* matching regions (80%) have a very low residual error rate below 1% and can hence be considered as exact audio transcripts. The automatic extension of exactly transcribed audio resources is a major contribution of our work towards automatically adaptable ASR systems. In future experiments the performance gain with automatically adapted acoustic models will be measured. Ergonomy for efficient correction of erroneous speech segments will be assessed. Information transfer of punctuations and NLP tags to automatic transcripts is a challenging research area.

9. References

Adda-Decker, M., Habert, B., Barras, C., Adda, G., Boula de Mareüil, P., Paroubek, P., (2003). A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. Proc. Disfluency in Spontaneous Speech, 67-70, 2003, Göteborg, Sweden, 5-8 September.

Adda-Decker M., Adda, G., Gauvain, J.L., Lamel, L. (1999), Large vocabulary speech recognition in French. Proc. IEEE ICASSP'99, Phoenix, AZ. vol.I, pp.45-48.

Barras, C., Geoffrois, E. Wu, Z. & Liberman, M. (2001), Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1). pp. 5-22.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254

Eickeler, S., Larson, M., Rüter, W., Köhler, J. (2002). Creation of an Annotated German Broadcast Speech Database for Spoken Document Retrieval. In Proceeding of the Third International Conference on Language Resources and Evaluation (pp. 334–33). Las Palmas, Spain.

Lamel L., Gauvain J.L., Adda G., (2003), Lightly supervised acoustic model training. Proc. ISCA-ITRW workshop ASR-2000, Paris. pp.150-154.

Lamel, L., Gauvain J.-L., 2003. Linguistic data: Fast transcription. oral presentation at RT-03S workshop, `http://www.nist.gov/speech/tests/rt/rt2003/spring/` Boston MA, May 19-20.