REPERE : premiers résultats d'un défi autour de la reconnaissance multimodale des personnes

Juliette Kahn[3] Aude Giraudel[1] Matthieu Carré[2] Olivier Galibert[3] Ludovic Quintard[3]

(1) DGA, 7 rue des Mathurins, 92 221 BAGNEUX Cedex

(2) ELDA, 57 Rue Brillat-Savarin 75013 Pariss

(3) LNE, 29, avenue Roger Hennequin 78197 TRAPPES Cedex

aude.giraudel@dga.defense.gouv.fr, nom@elda.org, prenom.nom@lne.fr

RÉSUMÉ

Le défi REPERE a pour objectif d'encourager les recherches et le développement de technologies dans le domaine de la reconnaissance des personnes par des indices multimodaux. Afin d'estimer la progression des solutions proposées, des campagnes annuelles d'évaluation autour de la reconnaissance multimodale des personnes sont organisées entre 2012 et 2014. Le corpus REPERE, un corpus [de 60h] de vidéos en français est développé à cette occasion. Quatre tâches correspondant aux quatre questions : Qui parle ? Qui voit-on ? De qui parle-t-on ? De qui le nom apparait à l'écran ? ont été définies et ce papier présente les premiers résultats obtenus lors du test à blanc de janvier 2012.

ABSTRACT

REPERE : preliminary results of a multimodal person recognition challenge

The REPERE Challenge aims at supporting researches on people recognition in multimodal conditions. To estimate the technology progress, annual evaluation campaigns on multimodal recognition of people in videos will be organized from 2012 to 2014. In this context, the REPERE corpus, a French videos corpus with multimodal annotation has been developed. The systems which participated to the dry run have to answer the following questions : Who is speaking ? Who is present in the video ? What names are cited ? What names are displayed ? This paper describes the corpus used during the january 2012 dry run and presents the first results.

MOTS-CLÉS : Corpus, Parole multimodale, Reconnaissance du locuteur, Campagne d'évaluation.

KEYWORDS: Corpora, Mutimodal conditions, Speaker recognition, Evaluation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 497–504,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Reconnaître une personne dans une vidéo est un défi qui connaît de nombreuses applications. Cette reconnaissance revient à extraire des informations pertinentes des deux flux visuel et acoustique et à les combiner afin de répondre à différentes questions comme qui parle à quel moment ou de qui parle-t-on.

Quelques campagnes comme TRECVid (Smeaton *et al.*, 2006) ou Biosecure Multimodal Evaluation Campaign (Ortega-Garcia *et al.*, 2010) ont déjà abordé en partie la reconnaissance multimodale des personnes en se fondant sur des corpus anglophones.

Le défi REPERE [1] a pour objectif d'encourager le développement de systèmes automatiques pour la reconnaissance de personnes en contexte multimoda en Français. Financé par l'Agence Nationale de la Recherche (ANR) et par la Direction Générale de l'Armement (DGA), ce projet a commencé en mars 2011 et se termine en mars 2014. Deux campagnes d'évaluation organisées par le LNE et ouvertes à toute la communauté, sont prévues aux débuts des années 2013 et 2014.

Ce papier présente les premiers résultats obtenus lors du test à blanc mené en janvier 2012. Dans une première partie, nous définissons précisément les tâches évaluées dans le cadre du défi REPERE. La seconde partie décrit la constitution du corpus produit par ELDA. La troisième partie revient sur les métriques utilisées. Après avoir présenté les premiers résultats en partie 4, nous proposons quelques perspectives.

2 Questions posées lors de l'évaluation

L'objectif du défi REPERE est d'encourager le développement de solutions automatiques pour la reconnaissance de personnes dans des vidéos. De chacune des vidéos, il est possible d'obtenir des images et un signal sonore d'où seront extraites les informations pertinentes. Le défi REPERE s'interesse donc à la reconnaissance de personnes dans un contexte multimodal. Les systèmes évalués lors du défi-REPERE doivent répondre à quatre questions élémentaires :

1. Qui est en train de parler ?
2. Qui apparaît à l'image ?
3. De qui le nom est cité oralement ?
4. De qui le nom apparaît à l'écran ?

