Towards Fully Automatic Annotation of Audiobooks for TTS

Olivier Boeffard, Laure Charonnat, Sébastien Le Maguer, Damien Lolive, Gaëlle Vidal

IRISA - University of Rennes 1, Lannion, France {olivier.boeffard,sebastien.le_maguer,damien.lolive}@irisa.fr,

{laure.charonnat,gaelle.vidal}@univ-rennes1.fr

Abstract

Building speech corpora is a first and crucial step for every text-to-speech synthesis system. Nowadays, the use of statistical models implies the use of huge sized corpora that need to be recorded, transcribed, annotated and segmented to be usable. The variety of corpora necessary for recent applications (content, style, etc.) makes the use of existing digital audio resources very attractive. Among all available resources, audiobooks, considering their quality, are interesting. Considering this framework, we propose a complete acquisition, segmentation and annotation chain for audiobooks that tends to be fully automatic. The proposed process relies on a data structure, ROOTS, that establishes the relations between the different annotation levels represented as sequences of items. This methodology has been applied successfully on 11 hours of speech extracted from an audiobook. A manual check, on a part of the corpus, shows the efficiency of the process.

Keywords: Audiobook, annotation, phone segmentation

1. Use of Large Corpora for Speech Synthesis

The use of statistical models, mainly from the speech recognition field, occurs in all the automatic language and speech processing disciplines. Learning such models on speech units needs a large number of observations and implies the use of large speech corpora. A direct consequence of the size of corpora is that conventional methods used to build, segment and annotate them have shown their limits. When we consider the field of speech synthesis, the unit generally used is the phoneme together with its precise linguistic and acoustic context. Representing in a database the whole set of the units generated by the combination of all the linguistic and acoustic features (about 50 features are used in HTS (Tokuda et al., 2000)) is impossible. Nonetheless, learning models of the units that are present in the target language can be done by the analysis of large natural speech corpora.

The increasing availability of digital resources enables research laboratories to have a great variety of recordings covering many speaking styles and recorded in varied conditions. In the framework of speech synthesis, we consider the use of audiobooks which enables to have recordings, read by a professionnal speaker and with a good acoustic quality, accompanied with the script. Audiobooks have a lot of benefits and can complete or even replace ad-hoc corpora. As an example, we can find different books read by the same speaker which increases the variety of speaking and literary styles, or we can find the same book recorded by different speakers and thus enabling studies using parallel data. In addition, a specificity of textual data associated to audiobooks is that they are nearly constant from one recording to another. However, possible variations may exist in the case of multiple editions (translations, unfinished manuscripts, etc.).

An important step to build a new voice is the annotation of the signal and the associated text. This step may appear simple if it is limited to the synchronisation of words onto the signal, but uses complex temporal relation when it comes to dealing with other levels of analysis, as semantic, lexical, grammatical, syntactical, phonetic or prosodic annotations. For each of these annotation levels, previous works found in the literature proposed automatic annotation systems that define labels to place on the text linked to the signal. Generally, these systems have been built separately and, consequently, use different description formats but not the same time scale. Their use on a corpus often requires a lot of conversions and generates a lot of heterogeneous and scattered files. In order to avoid desynchronized descriptions, we have recently proposed a solution based on relations between sequences, ROOTS (Barbot et al., 2011). This approach enables to define a minimal set of relations that exist between levels of description. Primitive relations are first defined and then all the necessary missing relations between two sequences of annotation are composed using algebraic rules.

The aim of this study consists in describing the process used to build speech corpora from audiobooks. This automatic annotation process is applied to several dozen hours of continuous speech. This study is restricted to mono-speaker recordings but the only constraint is to have an acoustic content and the corresponding text. The ouput of the system, according to the annotation levels required, is a set of ROOTS files, stored as XML files. These files put together give the structure of all the relations from text to signal.

The rest of the paper is organized as follows. In part 2, the proposed annotation process is described. The structure representing the information linked to the utterance is then introduced in part 3. This structure is used by the text splitting and alignment processes, detailed in section 4, and also to keep all the information obtained form the annotation step, section 5. Finally, part 6 illustrates the whole process with an application of the methodology.

