Analyse du corpus MATRICE-INA *: exploration et classification automatique d'archives audiovisuelles de 1930 à 2012

Antoine Laurent(2) Camille Guinaudeau(1,2) Anindya Roy(2)
(1) Université Paris-Sud, Rue du Château, 91400 Orsay
(2) LIMSI-CNRS, Rue John Von Neumann, 91400 Orsay
laurent@limsi.fr, guinaudeau@limsi.fr, roy@limsi.fr

RÉSUMÉ

Cet article décrit les méthodes mises en place pour permettre l'analyse d'un corpus composé de documents audiovisuels diffusés au cours des 80 dernières années : le corpus MATRICE-INA. Nous proposons une exploration des données permettant de mettre en évidence les différents thèmes et évènements abordés dans le corpus. Cette exploration consiste dans un premier temps à effectuer une analyse temporelle sur les notices documentaires produites manuellement par les documentalistes de l'Institut National de l'Audiovisuel et sur les transcriptions automatiques des documents. Puis, nous montrons, grâce à une technique de clustering automatique, que les transcriptions automatiques permettent également d'effectuer une analyse du corpus faisant émerger des thèmes cohérents avec les données traitées.

ABSTRACT

Analysis of the MATRICE corpus : exploration and automatic clustering of audiovisual archives from 1930 to 2012

This paper presents different approaches developed for the analysis of an audiovisual collection of documents broadcasted during the last 80 years, namely the MATRICE-INA corpus. We propose an exploration of the documents to highlight the different topics and events discussed in the corpus. This exploration consists first in a temporal analysis carried out on documentary records manually produced by the archivists of the French National Audiovisual Institute and on automatic transcripts of the documents. Then, we show, through an automatic clustering technique, that the automatic transcripts also allow an accurate topics detection that seems to be coherent.

MOTS-CLÉS : Fouille de données, clustering, transcription automatique de la parole, document audiovisuel.

KEYWORDS: Data mining, clustering, automatic speech recognition, audiovisual document.

1 Introduction

L'Institut National de l'Audiovisuel (INA), créé en 1974, est responsable du dépôt légal de la radio et de la télévision en France depuis la loi du 20 juin 1992. Ce travail donne lieu à une quantité de données très importante qui rend possible la mise en place de différentes études permettant

*. Ce projet bénéficie d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'investissements d'avenir EQUIPEX MATRICE.

notamment d'analyser la façon dont ont été traités des évènements historiques importants mais également les évolutions possibles des discours associés à ces évènements. Dans ce cadre, le projet transdiciplinaire MATRICE, regroupant 24 partenaires, a été mis en place dans l'objectif d'une part, de repérer et d'analyser les grands récits nationaux sur deux grands événements (la seconde guerre mondiale et les attentats du *World Trade Center* à New York le 11 septembre 2001) et d'autre part d'étudier les comportements des publics de mémoriaux.

Afin d'analyser le corpus MATRICE-INA, l'une des possibilités consiste à étudier les notices documentaires associées à chacun des documents par les documentalistes de l'INA. Cependant ces notices ne sont, d'une part, pas toujours disponibles [1] et ne contiennent, d'autre part, pas toujours d'informations sémantiques exploitables. Certaines d'entre elles ne contiennent, en effet, que des informations sur la structure de montage du document, par exemple « Plan large sur le général de Gaulle » ou « Panoramique sur le maréchal Pétain », *etc.* Pour pallier ces deux difficultés, il semble naturel de chercher à employer les transcriptions automatiques de la parole. Cette solution, cependant, reste problématique car le modèle de langue et le modèle acoustique du système de reconnaissance automatique de la parole employé ne sont pas toujours adaptés aux données à transcrire qui s'étalent sur 80 ans (évolution de la qualité des enregistrements, évolution linguistique en terme de discours et de diction, *etc.*). L'objectif de cet article consiste donc à présenter les premières expérimentations effectuées afin d'estimer la faisabilité d'une analyse basée uniquement sur des transcriptions automatiques de la parole prononcée dans des archives portant sur les huit dernières décennies.

