Accès aux connaissances orales par le résumé automatique

Benoît Favre [*,**] Jean-François Bonastre[**], Patrice Bellot[**], François Capman[*]

[*]Thales, Laboratoire MMP, 160 Bd de Valmy, 92700 Colombes,
benoit.favre@fr.thalesgroup.com,
francois.capman@fr.thalesgroup.com
[**]Université d'Avignon, LIA, 339 Ch des Meinajaries, 84000 Avignon,
benoit.favre@univ-avignon.fr
jean-francois.bonastre@univ-avignon.fr
patrice.bellot@univ-avignon.fr
http://www.lia.univ-avignon.fr

Résumé. Les données audio sont intimement liées au temps et la seule façon d'être sûr d'avoir capturé l'ensemble des informations d'un flux audio est de l'écouter dans son intégralité. Nous proposons dans cet article d'étudier les techniques issues du résumé automatique afin de réduire le temps d'écoute dans une interface d'accès à une base de données parlées. Un moteur de recherche, similaire à ceux utilisés pour naviguer dans le web, est secondé par une hiérarchisation thématique des résultats afin d'affiner les requêtes et une réduction de redondance par résumé automatique de parole. L'accent est mis sur les problématiques posées par le contenu audio dans le cadre spécifique de l'utilisation de données radio-diffusées.

1 Introduction

La parole est le moyen de communication qui nous est le plus naturel, les développements actuels de la téléphonie en sont la preuve. Les dernières avancées technologiques permettent d'aller plus loin que le transfert et l'archivage de la parole : la difficulté est alors de pouvoir capitaliser la mémoire audio de l'entreprise en permettant l'exploitation de ces fonds d'archives orales. Le cycle de vie du contenu audio était constitué jusqu'à présent de la capture, du transfert, de l'archivage et de la diffusion. En y ajoutant la structuration à travers une segmentation, une transcription et des traitements linguistiques, puis l'accès grâce à des moteurs de recherche et des interfaces utilisateurs adaptées, il devient possible de retirer tout le potentiel de la parole.

La grande différence entre les données audio et le texte ou l'image, est l'impossibilité de capturer rapidement le contenu du média. La seule façon d'assimiler toute l'information du message audio est de l'écouter. Nous proposons dans cet article de réduire la quantité d'information à écouter en supprimant la redondance du message et en améliorant la rapidité de l'accès à l'information. Cette réduction est obtenue dans le cadre d'un moteur de recherche sur des données audio à partir de techniques issues du résumé automatique et de la structuration thématique. Nous présentons dans une première partie les approches pour l'accès aux bases de données audio ainsi que quelques techniques issues du résumé automatique. Dans une seconde partie, l'architecture générale du système et le contexte de l'application sont développés, puis

analysés dans une troisième partie. Cette analyse est suivie des conclusions et perspectives de cet article.

2 Travaux similaires

Le lecteur trouvera de bons états de l'art sur les interfaces permettant d'améliorer l'accès au contenu audio, par rapport à une simple lecture, dans (Foote, 1997) et (Koumpis et Renals, 2005). Smoliar et al. (1996) expliquent que les approches à la recherche d'information peuvent être *expressives* (comparaison d'objets physiques, pour retrouver des sons et de la musique) ou *sémantiques* (extraction de descripteurs sémantiques et recherche en langue naturelle). Les travaux présentés dans cet article sont considérés comme appartenant aux approches *sémantiques*. Au delà des classiques moteurs de recherche sur un contenu audio, sont apparues des interfaces complètes de navigation dans un flux incorporant, en sus de l'audio, des informations sur les locuteurs, des vidéos et des métadonnées : Informedia (Hauptmann et Wtbrock, 1997) tire parti des sous-titres pour sélectionner des instants clé, le projet Video Mail Retrieval (Jones et al., 1995) indexe les sous-unités phonétiques pour réduire le problème des mots hors vocabulaire, enfin le Ferret Browser (Wellner et al., 2004) permet d'exploiter des enregistrements de réunions. Ces quelques exemples ont un point commun : quand la quantité de données à exploiter devient grande, il faut trouver de nouvelles approches, telles que le résumé automatique, pour satisfaire l'utilisateur.