2. Annotation Process

The annotation process has to fulfill a number of constraints that come from the use of annotated corpora and performance control. The first one concerns the text that has to be kept under its original form and the differences introduced by the reader have to be tagged. The second one concerns the corpus splitting. Our goal being to annotate a corpus on several levels, we have to manipulate speech fragments and text fragments of variable size. As an example, it is desirable for the syntactical analysis to have a complete sentence while the segmentation into phones is more efficient when applied on short extracts. The text and the recorded tracks will be split with a granularity sufficiently fine to enable working on short fragments that may be, if needed, merged together to build a sentence or just a longer fragment under the condition of keeping coherence of the text. Lastly, to garantee the annotation quality, a manual intervention is possible, and is guided by confidence indicators given by the different tools used during the process.

The annotation process, shown on figure 1, is composed of two main steps: the first one consists in splitting several hours of a long speech recording into short tracks and associating to each of them the corresponding text. This step, as time consuming as the signal cutting is thin, needs the use of a speech recognition system to find anchors of the corresponding textual transcription in the entire text. The extracts obtained are then merged to rebuild the sentences of the original text. The second step is the annotation of the text sentences and the signal fragments that have been put into correspondance during the first step. Data representation using ROOTS is done from the audiobook splitting step. It is enriched as data annotation is done by the addition of new description sequences and relations between them.

3. Corpus Representation

ROOTS is a software library designed to handle a coherent set of data structures built to describe annotated speech and linguistic contents(Barbot et al., 2011). This structure is used during the whole process and for each annotation state to build a unique representation of all annotating elements of the corpus.

3.1. Items, Sequences and Relations

The main invariant of the ROOTS structure is the time which seems to be the only valid referential concerning speech and justifies the use of item sequences. This axis can be materialized by time anchors or, in an abstract manner, by the order of the events (for example, a succession of syllables anchored on a succession of words).

Each type of annotation is described by a sequence of items. An item is an object whose nature may vary a lot, for example, a transcription of text, a label, an acoustic segment, etc. The only constraint to respect, to insure consistency of the sequences, is that all the items in one sequence are always of the same type.

Relations allow to connect sequence items in a n-to-m manner. ROOTS produces an XML file, that gives a complete description of the corpus, but may also provide files under well-known formats to guarantee the possible use of other tools to make analysis or manual verification like Wavesurfer or Transcriber.

Figure 2 shows an extract of ROOTS class diagram with the three main classes : Sequence, Relation and Utterance. The

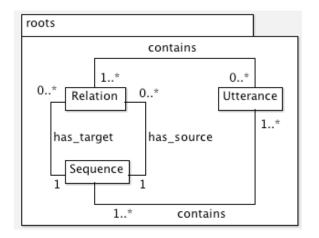


Figure 2: Class diagram of ROOTS main components: sequence, relation and utterance. The utterance is the entry point of the ROOTS architecture and can contain both sequences and relations. Sequences are generic since they are composed by an ordered set of items. The items can be as simple as a word or as complex as a syntax tree. Relations are built upon couples of sequences and put into relation the items contained by them. The mapping realized by a relation can be a *one-to-one*, *many-to-one* or *one-to-many* mapping.

last one, Utterance, is the main container that regroups all the sequences and relations corresponding to a piece of text.

3.2. Complete and Versatile Structure

A lot of different information can be kept in a ROOTS structure, going from linguistics (words, POS tags, syntax, named entities, fillers), or phonology (phonemes, syllables) to acoustics (acoustic segments, non speech sounds, allophones).

In addition, ROOTS has a modular structure which allows to keep several sequences of the same type simultaneously. For example, it is important to keep the original text, the spoken text (which may contain corrections of misspelled words or specific pronounciation) and finally the output of syntax analysis (for example, "j'avais" contains two words). For each previous type of text, a particular sequence is associated. Those three sequences are then linked together by relations, as shown on figure 3, which give a precise correspondance between items. Thus, this point of view enables to keep all the information untouched and to keep, in a unique structure, various utterance annotations. Furthermore, information can be split into several files while the coherence is also preserved.