Dans cet article, nous présentons tout d'abord le corpus MATRICE-INA ainsi que l'approche employée lors de la transcription automatique des documents qui le composent. Dans un second temps, nous proposons une analyse temporelle du corpus permettant la mise en évidence d'évènements marquants dans les documents étudiés. Finalement, nous montrons dans une troisième partie qu'il est possible de fournir une analyse pertinente des archives du projet en se basant uniquement sur les transcriptions automatiques et en utilisant des techniques de clustering automatique. Nous montrons notamment que les documents audiovisuels fournis dans le cadre du projet abordent des thèmes plus variés que ceux originellement prévus.

2 Le corpus MATRICE-INA

Le corpus, qui a terme doit contenir environ 100 000 documents, est composé pour l'instant de 30 000 documents audiovisuels. L'écoute d'une partie des documents (sélectionnée de façon aléatoire), nous a permis de constater que certains ne contenaient pas de parole francophone. Nous avons donc utilisé un système de détection automatique des langues sur chaque document et conservé ceux détectés comme ne contenant que du français, soit 23 000 documents. Ces 23 000 documents ont pour thème la seconde guerre mondiale et les attentats du 11 septembre 2001. Par ailleurs, tous les documents audiovisuels du corpus ne contiennent pas de la parole (présence de films muets). Le tableau 1 fournit des détails sur le corpus et montre notamment que la majorité des données étudiées dans le cadre du projet a été diffusée avant 1995. Pour chacun de ces 23 000 documents, il nous a été fournit à la fois des notices documentaires manuelles – composées d'une partie description et d'une liste de mots-clés appelés *descripteurs* – et l'audio correspondant aux canaux gauches et droits. En effet, à l'époque où les vidéos les plus anciennes ont été enregistrées, elles ne disposaient pas de timecode permettant de naviguer à l'intérieur

1. Le corpus actuel ne contient pour l'instant aucune notice traitant des attentats du 11 septembre 2001.

TABLE 1: Détails sur le corpus traité

	Durée	Durée de parole	Nombre de mots	Nombre de mots distincts
Avant 1995	10840h	8274h	96 millions	168k
Après 1995	2814h	2396h	29 millions	124k
Total	13654h	10670h	124 millions	179k

de l'enregistrement. Pour pallier cela, l'un des canaux servait à enregistrer l'horloge parlante, ce qui a nécéssité la mise en place de techniques adaptées pour transcrire efficacement le corpus MATRICE-INA. Dans cette section nous présentons donc, dans un premier temps, le système de reconnaissance automatique de la parole employé dans le cadre du projet puis nous décrivons les particularités de nos données ainsi que les stratégies mises en place pour gérer les spécificités de notre corpus.

2.1 Transcription automatique de la parole

Le travail du LIMSI dans ce projet consiste à adapter son Système de Reconnaissance Automatique de la Parole (SRAP) pour améliorer les transcriptions de ces 80 années d'enregistrements, réalisés dans des conditions très hétérogènes en terme de vocabulaire et d'acoustique.

Dans cette étude préliminaire, nous avons utilisé tel quel le SRAP du LIMSI (développé conjointement par le LIMSI et la société Vocapia [2]) pour obtenir un premier jeu de transcriptions. Ne disposant pas encore d'un corpus de développement exploitable, il nous est en effet impossible de procéder à l'adaptation des différents modèles.