Le résumé automatique de texte a été longuement étudié : il demande une analyse sémantique complète qui n'est envisagée actuellement que dans le cas de domaines restreints. En revanche, il a été observé que 70% des phrases d'un résumé apparaissaient au moins en partie dans le texte d'origine (Copeck et Szpakowicz, 2004). Les méthodes en découlant construisent un résumé en sélectionnant un sous ensemble de phrases du texte d'origine. Elles sont formulées par un problème d'optimisation : maximiser la couverture de l'information tout en minimisant sa redondance. Deux types d'approches dominent la sélection de phrases pour le résumé automatique : l'apprentissage des caractéristiques de phrases candidates à l'extraction sur des résumés manuels (Kupiec et al., 1995) et la résolution du problème d'optimisation sur des critères informationnels (Gong et Liu, 2001; Goldstein et al., 2000). Ces techniques ont été appliquées avec succès pour faire des résumé de réunions (Murray et al., 2005). Les hiérarchies thématiques, ou taxonomies, sont une autre forme de résumé mais elles apportent la possibilité d'une interaction rapide avec l'utilisateur. Les taxonomies créées manuellement ont des limites : leur couverture incomplète et leur inadéquation dans des contextes particuliers. Wordnet (Miller, 1995) par exemple, ne contient que très peu de noms propres liés à l'actualité. Les algorithmes capables de créer automatiquement une hiérarchie thématique appliquent généralement les critères du résumé automatique pour choisir les mots les plus fréquents comme étiquettes thématiques (Lawrie et Croft, 2003; Vaithyanathan et Dom, 2000; Kummamuru et al., 2004).

3 Approche

L'approche présentée dans ces travaux a été testée sur les données radiophoniques de la campagne ESTER (Galliano et al., 2005). Constituées d'émissions de radios à des tranches

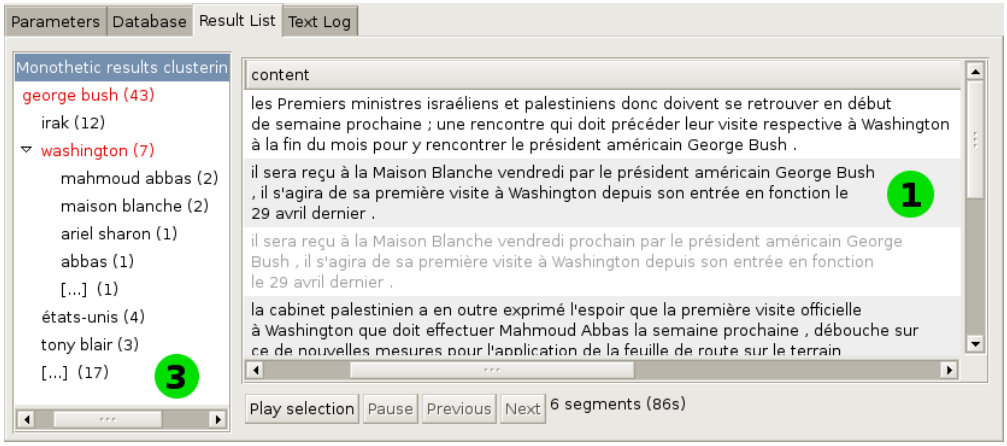

FIG. 1 – *Exemple de requête sur des données structurées manuellement*

horaires où l'information domine (7h-9h de France Inter, France Info, bulletins d'informations de la radio marocaine RTM), ces données présentent nombre de difficultés. Leur structuration a été évaluée lors de la campagne, début 2005. Le flux de parole est structuré grâce à l'extraction de descripteurs : segmentation en classes acoustiques (Fredouille et al., 2004), segmentation en locuteurs (Istrate et al., 2005), transcription de la parole (Nocera et al., 2004), segmentation thématique (basée sur la cohérence lexicale), et reconnaissance d'entités nommées (Favre et al., 2005).

Les descripteurs sémantiques sont exploités par un moteur de recherche implémentant le modèle vectoriel (Buckley et Walz, 1999). Les requêtes sont affinées par une hiérarchie thématique extraite automatiquement des résultats (Kummamuru et al., 2004). L'utilisation de taxonomies permet de résoudre les ambiguïtés des mots de la requête et de réduire le nombre de résultats à écouter. La redondance des résultats restants est supprimée grâce à l'algorithme *Maximal Marginal Relevance* (*MMR*) utilisé en résumé automatique pour la sélection de segments (Goldstein et al., 2000).

