Thematic Alignment Of Recorded Speech With Documents

Dalila Mekhaldi
Université de Fribourg
Chemin du Musée 3, 1700 Fribourg
+41.26.429.66.78
Dalila.Mekhaldi@unifr.ch

Denis Lalanne
Université de Fribourg
Chemin du Musée 3, 1700 Fribourg
+41.26.429.65.96
Denis.Lalanne@unifr.ch

Rolf Ingold
Université de Fribourg
Chemin du Musée 3, 1700 Fribourg
+41.26.300.84.66
Rolf.Ingold@unifr.ch

ABSTRACT

We present in this article a method for detecting similarity links between documents' content and speech recordings' content. This process, further called *thematic alignment,* is a novel research area that combines both document and speech analysis. This alignment will a) provide temporal indexes to documents, which are non-temporal data, and b) help discovering hidden thematic structures. This article first introduces a multi-layered document structure and quickly introduces the traditional speech structure. Further, it presents a simple similarity measure and various multi-level simple alignments between those two structures. Later, the meeting corpus is presented, as well as an evaluation of the implemented alignments. Finally, we present our future works on multi-alignments and thematic structure discovery.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**] *indexing methods*

H.3.3 [**Information Search and Retrieval**] *Clustering- Search process*

I.7.2 [**Document Preparation**] *Index generation- Multi media*

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Meeting recordings, multimodal analysis, thematic alignment, multi-layered structure, Document indexing and retrieval.

1. INTRODUCTION

Document alignment is an important research area in multilingual alignment [2], and in text to audio alignment [6], which uses external textual sources (Internet, teletext, etc.) in order to improve speech recognition. However, the temporal bi-modal alignment we propose in this article, between meeting document and speech recordings, has never been tackled in none of the recent multimodal meeting analysis project [4]. This bi-modal temporal alignment, that we propose, exploits features from both document annotations and speech transcription data. This alignment will a) associate temporal indexes to the document (When was it discussed?) b) help building document-based interfaces for retrieving multimedia meeting data (What was said about a part of a document?).

We briefly present in this paper the different document alignments we have discovered. We then focus our presentation on document content alignment and more specifically on *thematic* alignments. We detail our methods of segmentations and similarity measures. Finally, we conclude with the evaluation of the various thematic alignments we have implemented, and some aspects for future works that are deduced from the document/speech alignment.

2. DOCUMENT TEMPORAL ALIGNMENT

Document temporal alignment consists in associating temporal indexes with the document parts that must be represented at its various granularity levels. This association of temporal indexes with textual data will be expressed by the alignment process between this document and the recorded speech, a process that can be defined as detecting the thematic links between related units. The detected links can be classified into three categories [4], depending on its expression: a) *Citation alignments* are pure lexicographic matches between terms in documents and terms in the speech transcription (such as: "The author said << …>>"). *b) Reference alignments* establish links between printed documents and structured dialogs through the references made to documents in speech transcript (such as: "the caption on the right side", etc.) c) *Thematic alignments* are content-based similarity links between document units and the dialog structure of speech.

3. METHODOLOGY

Determining the existing relations between documents and speech transcript consists in detecting the links between their respective units. For this reason, the documents and the speech transcripts must be first segmented in various structures. We present in this section various document's structures, that could be integrated in a single multi-layered representation, and briefly introduce the standard structure of the speech transcript. Further in this section, we present various alignment strategies and a method for measuring similarities between units.

3.1 Documents and Speech transcript segmentation

3.1.1 A multi-layer document structure

A document can be represented in various levels of structures, such as *physical*, *logical, thematic*, or *syntactical* structure. The *physical* level is often designated as the page analysis level [5], which dictates that the document is composed of a set of interconnecting rectangular printed regions. The *logical* structure

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
DocEng'03, November 20–22, 2003, Grenoble, France.
Copyright 2003 ACM 1-58113-724-9/03/0011…$5.00.

is a symbolic description of the document's structure and contents, e.g. title, author name, etc. [5] Currently, we are extracting the document's logical structure manually. The *thematic* structure is the text's organization into themes. The document's *thematic* structure we extracted, using *TextTiling* [3], is not yet satisfactory for the document type we are handling. This initial evaluation thus focuses on other document structures. In the future, treating in parallel the text-tiling and alignment should improve both processes. Finally, the *syntactical* document structure is its description as a sequence of textual components, e.g. words, sentences, paragraphs, etc., without the concern about geometrical and typographical properties.