4. Speech Signal Split and Alignment with the Text

Many papers are dealing with the alignment of large texts and their corresponding signal. (Braunschweiler et al., 2010) has proposed a system which automatically aligns text parts of an audiobook for TTS applications. Other systems deal with approximate transcriptions (Tao et al., 2010) or align the text without splitting it (Moreno and Alberti, 2009)(Prahallad et al., 2007). In our process, the text is

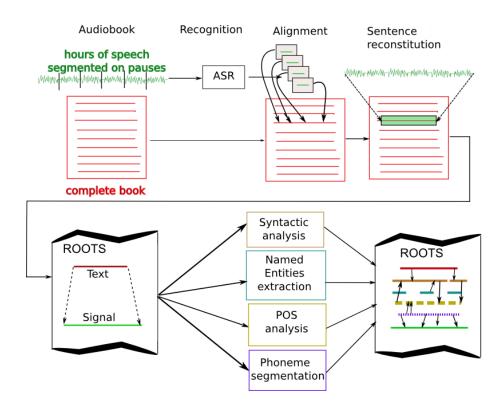


Figure 1: This figure describes the audiobook annotation process which begins with a complete text and the corresponding recording. This process is based on (1) the use of ASR to split the complete text and signal into smaller pieces (of the size of a sentence) and (2) the annotation of the text using the ROOTS structure.

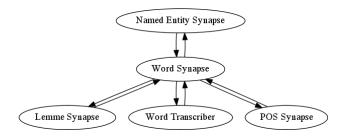


Figure 3: Example of relation graph obtained from relations contained in ROOTS files. This graph shows 4 sequences and their relations as arcs between sequences. Relations can be composed to move from one sequence to another in the graph. Here, the sequence named *Word Synapse* can be considered as a pivot in the graph.

segmented to be annotated and manipulated more easily, especially for human operations.

We have chosen to realize the text alignment with the signal in three steps like illustrated in the figure 1:

- 1. Split the signal on large acoustic pauses,
- 2. Find the corresponding text description of the signal by using Automatic Speech Recognition (ASR) System,
- 3. Align the recognized text with the original one.

The main goal of this alignment is to put text anchors in the signal.

Rought cutting of a speech signal is based on the observation of energy levels and lengths of silent pauses. Thresholds are determined and adjusted by the current speaking rate and recording levels. The goal is to produce segments that represent units smaller than a sentence.

Text recognition of each signal fragment is achieved by using the Nuance ASR system (Nuance, 2010). Models are speaker independent and the language model is learnt on the text extracted from the whole analyzed audiobook. By comparing the original text with the recognized one, we can determined word error rates. These errors can result of wrong recognition achieved by the system or a mistake realized while the speaker reads the text (like a bad pronounciation or a text replacement). In our experiments, the analysis of the audiobook we use, which duration is more than 11 hours, leads to a word error rate of 5.2%. This rate can be reduced by adapting the speaker models with a small amount of sentences extracted from the original audiobook.

The text provided by the recognition system is aligned with the original text using a Levenstein distance computed at a word level. When the end position of a text segment is found, the original text segment is associated to the signal and deleted from the original text. This process is repeated until text end.

Finally, to preserve as much as possible the text structure, segments are brought together into sentences by using major punctuation marks. When a segment is not ended by a terminal punctuation mark, it is grouped with the next one. Nevertheless, the boundary position between the segments is memorized to help later processing. The segmented text and signal are then adapted for Transcriber (Barras et al., 2001). A speech turn is composed of a single sentence or a set of undivided adjacent sentences. This prevents dividing

segments too early. Synchronization points define boundaries between two audio segments. The use of transcriber allows manual verification before the annotation process: differences between the original and the recognized text are highlighted by specific tags to be manually checked. The alignment of the segmented text and signal is automatically converted into the ROOTS format.

5. Data Annotation

In a second stage, ROOTS provides acoustic and textual data in an appropriate format to the annotation tools that return their own analysis. Currently, annotation levels used for our studies are the following:

- named entities extraction,
- syntactic annotation,
- part of speech tagging,
- phone segmentation,
- prosodic prominence extraction.

The grammatical and syntactical analyses are produced using a Synapse Developpement software, (Synapse, 2011) (Developpement, 2011) while acoustic analysis is done with our own tools. The information provided by each annotation system is included into the ROOTS object thereby refining the corpus description and allowing the generation of new relations between items of different sequences. For example, after a complete analysis, one can define, for a given phoneme, to which word it belongs and what is its corresponding POS tag.

