Recherche d'information personnalisée

Patrice BELLOT

7 juin 2011

Table des matières

PREMIÈRE PARTIE

NAVIGATION

Chapitre 1

Navigation dans les documents audio par le résumé automatique

1.1. Introduction

Avec la facilité d'enregistrer et de stocker des données , il devient urgent de pouvoir manipuler ces données avec la même facilité que pour des données textuelles. L'avènement des baladeurs numériques, par exemple, a fait émerger l'écoute d'émissions de radio-amateurs (*podcasts*), et de livres lus, disponibles à la demande sur Internet. Même si ces documents sont souvent consommés comme des émissions de radio, leur archivage est généralisé et il n'existe pas de solution pour les retrouver par leur contenu. Seules des métadonnées créées par leurs auteurs permettent d'y accéder.

Dans de nombreux domaines, des conversations sont enregistrées et archivées. Les services client par téléphone, par exemple, étudient *a posteriori* le contenu des conversations entre agents et usagés pour améliorer leur services. Dans les domaines légaux et financiers, de nombreuses conversations sont enregistrées pour assurer une traçabilité des décisions. Toute réunion de travail peut être potentiellement enregistrée pour permettre aux participants de retrouver une information orale, ou à d'autres de se tenir au courant de l'avancement des sujets discutés. Bien que l'enregistrement et l'archivage de documents audio soient très développés, il n'existe que peu de moyens de structurer, indexer et retrouver l'information qu'ils contiennent.

La dans les documents audio est un problème omniprésent dû à la nature éphémère du son. En effet, la lecture du son est continue dans le temps et alors que l'on peut

---

Chapitre rédigé par Benoit FAVRE.

identifier un objet en y jetant un coup d'œil, il faut écouter un son dans son intégralité pour l'identifier. Il semble plus difficile de localiser des événements dans le temps que d'utiliser le retour continu de la vision pour localiser des objets dans l'espace. Il en résulte une difficulté à développer des interfaces efficaces pour accéder au contenu de documents audio.

Il a été vu au chapitre **??** de cet ouvrage qu'il était possible de retrouver la réponse précise à une question dans un enregistrement audio en utilisant des méthodes de QAst (*Question Answering on Speech Transcripts*). Il est possible de retrouver des éléments factuels et définitoires mentionnés explicitement dans des enregistrements de parole préparée ou spontanée. Toutefois les questions plus complexes ne sont pas couvertes par ce paradigme et il serait souhaitable de traiter des requêtes de l'utilisateur dont la réponse n'est pas explicite ou nécessite un traitement élaboré de l'information.

La navigation peut être vue comme une extension de la recherche d'information, enrichissant les moyens de capturer le besoin de l'utilisateur. D'une part, une requête de quelques mots comme il en est donné aux moteurs de recherche web actuels ne permet pas forcément de cerner complètement le besoin de l'utilisateur. L'historique des interactions, mis en rapport avec celui des autres utilisateurs, offre par exemple un moyen beaucoup plus riche pour capturer ce besoin. D'autre part, il existe de nombreux intermédiaires entre rapporter une liste de documents et rapporter des réponses ciblées comme le fait la QAst. La navigation permet à l'utilisateur d'explorer rapidement de grands documents et de conceptualiser une réponse courte, érigeant un pont entre la recherche documentaire traditionnelle et la problématique question-réponse.

Les types de besoins utilisateur visés par la navigation sont les suivants :

– un besoin exploratoire : l'utilisateur n'a pas de besoin précis mais souhaite capturer les différentes facettes des documents pertinents d'un thème donné ;

– un besoin de navigation : retrouver une information précise dont il connais l'existence mais à laquelle il ne sait pas accéder ;

– un besoin d'investigation : l'utilisateur veut comprendre comment une décision a été prise, comment des événements ont abouti à une situation.

Ces besoins ne peuvent être remplis que par une exploitation profonde du contenu des documents. Il faut donner à l'utilisateur les moyens de retrouver ce qui l'intéresse rapidement et avec précision.

La modalité audio apporte son lot de complexités spécifiquement dues à la nature du média traité. Une interface utilisateur facilitant la navigation dans des documents audio devrait :

– convertir la localisation temporelle en localisation spatiale pour que l'utilisateur puisse utiliser ses capacités à situer des objets dans l'espace ;

– établir la structure de l'information, sous une forme hiérarchique, par exemple ;

– diriger l'utilisateur vers l'information importante afin de limiter le temps qu'il passe à écouter des éléments non pertinents.

Le vise à produire une version concise d'un ou plusieurs documents documents ne contenant que l'information la plus importante, répondant éventuellement à un besoin utilisateur. Idéalement, la recherche d'information devrait faire appel au résumé automatique pour présenter des résultats synthétiques, demandant un temps minimal pour être appréciés par l'utilisateur. Même si le manque de robustesse des approches actuelles ne permet pas une telle application, les méthodes de résumé par extraction tendent à sélectionner l'information la plus pertinente tout en pénalisant la , une idée que nous allons développer pour la navigation.

Dans ce chapitre, nous allons tout d'abord lister l'état de l'art de la navigation et du résumé dans les documents audio, puis nous détaillerons une expérience prouvant l'utilité du résumé de parole. Deux applications seront alors explicitées pour illustrer une meilleure capture du besoin utilisateur à l'aide de mots-clés et une navigation dans des documents s'étalant sur une grande durée temporelle.

1.2. Recherche d'information parlée

Nous détaillons ici les spécificités de la recherche documentaire dans des données audio, et les approches pour résumer automatiquement ce média.

1.2.1. *Recherche documentaire*

Les progrès en de la parole ont permis d'imaginer l'application aux documents audio de méthodes de recherche d'information développées pour le texte. Ces méthodes ont dû être adaptées pour faire face aux problèmes apportés par l'automatisation de la transcription :

– les documents sont bruités à cause des erreurs de transcription, jusqu'à 40 % des mots sont faux ;

– la transcription se fait sur un vocabulaire de taille limitée, laissant de côté les mots peu fréquents comme les noms propres.