Le système est formé d'un partitioneur parole/non-parole et d'un décodeur de mots. Le décodeur de mots utilise un modèle acoustique HMM et un modèle de langage bi-gramme pour construire un treillis de mots qui est ensuite réévalué à l'aide d'un modèle de langue 4-gram. L'apprentissage des modèles acoustiques est réalisé en alignant une transcription orthographique exacte sur le signal de parole au moyen d'un jeu de modèles phonétiques et d'un lexique de prononciation. Ces modèles sont entraînés à partir de paramètres MLP (Multi-Layer Perceptron) concaténés avec des paramètres cepstraux standards (Perceptual Linear Prediction – PLP) et la fréquence fondamentale (F0). Ils ont été estimés à l'aide d'un apprentissage discriminant (MMIE) avec un apprentissage adaptatif aux locuteurs (SAT). Une adaptation par blocs diagonaux a été utilisée pour les paramètres MLP et PLP+F0. Le vecteurs de paramètres cepstraux est composé de 81 paramètres ($39MLP + 39PLP + F0 + \Delta F0 + \Delta\Delta F0$). Les modèles acoustiques sont des HMMs (Hidden Markov Models) composés d'un mélange de gaussiennes. Ils sont à états liés, dépendants du contexte et du sexe du locuteur. Le silence est modélisé par un seul état composé de 1024 gaussiennes. Le corpus d'apprentissage est composé de 866 heures de paroles, à partir desquelles 19 679 modèles contextuels ont été entraînés, partageant 11 517 états, pour un total de 370k gaussiennes. Les modèles de langue génériques ont été estimés sur un corpus d'environ 2.9 milliards de mots. Les modèles sont construits autour d'un vocabulaire de 200K mots. Les prononciations du dictionnaire sont obtenues par une phonétisation automatique à base de règles.

2.2 Quantification des données exploitables

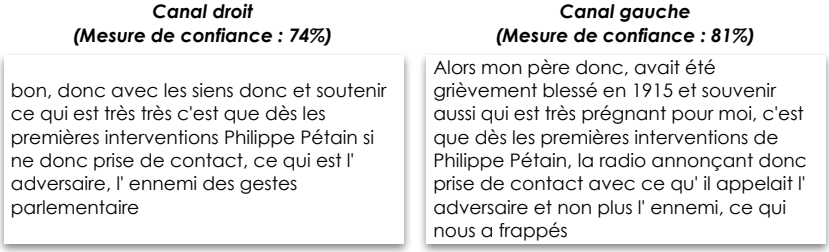
À partir de la sortie du système de reconnaissance automatique de la parole, nous avons cherché à éliminer le canal contenant l'enregistrement de l'horloge parlante. Pour ce faire, une recherche de

2. http ://www.vocapia.com/

TABLE 2: Nombre de documents par canal (G=Gauche, D=Droit)

	G et D identiques	Canal G et D différents		G ou D vide	G et D vide	G et D horloge
		Horloge parlante	Autres			
Nombre	19871	3853	2479	366	2638	6

FIGURE 1: Différence de transcription entre canal droit et canal gauche



différents mots clés dans la transcription automatique (par exemple « au top il sera exactement ») et une analyse des durées des segments de parole a été réalisée (les segments des fichiers contenant l'horloge parlante durent entre 3 et 4 secondes).

Lorsque l'audio est enregistré sur les deux canaux, il existe pour certains documents une importante différence de qualité d'enregistrement entre le canal droit et le canal gauche. Dans ce cas nous avons considéré le canal permettant de maximiser la mesure de confiance du système de reconnaissance automatique de la parole. Pour chacun des documents un taux d'erreur mot a été calculé entre le canal droit et le canal gauche pour déterminer s'ils étaient identiques ou non. La figure 1 présente un exemple où la transcription sur les deux canaux est différente. Il s'agit d'un extrait du décodage de la même zone audio sur les deux canaux. Il apparaît clairement que la transcription sur le canal gauche est plus proche de ce qui a réellement été prononcé. A l'écoute, le canal droit est vraiment très bruité. En tout, 237h d'audio (films muets, horloge parlante sur les deux canaux et documents pas en Français) ne sont pas utilisables. Le nombre de documents correspondant à chaque configuration est présenté dans le tableau 2.

