Fusion d'espaces de représentations multimodaux pour la reconnaissance du rôle du locuteur dans des documents télévisuels

Sebastien Delecraz, Frédéric Béchet, Benoit Favre, Mickael Rouvier

Fusion d'espaces de représentations multimodaux pour la reconnaissance du rôle du locuteur dans des documents télévisuels

Sebastien Delecraz, Frederic Bechet, Benoit Favre, Mickael Rouvier
Laboratoire d'Informatique Fondamentale de Marseille
UMR 7279 CNRS / Aix-Marseille Université
163 avenue de Luminy, 13288 Marseille Cedex 9, FRANCE
{firstname.lastname}@lif.univ-mrs.fr

RÉSUMÉ

L'identification du rôle d'un locuteur dans des émissions de télévision est un problème de classification de personne selon une liste de rôles comme présentateur, journaliste, invité, etc. À cause de la non-synchronie entre les modalités, ainsi que par le manque de corpus de vidéos annotées dans toutes les modalités, seulement une des modalités est souvent utilisée. Nous présentons dans cet article une fusion multimodale des espaces de représentations de l'audio, du texte et de l'image pour la reconnaissance du rôle du locuteur pour des données asynchrones. Les espaces de représentations monomodaux sont entraînés sur des corpus de données exogènes puis ajustés en utilisant des réseaux de neurones profonds sur un corpus d'émissions françaises pour notre tâche de classification. Les expériences réalisées sur le corpus de données REPERE ont mis en évidence les gains d'une fusion au niveau des espaces de représentations par rapport aux méthodes de fusion tardive standard.

ABSTRACT

Multimodal embedding fusion for robust speaker role recognition in video broadcast

Person role recognition in video broadcasts consists in classifying people into roles such as anchor, journalist, guest, etc. Existing approaches mostly consider one modality, either audio (speaker role recognition) or image (shot role recognition), firstly because of the non-synchrony between both modalities, and secondly because of the lack of a video corpus annotated in both modalities. Deep Neural Networks (DNN) approaches offer the ability to learn simultaneously feature representations (embeddings) and classification functions. This paper presents a multimodal fusion of audio, text and image embeddings spaces for speaker role recognition in asynchronous data. Monomodal embeddings are trained on exogenous data and fine-tuned using a DNN on 70 hours of French Broadcasts corpus for the target task. Experiments on the REPERE corpus show the benefit of the embeddings level fusion compared to the monomodal embeddings systems and to the standard late fusion method.

MOTS-CLÉS : Identification du rôle du locuteur, fusion multimodale, émissions de télévision.

KEYWORDS: Speaker role recognition, multimodal speaker embeddings, broadcast news.

1 Introduction

L'identification du rôle d'une personne dans des émissions de télévisions consiste en un problème de classification de personne (parlant et/ou visible) selon une liste de rôles comme présentateur principal, journaliste, invité, etc. Dans ce contexte, les modalités audio et image sont naturellement complémentaires puisque l'on retrouve les caractéristiques du rôle dans le signal audio, la transcription écrite de la parole et les caractéristiques des scènes. De nombreuses approches proposées jusqu'à présent pour l'identification du rôle de la personne ne prennent en compte qu'une seule des modalités et ce pour deux principales raisons : premièrement la présence d'une personne n'est pas toujours synchronisée sur les différentes modalités. En effet, les locuteurs ne sont pas toujours visibles et tous les visages visibles à l'écran ne sont pas tous en train de parler. De plus, le manque de données multimodales annotées limite les possibilités pour réaliser un apprentissage joint de systèmes multimodaux qui supposent généralement une parfaite synchronie entre les modalités.

Les approches récentes basées sur les réseaux de neurones profonds (*Deep Neural Networks*, DNN) ont atteint des performances état-de-l'art pour de nombreuses tâches du traitement de l'audio et de l'image. Le principal avantage de ces techniques est d'apprendre simultanément des caractéristiques de représentations et des fonctions de classification. L'initialisation des caractéristiques de représentations peut être effectuée sur de grands corpus de données génériques pas nécessairement liés à la tâche cible pour plonger les données dans des espaces de représentations (dénommés *embeddings* en anglais) qui pourront être ajustés à la tâche cible de façon jointe. Cette approche a été proposée pour la tâche synchrone de détection et identification d'activité labiale (Ngiam *et al.*, 2011).