3.1.2 Speech transcript structure

We used for the software *Transcriber* tool [1] for manually transcribing the speech. It describes the speech structure as a sequence of *thematic episodes* (sections), which involve one or more speakers engaged in a dialog about a specific topic. Each *episode* is composed of *turns* where each *turn* can be decomposed into *utterances* (a small coherent part of one speaker's speech) [1]. The following example shows an extract of a speech transcript:

```
<Section>
<Turn speaker="Rim" startTime="11.81" endTime="15.84">
 <Sync time="11.81"/>  Voilà, alors les surprises du procès Elf.
 <Sync time="12.77"/> La semaine du procès Elf commence…
</Turn>
</Section>
```

3.2 Document and Speech alignment

3.2.1 Alignment strategies

First of all, our alignment technique is oriented. A *source* file is aligned with a *target* file. For each unit of the *source* file, a most similar unit in the *target* file is considered. Therefore, the alignment is asymmetrical; if a unit *u1* from a document *D1* is aligned with a unit *u2* from a document *D2*, it does not mean that *u2* will be aligned with *u1*. Thus there are two alignment orientations to take into consideration: a) from documents to speech transcript and b) from speech transcript to documents.

Many alignment strategies can be explored considering the numerous segmentations available for documents and speech transcripts and the two different directions of alignment. However, most of them do not provide any significant benefit to the alignment process. For example, the *physical* segmentation is mainly useful when browsing the alignment's results. Further, aligning documents' *logical* blocks with speech *utterances* will not be informative because only one best *utterance* will be found for each *logical* block, which imposes that the *source* units should be smaller or equal to the *target* units. This limitation could be easily solved by not only considering the best alignment for each source unit but all the alignments that overcome a certain threshold. We could also consider other metrics for comparing units, such as *membership* and *ownership*. However, we start by evaluating simple alignments by the use just of the *similarity* method. The following units are thus considered for the source file: a) *utterances* and *turns* in the speech transcript and b) *sentences* in the document.

Alignment requires a common representation format, so that each level of alignment can be combined at the end. We are currently representing both document and speech transcript as streams of characters and their various annotations point on these streams.

3.2.2 Similarity measure

Assuming that every document unit and speech transcript unit is represented as a bag of weighted terms, it is possible to compute pairwise *similarity* between units, which is based on the co-occurrences of terms in the respective units (*cosine* measure). For two vector representations *x* and *y*, and *n* distinct terms, where $w_{t,v}$ is the weight assigned to a term *t* in vector *v*:

$$\boldsymbol{cos(x, y) = \sum\nolimits^{n}_{t=1} w_{t,x}\, w_{t,v} / \sqrt{\sum\nolimits^{n}_{t=1} w^{2}_{t,x} \sum\nolimits^{n}_{t=1} w^{2}_{t,v}}}$$

The evaluation, presented at the end, uses only the similarity measure. However, we observed the need for two other measures: 1) *membership* (is part of) and 2) *ownership* (contain).

4. EVALUATION

4.1 Test Data

The first step for validating the integration of documents into multimedia archives, and to measure the document/speech alignments, is to build corpuses of meeting recordings based on scenarios where participants have a high interaction with documents. We have decided to focus our efforts on press reviews, i.e. meetings where participants discuss the cover page of the daily newspapers, which contain several small articles with heterogeneous topics. Thus, press reviews follow a structured agenda that should fit well document temporal alignment through document content alignment with speech transcripts. In the next sections, we present the results of diverse alignments, at some fixed levels of the document and speech transcript structures. We studied in particular eight meetings, with a total of 228 *turns* and 572 *utterances*. The newspaper's cover pages studied, are composed of 90 *logical units* (newspaper articles) and 1409 *sentences*. Among the 8 meetings tested, two of them treat several documents, such is the case in real meetings. In this case, we have grouped all the documents in a single collection.

4.2 Metrics for evaluating alignments

The *Recall* and *Precision* notions help evaluating the quality of a given alignment in respect to a prepared manual *ground-truth*, that contains all the possible alignments, where null alignments are not considered, as well as units containing only stop words.

Recall= Number of correct alignments found/ Number of correct alignments that should be found

Precision= Number of correct alignments found/Number of alignments found.

Efficiency measure F=2*(*Precision***Recall*)/(*Precision* +*Recall*)

4.3 Alignment results

In most of the incorrect alignments generated, that should have been null according to the *ground-truth*, the *similarity* value was inferior to (0.1). For this reason, we have fixed this value as a threshold. However, as the *similarity* value is based on terms weight, in respect to their frequency in their units, it is heavily influenced by the units' size. Thus, the threshold must be calculated according to various variables (units' size, *membership* and *ownership* values, etc.).

4.3.1 Aligning documents with the speech transcript

In this alignment direction, we have considered the document's *sentences* as the units to be matched with the speech *utterances*, and then with the speech *turns*, as showed in Table 1.

Table 1: Aligning document's sentences with a) speech utterances and b) speech turns

Alignment pairs	Units number	R	P	F
Sentence/Utterance	1409 (8 meetings)	0.87	0.51	0.63
Sentence/Turn	1409 (8 meetings)	0.78	0.60	0.67

Precision values are relative low. This is mainly due to the two meetings where several documents were used, which multiply the possible matches of *sentences* with *utterances* and *Turns*. When matching *sentences* with *turns*, the *similarity* threshold is too much filtering, and *membership* measure should be considered in order to avoid the correct alignment elimination.