To complete the data structure in a speech synthesis context, other information are essential (pitchmarks, fundamental frequency contours, etc.) but are not part of the current study.

6. Application

6.1. Material

The complete annotation process was experimented on a Marcel Proust's literary work called "Albertine disparue". This book contains around 120,000 words and its duration is 11 hours and 43 minutes. Since Proust had rewritten some parts of the book, the version of the text has been checked.

The process described on figure 1 is divided into two main phases: block alignment and then annotation. The rest of the section describes the results obtained for those phases. A special focus is made on the segmentation process considering the manual verification we have done.

6.2. Phase 1: block alignment

6.2.1. Text Splitting

Tracks were splitted on silences whose duration were longer than one second. The split step has produced 11, 693 files with an average duration of 3 seconds. Then, files were analyzed by the ASR system. The average word error rate between the original and recognized texts is about 5.21%. This rate includes misspelled words and grammatical agreements. As we mentioned before, this error rate can be improved by speaker adaptation.

6.2.2. Text Alignment

Text obtained in output of the ASR system is then used to split the original text to match the signal tracks. When a conflict is detected on a starting segment or an ending segment word, particularly when the conflict is an insertion or a deletion, the alignment does not garantuee a suitable split of the text for this word. In this case, we indicates that a human operation is needed. This problem appeared 969 times out of the 11, 693 files which represent 8,3% of the segments. A manual check established that a real splitting error occured in 8% of the alerts (78 files). The high number of false alert comes from the fact that all the disagreements occuring on a frontier are signaled (insertion, elision, substitution) and a lot of substitutions have no impact on the alignment.

6.2.3. Sentence Reconstruction

Resultant files are then grouped into speech turns, using punctuation, in Transcriber files (one file per CD track which represents, generally, fifteen minutes of speech). We obtain a total of 3, 340 turns whose durations are between 660ms and 1min22s. One turn contains around 3.5 segments and the anchors of these segments are kept.

After the Transcriber file creation, all the conflicts between the original text and the ASR output are identified by using a specific tag. The operator can then add new tags to indicate reading mistakes or bad pronounciation. In our case, 86 lexical tags (word or expression replaced by another), 213 pronounciation tags were added and 78 corrections were realized.

6.3. Phase 2: Annotation

The next step is to generate a ROOTS structure from the splitted and aligned corpus in order to use various tools to get the levels of annotation.

Informations from the annotation tools are then integrated directly into the ROOTS structure. Using relations, it is possible to find, for each text or audio segments, all of its associated features. Figure 4 shows an example of a sentence generated from ROOTS. We can see that different texts are associated for this utterance: the transcriber text gives segmentation according to space characters while synapse text is separating elements according to their nature.

Actually, linguistic analysis has not been validated but the phone segmentation step was manually corrected on a 2 hours and 16 minutes long part of the corpus. Consequently, we study the results of the segmentation process considering the use of the text preparation process presented in this paper.

6.3.1. Phone Segmentation Process

The automatic segmentation is achieved by using Markov models (a model for each phone, a model for the pause and a model for the inspiration) on a phonetisation graph. Models are context-independent and they are trained on observation vectors of 39 coefficients (12 MFCC, the first order and second order derivative, and energy). These vectors are computed on 30ms frames using a 10ms shift. The phonetisation graph is obtained by the text phonetisation achieved by Liaphon (Bechet, 2001) which has been completed by variations of some phonemes: /ø/ and /ə/ are optional for a

majority of words, connections are optional too and may be preceded or followed by short silences. Finally, pauses can be replaced by inspirations. During the segmentation step, the speech segments we use are as short as possible in order to take into account variations directly in the graph without reducing the efficiency of the alignment algorithms. The text used by the segmentation process does not take into account the lexical and pronunciation tags inserted in the Transcriber files, but takes into account the manual corrections of text splitting.