Dans cet article, nous voulons classifier les locuteurs en quatre rôles en utilisant les modalités audio, image et texte :

- R1 : les présentateurs, caractérisés par leur présence tout au long de l'émission ;
- R2 : les journalistes, des professionnels du monde de la télévision qui apparaissent une fois ou plus au cours d'une émission ;
- R3 : les reporters, proches du rôle R2, ce sont les correspondants qui couvrent les événements en dehors du plateau de l'émission ;
- R4 : les invités et autres, invités pour parler de l'actualité en raison de leur expertise ou leur renommée, ne prenant pas part à l'organisation et n'étant pas les leaders des débats ; les autres sont toutes les personnes qui peuvent apparaître, comme les personnes interviewées lors d'un reportage.

Nous présentons dans cet article [1] une alternative au paradigme de fusion tardive standard basée sur des *embeddings* multimodaux ajustés pour la tâche d'identification du rôle du locuteur (*Speaker Role Recognition*, SRR). La principale nouveauté de notre approche est une fusion au niveau des *embeddings* qui caractérise une information multimodale sans qu'une synchronie entre les modalités ne soit requise. Les expériences sur le corpus français REPERE met en évidence le gain de cette approche au regard de stratégies monomodales et des méthodes de fusion tardive.

2 Travaux connexes

L'identification automatique du rôle du locuteur (SRR) admet que les rôles soient identifiables par des caractéristiques spécifiques acoustiques, visuelles et textuelles comme le style de langage ou la

1. Partiellement traduit à partir de l'article (Rouvier *et al.*, 2015b).

prosodie. Dans la littérature, les méthodes de SRR ont été étudiées dans le but de catégoriser des documents audio-visuels (émissions de débats et d'informations). Les méthodes existantes sont séparées suivant les caractéristiques extraites (audio et/ou texte) ; le niveau de décision : pour chaque tour de parole (Liu, 2006) ou sur l'ensemble des tours de parole pour un locuteur donné (Dufour *et al.*, 2011; Hutchinson *et al.*, 2010; Zhang *et al.*, 2010; Wang *et al.*, 2011) ; et des techniques de classification supervisée (Dufour *et al.*, 2011; Bigot *et al.*, 2010; Liu, 2006) ou non-supervisée (Hutchinson *et al.*, 2010; Zhang *et al.*, 2010).

Dans (Dufour *et al.*, 2011), basé sur l'hypothèse que la classification de discours spontanés est un indice pour la tâche de SRR, les auteurs proposent une application de la détection de discours spontanés pour la tâche de SRR. Dans (Hutchinson *et al.*, 2010) les auteurs ont proposé un système non-supervisé de regroupement de locuteurs suivant leur rôle basé sur des caractéristiques structurelles et lexicales. Dans (Bigot *et al.*, 2010), les auteurs utilisent des caractéristiques temporelles, acoustiques et prosodiques pour classifier les rôles au niveau d'un groupe de locuteurs.

(Liu, 2006) classifie les rôles des locuteurs en utilisant des modèles de Markov cachés (*Hidden Markov Model*, HMM) et des classifieurs à maximum d'entropie (MaxEnt) avec une reconnaissance automatique de la parole (*Automatic Speech Recognition*, ASR) et une segmentation des tours de paroles des locuteurs réalisée manuellement. (Damnati & Charlet, 2011) ont présenté un système multimodal basé sur des caractéristiques lexicales et acoustiques pour la reconnaissance du rôle.

Pour ce qui est de la modalité visuelle, il n'y a pas à notre connaissance de travaux reposant sur l'image pour la détection du rôle des locuteurs. Les recherches portent surtout sur la reconnaissance de type de plan, comme par exemple dans (Feng *et al.*, 2014).

3 Approche

L'approche que nous proposons consiste en la création de représentation pour chaque modalité adaptées à la tâche de SRR. Ces représentation sont utilisées en entrée d'un classifieur multimodal qui peut tirer avantage de caractéristiques cross-modales tirées de la concaténation des représentations monomodales. La Figure 1 illustre cette approche.