4.3.2 Aligning the speech transcript with documents

We finally tried to align documents and speech in the reverse order. We have considered in this case the smallest speech transcript unit, *utterance*, and matched it with two document units: a) *sentences*, and b) *logical units*. Table 2 shows the results:

Table 2: Aligning speech utterances with a) document's sentences and b) document's logical structure

Alignment pairs	Units number	R	P	F
Utterance/Sentence	572 (8 meetings)	0.83	0.71	0.77
Utterance/Logical unit	572 (8 meetings)	0.84	0.77	0.80

When trying to align *utterances* with document units, most of the inexistent alignments are detected because of terms' co-occurrence, even thought the topic discussed is different, this require to ignore the terms that are equitably frequent in overall collection, and that alter the similarity calculation. Other *utterances* are imperfectly aligned because of the likeness of topics between some document's units. This problem especially appears when discussing about different documents having a similar content. This conflict can be avoided by considering more than one similar unit for each *utterance*. When aligning *turns* with *sentences*, more than one *sentence* can be matched with a specific *turn*, mainly because the *source* unit is larger than the *target* unit. Further, a *turn* can contain several topics and could be as well aligned with several paragraphs or *logical units*. As seen previously, this problem can be resolved by considering more than one pertinent unit in each alignment.

Table 3: Aligning speech turns with a) document's sentences and b) document's logical structure

Alignment pairs	Units number	R	P	F
Turn /Sentence	228(8 meetings)	0.86	0.69	0.77
Turn/Logical unit	35 (2 meetings)	0.88	0.81	0.85

4.4 Remarks

The first remark is that this simple alignment gives back good results, partially because the first meetings captured are stereotyped and follow very closely the document structure and content. However we realized that we should consider all similar units whose similarities overcome a certain threshold, especially when the *source* unit's size is higher than the *target* one (*Turns* vs. *sentences*, *logical units* vs. *utterances*), this would make the alignment symmetrical. Indeed the relationships detected that overcome the threshold, will be the same in both alignment directions. Nevertheless, we need a proper *ground-truth* for evaluating those multiple alignment, which is a complex task due to the subjectivity and multiplicity of alignments. Further, the alignability of two units must consider the *similarity*, the units' size, the *ownership* and the *membership* values. Correct thresholds and heuristics will have to be defined either with empirical studies or statistical methods.

An important aspect of this symmetrical alignment is that it can help discovering the *thematic* segmentation of both speech transcript and documents. Considering that two trees represent the document and the speech transcript, aligning these trees will consist in finding *similarities* between their nodes (double oriented links). Detecting the most connected regions of the whole bipolar graph can then help discovering *thematic regions* (see Figure 1a). Furthermore, this symmetrical alignment provides a solid framework for merging the individual alignments at various levels of both sources trees (Figure 1b).

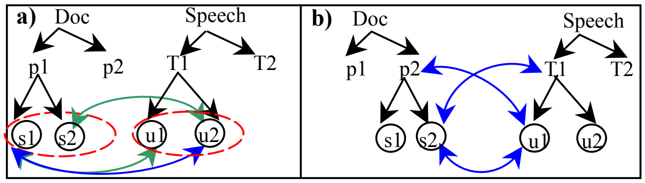

Figure1: a) Thematic regions discovering b) merging the individual alignments.

5. CONCLUSION

We proposed in this article a method for integrating non-temporal documents to multimedia meeting archives. We described strategies for aligning their content and structures, with the speech transcription of the meetings. In this preliminary study, we have noticed that document alignment is closely related to the preceding segmentation phase, which we plan to reunify in a single proceeding loop. We have also discovered that alignments can help discovering hidden document structures, such as the *thematic* structure, which will constitute our future work. Finally, we plan to evaluate our methods with other types of documents and document-oriented meetings.

6. REFERENCES

[1] Barras C., et al., Transcriber: development and use of a tool for assisting speech corpora production Speech Communication, Speech Communication, vol. 33, 2001.

[2] Ghorbel H., Alignement Multicritères des Textes Appliqué aux Documents Médiévaux: Critères Linguistiques et Structurels, PhD thesis, Swiss Federal Institute of technology (EPFL), 2002, Lausanne.

[3] Hearst M., Multi-Paragraph Segmentation of Expository Text, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics 1994.

[4] Lalanne D., Sire, S., Ingold, R., Behera, A., Mekhaldi D. and von Rotz, D. A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings, 3rd International Workshop on Multimedia Data and Document Engineering, in conjunction with VLDB 2003.

[5] Niyogi D. and Srihari S.N. Knowledge-based derivation of document logical structure, ICDAR'95.

[6] Roy, D. and Malamud, C., Speaker identification based text to audio alignment for an audio retrieval system, ICASSP'97.