Towards A Versatile Multi-Layered Description of Speech Corpora Using Algebraic Relations

Nelly Barbot[1], *Vincent Barreaud*[1], *Olivier Boëffard*[1], *Laure Charonnat*[1], *Arnaud Delhay*[1], *Sébastien Le Maguer*[1], *Damien Lolive*[1]

[1]IRISA - ENSSAT University of Rennes I, Lannion, France

{nelly.barbot,vincent.barreaud,olivier.boeffard,laure.charonnat}@irisa.fr

{arnaud.delhay,sebastien.le_maguer,damien.lolive}@irisa.fr

Abstract

This paper presents a software library, namely ROOTS for Rich Object Oriented Transcription System, that helps to describe spoken messages in a coherent manner linking sequences of items on numerous levels (linguistic, phonological, or acoustic). The proposed representation is incremental and can thus describe any or all parts of an utterance. In order to link different levels of description, algebraic relations are used. Instead of relying solely on fixed, pre-determined relations, algebraic composition operators are proposed that can create a missing relation on demand. In terms of software architecture, object classes are defined based on a well-grounded theoretical representation of speech (text, syntax, phonology and acoustics), without particular dependences on an annotation system (e.g. IPA is fully implemented). The API documentation for this software is available online [7].

Index Terms: Speech annotation, Algebraic relations, Software library.

1. Introduction

In speech processing, the study of phenomena related to various aspects of speech (linguistics, phonetic, etc.) requires the use of descriptive corpora. As an example, a text to speech system uses features on different descriptive levels in order to accurately predict parameters such as melody, phone durations, segmental acoustic units, etc. Moreover, such a system usually needs to build up accurate lists of candidate units in order to output the best speech quality as possible. For example, considering a sentence to synthesize, the system may have to answer to complex queries as "find the candidate units that correspond to open syllables ending with sound [i] that are located at beginning of a word itself at the end of a sentence". This request can also be extended to seek the 10 closest candidates of a target sound characterized by a specific spectral profile. To solve these queries, the system needs to know a set of complex relations from the acoustic signal to its linguistic description. Since all of these levels of description are entangled, syllables are related to words, phonemes to syllables, part of speech to words and syntax, one difficulty is to maintain consistency in structuring this multi-layered data.

In the past years, few systems have been proposed to cope with the problem of representing several levels of speech annotations in a unique and coherent structure. Bird *et al.* [1] have proposed a representation by a direct acyclic graph (Annotation Graph). HRG (Heterogeneous Relation Graph) has been proposed by Taylor *et al.* [2] in order to build a fast and efficient system for speech synthesis purposes. In HRG, linguistic information is represented by a graph where nodes are linked to linguistic items (words, syllables, phonemes, etc) and where edges define the relationship between theses items. HRG avoids replication of information in order to eliminate inconsistencies especially with respect to time. Cassidy *et al.* [4] have developed a tool called Emu for manipulating speech corpus annotations. More recently, Veaux *et al.* [5] have organized the annotation representations into two classes: the first one gives annotation information and the second one the hierarchical and/or sequential relations between annotations. All of these approaches offer a coherent way to structure a hierarchical information. However, it should be noted that these various relations between levels of information have to be defined in advance. Given the complexity of certain speech processing tasks, it seems difficult to predict all possible relations between different levels of description. Moreover having a great number of relations for the same utterance is not necessarily advantageous as the more relations between sequences we have, the greater is the risk of errors during updating features.

Our proposed system consists in defining first a minimum number of relations between layers of description (for example, a few well-grounded relations such as words to phonemes, words to syntax, phones to acoustic segments, etc). Next, new relations can be defined by algebraic composition of these pre-existing relations. These compositions are purely deductive, and therefore do not interfere with data consistency. For example, syncing words to the speech signal can be inferred by the following composition of relations: words to phonemes, phonemes to phones and phones to acoustic segments.

Offering different layers of description for the same statement also requires the use of different annotation tools (it may be an automatic system or a simple editor). It is hardly conceivable to have a unique tool to cover the different theoretical areas underlying the observed speech phenomenon. The use of such a combination of different annotation systems raises the problem of data consistency both in different levels of description and over time. The description system we propose takes into account this issue by providing mechanisms (notably time stamping) that will, in turn, enable a human expert or an automatic processing tool to determine the obsolescence of one piece of information compared to another.