Chaque représentation monomodale est entraînée sur un grand corpus monomodal qui n'est pas nécessairement en lié à la tâche de SRR. Le corpus multimodal annoté est utilisé seulement lorsque l'on entraîne la fusion. Cette méthode nous permet de tirer parti en même temps des fusions précoces et tardives : nous pouvons utiliser un grand corpus de données monomodales pour lequel nous n'avons pas une synchronie d'annotation dans les autres modalités comme cela peut être fait en fusion tardive ; nous entraînons des classifieurs multimodaux qui permettent la construction de caractéristiques multimodales directement depuis chaque modalité comme la fusion précoce.

Pour la modalité texte nous entraînons un réseau de neurones convolutionnel (*Convolutional Neural Network*, CNN) qui commence par l'apprentissage d'*embeddings* de mot sur un grand corpus de textes. La modalité audio utilise une représentation extraite d'un DNN modélisé à partir d'un système d'identification du locuteur, mais entraîné sur la tâche de SRR. La modalité image repose sur une représentation extraite d'un réseau de neurones d'identification de concepts visuels *ImageNet*, et raffiné pour la tâche cible. La fusion consiste en la concaténation des couches cachées de ces systèmes monomodaux à laquelle sont ajoutées des couches de neurones complètement reliées (*fully-connected*) pour créer un espace de représentation multimodal dans lequel sont fusionnées les caractéristiques

monomodales nécessaires à la prise de décision. La section suivante détaille l'architecture des composants monomodaux et multimodaux.

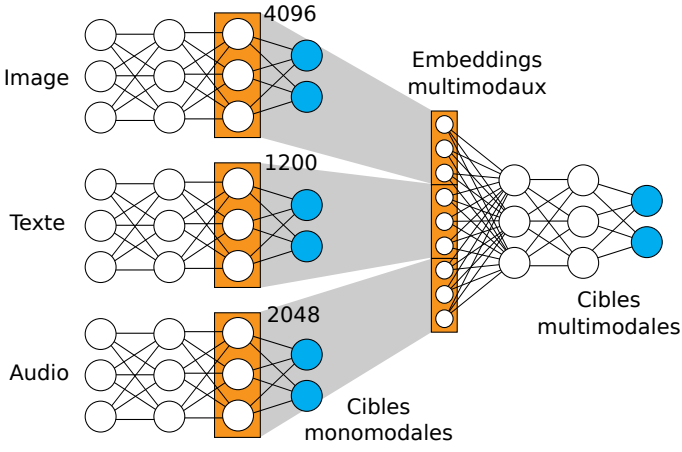

FIGURE 1 – Illustration de l'approche de fusion d'*embeddings*. Les systèmes monomodaux sont d'abord entraînés de manière indépendante, puis les activations de leurs couches cachées sont concaténées pour servir d'entrée au réseau multimodal. La taille des *embeddings* est donnée pour chaque modalité.

4 Modalité texte

De récents travaux ont montré que les CNN sont très performants dans le Traitement Automatique du Langage Naturel (TALN) pour des problèmes de classification (Collobert, 2011). Un CNN est un réseau de neurones profonds possédant plusieurs couches de convolution et de *max-pooling* suivis par un simple classifieur (souvent un Perceptron multicouche). Le principal avantage d'utiliser la convolution est la capacité du modèle de traiter des entrées de dimension variable (des phrases dans notre cas). De plus, les multiples filtres convolutionnels extraient des N-grammes lexicaux de différentes granularité alors que les couches de *pooling* extraient des caractéristiques sémantiques globale de l'entrée. Dans nos travaux, nous utilisons la transcription du discours du locuteur courant pour la tâche de SRR.

Premièrement, chaque mot est représenté par un vecteur de dimension 300 à valeurs réelles appelé *word embeddings*. Dans nos expériences, les *word embeddings* [2] ont été entraînés sur Wikipedia en utilisant le modèle *skip-gram* (5 itérations avec une fenêtre de taille 7). Cette stratégie nous permet de caractériser des associations grammaticales et sémantiques entre les mots.

Puis, les *word embeddings* pour les mots du tour de parole courant sont passé au travers de trois filtres convolutionnels qui sélectionnent les meilleurs 3-grammes, 4-grammes et 5-grammes. Ils sont combinés avec un une couche de *max-over-time pooling* (dimension 400) puis une couche de *soft-max fully-connected*. Nous utilisons du *dropout* pour désactiver 40% des neurones à chaque itération, agissant comme une régularisation.