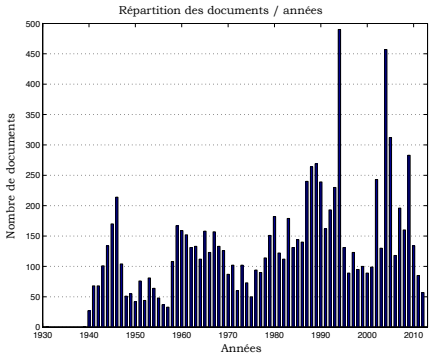
Un système de détection de la langue (LID) (BenZeghiba *et al.*, 2012) a été utilisé pour détecter les enregistrements contenant des données non francophones. En tout nous avons détecté 7 langues différentes en plus du Français.

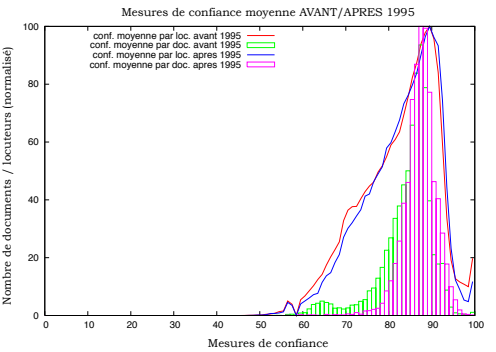
La figure 2 présente la mesure de confiance moyenne des transcriptions par documents et par locuteur pour les enregistrements d'avant et après 1995[3]. On constate que la plupart des mesures de confiance se situent aux alentours de 85% pour les données d'avant et après 1995, le nombre de documents ayant des mesures de confiance faible étant cependant plus important pour les enregistrements d'avant 1995. Cette observation s'explique par le fait que les conditions acoustiques des enregistrements et le vocabulaire utilisé sont plus éloignés des données d'apprentissage du SRAP du LIMSI. Cette figure présente également le nombre de documents dont nous disposons pour chaque année.

Finalement, le lien entre les notices documentaires et les documents audiovisuels est souvent

3. La normalisation a été faite en ramenant à 100 le nombre maximum de documents/locuteurs rencontrés pour une classe de mesure de confiance.

FIGURE 2: Mesures de confiance moyenne avant/après 1995, répartition des documents par années





difficile à faire car les notices ne contiennent pas toujours le nom du fichier audiovisuel qu'elles décrivent. De plus, bien que nous ayons le même nombre de documents audio et de notices, aucune notice ne semble être associée aux documents abordant le thème du 11 septembre, alors même qu'il existe des documents abordant ce thème dans notre corpus. Pour ces raisons, nous avons choisi de nous concentrer, dans la suite de cet article, sur un sous-corpus composé uniquement de documents portant sur la seconde guerre mondiale et pour lesquels chaque transcription est associée de manière certaine à une notice (même titre de programme et d'émission et même date de diffusion)

Sur ce sous-corpus, composé de 12 592 documents, nous cherchons à évaluer dans quelle mesure les transcriptions automatiques imparfaites peuvent nous permettre de manipuler le corpus MATRICE-INA pour en faire ressortir des informations intéressantes. Pour cela, nous étudions dans un premier temps (Section 3) la qualité de l'analyse temporelle effectuée à partir des transcriptions automatiques, en comparant les mots apparaissant les plus fréquemment dans les documents traitant du général de Gaulle avec ceux obtenus à partir des notices documentaires. Puis nous estimons dans un second temps, la capacité de techniques de clustering appliquées sur les transcriptions automatiques à retrouver des thèmes extraits manuellement à partir des notices documentaires de qualité (*cf.* Section 4).

3 Analyse temporelle des documents

Dans cette partie, nous analysons les cooccurrences de termes les plus fréquentes au sein d'un thème particulier. Nous cherchons notamment à retrouver le vocabulaire couramment associé au Général de Gaulle afin d'évaluer si les transcriptions permettent de faire ressortir les mêmes évènements marquant que ceux observés grâce aux notices documentaires.