This article addresses the problem of generating a coherent description on multiple levels including linguistics, phonology, and acoustics. In our opinion, an ideal annotation or transcription system for speech utterances should fulfill two basic requirements: to propose a functional description based on theoretical principles that ensure the fundamentals of speech (acoustics, phonetics, phonology, grammar and so on) and to relegate

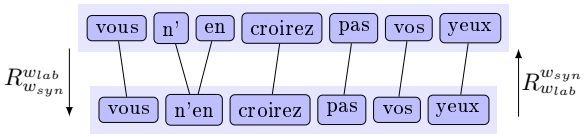

Figure 1: Example of Sequential Relations between two word sequences in French: w_{syn} at the top, and w_{lab} at the bottom. The direction of each relation is indicated by an arrow.

the semantic of these dependencies to the experimental settings or to the application that might use this corpus. Accordingly, on the phonological level, we have chosen to make an extensive use of the IPA system, on which a vowel such as the phoneme /a/ is defined by a whole set of descriptors including tongue position, aperture level or lips rounding gesture. This choice helps for dealing with various languages and authorizes well defined phonological requests regardless of the particular language (as an application example, asking if the next phoneme in a sequence is palatal or not).

This paper is organized as follows. Section 2 details the formulation of our model describing relations between sequences. Section 3 presents the main software components of the complete description system. Before concluding, an illustrative example is provided in Section 4 showing implementation of the description of an utterance.

2. Algebraic formulation of Relations

Within this framework of description, a sequence of items comprises one layer of description (it could be a sequence of words, a sequence of parts-of-speech, a sequence of MFCC coefficients, etc). Sequencing items is justified as the time axis is considered as a strong invariant in speech processing. This time axis seems to be a suitable referential for speech and may be concretely materialized by time indicators (e.g. a succession of syllables anchored on the acoustic signal) or in a more abstract manner by considering the order of events in a sequence (e.g. a succession of syllables linked to the word sequence where the acoustic signal is missing).

A relation is designed to link items between two sequences. Contrary to item sequences that are homogeneous, a ROOTS relation is generally heterogeneous by definition, that is to say it corresponds to a binary relation from a "source" item set to a "target" one, where the types of the source and target items can be different.

A relation from a source item sequence (a_i) to a target one (b_j) is stored as a matrix R^b_a. The (i,j)-th entry of R^b_a is a boolean number equal to 1 if a_i is related to b_j. The ordering of the sequence indices with respect to the time axis constrains the structure of these matrices: entries equal to 1 are organized as time-oriented blocks. The reciprocal relation from (b_j) to (a_j) is easily derived by the transposed matrix $\left(R^b_a\right)^T$.

In simple cases, a ROOTS relation can be interpreted as a function or an application: each "source" item is related to at most one "target" item. The associated matrix is then characterized by the presence of at most one (resp. only one) element equal to 1 in a matrix line for a function (resp. an application). Fig. 1 describes an example of relations between two word sequences, called w_{syn} and w_{lab}, derived from the French sentence "Vous n'en croirez pas vos yeux." (*"You won't believe it."*). The first word sequence, w_{syn}, corresponds to the tokenization derived from the syntactic analysis, whereas the second one, which puts together *n'* and *en* in one item (*n'en*), is more convenient to describe the syllabic construction of that sentence. Let us observe that w_{syn} chunks *n'* and *en* are related to w_{lab} item *n'en*. $R^{w_{lab}}_{w_{syn}}$ is an application and its matrix is given by:

$$R^{w_{lab}}_{w_{syn}} = \begin{array}{c} \begin{matrix} \text{vous} & \text{n'en} & \text{croirez} & \text{pas} & \text{vos} & \text{yeux} \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} \text{vous} \\ \text{n} \\ \text{en} \\ \text{croirez} \\ \text{pas} \\ \text{vos} \\ \text{yeux} \end{matrix} . \quad (1)$$

The reciprocal relation $R^{w_{syn}}_{w_{lab}}$ turns out to be more complex since item *n'en* is related to two target items. Considering this sentence again, it is composed of 7 syllables $s_1, \ldots, s_7$. The structure of syllables and the associated sequence, called syl, are detailed in the next section and described in Fig. 3. Then, another relation example can be built from w_{lab} to syl by the following matrix $R^{syl}_{w_{lab}}$:

$$\begin{array}{c} \begin{matrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} \text{vous} \\ \text{n'en} \\ \text{croirez} \\ \text{pas} \\ \text{vos} \\ \text{yeux} \end{matrix} . \quad (2)$$