2. Nous avons utilisé les toolkit *Word2Vec*.

5 Modalité audio

Dans nos travaux précédents, il a été proposé d'apprendre des caractéristiques acoustiques de haut niveau pour l'identification du locuteur (Rouvier *et al.*, 2015a), appelées *speaker embeddings*. Dans la même idée, nous proposons d'apprendre des caractéristiques du rôle du locuteur de haut niveau, appelées "*audio embeddings*", en utilisant des modèles profonds entraînés pour la tâche de SRR.

Les *audio embeddings* sont appris de la façon suivante : premièrement, un vecteur de caractéristiques acoustique de dimension 60 est extrait pour chaque tour de parole [3] avec un taux d'échantillonnage de 10ms (19 MFCCs, log énergie et deltas du premier et second ordre). Puis les statistiques du premier ordre, centrées-normalisées, obtenues depuis un modèle du monde (*Universal Background Model*, UBM) sont générées. Le modèle du monde est un GMM diagonal à 1024 composantes indépendant du canal (calculé avec le toolkit *Kaldi* [4]). Ensuite, les statistiques du premier ordre sont utilisées comme entrées d'un DNN avec deux couches cachées de 2048 neurones. Les fonctions d'activation de ces couches sont des unités linéaires rectifiées (ReLU) et la sortie est un *soft-max*. L'entraînement est effectué en minimisant l'entropie croisée sur les données d'apprentissage. Dans nos expériences, tous les hyper-paramètres des DNN sont déterminés sur un corpus de développement. La taille de mini-batch est de 512, 8 époques ont été effectuées et le taux d'apprentissage de 0.04 est réduit à 0.004 lors de la convergence. *In fine*, les *embeddings* audio de taille 2048 sont extraits à partir de la dernière couche cachée du DNN et utilisés pour la fusion multimodale.

6 Modalité visuelle

Les grammaires visuelles et les émissions de débats et d'informations sont une vraie source d'informations pour l'identification du rôle du locuteur.Nous utilisons des *embeddings d'image* issue de DNN qui se sont montrés extrêmement performants lors des campagnes d'évaluation *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC). Nous affinons le CNN nommé *AlexNet* (Krizhevsky *et al.*, 2012) entraîné sur le corpus de données ILSVRC-2012 pour la classification d'image. L'architecture du réseaux se compose de cinq couches de convolution, trois couches *fully-connected*, des couches de *max-pooling* et de normalisation. Elles utilisent la fonction *ReLU* (*Rectified Linear Unit*) comme fonction d'activation pour accélérer l'apprentissage et du *dropout* après les deux premières couches de neurones *fully-connected* pour éviter le sur-apprentissage. Ce modèle prend en entrée des images redimensionnées en (256×256) et normalisées en mini-batches de taille 512. Les poids sont mis à jour suivant les règles décrites dans (Krizhevsky *et al.*, 2012).

Nous avons adapté l'architecture *AlexNet* en changeant la dernière couche de neurones *fully-connected* pour prédire seulement quatre classes et ajusté les poids déjà appris du modèle *AlexNet* pour obtenir un nouveau CNN pour la tâche de SRR. Nous augmentons le taux d'apprentissage sur les couches de neurones *fully-connected* (dix fois le taux d'apprentissage global) afin de régulariser l'affinage de l'entraînement. Nous avons entraîné le réseau pendant 270 époques sur 19k images en utilisant *Caffe* [5] sur des GPU. Pour terminer, les *embeddings d'images* sont extraits depuis la seconde couche de neurones *fully-connected*, ce qui nous fournit des vecteurs de dimension 4096 qui seront utilisés dans le système de fusion multimodale.

3. La durée d'un segment moyen est de 7,8 seconde dans le corpus d'apprentissage.
4. http ://kaldi-asr.org/
5. http ://caffe.berkeleyvision.org/

7 Fusion multimodale

Il existe deux approches utilisées communément pour la fusion multimodale : les fusions précoce et tardive. La fusion tardive considère que les modalités sont indépendantes en appliquant d'abord des classifieurs séparés pour chaque modalité puis en fusionnant leurs sorties dans un classifieur de haut niveau. Malheureusement, le classifieur ne peut pas modéliser les corrélations entre les modalités et a seulement accès aux décisions des systèmes monomodaux. L'approche par fusion précoce contourne ce problème en apprenant des caractéristiques et des relations entre les classes pour modéliser les interactions entre les modalités. Toutefois, cette approche nécessite une synchronie entre les modalités.