Pour mener à bien cette comparaison, nous sélectionnons tout d'abord dans notre corpus uniquement les couples notices-transcriptions pour lesquels les notices contiennent le descripteur *de Gaulle* et nous créons pour chaque type de données cinq sous-corpus correspondant aux cinq décennies entre la fin des années 1950 et les années 2000. Puis, nous traitons de la même façon le contenu des notices (c'est-à-dire la description du document, privée des descripteurs) et les transcriptions. Le contenu des notices et les transcriptions sont lemmatisés grâce à l'outil *TreeTagger*(Schmid, 1994) et une liste de *stop words* est définie afin de supprimer les termes

TABLE 3: Cooccurrences par décénnie

Notices					Transcriptions				
60s	70s	80s	90s	00s	60s	70s	80s	90s	00s
visite	intérieur	1er	ensemble	reportage	traverser	forces	épisode	madame	recherche
foule	vie	estaing	image	presse	auteur	cercueil	basque	écrire	milan
algérie	pompidou	différent	porter	mairie	nord	découvrir	bâton	fédéral	dénicher
arrivé	dernier	plan	président	églises	parlement	exister	formidable	église	oiseau
madame	chaban	libération	rue	csa	ouverture	préfet	larme	reprocher	malade
chrono	paris	foule	jour	geneviève	carrière	main	pension	recours	gueule
enfant	jean	appel	colombey	déplorer	ligne	taille	allemagne	anticyclone	76e
ville	président	image	rpr	académicien	algériens	valable	lecture	explosion	patrimonial
homme	cimetière	archives	jacques	consécration	discuter	littérature	ramener	cellule	hastings
gaulle	colombey	anniversaire	émission	tenir	puissant	génie	grève	défenseur	laure

vide de sens qui peuvent apparaître dans notre collection de documents. Certains de ces *stop words* sont très spécifiques à nos données, c'est le cas par exemple des termes « pano », « gp », « ps » (abbréviations respectives de « panoramique », « gros plan » et « plan serré »), présents dans les notices et employés pour décrire la structure de montage du document. Chaque mot des transcriptions et des notices est ensuite associé à son score *tf-idf*[4] et les 10 termes associés aux valeurs *tf-idf* les plus élevées sont conservés.

Le tableau 3 liste pour chaque décennie les termes les plus représentatifs des documents traitant du général de Gaulle. On peut voir dans ce tableau que les mots clés retrouvés pour les notices et pour les transcriptions sont à peu près équivalents pour les années 1960 et 1970. Pour les années 1960, le thème de la guerre d'Algérie est présent à la fois dans les notices, avec le mot « algérie », et dans les transcriptions, avec le mot « algérien ». De la même manière, on voit apparaître le thème de la mort du général de Gaulle dans la décénnie 70 à la fois grâce aux notices (« cimetière », « colombey ») et grâce aux transcriptions (« cercueil »). Cependant, à partir des années 1980, on peut constater que les mots clés obtenus grâce aux transcriptions sont plus flous et il devient moins facile de faire le lien entre ces mots clés et le général de Gaulle alors que ceux extraits des notices gardent un lien avec la politique et le décès du général (« églises » dans les années 2000 fait ainsi référence à Colombey les deux églises, lieu de décès de Charles de Gaulle). Selon nous, cette observation peut s'expliquer par le fait que les documents audiovisuels fournis par l'INA ne sont peut être pas thématiquement homogènes et qu'ils abordent probablement d'autres thèmes que celui de la seconde guerre mondiale. Afin de valider cette hypothèse, nous effectuons dans la section suivante un clustering automatique des transcriptions des documents afin de faire émerger les thèmes les plus importants de notre corpus.

4 Détection automatique de thèmes

Afin de faire émerger les thèmes de notre corpus, nous appliquons des méthodes de clustering automatique sur le corpus composé des 12 592 couples notice/document abordant tous les thèmes de la seconde guerre mondiale.

4. La notation *tf-idf* fait référence à la multiplication de la valeur de la fréquence d'apparition d'un terme dans un document, *tf*, par l'inverse de sa fréquence d'apparition dans l'ensemble des documents, *idf*. Ici, la valeur de l'*idf* est estimée sur chaque sous-corpus (transcriptions *vs* notices) séparemment.