Pratiquement, des images clé sont extraites toutes les 10 secondes. Pour chacune de ces images, le système fourni la liste des personnes qui parlent (Question 1), qui apparaissent à l'écran (Question 2), des noms de personnes qui sont cités oralement (Question 3) et des noms de personnes qui apparaissent à l'écran (Question 4). La tâches principale du défi est de lister, pour chaque image clé, qui apparaît à l'écran ou parle.

Cette tâche principale peut être réalisée en mode supervisée (les systèmes peuvent alors avoir des modèles de voix et/ou de visages des personnes *a priori*) et en mode non-supervisé (les systèmes ne peuvent utiliser que les informations présenes dans la vidéo)

1. Pour plus d'information consultez le site www.defi-repere.fr

Pour répondre à ces questions élémentaires, plusieurs briques techologiques peuvent être envisagées. Quelques unes d'entre elles sont également évaluées lors du défi REPERE (suivi de têtes et textes, Segmentation en locuteurs, Segmentation des textes, Segmentation des têtes, Transcription de la parole, Transcription des textes inscrustés). Afin de développer des solutions pour répondre à ces différentes questions, un corpus, décrit dans la prochaine section, est produit par ELDA.

3 Corpus REPERE

3.1 Sélection des données

La première partie du corpus REPERE, dédiée au test à blanc, regroupe six heures de vidéos extraites de différents programmes de chaînes de télévisions d'information françaises (BFM TV et LCP). En fin de projet, le corpus comportera soixante heures de vidéo. Les émissions, dont la répartition est accessible dans le Tableau 1, sont des journaux télévisés et des débats pour lesquelles ELDA a conclu des accords permettant leur utilisation légale. Les futures données

Emissions	Chaîne	Durée (minutes)
BFM Story	BFM TV	60
Planète Showbiz	BFM TV	15
Ca vous regarde	LCP	15
Entre les lignes	LCP	15
Pile et Face	LCP	15
LCP Info	LCP	30
Top Questions	LCP	30

TABLE 1 – Emissions télévisuelles présentes dans le corpus REPERE (6H)

collectées respecteront également les mêmes répartitions. Elles proviennent des six émissions suivantes :

– *Top Questions* est une retransmission des questions au gouvernement de l'Assemblée Nationale. Les prises de parole dans cette émission correspondent dans leur grande majorité à de la parole préparée. Les vidéos sont composées de nombreux travelling sur l'ensemble de l'Assembléee Nationale.
– *Ça vous regarde*, *Pile et Face* et *Entre les lignes* sont des émissions de débats politiques qui incluent à la fois de la parole préparée et de la parole spontanée. Il s'agit pour la grande majorité de ces émissions de débats en plateau.
– *LCP Info* et *BFM Story* sont des journaux d'information avec un nombre réduit de présentateurs et de nombreux reporters spéciaux. Ces émissions donnent lieu à de multiples interviews qui sont agrémentées de reportages illustratifs.
– *Planète Showbiz* est un magazine people commenté principalement en voix-off. De nombreuses personnes inconnues sont filmées et il s'agit en grande majorité de parole spontanée.

Les vidéos sont sélectionnées afin d'obtenir une grande diversité de situations aussi bien au niveau du son que de l'image. Un premier critère de sélection est d'équilibrer la répartition entre parole spontanée et parole préparée afin de pouvoir mesurer, dans un second temps, leur impact

sur les systèmes. Au niveau des images, nous avons cherché à obtenir des vidéos où les têtes sont filmées de manières différentes afin d'assurer la diversité des cas possibles (luminosité, taille des tête, angle de la caméra...). Par exemple, la taille des têtes de personnes annotées varie de 936 pixels^2 à 192 702 pixels^2. Des exemples d'images sont donnés en figure 1.