Le premier problème a surtout été étudié à travers la tâche de recherche de documents parlés () qui a fait l'objet d'évaluations lors des campagnes [GAR 00] et [PEC 07]. Lors de la tâche de TREC, les participants devaient effectuer des recherches documentaires dans 500 heures de journaux radiophoniques et télévisés, transcrits à

différents taux d'erreur mot[1] (de 15 à 50 % dans cette tâche). Les erreurs de transcription provoquent une baisse des performance en recherche documentaire à cause du bruit introduit par les erreurs dans les mesures de similarité entre documents et la requête utilisateur. Dans ce cadre, Johnson remarque que l' (ajouter à la requête des termes apparaissant dans les meilleurs documents retrouvés) et l' (même opération avant l'indexation en utilisant des données parallèles non bruitées) permettent de limiter l'impact des erreurs de transcription [JOH 00]. Toutefois, les mots importants sont souvent des noms propres peu fréquents, inconnus des systèmes de transcription automatique, et donc impossibles à retrouver. Cet effet est relativisé dans les campagnes TREC car les requêtes, peu réalistes par rapport aux usages actuels de moteurs de recherche du Web, font souvent plus de dix mots. Des systèmes de recherche de documents parlés calqués sur les moteurs de recherche Web sont décrits dans [HIR 99, VAN 00].

Le second problème, le traitement des mots inconnus, a été étudié à travers la tâche de détection de termes parlés (), par exemple lors de la campagne d'évaluation NIST [FIS 07]. Cette tâche consiste en la recherche de mots ou séquences de mots dans des fichiers audio. Afin de limiter les problèmes dus à la transcription automatique, l'index permettant les futures recherches n'est pas créé sur la meilleure hypothèse de transcription, mais sur un treillis d'hypothèses plausibles fonction des scores de confiance de transcription. L'index couvre alors plus de solutions possibles ce qui permet d'augmenter le rappel. Cette idée est poussée encore plus loin en relâchant les contraintes de mots : seul un réseau de ou de est conservé, autorisant la recherche de mots qui n'étaient pas connus du moteur de transcription [MAM 07, AKB 08].

La problématique question-réponse (Q/A), qui consiste à retrouver la réponse exacte et précise à une question de l'utilisateur au lieu de lui présenter un document contenant cette réponse, touche aussi au domaine de la parole. Ici, l'intérêt est porté au traitement des requêtes parlées qui sont capturées de manière plus précise à l'aide d'un dialogue homme-machine [VAR 08]. Néanmoins, les systèmes de question-réponse sont souvent cantonnés aux questions factuelles, ou aux questions dont la réponse se trouve exactement dans les données. Le résumé automatique, lorsqu'il est orienté par un besoin utilisateur, recherche des réponses aux questions plus difficiles, comme « pourquoi » et « comment ».

1.2.2. *Résumé automatique*

Le résumé automatique de parole tire ses origines de l'adaptation de méthodes développées pour le résumé de texte. L'idée est de créer une version synthétique du ou

1. Le taux d'erreur mots est le nombre d'insertions, de substitutions et de délétions dans la transcription automatique par rapport au nombre de mots prononcés.

des documents d'origine contenant une information pertinente et non redondante, en imposant une qui se traduit en général par un nombre maximum de mots, mais pouvant se rapporter à la quantité de temps que l'utilisateur souhaite passer pour assimiler les documents (temps d'écoute par exemple).

Les systèmes de résumé automatique de texte sont évalués lors de campagnes internationales annuelles, comme () organisée par NIST. Cette dernière requiert que les participants génèrent automatiquement des résumés d'une centaine de mots à partir d'une dizaine d'articles de journaux sur un sujet donné. L' est faite tout d'abord par des humains qui jugent la qualité du fond et de la forme des soumissions, puis automatiquement, par comparaison à plusieurs résumés écrits à la main. La mesure , la couverture en n-grammes de mots des résumés de référence, est la plus répandue pour l'évaluation automatique malgré ses nombreuses faiblesses comme la possibilité de générer des résumés de mauvaise qualité avec de très bons scores. On peut évaluer le résumé de parole de la même façon en gardant à l'esprit que comme pour la recherche documentaire, les erreurs de transcription ont un impact sur les méthodes de résumé et sur les méthodes d'évaluation.

Le résumé automatique de parole a été appliqué à différent domaines comme les journaux radio et télédiffusés [HOR 02, CHR 04, INO 04, ZHA 07, MAS 06a, MRO 05], les conversations téléphoniques [ZEC 02, ZHU 06], les cours magistraux [MRO 05, FUR 04] ou les réunions [MUR 05, LIU 08, RIE 08]. Chacun de ces domaines amène des problèmes différents et nécessite des approches spécifiques. Par exemple, alors que le résumé de journaux radiodiffusés est très semblable au résumé de journaux écrits (structure du contenu, phrases préparées), le résumé de conversations pose le problème du traitement des interactions entre plusieurs locuteurs.

Le résumé automatique de parole est principalement abordé comme un problème de à partir du contenu transcrit. Ceci étant, certaines approches font appel à la et suppriment des mots à l'intérieur des phrases pour les rendre plus courtes [HOR 02, FUR 04, LIU 09]. Les systèmes de résumé par sélection peuvent être catégorisés selon s'ils font appel à des méthodes supervisées ou non supervisées.

Les premières requièrent un corpus de documents audio accompagnés d'une annotation manuelle des phrases. Cela permet d'entraîner un à séparer les phrases importantes des phrases qui sont superflues, en fonction de caractéristiques textuelles, structurelles et acoustiques. Les caractéristiques textuelles sont fondées sur des mesures du type $tf.idf$ provenant du domaine de la recherche documentaire qui jugent de l'importance des mots en fonction de leur fréquence dans le document (voir par exemple [CHR 04]). Au niveau structurel, la position des phrases et leur longueur, le rôle des locuteurs ou la structure de l'argumentation donnent des indices précieux pour déterminer si une phrase est importante [MUR 05]. Enfin, les caractéristiques acoustiques regroupent des statistiques sur l'enveloppe de la fréquence fondamentale et de l'énergie

du signal acoustique, la vitesse de locution, la présence de pauses et de défauts de locution (hésitations, interruptions, faux départs...) [MAS 06a, ZHU 06, INO 04, XIE 09].