Dans un premier temps, nous utilisons l'algorithme *Mean Shift* – proposé dans (Fukunaga et Hostetler, 1975) et qui ne nécessite pas de connaissances a priori sur le nombre de clusters à définir – afin de détecter automatiquement le nombre de thèmes présents dans notre corpus. Appliqué sur les notices documentaires, il nous permet de fixer à 6 le nombre de thèmes principaux présents dans nos données.

Dans un second temps, afin de caractériser automatiquement les différents thèmes présents dans les transcriptions du corpus, nous avons employé une factorisation de matrices non négatives NMF (Lee et Seung, 2001) qui consiste à factoriser une matrice V en deux sous-matrices W et H telles que les trois sous matrices soient composées uniquement d'éléments positifs (ou égaux à 0). Cette technique, couramment utilisée dans l'extraction automatique d'informations à partir de données textuelles mais également dans le contexte du regroupement de documents en thèmes (Shahnaz *et al.*, 2006), permet, grâce au principe de non-négativité, de manipuler des matrices plus faciles à analyser. En effet, dans le cadre de la NMF, le nombre de colonnes de la matrice W et le nombre de lignes de H sont choisis de façon à ce que le produit WH soit une approximation de V. L'obtention de matrices W et H de plus petite taille que V les rend plus facilement manipulables et analysables. Cela permet également de représenter les éléments de W par une quantité de données beaucoup plus faible.

Dans notre cas, la matrice V, de taille N (égal au nombre de documents) $\times$ D (correspondant au nombre de lemmes considérés dans notre corpus), contient les valeurs des scores *tf-idf* de chacun des lemmes dans les documents. La matrice W, de taille d – le nombre de thèmes, fixé à priori à 6 dans notre cas – $\times$ D, va quant à elle contenir des valeurs non négatives traduisant l'appartenance de chaque lemme à un thème. C'est cette matrice qui nous permet de fournir la caractérisation des thèmes présentée dans le tableau 4. Finalement, la matrice H, de taille $N \times d$ contient, pour chaque document, les thèmes activés.

Cette approche nous permet d'extraire automatiquement du corpus les 6 thèmes reportés dans le tableau 4. Chaque thème est caractérisé par ses 5 mots les plus représentatifs grâce à la matrice W. Dans la partie droite du tableau, nous voyons apparaitre les thèmes et les caractérisations extraits des notices. Ces thèmes sont majoritairement corrélés avec la thématique globale du corpus. Seul le thème **B** diffère du thème de la seconde guerre. Il est très spécifique au corpus que constituent les notices documentaires et correspond aux notices qui décrivent précisément la structure de montage des émissions. Concernant le clustering effectué sur les transcriptions automatiques des documents audiovisuels, 4 thèmes sur 6 semblent aborder précisèment la seconde guerre mondiale, à travers le débarquement (thème **2**), le général de Gaulle (thème **4**), le procès de Maurice Papon (thème **5**) et l'holocauste (thème **6**). Les caractérisations proposées pour les 2 autres thèmes sont cependant plus floues et semblent aborder des thèmes plus variés. Cette différence entre notices et transcriptions peut être liée, d'une part, au fait que les notices, plus synthétiques, ne contiennent que des élèments en rapport avec le thème de la seconde guerre mondiale. D'autre part, nous pensons, comme nous l'avons mentionné dans la section 3, que les documents audiovisuels transcrits ne sont pas thématiquement homogènes.