Based on this algebraic concept, ROOTS enables modeling of various types of sequence relationships and provides the user with the definition of the semantic according to the considered application. Furthermore, due to the matrix form of the relation, composing relations is possible and easy to calculate. For the given item sequences a, b and c stemming from the same utterance, if relation R^b_a from a to b corresponds to a function, it can be composed with relation R^c_b from b to c, using matrix product $R^b_a R^c_b$, in order to provide a relation from a to c. For instance, relation $R^{w_{lab}}_{w_{syn}}$ detailed in (1) can be composed with relation $R^{syl}_{w_{lab}}$ described by (2). The following relation from w_{syn} to syl is then obtained:

$$R^{syl}_{w_{syn}} = \begin{array}{c} \begin{matrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} \text{vous} \\ \text{n} \\ \text{en} \\ \text{croirez} \\ \text{pas} \\ \text{vos} \\ \text{yeux} \end{matrix}$$

A graph of linked sequences can be built upon the set of relations. A relation corresponds to an arc between two nodes in the graph. Then, when a relation composition is possible, it turns out that the resulting relation depends on the path choosen for this graph traversal. More precisely, for the item sequences a, b, c and d, R^c_a can be different from $R^b_a R^c_b$ or $R^d_a R^c_d$ even if these (composed) relations are defined. It is due to the lack of "reversibility" of a relation: product $R^b_a (R^b_a)^T$ does not equal the identity matrix in general. The interpretation of this is that

composing relations can reveal uncertainty in the structure of relations. Consequently, one can use this propriety, either automatically or manually, as a help for choosing a path to compute a new relation. One major benefit of ROOTS relations is the focus on essential relations (word to syntax / word to phonology / phonology to acoustics, etc) and ROOTS delegates the inference of any missing relations as needed (like word to acoustics).

3. Software description and implementation

The software development of ROOTS is based on an object-oriented model: each type of annotation is described in a class and each annotation corresponds to an instance of that class.

Within a class, each instance is stored in a sequence and is denoted as an "item" afterwards. ROOTS affords, for each of its objects, a serialization operator towards an XML external description. Each object is then responsible for its own external descriptors and is capable of loading such an XML description when it is created. This serialization mechanism guarantees the encapsulation hierarchy for complex objects. In the same way, we have added a graphical output mechanism in order to build figures in the LaTeX/PGF format. This visualization tool is very convenient to analyse and illustrate (this is the case of figures presented in this article). At the same time, an import/export mechanism towards non XML files is provided. It relies on the most common formats, for example the HTK format in order to describe acoustic segments and labels and to guarantee the possible usage of common tools like Wavesurfer or Transcriber.

The heart of ROOTS is based on two essential classes: *Sequence* that contains items and *Relation* that enables linking of items of two sequences.

The generic aspect of the representation enables keeping a coherent semantic for sequences by guaranteeing the homogeneity of items they contain, but it also enables diversifying them by specializing their classes. For example, a sequence of acoustic elements may use raw signal segments, represented by a sequence of samples, and also segments represented by any kind of analysis by synthesis parametric model. This example points out the fact that coherence is insured regarding the nature of items even if their representations are mixed in the sequence.

In an automatic speech processing framework, the type of items one could encounter would entail the different speech analysis levels from linguistic aspects (syntax, words, parts-of-speech, ...), phonological aspects (phoneme, syllable, ...) to acoustic aspects (signal, F_0, ...). Each level is represented by one or more sequences allowing for several representations.

The class Relation factorizes methods to access indices within the related sequences as well as to check the existence of a relation between two items. It gives a common interface to concrete classes that inherit from it. An object of class Relation also stores the references of the related sequences. This object may be constructed from external information (e.g. a file given the positions of phonetic tags on the signal) or by operations on the sequences (e.g. an alignment of two word sequences).