Les méthodes non supervisées ne requièrent pas de données d'apprentissage, mais estiment l'importance des phrases uniquement en fonction de caractéristiques statistiques des documents résumés. Elles regroupent des approches portées du résumé textuel, comme () [GOL 00] qui sélectionne itérativement les phrases les plus pertinentes tout en étant le moins redondantes avec le reste de la sélection [MUR 05, RIE 08]. L'algorithme MMR repose sur une représentation dans le modèle vectoriel du contenu des phrases en entrée et du besoin de l'utilisateur. La pertinence d'une phrase est calculée comme sa avec le besoin de l'utilisateur exprimé sous forme vectorielle. Dans le cas d'un résumé générique, le besoin de l'utilisateur est remplacé par le vecteur créé par le sac de mots du document ou de l'ensemble de documents. La redondance est mesurée par la similarité *cosine* entre deux phrases sélectionnées pour le résumé. L'algorithme MMR est détaillé dans la figure 1.

Soit $D = \{p_0 \dots p_n\}$ l'ensemble des phrases disponibles ;
Soit L la limite de longueur ;
Soit $S = \emptyset$ l'ensemble des phrases sélectionnées ;
while $\sum_{p \in S} longueur(p) \leq L$ **do**
  $\hat{p} = argmax_{p \in D}\, \lambda pertinence(p) - (1 - \lambda)\, max_{q \in S}\, redondance(p, q)$;
  if $longueur(\hat{p}) + \sum_{p \in S} longueur(p) < L$ **then**
    $S = S \cup \{\hat{p}\}$;
  end
  $D = D \cap \{\hat{p}\}$;
end
Retourner S.

Algorithm 1: *Maximum Marginal Relevance*, algorithme glouton qui maximise la pertinence et minimise la redondance dans le résumé.

Afin de modéliser la redondance plus finement, la méthode à base de concepts [GIL 09] cherche une sélection de phrases qui maximise le nombre de concepts dans le résumé en ne comptant qu'une seule de leurs occurrences. Contrairement à MMR, cette approche recherche une sélection de phrases selon un critère d'optimisation global, pouvant être formulé comme un programme en nombre entiers, résolu par des implémentations classiques.

Il existe aussi des travaux pour créer des résumés sans effectuer de transcription automatique. Par exemple, Maskey construit une approche pour détecter les phrases importantes n'utilisant que des indices structuraux et prosodiques, ainsi qu'un modèle de séquence [MAS 06b]. Zhu recherche des motifs fréquents dans l'acoustique qui représentent des séquences mots, sans nécessairement identifier des mots [ZHU 09].

1.3. Interactions avec l'utilisateur

Les interactions avec l'utilisateur sont détaillées selon deux axes : la navigation locale dans un document de courte durée, et la navigation globale dans une collection de documents.

1.3.1. *Navigation locale*

Pour l'exploration à courte distance dans des documents audio, l'interface de navigation triviale est la barre de lecture. Elle permet de positionner la lecture à un instant t mais n'indique rien sur le contenu autour de cet instant. Plusieurs approches ont essayé d'améliorer les propriétés de cet outil.

Les travaux d'Arons [ARO 97] se concentrent sur la navigation à courte distance dans un flux audio. Il s'intéresse aux méthodes de lecture rapide (et retour rapide) qui donnent des indications sur le contenu du flux audio. Il classifie les méthodes de lecture rapide offrant un rapport de vitesse jusqu'à cinq fois la vitesse normale de lecture :

– la lecture des échantillons audio à une fréquence supérieure, ce qui a pour effet d'augmenter la fréquence fondamentale de la voix ;

– la suppression de segments à intervalles réguliers ou leur dissociation entre les canaux gauche et droit ;

– la suppression des pauses et la détection de segments de parole ;

– la détection de mots importants grâce à une analyse de la prosodie.

Arons propose une interface de navigation, SpeechSkimmer, qui combine ces propriétés pour naviguer rapidement en avant et en arrière dans des documents audio. L'interface prend la forme d'une télécommande avec des boutons pour activer diverses vitesses de lecture (figure 1.1). Des fonctions permettent d'apposer des signets sur le flux et d'y revenir.

L'approche consistant à dissocier le son sur plusieurs canaux est étendue par Schmandt [SCH 95] pour permettre la diffusion de plusieurs flux audio en même temps. Grâce à l'effet de focalisation de l'attention qui permet d'écouter quelqu'un au milieu d'une foule bruyante, l'utilisateur écoute plusieurs conversations diffusées à différents endroits dans l'espace. Kobayashi [KOB 97] réutilise cette idée pour permettre une navigation : un flux audio est spatialisé pour donner l'impression qu'il tourne autour de l'auditeur ; ce dernier se souvient de la position spatiale liée à un thème et peut y revenir en pointant dans cette direction (figure 1.2). Ainsi, le défaut de localisation temporelle est adoucie par le recours à la localisation spatiale.

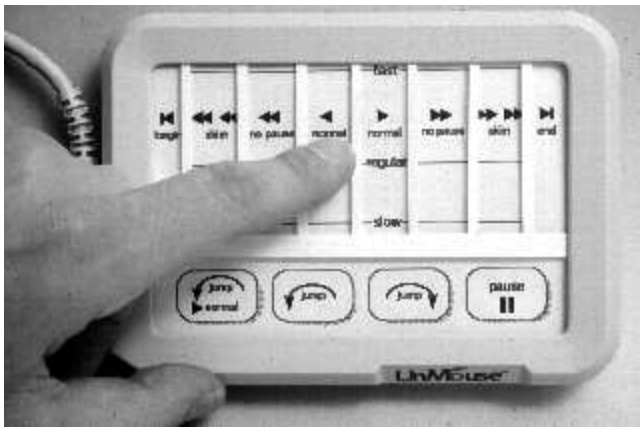

Figure 1.1. *L'interface SpeechSkimmer propose des boutons pour les différentes vitesses de lecture avant et arrière. Source :* `xenia.media.mit.edu/ barons/html/tochi97.html`

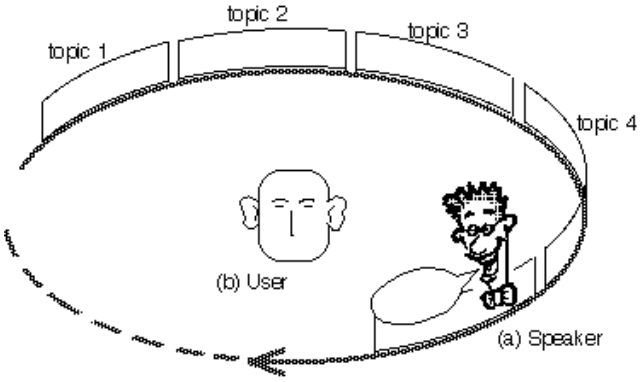

Figure 1.2. *Dynamic soundscapee utilise un ensemble de haut-parleurs pour positionner le son dans l'espace. Ceci permet d'associer, par exemple, un thème et une position dans l'espace. Source :* `www.sigchi.org/chi97/proceedings/paper/kob.htm`