Nous proposons une approche par fusion précoce basée sur des DNN où les entrées sont les *embeddings* de toutes les modalités pour la tâche spécifiée. Premièrement, les DNN sont entraînés indépendamment pour chaque modalité pour pouvoir en extraire des représentations monomodales générales (*embeddings* de texte, d'audio et d'image). Puis, ces *embeddings* sont utilisés en entrées d'un nouveau DNN entraîné pour apprendre à classifier le rôle des locuteurs à l'aide de caractéristiques multimodales. Contrairement à la fusion tardive, notre méthode peut tirer avantage de sous-espaces de caractéristiques pertinents (*embeddings*) de toutes les modalités.

Dans nos expériences, le DNN utilisé pour la fusion précoce est composé de deux couches cachées de dimension 1024. La non-linéarité de ces couches cachées est corrigée par une fonction d'activation ReLU. Les poids sont mis à jour en utilisant des mini-batch de taille 512, entraînés pendant 6 époques. Le taux d'apprentissage initialement de 0.01 est réduit à 0.001 lors de la convergence.

8 Expériences

Nous présentons les expériences réalisées sur le corpus multimodal REPERE (Giraudel *et al.*, 2012). La segmentation en locuteurs est effectuée par le système du LIUM (Rouvier *et al.*, 2013) qui obtient un Diarization Error Rate (DER) de 12.03% sur ce corpus. La transcription automatique est générée par *Kaldi* [4] et obtient un taux d'erreur mot (WER) de 19.67%. Le système est décrit en détail dans (Rouvier & Favre, 2014). Dans les émissions TV, les locuteurs apparaissent à l'écran seulement 60% du temps et les têtes visibles ne parlent que 30% du temps. Cet asynchronisme nous pousse à choisir une seule image pour représenter chaque tour de locuteur. Pour un tour de parole donné, il s'agit de l'image médiane de la plus grande intersection avec la segmentation en plan.

Les expériences sont menées sur le corpus REPERE (Giraudel *et al.*, 2012) rassemblant 70 heures d'émissions TV de 9 chaînes françaises. Chaque tour de parole est manuellement annoté avec la transcription, l'identité des locuteurs, et les rôles des locuteurs selon les classes suivantes : présentateur (R1), journaliste / chroniqueur (R2), reporter terrain (R3) et invité / autre (R4 [6]). Le corpus est divisé en ensembles d'apprentissage (18951 tours de parole), de développement (1402 tours de parole) et de test (4627 tour de parole) utilisés respectivement pour l'apprentissage des systèmes, la validation de la structure des réseaux de neurones et des hyper-paramètres des classifieurs, et l'évaluation des résultats. L'ensemble de test contient des types d'émissions qui se trouvent aussi dans les ensembles d'apprentissage et de développement (c'est le jour des émissions qui diffère), aussi bien que des nouveaux types d'émission qui sont inconnus des modèles entraînés sur l'ensemble d'apprentissage et

6. Fusion des rôles R4 et R5 du corpus original car cette paire de classes a un accord inter-annotateur faible.

de développement, permettant de vérifier la capacité de généralisation de notre modèle. La répartition des rôles dans l'ensemble de test et la suivante : 23,34% de présentateurs, 11,28% de journalistes, 14,22% de reporters et 51,16% d'autres.

Tous les résultats sont donnés en utilisant la précision (le nombre de rôles correctement identifié) et le *Diarization Error Rate* (DER). Le DER consiste à calculer les erreurs de SRR au niveau des trames, la même métrique étant utilisée pour les tâches de segmentation de locuteur. Le principal avantage de cette métrique est qu'elle nous permet de comparer deux sorties de systèmes de SRR différentes avec une segmentation en locuteur différente. Dans les résultats, DER-Man correspond au DER calculé sur la segmentation et transcription manuelles, alors que DER-Auto correspond à la segmentation et transcription automatiques.