6.3.2. Phone Segmentation Accuracy

On 82,936 phonetic units manually checked, 94.17% are correctly labeled, 2.52% are missing and 3.31% are replaced by another label. Furthermore, we can mention that 2.77% of phone insertion occurs. A great part of elisions involves $/\phi/$ (46%), this phoneme is generally optional, and 24% of its occurrences were not detected. The majority of substitutions (41%) concerns the substitution of /e/ by $/\epsilon/$ which is the result of a specific pronounciation. The second important part of substitutions (19%) deals with pauses and inspirations. Insertions are mainly composed of pauses and inspirations (72%).

These results are quite good but they can be improved by at least two means:

- Variants taking into account the speaker specific pronunciation may be added,
- Post-processing can be done on pauses and inspirations.

The alignment of phones leads to a good score: 86.34% of the phone boundaries are located at less than 20ms of manual positions, the average distance is about 8.7ms which is less than a shift between two frames. These measures are also collected before a manual correction of the text alignment. In that case the proportion of correct phone labels is 91.12% and 86.24% of phone boundaries were placed at less than 20ms from manual marks.

7. Conclusion

The automation of the whole process that converts an audiobook to a corpus annotated at several levels gives a quick access to large corpora. The time reduction gained with this methodology to build new speech synthesis voices is considerable when we take into account the time saved for both corpus recording and corpus annotation. The corpus representation generated by ROOTS removes the problems related to the heterogeneity of the files provided by each annotation systems and simplifies its handling. However, the annotation process has to be reinforced by confidence indicators for each annotation step, either to remove uncertain areas, or to enable a supervised modality. In that case, the indicators would be given to an operator to guarantee a fast intervention as well as data in an appropriate format for the validation tools.

8. References

N. Barbot, V. Barreaud, O. Boeffard, L. Charonnat, A. Delhay, S. Le Maguer, and D. Lolive. 2011. Towards a versatile multi-layered description of speech corpora using algebraic relations. In *Proceedings of the International Conference on Speech Communication and Technology* (*Interspeech*), pages 1501–1504.

- C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.
- F. Bechet. 2001. Liaphon un systeme complet de phonetisation de textes. *Traitement Automatique des Langues* (*T.A.L.*) edition Hermes, 42(1).
- N. Braunschweiler, M.J.F. Gales, and S. Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Eleventh Annual Conference of the International Speech Communication Association*.
- Synapse Developpement. 2011. Documentation Technique: Extraction d'entités nommées.
- P.J. Moreno and C. Alberti. 2009. A factor automaton approach for the forced alignment of long speech recordings. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4869–4872. IEEE Computer Society.
- Nuance. 2010. Dragon Naturally Speaking SDK Server Edition version 10.
- K. Prahallad, A.R. Toth, and A.W. Black. 2007. Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases. In *Proc. of Interspeech*. Citeseer.
- Synapse. 2011. Documentation technique: Composant d'étiquetage et lemmatisation.
- Y. Tao, L. Xueqing, and W. Bian. 2010. A dynamic alignment algorithm for imperfect speech and transcript. *Computer Science and Information Systems*, 7(1):75–84.
- K. Tokuda, H. Zen, and A.-W. Black. 2000. An hmmbased speech synthesis system applied to english. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*

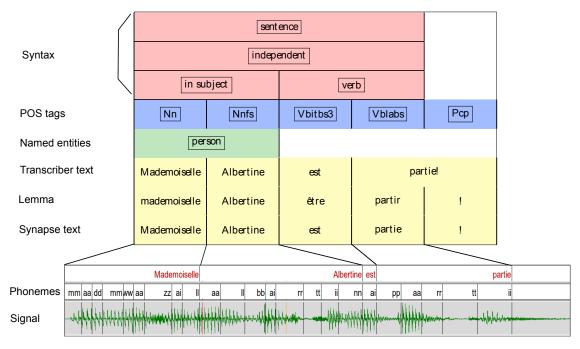


Figure 4: Illustration of an annotated utterance extracted from the corpus. The different levels of information are represented on the figure: syntax, POS tags, named entities, text (Transcriber, Synapse), lemmas, phonemes and signal segments. The vertical bars represent synchronisations between items of the differents sequences. For example, we can see that "est partie" is the verb of the sentence, is composed of two words and has two POS tags. The first one, which is "Vbitbs3" means it is a verb in the indicative present, is finite, singular at the third person.