Il y a d'autres travaux qui utilisent une annotation implicite de l'audio dans le cadre de la prise de notes. Whittaker [WHI 94] et Stifelman [STI 96] utilisent la coindexation de l'audio et de l'écriture, donnant accès à l'audio qui a été enregistré au moment où les notes ont été prises. Ainsi, l'utilisateur peut avancer dans le flux audio en tournant les pages du cahier de notes et en pointant les notes elles-mêmes (figure 1.3).

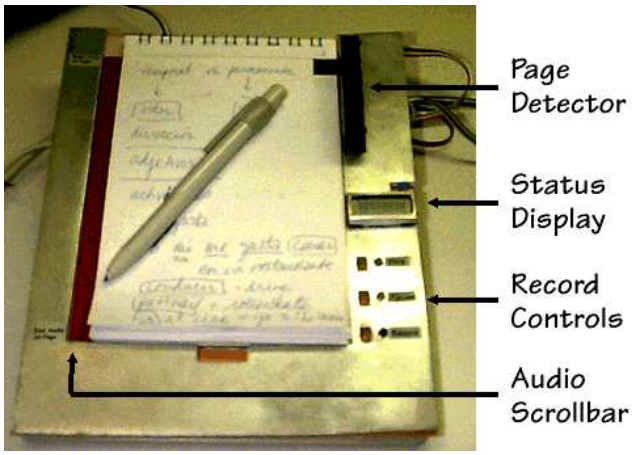

Figure 1.3. *Le son enregistré est synchronisé avec le moment où l'on tourne les pages de façon à pouvoir accéder rapidement à ce qui a été dit lorsque l'on prenait des notes sur une page donnée (Audio Notebook, source :* `web.media.mit.edu/ nitin/msthesis/www/thesis.html`*).*

Ces approches autorisent la navigation à courte distance dans un flux audio. Elles trouvent leur limite lorsque la taille des documents ou la collection de documents deviennent grands. C'est le cas par exemple des archives de l'INA qui totalisent des millions d'heures de documents audio-vidéo.

1.3.2. *Navigation orientée contenu*

Il serait souhaitable d'utiliser les résultats de transcription automatique pour accéder à des documents audio à travers leur contenu. Par exemple, selon [TUC 08], la compression temporelle (accélération de la lecture) est plus fatigante pour un auditeur que la suppression de phrases peu pertinentes. Les utilisateurs préfèrent écouter une élocution normale plutôt que d'avoir à assimiler un rythme de parole plus élevé. En revanche, il est observé par [HE 00] que la transcription exacte d'un document audio est difficile à utiliser car le langage qu'elle contient est relativement différent du texte écrit à cause des hésitations, agrammaticalités et réparations typiques du discours oral. Il est possible de changer le style du discours pour qu'il ressemble plus à du texte écrit, par exemple par traduction automatique de la langue orale vers la langue écrite [HAJ 08]. Malgré l'inadéquation des transcriptions erronées, les auteurs de [WHI 02] rapportent que les utilisateurs de leur interface de gestion de messages vocaux font trop confiance à la transcription : ils n'écoutent guère les données acoustiques, ce qui peut poser problème en cas d'introduction d'erreurs par la transcription

automatique. Malgré un attrait vers l'utilisation de la sémantique par cette transcription, il ne s'agit pas de proposer seulement un document textuel à l'utilisateur, mais de lui apporter les moyens d'en tirer le meilleur parti possible.

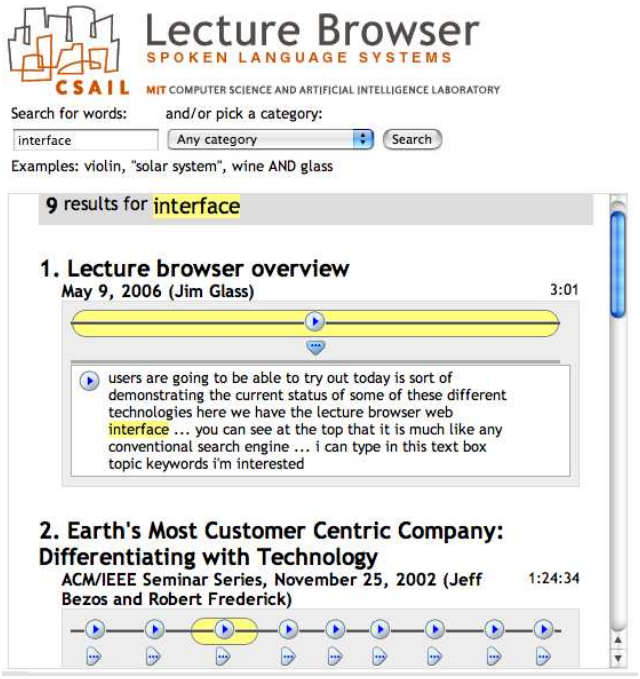

Figure 1.4. *Capture d'écran de l'interface d'accès aux cours archivés du MIT. Cette interface peut être essayée sur* `web.sls.csail.mit.edu/lectures/`.

De nombreuses interfaces d'accès aux enregistrements de cours magistraux ont été développées. Par exemple, le projet *MIT Lecture Browser* [GLA 07] entend fournir aux étudiants un accès aux cours enregistrés au MIT (figure 1.4). Un moteur de recherche web permet de retrouver des segments de cours à partir de leur transcription, puis d'en diffuser la vidéo. Les segments jugés pertinents sont présentés sur une ligne temporelle sous la forme de boutons lecture. Cliquer sur l'un de ces boutons montre la transcription et diffuse la vidéo tout en mettant en évidence les mots prononcés. Une limite de cette interface réside dans la mauvaise qualité des transcriptions automatiques, avec un entre 30 % et 50 %. Il serait souhaitable d'améliorer la qualité de ces transcriptions pour un meilleur confort de lecture lors de l'utilisation de l'interface.