Premièrement, la Table 1 compare les approches de bases et celles d'apprentissage profond monomodales. Parmi les approches de bases nous avons : *Majorité* où l'on choisit le rôle le plus fréquent pour chaque trame ; *Adaboost*, un classifier à base de *boosting* [7] sur des n-grammes de mots (modalité texte) ; *JFA* qui entraîne des modèles JFA sur les MFCC (*Joint-Factor-Analysis*) (Kenny *et al.*, 2005) (modalité audio) ; et *SVM-HOG* un classifieur à base de SVM sur des histogrammes de gradient (modalité image). Les résultats de la Table 1 indiquent clairement que les approches qui utilisent des DNN surpassent les systèmes de base. De plus, la modalité audio montre les classifieurs monomodaux les plus performants.

Système	**Modalité**	**Acc-Man**	**DER-Man**	**DER-Auto**
JFA	Audio	26,76	37,48	42,54
DNN-Audio	Audio	**77,52**	**19,79**	**25,43**
Adaboost	Texte	62,13	28,80	34,33
CNN-Texte	Texte	67,50	29,11	32,66
SVM-HOG	Image	62,76	36,97	42,04
CNN-Image	Image	70,48	25,69	35,25

TABLE 1 – Scores de précision et DER monomodaux sur l'ensemble de test.

Dans la suite, nos expériences montrent les résultats d'une fusion tardive basée sur un classifieur SVM. Toutes les probabilités données pour chaque modalité sont regroupées dans un vecteur, et un SVM linéaire est entraîné sur ces vecteurs de probabilités pour prédire le rôle du locuteur. Les résultats présentés dans la Table 2 montrent qu'une fusion des décisions au niveau des *embeddings* est plus performante qu'une fusion tardive. Ils justifient aussi l'utilisation de modèles multimodaux pour la tâche : le gain des performances des systèmes multimodaux comparé aux systèmes monomodaux est très important. Le meilleur score de DER-Man était de 19,79 (respectivement 25,43 pour le DER-Auto) pour les systèmes monomodaux alors qu'il est seulement de 13,84 (respectivement 19,9) pour les systèmes multimodaux. Nous pouvons aussi observer que c'est le modèle qui utilise les trois modalités qui donne les meilleurs résultats.

Afin d'étudier la robustesse de nos méthodes, la Table3 montre la précision et le DER sur un sous-ensemble du corpus de test qui correspond à des conditions d'émissions inconnues (le format des émissions est différent). Le système basé sur des *embeddings* de texte est robuste dans ces conditions alors que les résultats des modalités audio et image diminuent sur ces nouvelles émissions. C'est particulièrement vrai pour la modalité image qui passe d'un DER de 25,69 sur tout le corpus de test à 43,29 sur les émissions inconnues seulement. Dans ces conditions il n'est pas surprenant que les

7. https ://code.google.com/archive/p/icsiboost/

Système	Modalité	Acc-Man	DER-Man	DER-Auto
Majorité	-	51,16	39.77	44.54
Tardive	A+T	78,49	18,67	24,11
Tardive	A+I	80,98	17,26	22,98
Tardive	I+T	78,02	21,16	27,60
Tardive	A+I+T	82,36	15,37	20,97
Embedding	A+T	80,16	15,90	21,82
Embedding	A+I	82,16	15,45	2065
Embedding	I+T	76,01	22,83	28.60
Embedding	A+I+T	**85,28**	**13.84**	**19.79**

TABLE 2 – Scores de précision et DER pour les fusions tardives à posteriori et au niveau des *embeddings* (**A**udio, **I**mage, **T**exte).

méthodes de fusion ne donnent pas de meilleurs résultats que la meilleure des modalités seule. Ces résultats pointent une des faiblesses de notre approche quand toutes les modalités n'ont pas la même capacité de généralisation sur des événements inconnus.

Système	Modalité	Acc-Man	DER-Man	DER-Auto
CNN-Texte	T	**70,07**	**26,65**	**28,42**
DNN-Audio	A	65,69	29,88	37,31
CNN-Image	I	51,09	43,29	46,47
Tardive	A+I+T	70,07	27,77	34,61
Embedding	A+I+T	66,42	32,80	34,06

TABLE 3 – Résultats dans les conditions d'émission inconnue (5% de l'ensemble de test).