One important aspect of ROOTS is its ability to represent multilingual data by using ISO standards. For example, at the phonological level, the full IPA alphabet has been implemented in ROOTS. A specialized class, that inherits from the class IPA, has been written in order to restrict the IPA alphabet to French and also to provide methods to convert the IPA label to other French alphabets (for example, the one from *liaphon* [6]). Those methods enable having a purely symbolic and readable representation of an allophone and are easily extensible to other languages.

Fig. 2 shows an example, also discussed in section 2, of sequences derived from the French sentence "Vous n'en croirez pas vos yeux." (*"You won't believe it."*). The grammatical tags of the Parts-Of-Speech (POS) sequence have been obtained using a syntactic analysis software for French, *Synapse Cordial Analyser*, and their semantic is as follows: "Pnp", personal pronoun; "Av", adverb; "Vb", verb; "Dsp", singular possessive determinant and "Nnmp", male plural noun. In this example, there are two sequences of words. The first one corresponds to the tokenization in relation with the syntactic analysis, whereas the second one, which puts together *n'* and *en* in one item (*n'en*), is more convenient to describe the syllabic construction of that sentence.

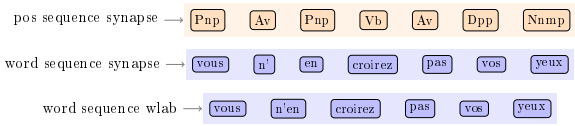

Figure 2: Sequences of items (POS and words) related to the utterance "Vous n'en croirez pas vos yeux".

Some items like syllables and syntactic trees are composed and based on other items, namely phones and words, respectively. We introduce the notion of composed items in order to deal with this new kind of representation. A composed item cannot exist alone within ROOTS. Indeed, the atomic elements that it is composed of are nothing but references of items of another sequence. So, a composed item qualifies an item (of another sequence) by declaring it as part of its structure. These composed items should not be confused with a true relation between sequences as explained in section 2.

Fig. 3 represents an allophone sequence (bottom) which is qualified by a sequence of syllables (top). A syllable item is structured as a tree. This structure is determined by phonological considerations. The leaves identify the atomic parts of a syllable, these are: onset (O), nucleus (N) and coda (C). The node that regroups nucleus and coda defines the rime. All the leaves contain references to items of the sequence linked to the tree (here, the phoneme sequence).

Considering Sequence and Relation classes, an additional class, called *Utterance*, can be defined in order to unify the multi-tier interpretation of a single speech realization. Thus, an Utterance contains item sequences and their associated relations. Fig. 4 illustrates the abstract representation of our system model based on Sequence, Relation and Utterance classes.

4. Illustrative example

In order to illustrate an object of class Utterance, Fig. 5 presents an excerpt corresponding to the words "Vous n'en" extracted from the sentence "Vous n'en croirez pas vos yeux". The Utterance contains several sequences of items (POS, w_{lab} words, w_{syn} words, syllabes, allophones, segments) and relations between them.

To get a specific piece of information associated to a segment, only sequences and relations are necessary to move from one level of description to another. A relation can be seen as a way to build a path through a graph. Let us consider the allophone [n] from the allophone sequence. Leaving the composed items problem aside, we can observe on Fig. 5, that this allophone is involved in the onset part of syllable S_2. If we want to get information about this particular allophone, like the word

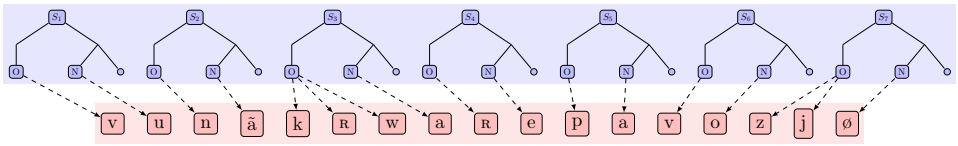

Figure 3: Mapping between a syllable sequence (top) and its corresponding allophone sequence (bottom).