Dans cette optique, [MUN 08] propose une interface d'accès aux enregistrements de présentation supportant l'édition collaborative (figure 1.5). Les utilisateurs du système peuvent corriger les transcriptions afin que d'autres puissent en profiter. Il est

alors possible de prendre en compte les corrections pour améliorer la transcription sur d'autres segments non corrigés, en ajoutant, par exemple, les mots inconnus au modèle de langage du système de transcription.

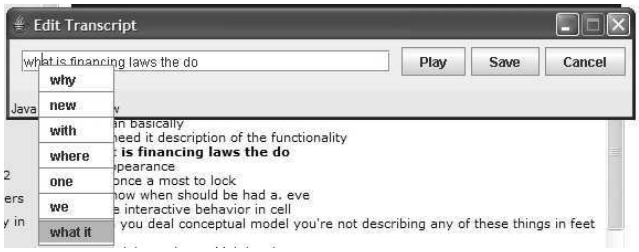

Figure 1.5. *Édition collaborative de transcriptions. Source : [MUN 08].*

Une autre limite de ces interfaces est qu'elles ne donnent pas d'aperçu des résultats : on est obligé de débuter la lecture pour déterminer si un passage est pertinent ou non. Pour y remédier, les travaux de Kong [KON 09] et de Togashi [TOG 08] proposent de résumer les résultats soit en générant des listes de mots-clés représentatifs, soit en sélectionnant des phrases représentatives du contenu des documents.

Malgré les progrès récents en résumé automatique de la parole, l'utilité de tels résumés reste à prouver car ils combinent les erreurs de la méthode de résumé et les erreurs de transcription. Afin de mesurer l'utilité du résumé de parole pour la recherche d'information, Murray *et al.* ont demandé à des utilisateurs de remplir un besoin en information complexe à partir d'enregistrements d'une série de réunions sur un thème donné [MUR 09]. Les utilisateurs devaient, dans un temps limité, retrouver comment les participants aux discussions étaient arrivés à une décision majeure, ce qui n'est pas possible à l'aide d'une recherche de mots-clés. Ils avaient à leur disposition les enregistrements des réunions, les transcriptions automatiques ou manuelles, et des résumés créés automatiquement ou manuellement sur ces transcriptions. Les phrases des résumés étaient liées aux phrases les supportant dans les enregistrements afin de permettre une navigation dans le contenu. Le résultat de cette étude est que les sujets utilisant les résumés par extraction créés à partir de transcriptions automatiques ne sont pas significativement plus mauvais pour remplir le besoin en information que s'ils avaient utilisé des résumés écrits manuellement. Cette expérience justifie donc l'utilisation du résumé automatique pour la navigation.

EXEMPLE.– Exemple de dialogue lié au choix de la couleur d'un produit.

Locuteur 1 : Alors maintenant on doit se décider sur la couleur et la texture de la télécommande.

Locuteur 2 : Euh oui, donc on avait le choix entre du gris et du blanc.

Locuteur 3 : Ah, ces deux couleurs sont beaucoup trop classiques, il nous faut quelque chose d'éclatant : je dis du rose.

Locuteur 1 : Du rose, mais c'est trop voyant !

Locuteur 2 : Il a raison. Notre cible est une clientèle particulière qui recherche justement ce genre d'extravagance.

Locuteur 3 : Vous voyez, du rose, c'est exactement ce qu'il nous faut.

Locuteur 2 : En plus je dois ajouter que ça permettrait de nous démarquer de nos concurrents.

Locuteur 1 : Donc c'est sans équivoque, nous nous décidons pour du rose.

De ce dialogue, on peut déduire qu'il y a eu plusieurs propositions, des arguments pour ou contre chacune de ces propositions et une décision finale.

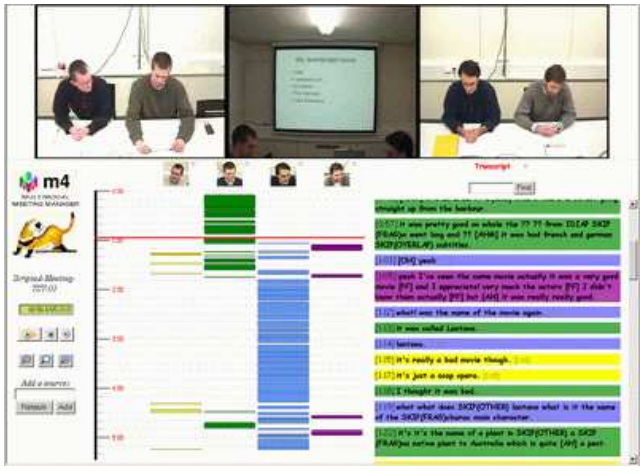

Figure 1.6. *Ferret Browser : visualisation de la transcription, des locuteurs et de la vidéo. Source :* `mmm.idiap.ch/demo/`.

Un problème récurrent dans le traitement de conversations est le fait que les participants se réfèrent à des événements qui ont lieu en dehors de ces conversations ou qui ne sont pas capturés à l'audio. Dans ce cadre, les travaux sur l'accès aux enregistrements de réunions comme le *Ferret meeting browser* [JAI 07] ou le *CALO meeting browser* [TUR 10] tentent d'introduire le contexte des réunions dans la recherche d'information. Ce dernier s'inscrit dans un environnement complet pour le travail collaboratif, offrant les transcriptions de chaque participant, une segmentation en thèmes, la synchronisation avec les notes prises par les participants avec des stylos spécialisés et l'intégration au partage de documents, de courriels, etc. Une telle approche globale à la recherche d'information permet de proposer à l'utilisateur une réponse beaucoup plus complète à son besoin. Toutefois, cette réponse doit être non redondante et focalisée pour qu'il ne se retrouve pas enseveli sous une grande quantité d'information. Un autre problème, récurrent pour l'accès à des données audio, est de fournir un aperçu des données pour guider un utilisateur qui n'aurait pas participé aux réunions.