9 Conclusion

Dans cet article, nous avons introduit un système d'identification du rôle du locuteur basé sur la fusion d'espaces de représentations multimodaux pour des données asynchrones. Les expériences sur le corpus REPERE en utilisant une segmentation en locuteur et une transcription manuelles ou automatiques ont montré que la fusion de caractéristiques textuelles, audio et visuelles améliore considérablement les performances pour la classification en rôles des locuteurs au regard d'approches monomodales. Nos *embeddings* multimodaux ont permis de capturer les caractéristiques du rôle du locuteur sous plusieurs points de vues et l'utilisation d'une fusion au niveau des *embeddings* permet d'obtenir les meilleurs résultats avec 19,79% de DER sur une segmentation en locuteur automatique. Notre méthode tire avantage à la fois des fusions tardive et précoce en même temps : nous pouvons utiliser une grande quantité de données monomodales pour lesquelles nous n'avons pas d'annotations synchrones dans les autres modalités comme il se fait en fusion tardive ; nous entraînons des classifieurs multimodaux pour construire des caractéristiques multimodales directement depuis chaque modalité comme en fusion précoce.

Cependant, un des inconvénient de cette méthode est son manque de généralisation des modèles audio et image dans des conditions de contenus inconnus. L'augmentation de la robustesse dans ces conditions ainsi que l'application de notre modèle à d'autres tâches comme l'identification du locuteur sont des pistes intéressantes suite aux résultats obtenus.

Remerciements

Ces travaux ont été réalisés grâce à l'appui du projet A*MIDEX (n° ANR-11-IDEX-0001-02) financé par le programme du Gouvernement français "Investissement d'Avenir" et dirigé par l'Agence Nationale de la Recherche (ANR) ; ainsi qu'au soutien financier apporté par la Direction Générale de l'Armement (DGA) en partenariat avec Aix-Marseille Université dans le cadre du *Club des partenaires Défense*.

Références

BIGOT B., FERRANÉ I., PINQUIER J. & ANDRÉ-OBRECHT R. (2010). Speaker role recognition to help spontaneous conversational speech detection. In *SCSS*.

COLLOBERT R. (2011). Deep learning for efficient discriminative parsing. In *AISTATS*.

DAMNATI G. & CHARLET D. (2011). Multi-view approach for speaker turn role labeling in TV broadcast news shows. In *InterSpeech*.

DUFOUR R., ESTEVE Y. & DELÉGLISE P. (2011). Investigation of spontaneous speech characterization applied to speaker role recognition. In *Interspeech*.

FENG B., BAI J., CHEN Z., HUANG X. & XU B. (2014). Anchor shot detection with deep neural network. In *PCM*.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *LREC*.

HUTCHINSON B., ZHANG B. & OSTENDORF M. (2010). Unsupervised broadcast conversation speaker role labeling. In *ICASSP*.

KENNY P., BOULIANNE G., OUELLET P. & DUMOUCHEL P. (2005). Factor analysis simplified. In *ICASSP*.

KRIZHEVSKY A., SUTSKEVER I. & HINTON G. E. (2012). Imagenet classification with deep convolutional neural network. In *NIPS*.

LIU Y. (2006). Initial study on automatic identification of speaker role in broadcast news speech. In *NAACL*.

NGIAM J., KHOSLA A., KIM M., NAM J., LEE H. & NG A. Y. (2011). Multimodal deep learning. In *ICML*.

ROUVIER M., BOUSQUET P.-M. & FAVRE B. (2015a). Speaker diarization through speaker embeddings. In *EUSIPCO*.

ROUVIER M., DELECRAZ S., FAVRE B., BENDRIS M. & BECHET F. (2015b). Multimodal embedding fusion for robust speaker role recognition in video broadcast. In *ASRU*.

ROUVIER M., DUPUY G., GAY P., KHOURY E., MERLIN T. & MEIGNIER S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *InterSpeech*.

ROUVIER M. & FAVRE B. (2014). Speaker adaptation of dnn-based asr with i-vectors : Does it actually adapt models to speakers ? In *InterSpeech*.

WANG W., YAMAN S., PRECODA K. & RICHEY C. (2011). Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *ICASSP*.

ZHANG B., HUTCHINSON B., WU W. & OSTENDORF M. (2010). Extracting Phrase Patterns with Minimum Redundancy for Unsupervised Speaker Role Classification. In *NAACL*.