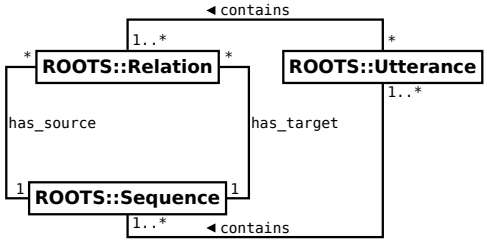

Figure 4: Class diagram showing the components of an utterance. The utterance is the entry point of the ROOTS architecture for describing speech and contains both sequences and relations. Sequences are generic since they are composed by an ordered set of items. The items can be as simple as a word or as complex as a syntax tree. Relations are built upon couples of sequences.

it is part of or the grammatical tag associated with it, we can compose the following relations: $R^{syl}_{w_{lab}}$, $R^{w_{lab}}_{w_{syn}}$ and $R^{pos}_{w_{syn}}$. Formally, relation R^{pos}_{syl} can be computed as:

$$R^{pos}_{syl} = \left(R^{w_{syn}}_{pos} R^{w_{lab}}_{w_{syn}} R^{syl}_{w_{lab}}\right)^T$$

where

$$R^{w_{syn}}_{pos} = \begin{array}{c} \begin{matrix} \text{vous} & \text{n'} & \text{en} & \text{croirez} & \text{pas} & \text{vos} & \text{yeux} \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} \text{Pnp} \\ \text{Av} \\ \text{Pnp} \\ \text{Vb} \\ \text{Av} \\ \text{Dpp} \\ \text{Nnmp} \end{matrix} .$$

Practically, the user can compute and store this relation once or can use the methods of the different relations when necessary. In the first case, the computation of the matrix product may be done for relations that may be used often. On the contrary, the second method adopts an object-oriented point of view. The user is always responsible for the semantic of the relation he obtains by composition. In this example, the composition associates two different grammatical tags to a unique syllable. This can be explained by the fact that in spoken language two words may be contracted into one syllable (as it is the case for words "n'" and "en"). Then the user/tool is in charge of solving this issue according to the information he wants to keep.

5. Conclusion

In this article, we have proposed an original method for representing several levels of description of a speech utterance. The software framework, ROOTS, is well adapted for describing

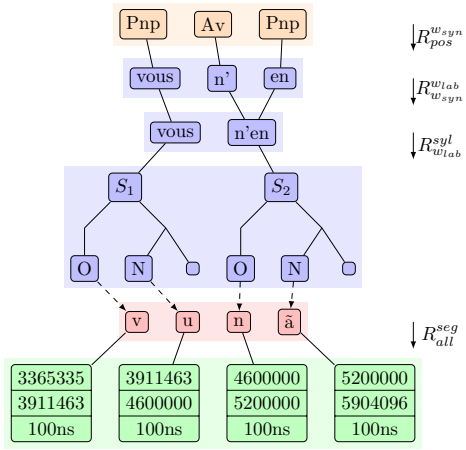

Figure 5: Excerpt extracted from the French sentence "Vous n'en croirez pas vos yeux". From top to bottom, this figure represents POS sequence, words sequences w_{lab} and w_{syn}, and syllables, allophones and acoustic segment sequences.

speech corpora and is a priori independent of the language. Two main notions underly the software architecture: first, sequences guarantee a time-based relation order, and second, relations allow for connection of items of sequences in a n-to-m manner. One originality of ROOTS lies in algebraic modeling of relations between sequences. Such a modeling allows for the composition of new missing relations very efficiently. As for the capabilities of the description, we are currently developing a comprehensive multilingual syntactic level (following the ITU standard). In terms of software development, we are working on a set of tools to interact graphically with the XML representation.

6. References

[1] Bird, S. and Liberman, M., "A formal framework for linguistic annotation", Speech Communication, 33(1-2):23–60, 2001.

[2] Taylor, P. and Black, A.W. and Caley, R., "Heterogeneous relation graphs as a formalism for representing linguistic information", Speech Communication, 33:153–174, 2001.

[3] Rojc, M. and Kai, Z., "Time and space-efficient architecture for a corpus-based text-to-speech synthesis system", Speech Communication, 49(3):230–249, 2007.

[4] Cassidy, S., "Multi-level annotation in the Emu speech database management system", Speech Communication, 33(1-2):61–77, 2001.

[5] Veaux, C. and Beller, G. and Rodet, X., "IrcamCorpusTools: an extensible platform for speech corpora exploitation", Proceedings of the LREC Conference, 2008.

[6] Bechet, F., "LIAPHON - Un systeme complet de phonetisation de textes", Traitement Automatique des Langues (T.A.L.) edition Hermes, 42(1), 2001

[7] ROOTS homepage, http://www.irisa.fr/cordial/roots, 2011.