Afin de vérifier la présence d'autres thèmes dans les documents transcrits, nous avons effectué un second clustering grâce à la NMF en fixant le nombre de thèmes à 20 afin d'étudier les caractérisations obtenues et présentées dans le tableau 5. Grâce à ce tableau, nous voyons apparaitre des thèmes relatifs à la seconde guerre mondiale, tel que les thèmes numéro **5**, **6** ou **13**, mais également des thèmes moins corrélés avec la thématique globale du corpus. Les thèmes numéro **8**, **14** et **15**, par exemple, semblent respectivement liés au football, à la météo et à la

TABLE 4: Caractérisation des 6 thèmes détectés automatiquement dans les notices et les transcriptions automatiques correspondantes

Transcriptions	**Notices**
1 - jean pourcent paris pays foi	**A** - gaulle général foule discours pompidou
2 - débarquement juin plage soldat vétéran	**B** - plan différent ensemble interview blanc
3 - gens penser livre vie film	**C** - ficher chrono fichier détectage juif
4 - gaulle général politique président république	**D** - mitterrand françois chirac cérémonie gerbe
5 - papon maurice procès avocat prison	**E** - soldat allemand débarquement char avion
6 - juif camp nazi guerre barbie	**F** - jean barbie interview guerre film

TABLE 5: Caractérisation des 20 thèmes détectés automatiquement dans les transcriptions

1 - problème politique gouvernement pays question	**11** - barbie klaus procès moulin lyon
2 - livre penser gens film écrire vie	**12** - famille passer maison femme arriver
3 - gaulle général colombey juin charles	**13** - résistance guerre paris allemands général
4 - président république ministre mitterrand état	**14** - paris pourcent ouest température nuage
5 - papon maurice procès avocat bordeaux	**15** - église pape catholique jean paul vatican
6 - juif camp auschwitz déporté nazi	**16** - soviétique moscou communiste staline pays
7 - touvier paul procès crime milice	**17** - sarkozy nicolas chirac cérémonie jacques
8- match équipe championnat champion monde	**18** - allemand allemagne hitler europe berlin
9 - parti politique socialiste candidat gauche	**19** - palestinien israélien israël irakien irak
10 - algérie algérien alger musulman algériens	**20** - débarquement vétéran juin plage normandie

religion. Nous concluons de cette observation que certains documents abordent plusieurs thèmes potentiellement très différents, ce qui n'apparaissait pas dans le notices documentaires fournies par l'INA.

5 Conclusion et perspectives

Dans cet article, nous avons montré qu'il était possible de faire une analyse pertinente des archives du corpus MATRICE-INA pour en extraire les thèmes mais également les principaux évènements en se basant uniquement sur les transcriptions automatiques obtenues par un système non optimisé pour ces données. Nous avons notamment mis en évidence les thèmes principaux abordés dans les archives grâce à une factorisation de matrices non négatives NMF. Au cours de notre analyse, nous avons également constaté l'émergence de certaines thématiques très différentes du thème de la seconde guerre mondiale. Dans la suite du projet MATRICE, il nous semble donc nécessaire de nous attacher dans un premier temps à la mise en place d'étapes de pré-traitement de notre corpus et notamment une étape de segmentation thématique des transcriptions des documents audiovisuels afin de pouvoir nous concentrer uniquement sur les parties du corpus les plus pertinentes pour les problématiques soulevées dans le cadre du projet. Cette segmentation thématique des documents doit également nous permettre de mettre en place une adaptation des modèles de langue. Par ailleurs, nous souhaitons adapter de façon non supervisée les modèles acoustique pour les enregistrements les plus anciens (comme présenté dans (Lamel *et al.*, 2002)), et envisageons d'ajouter de nouveaux « fillers » pour modéliser des bruits très présents dans les actualités diffusées en début de films au cinéma dans les années 50, par exemple le bruit des bombes.

Références

BENZEGHIBA, M., GAUVAIN, J. et LAMEL, L. (2012). Phonotactic language recognition using mlp features. *In Interspeech*, Portland, USA.

FUKUNAGA, K. et HOSTETLER, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40.

LAMEL, L., GAUVAIN, J.-L. et ADDA, G. (2002). Lightly supervised and unsupervised acoustic model trainings. *Computer Speech and Language*, 16:115–129.

LEE, D. D. et SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems 13 : Proceedings of the 2000 Conference*, pages 556–562.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

SHAHNAZ, F., BERRY, M. W., PAUCA, V. et PLEMMONS, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373 – 386.