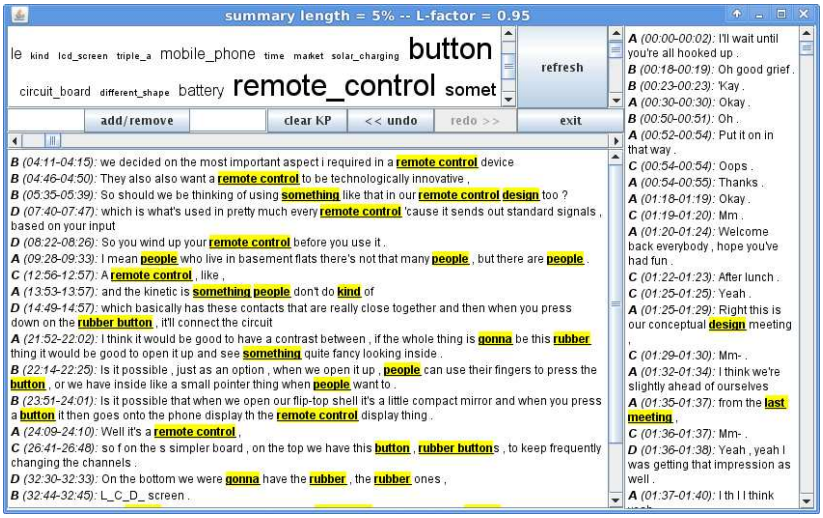

Figure 1.7. *Génération d'un résumé à partir de mots-clés. L'utilisateur peut changer le poids des mots clé, en ajouter et en supprimer. Puis, il peut générer une sélection de phrases les mettant en contexte, présentée dans la partie centrale de l'interface. Des ascenseurs règlent les paramètres de l'algorithme, comme le nombre de phrases sélectionnées. La partie droite montre la transcription du contenu parlé de la réunion.*

Des solutions à ces problèmes sont ébauchées dans [RIE 08], en générant conjointement une liste de mots-clés représentatifs et une sélection de phrases les remettant en contexte. Les mots-clés sont découverts à partir de groupes nominaux fréquents correspondant à des motifs d'étiquettes morpho-syntaxiques. La sélection de phrases est effectuée à l'aide de l'algorithme MMR [GOL 00] dans l'espace des mots-clés, pondéré par leur fréquence. Ceci revient à conserver les phrases qui contiennent le plus de mots-clés de fort poids en omettant les phrases redondantes. L'interface utilisateur, présentée dans la figure 1.7, donne dans un premier temps un ensemble générique de mots-clés et le résumé correspondant, puis l'utilisateur peut ajouter, retirer des mots-clés, et modifier leur poids. La sélection de phrases mise à jour en fonction de ces nouveaux mots-clés peut être très similaire à la sélection précédente, ce qui est montré dans l'interface en grisant les phrases déjà vues par l'utilisateur. En cliquant sur une phrase, celle-ci est remise en contexte dans la transcription de la réunion. On reproche souvent aux méthodes de résumé par sélection la perte du contexte : le fait de pouvoir naviguer jusqu'à la transcription d'origine évite en partie cet inconvénient.

Lorsque le nombre de documents ou la taille des documents augmentent, saisir l'ensemble de l'information pertinente à une requête devient difficile. Cet effet est le même que celui qui affecte les moteurs de recherche web qui montrent des dizaines de pages de résultats. [FAV 07] propose une interface pour accéder à des archives audio

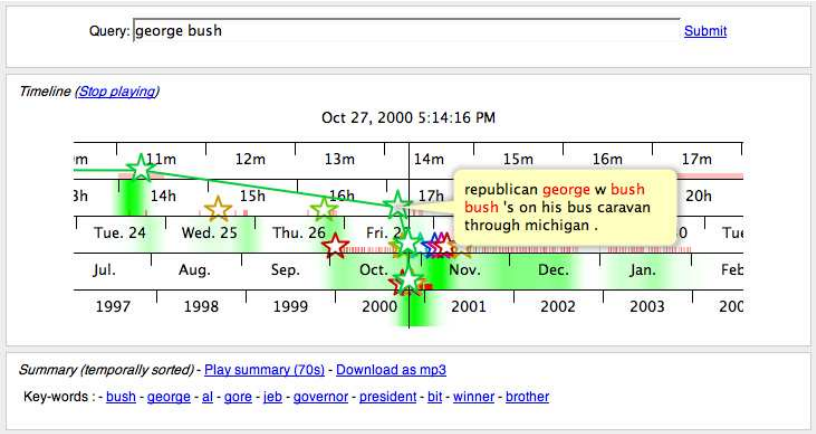

Figure 1.8. *Interface d'accès à une grande base de données audio. La barre de lecture multi échelles facilite la navigation dans le temps. Lorsque l'utilisateur saisit une requête, les réponses à celle-ci sont affichées sous la forme d'une densité d'information sur chaque échelle temporelle, et sous la forme de points d'ancrage, passages les plus représentatifs du contenu pertinent.*

s'étalant sur plusieurs années en situant les résultats dans le temps. Le problème est ici d'obtenir la bonne granularité pour pouvoir naviguer dans le signal audio, à la fois à l'échelle des années (navigation globale) et celle des secondes (navigation locale). La figure 1.8 montre une barre de lecture où sont représentées plusieurs échelles (années, mois, jours, heures, minutes) synchronisées selon un point de référence qui sert de curseur de lecture. Cette barre de lecture permet aussi de représenter les informations intéressant l'utilisateur sur chaque échelle de temps. Deux modes de représentation sont montrés sur cette figure : une densité d'information et des points d'ancrage pour la navigation. La densité est calculée comme le score moyen de pertinence estimé par un modèle de recherche documentaire sur une unité de temps affichée. Les points d'ancrage sont déterminés en appliquant une méthode de résumé automatique pour sélectionner les passages les plus importants tout en étant le moins redondants. Ces points d'ancrage sont représentés dans toutes les échelles à l'aide de marqueurs différents pour en faciliter le repérage.

1.4. Conclusion

Nous avons vu dans ce chapitre un ensemble de méthodes pour faciliter l'accès aux collections de documents audio. La plupart d'entre elles reposent sur une transcription automatique du discours parlé alors que d'autres tentent de localiser l'information importante en analysant directement l'acoustique.

Que ce soit à cause d'une transcription imprécise ou à cause de la forme impropre à la lecture du langage spontané, le meilleur moyen de capturer l'information d'un document audio reste de l'écouter. Afin de servir au mieux l'utilisateur, il faut le guider vers l'information qui l'intéresse, tout en évitant de lui proposer une information à laquelle il a déjà été soumis. Ceci mène naturellement à étudier la navigation et le résumé automatique comme structurants de cette navigation.

