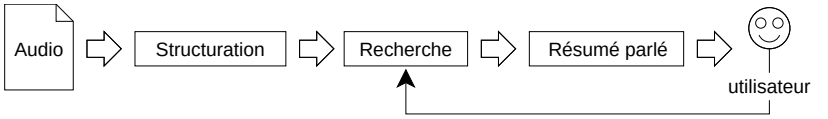
Accès aux connaissances orales par le résumé automatique

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

[*]Thales, Laboratoire MMP, 160 Bd de Valmy, 92700 Colombes,
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

Le temps nécessaire pour écouter un flux audio est un facteur réduisant l'accès efficace à de grandes archives de parole. Une première approche, la structuration automatique des données, permet d'utiliser un moteur de recherche pour cibler plus rapidement l'information. Les listes de résultats générées sont longues dans un souci d'exhaustivité. Alors que pour des documents textuels, un coup d'oeil discrimine un résultat interessant d'un résultat non pertinant, il faut écouter l'audio dans son intégralité pour en capturer le contenu. Nous proposons donc d'utiliser le résumé automatique afin de structurer les résultats des recherches et d'en réduire la redondance.



Les données radiophoniques exploitées pour cette approche sont issues de la campagne ESTER (Galliano et al., 2005), évaluatrice de la structuration automatique d'émissions et de bulletins à caractère informatif. Le processus de structuration de notre système est le suivant : 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 (Sitbon et Bellot, 2004), et reconnaissance d'entités nommées (Favre et al., 2005). Grâce à cette structuration, un moteur de recherche basé sur le modèle vectoriel permet de présenter à l'utilisateur la liste des segments correspondant à son besoin en information.

Fondé sur l'observation que 70% des phrases d'un résumé écrit manuellement proviennent des textes d'origines, le résumé par extraction est l'approche la plus utilisée actuellement en domaine ouvert pour le texte. En prenant pour hypothèse que cette observation est similaire pour la parole (les titres des journaux radiodiffusés), nous l'appliquons à la fois pour extraire des étiquettes thématiques structurant hiérarchiquement les résultats et pour extraire les segments les plus représentatifs du contenu des résultats.

L'algorithme *Maximal Marginal Relevance* (MMR), proposé par (Goldstein et al., 2000) pour sélectionner les segments maximisant la couverture en information tout en minimisant sa redondance, peut être appliqué pour sélectionner des mots-clés comme étiquettes thématiques dont on obtient une hiérarchie en faisant varier la granularité. Le critère de sélection par gain en