4 Implémentation et Analyse

L'application de méthodes issues du résumé automatique pour réduire le nombre de résultats écoutés par l'utilisateur amène plusieurs commentaires : certains phénomènes intrinsèquement liés à la parole affectent le système quelle que soit la qualité de la structuration ; d'un autre côté, la structuration automatique est incertaine et ses erreurs ont un impact pour l'utilisateur. Les phénomènes intrinsèques sont les suivants :

– **Les problèmes d'élocution** (bégaiement, coupures, reprises, pauses remplies). Ils perturbent le contenu et nuisent à la compréhension. Un étiquetage morpho-syntaxique permet de les supprimer de la transcription, mais peut-t-on en faire de même dans le flux audio ? Une solution serait de les minimiser en même temps que la redondance dans la mesure *MMR*.

– **Le manque de grammaticalité de l'oral** et notamment la notion toute relative de phrase. Les méthodes linguistiques crées pour le texte et appliquées à l'audio sont dégradées lorsqu'elles font appel à des éléments grammaticaux.
– **La cohérence des dialogues**. Les tours de parole des locuteurs sont considérés comme une bonne base de segmentation, or il ne faut pas sélectionner une question sans sa réponse.
– **La disparition du contexte** (voir figure 1, point 1). La sélection de segments peut les séparer de leur contexte, ainsi toutes les références (pronoms, titre, ...) à des entités citées précédemment perdent leur sens : les chaînes de coréférences sont brisées. Une résolution des coréférences est la solution appliquée en résumé textuel, mais encore une fois, peut-t-on remplacer un pronom par son référant dans un flux audio ? C'est une des limites du résumé par extraction.
– **L'identité des locuteurs**. Elle joue un rôle différent s'ils prennent parti (les hommes politiques par exemple) ou s'ils présentent une information objective (les journalistes). Dans le domaine d'application particulier des journaux radiodiffusés, les journalistes sont connus à l'avance et le rôle des autres locuteurs peut être déterminé automatiquement (invité, auditeur). L'identité pourra être utilisée pour présenter plusieurs points de vue, mais cela reste un problème ouvert.

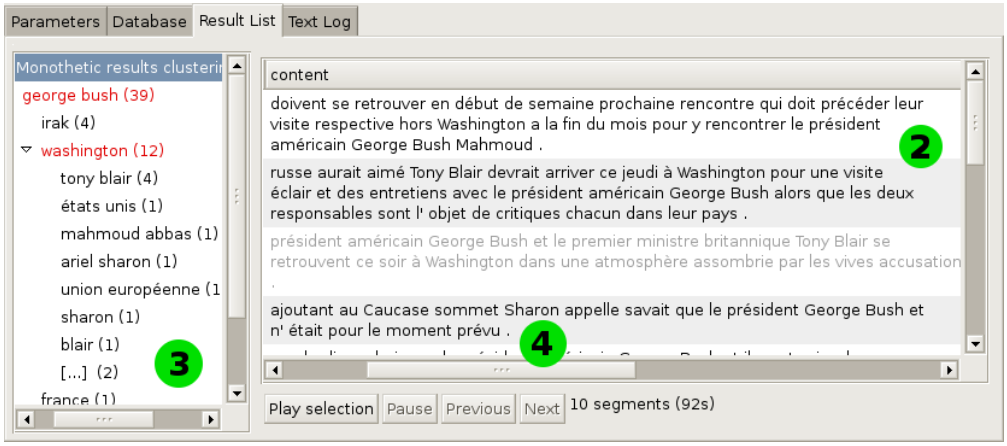

FIG. 2 – *Exemple de requête sur des données structurées automatiquement*

Par contre, chaque phase de la structuration automatique amène des erreurs :
– **Les erreurs de transcription** (voir figure 2, point 4). Le système de transcription a un vocabulaire limité aux mots les plus fréquents. Or, les noms propres appartiennent à une classe ouverte et dynamique qui ne peut être couverte par les modèles de transcription. Mots inconnus, ils sont remplacés par une séquence de mots acoustiquement pertinents mais dépourvus de sens. Heureusement, les erreurs sont considérées comme un bruit qui n'a que peu d'impact sur les analyses fréquentielles caractérisant l'information importante.
– **Les différentes erreurs de segmentation** (classes acoustiques, locuteurs, thèmes). La dégradation n'est pas mesurable directement. Par contre, la segmentation en pseudo-

phrases (unités utilisées par le moteur de recherche) a une grande importance dans le sens ou elle peut générer des coupures dans les composants grammaticaux et prosodiques, impliquant une forte réduction de la cohérence sémantique à l'écoute du contenu (voir figure 2, point 2).

– **Les concepts utilisés** (comparer figure 1 et 2, point 3). Les entités nommées sont très adaptées au domaine d'application, notamment pour la génération de hiérarchies thématiques afin d'affiner les résultats ; il est évident que la qualité de leur extraction est essentielle à l'exploitation des données. Enrichir les documents sur des corpus textuels permet de rattraper une partie des erreurs en tirant parti des co-occurrences.