Nous avons abordé les aspects spécifiques de ces deux thèmes lorsqu'ils s'appliquent à un contenu parlé. Toutefois, même s'ils contribuent à améliorer l'expérience des utilisateurs face à des documents audio, il reste un certain nombre de sujets pour lesquels il faut continuer les recherches :

– une transcription de meilleure qualité dans toutes les conditions y compris lorsque le champ lexical du domaine n'est pas connu à l'avance et donc un grande partie des mots importants sont inconnus du système de transcription automatique ;

– l'extraction et l'utilisation de descripteurs sémantiques à partir de la parole comme d'une part le résultat de l'extraction d'entités et de relations, ou le résultat de la détection de cadres sémantiques et leurs arguments ;

– la génération d'étiquettes plausibles pour représenter l'acoustique, servant de titre pour un segment de document ou un groupe de plusieurs documents ;

– une modélisation et une représentation des interactions dans les conversations et de ce qui se passe autour de l'enregistrement (que font les participants), similaires à des légendes d'illustrations décrivant les éléments mis en scène ;

– l'intégration avec les données provenant d'autres sources, médias et capteurs, comme les téléphones mobiles des participants capables de capturer des images, mais aussi des informations de géolocalisation.

Ces pistes suggèrent une convergence des différents domaines liés à la recherche d'information, de la fouille de données, de la structuration de l'information, de la navigation et de la modélisation de l'utilisateur, et serviront probablement de manière significative les média non audio comme le texte ou l'image.

1.5. Bibliographie

[AKB 08] AKBACAK M., VERGYRI D., STOLCKE A., « Open-vocabulary spoken term detection using graphone-based hybrid recognition systems », *ICASSP 2008*, 2008.

[ARO 97] ARONS B., « SpeechSkimmer : a system for interactively skimming recorded speech », *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 4, n° 1, p. 3–38, ACM New York, NY, USA, 1997.

[CHR 04] CHRISTENSEN H., KOLLURU B., GOTOH Y., RENALS S., « From Text Summarisation to Style-Specific Summarisation for Broadcast News », *Lecture Notes in Computer Science*, vol. 2997, p. 223–237, Springer, 2004.

[FAV 07] FAVRE B., BONASTRE J.-F., BELLOT P., « An Interactive Timeline for Speech Database Browsing », *Interspeech 2007, Antwerp (Belgium)*, 2007.

[FIS 07] FISCUS J., AJOT J., GAROFOLO J., DODDINGTION G., « Results of the 2006 spoken term detection evaluation », *Searching Spontaneous Conversational Speech*, page 51, 2007.

[FUR 04] FURUI S., KIKUCHI T., SHINNAKA Y., HORI C., « Speech-to-text and speech-to-speech summarization of spontaneous speech », *Speech and Audio Processing, IEEE Transactions on*, vol. 12, n°4, p. 401–408, 2004.

[GAR 00] GAROFOLO J., AUZANNE C., VOORHEES E., « The TREC spoken document retrieval track : A success story », *NIST SPECIAL PUBLICATION SP*, p. 107–130, Citeseer, 2000.

[GIL 09] GILLICK D., RIEDHAMMER K., FAVRE B., HAKKANI-TÜR D., « A Global Optimization Framework for Meeting Summarization », *Proc. ICASSP 2009, Taipeh, Taiwan*, 2009.

[GLA 07] GLASS J., HAZEN T., CYPHERS S., MALIOUTOV I., HUYNH D., BARZILAY R., « Recent progress in the MIT spoken lecture processing project », *Proc. Interspeech*, vol. 3, 2007.

[GOL 00] GOLDSTEIN J., MITTAL V., CARBONELL J., KANTROWITZ M., « Multi-document summarization by sentence extraction », *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, p. 40–48, 2000.

[HAJ 08] HAJIC J., CINKOVÁ S., MIKULOVÁ M., PAJAS P., PTÁCEK J., TOMAN J., URESOVA Z., « PDTSL : An annotated resource for speech reconstruction », *IEEE Spoken Language Technology Workshop, 2008. SLT 2008*, p. 93–96, 2008.

[HE 00] HE L., SANOCKI E., GUPTA A., GRUDIN J., « Comparing presentation summaries : slides vs. reading vs. listening », *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM New York, NY, USA, p. 177–184, 2000.

[HIR 99] HIRSCHBERG J., WHITTAKER S., HINDLE D., PEREIRA F., SINGHAL A., « Finding information in audio : A new paradigm for audio browsing/retrieval », *Proceedings of the ESCA workshop : Accessing information in spoken audio*, p. 117–122, 1999.

[HOR 02] HORI C., FURUI S., MALKIN R., YU H., WAIBEL A., « Automatic Speech Summarization Applied to English Broadcast News Speech », *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 1, IEEE ; 1999, 2002.

[INO 04] INOUE A., MIKAMI T., YAMASHITA Y., « Improvement of Speech Summarization Using Prosodic Information », *Speech Prosody 2004, International Conference*, ISCA, 2004.

[JAI 07] JAIMES A., BOURLARD H., RENALS S., CARLETTA J., « Recording, Indexing, Summarizing, and Accessing Meeting Videos : An Overview of the AMI Project », *Image Analysis and Processing Workshops, 2007. ICIAPW 2007. 14th International Conference on*, p. 59–64, 2007.

[JOH 00] JOHNSON S., JOURLIN P., MOORE G., JONES K., WOODLAND P., « Audio indexing and retrieval of complete broadcast news shows », *Proc. RIAO*, p. 1163–1177, 2000.

[KOB 97] KOBAYASHI M., SCHMANDT C., « Dynamic Soundscape : mapping time to space for audio browsing », *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM New York, NY, USA, p. 194–201, 1997.

[KON 09] KONG S., WU M., LIN C., FU Y., LEE L., « Learning on demand-course lecture distillation by information extraction and semantic structuring for spoken documents », *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing-Volume 00*, IEEE Computer Society, p. 4709–4712, 2009.

[LIU 08] LIU Y., XIE S., « Impact of Automatic Sentence Segmentation on Meeting Summarization », *Proc. ICASSP, Las Vegas, USA*, 2008.

[LIU 09] LIU F., LIU Y., « From Extractive to Abstractive Meeting Summaries : Can It Be Done by Sentence Compression ? », *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore, Association for Computational Linguistics, p. 261–264, August 2009.

[MAM 07] MAMOU J., RAMABHADRAN B., SIOHAN O., « Vocabulary independent spoken term detection », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, page622, 2007.

[MAS 06a] MASKEY S., HIRSCHBERG J., « Summarizing Speech Without Text Using Hidden Markov Models », *Proc. NAACL*, p. 89–92, 2006.

[MAS 06b] MASKEY S., HIRSCHBERG J., « Summarizing speech without text using hidden markov models », *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers on XX*, Association for Computational Linguistics, p. 89–92, 2006.

[MRO 05] MROZINSKI J., WHITTAKER E., CHATAIN P., FURUI S., « Automatic sentence segmentation of speech for automatic summarization », *Proceedings of ICASSP*, n° 51, page 12, 2005.

[MUN 08] MUNTEANU C., BAECKER R., PENN G., « Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts », *CHI '08 : Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, New York, NY, USA, ACM, p. 373–382, 2008.

[MUR 05] MURRAY G., RENALS S., CARLETTA J., « Extractive Summarization of Meeting Recordings », *Ninth European Conference on Speech Communication and Technology*, ISCA, 2005.

[MUR 09] MURRAY G., KLEINBAUER T., POLLER P., BECKER T., RENALS S., KILGOUR J., « Extrinsic summarization evaluation : A decision audit task », *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 6, n°2, p. 1–29, ACM, 2009.

[PEC 07] PECINA P., HOFFMANNOVA P., JONES G., ZHANG Y., OARD D., « Overview of the CLEF-2007 cross-language speech retrieval track », *Working Notes of the CLEF-2007 Evaluation*, Springer, 2007.

[RIE 08] RIEDHAMMER K., FAVRE B., HAKKANI-TÜR D., « A Keyphrase Based Approach to Interactive Meeting Summarization », *Spoken Languge Technologies (SLT), Goa (India)*, 2008.

[SCH 95] SCHMANDT C., MULLINS A., « AudioStreamer : exploiting simultaneity for listening », *Conference on Human Factors in Computing Systems*, ACM New York, NY, USA, p. 218–219, 1995.

[STI 96] STIFELMAN L., « Augmenting real-world objects : A paper-based audio notebook », *Conference on Human Factors in Computing Systems*, ACM New York, NY, USA, p. 199–200, 1996.

[TOG 08] TOGASHI S., NAKAGAWA S., « A Browsing System for Classroom Lecture Speech », *In Proceedings of Interspeech*, 2008.

[TUC 08] TUCKER S., WHITTAKER S., « Temporal compression of speech : An evaluation », *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, n°4, p. 790–796, 2008.

[TUR 10] TUR G., STOLCKE A., VOSS L., DOWDING J., FAVRE B., FERNANDEZ R., FRAMPTON M., FRANDSEN M., FREDERICKSON C., GRACIARENA M., HAKKANI-TÜR D., KINTZING D., LEVEQUE K., MASON S., NIEKRASZ J., PETERS S., PURVER M., RIEDHAMMER K., SHRIBERG E., TIEN J., VERGYRI D., YANG F., « The CALO Meeting Assistant System », *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[VAN 00] VAN THONG J., GODDEAU D., LITVINOVA A., LOGAN B., MORENO P., SWAIN M., « Speechbot : a speech recognition based audio indexing system for the web », *Proc. of the 6th RIAO Conference*, 2000.

[VAR 08] VARGES S., WENG F., PON-BARRY H., « Interactive question answering and constraint relaxation in spoken dialogue systems », *Natural Language Engineering*, vol. 15, n°01, p. 9–30, Cambridge Univ Press, 2008.

[WHI 94] WHITTAKER S., HYLAND P., WILEY M., « Filochat : Handwritten notes provide access to recorded conversations », *Proceedings of the SIGCHI conference on Human factors in computing systems : celebrating interdependence*, ACM New York, NY, USA, p. 271–277, 1994.

[WHI 02] WHITTAKER S., HIRSCHBERG J., AMENTO B., STARK L., BACCHIANI M., ISENHOUR P., STEAD L., ZAMCHICK G., ROSENBERG A., « SCANMail : a voicemail interface that makes speech browsable, readable and searchable », *Proceedings of the SIGCHI conference on Human factors in computing systems : Changing our world, changing ourselves*, ACM, page282, 2002.

[XIE 09] XIE S., HAKKANI-TÜR D., FAVRE B., LIU Y., « Integrating Prosodic Features in Extractive Meeting Summarization », *ASRU*, 2009.

[ZEC 02] ZECHNER K., « Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres », *Computational Linguistics*, vol. 28, n°4, p. 447–485, MIT Press, 2002.

[ZHA 07] ZHANG J., CHAN H., FUNG P., « Improving lecture speech summarization using rhetorical information », *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, p. 195–200, 2007.

[ZHU 06] ZHU X., PENN G., « Utterance-Level Extractive Summarization of Open-Domain Spontaneous Conversations with Rich Features », *Multimedia and Expo, 2006 IEEE International Conference on*, p. 793–796, 2006.

[ZHU 09] ZHU X., PENN G., RUDZICZ F., « Summarizing multiple spoken documents : finding evidence from untranscribed audio », *ACL/IJCNLP, Suntec, Singapore*, p. 549–557, 2009.

Fiche pour le service de fabrication

Auteurs :

Patrice BELLOT

Titre du livre :

Recherche d'information personnalisée

Titre abrégé :

RI personnalisée

Date de cette version :

7 juin 2011

Contact :

– téléphone : 04 92 94 27 48
– télécopie : 04 92 94 28 96
– Mél : rr@unice.fr

Logiciel pour la composition :

– LaTeX, avec la classe `ouvrage-hermes.cls`,
– version 1.3, 17/09/2001.
– traité (option treatise) : Oui *(chapitres avec différents auteurs)*
– livre en anglais (option english) : Non *(par défaut en français)*
– tracé des limites de page (option cropmarks) : Non *(par défaut)*
– suppression des en-têtes de page (option empty) : Non *(par défaut)*
– impression des pages blanches (option allpages) : Oui
– césures actives : voir la coupure du mot signal dans le fichier .log