5 Conclusions et perspectives

Après avoir détaillé les différents moyens d'accéder à un contenu audio, nous avons vu quelles techniques issues du résumé automatique pouvaient aider à améliorer un moteur de recherche sur des données orales. L'approche permet à la fois d'affiner les requêtes des utilisateurs et de réduire la redondance des résultats qui leur sont présentés. Une analyse de l'impact d'un contenu constitué d'émissions radio a mis au jour un certain nombre de problématiques et donné des pistes pour les résoudre. Plus précisément, il faut en priorité résoudre le problème des coréférences non résolues et établir une segmentation en phrases utile à notre application. Cela apportera les performances nécessaires avant de mettre en place une évaluation complète auprès d'utilisateurs pour caractériser la dégradation des performances provoquées par les systèmes automatiques. Deux finalités concrètes de ces recherches sont la création automatique de flash d'information radio en poussant plus loin le concept de résumé automatique de parole et l'intégration de l'approche à un système de dialogue pour constituer un moteur de recherche entièrement vocal.

Références

Buckley, C. et J. Walz (1999). Smart in trec 8.

Copeck, T. et S. Szpakowicz (2004). Vocabulary agreement among model summaries and source documents. In *DUC 2004 Workshop*.

Favre, B., F. Béchet, et P. Nocéra (2005). Robust named entity extraction from large spoken archives. In *HLT-EMNLP'05*.

Foote, J. (1997). An overview of audio information retrieval.

Fredouille, C., D. Matrouf, G. Linares, et P. Nocera (2004). Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ester. In *JEP'04*.

Galliano, S., E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier (2005). The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Proc. Eurospeech'05*.

Goldstein, J., V. Mittal, J. Carbonell, et J. Callan (2000). Creating and evaluation multi-document sentence extract summaries. In *CIKM 2000 - ACM, McLean, VA USA*.

Gong, Y. et X. Liu (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proc. ACM SIGIR*, pp. 19–25.

Hauptmann, A. et M. Wtbrock (1997). Informedia : News-on-demand. multimedia information acquisition and retrieval. In *Intelligent Multimedia Information Retrieval*.

Istrate, D., N. Scheffer, C. Fredouille, et J.-F. Bonastre (2005). Broadcast news speaker tracking for ester 2005 campaign. In *Eurospeech'05*.

Jones, G. J. F., J. T. Foote, K. S. Jones, et S. J. Young (1995). Video mail retrieval : the effect of word spotting accuracy on precision. In *ICASSP'95*, Volume 1, pp. 309–312.

Koumpis, K. et S. Renals (2005). Content-based access to spoken audio.

Kummamuru, K., R. Lotlikar, S. Roy, K. Signal, et R. Krishnapuram (2004). A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *ACM WWW'04*.

Kupiec, J., J. Pederson, et F. Chen (1995). A trainable document summarizer. In *ACM SIGIR'95*, pp. 68–73.

Lawrie, D. et W. Croft (2003). Generating hierarchical summaries for web searches. In *Proc. of SIGIR*, pp. 457–458.

Miller, G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM 38*(11), 39–41.

Murray, G., S. Renals, et J. Carletta (2005). Extractive summarization of meeting recordings. In *Proc. Interspeech*.

Nocera, P., C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, et F. Béchet (2004). The LIA's french broadcast news transcription system. In *SWIM : Lectures by Masters in Speech Processing, Maui, Hawaii*.

Smoliar, S. W., J. D. Baker, T. Nakayama, et L. Wilcox (1996). Multimedia search : An authoring perspective. In *Proceedings of the First International Workshop on Image Databases and Multimedia Search*, pp. 1–8.

Vaithyanathan, S. et B. Dom (2000). Model-based hierarchical clustering. In *Proc. of Uncertainty in Artificial Intelligence*, pp. 599–608.

Wellner, P. D., M. Flynn, et M. Guillemot (2004). Browsing recordings of multi-party interactions in ambient intelligence environments. In *Proc. CHI Workshop Lost in Ambient Intelligence ?*

Summary

Spoken data are bound to time and the only way to capture an audio stream content in its integrality is to listen to it. We propose in this article to study approches used in automatic summarization in order to reduce listening time in spoken archives access interfaces. It is achieved in the context of an audio search engine using a topic hierarchy extracted from results and an anti-redondancy algorithm. The problematics of spoken content are analyzed in the specific domain of broadcast news.