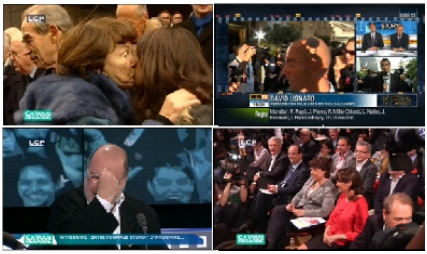

FIGURE 1 – Exemples d'images extraites des vidéos traitées

3.2 Annotations

Les annotations effectuées sur le corpus concernent à la fois le signal sonore et les images. Les annotations du signal de parole ont été effectuées à l'aide de Transcriber (Barras *et al.*, 2000) [2] et sont disponibles au format *trs*. Elles s'appuient sur le guide d'annotation élaboré pour la campagne ESTER2 [3] (Galliano *et al.*, 2005) et incluent les éléments suivants :
– La segmentation du signal en tours de parole.
– Le nommage des locuteurs.
– La transcription de la parole en indiquant les disfluences et les hésitations.
– Le balisage des citations de noms de personnes dans la transcription.
L'annotation des personnes dans les images a donné lieu à la création d'un guide d'annotation spécifique accessible sur le site du défi REPERE [4]. Plusieurs éléments ont été annotés à l'aide de VIPER-GT [5] après modification des sources afin d'assurer le cohérence des index audio et video (i.e. correspondance de chaque image avec un temps audio précis). L'annotation se concentre sur six types d'information :
– La segmentation des têtes consiste à détourer les têtes susceptibles d'être reconnues par les systèmes. Ainsi, les plus petites têtes ne sont pas détourées, mais simplement signalées comme étant présentes à l'image. Les têtes détourées sont celles dont la surface est supérieure à un seuil donné (dans notre cas, 2 500 pixels^2). Un exemple est donné en Figure 2. Il est à noter qu'il s'agit d'annotations de têtes et non d'annotations de visages. Ainsi, par exemple, les têtes de profil sont annotées.
– La description de tête consiste à décrire des caractéristiques physiques de la tête comme le fait de porter des lunettes ou d'avoir une moustache, mais aussi son orientation (face, profil, dos)

2. http ://trans.sourceforge.net/
3. http ://www.afcp-parole.org/camp_eval_systemes_transcription
4. http ://www.defi-repere.fr/
5. http ://viper-toolkit.sourceforge.net/

et si la tête n'est pas partiellement cachée par un autre objet. Cette description pourra être utilisée pour analyser les erreurs des systèmes.
– L'identification des personnes consiste à annoter le nom des personnes présentes à l'écran. L'annotation est faite à l'aide des renseignements présents dans la vidéo. Les personnes non-citées se voient attribuées un identifiant unique.
– Le détourage et la transcription du texte présent dans l'image consistent à repérer toutes les zones de texte présentes à l'écran et à les transcrire. Le détourage se fait à l'aide de rectangles. Le texte est segmenté en blocs cohérents et il est indiqué si il s'agit d'un texte complet, incomplet ou illisible. La transcription respecte la typographie des caractères présents à l'écran. Un exemple de détourage de texte est donné par la figure 2.
– Le balisage des noms de personnes présents dans le texte.
– Le repérage des moments d'apparition et de disparition qui indique l'intervalle temporel de présence des textes et têtes à l'image.

FIGURE 2 – Exemple de segmentation

3.3 Répartition des annotations et des données dans le corpus du test à blanc

Un test à blanc a été mené en janvier 2012. Le Tableau 2 résume les annotations effectuées sur les six premières heures de corpus ainsi que le nombre de personnes qu'il est possible de trouver à partir des indices sonores ou visuels. Les trois premières heures du corpus ont constituées le corpus de Dev tandis que les trois autres ont servi de corpus de test. Au niveau du corpus de développement, il est à noter que 45% des personnes à trouver ont leur nom qui apparaît à l'écran et que 55% des personnes à trouver ont leur nom cité dans le signal sonore. Par ailleurs, 33% des personnes à trouver ne sont citées ni oralement ni par écrit . Ainsi, en apprentissage non supervisé, il n'est possible de trouver que 67% des personnes. Enfin, 51% des personnes apparaissent à l'écran et parlent, 40% des personnes apparaissent à l'écran sans parler et 9% des personnes ne peuvent être repérées que par le signal sonore. La reconnaissance des têtes es donc un élément clé d'un sysème performant.

Ces tendances se retrouvent au niveau du corpus de test même si les proportions ne sont pas tout à fait les mêmes. 49% des personnes à trouver ont leur nom qui apparaît à l'écran et 69% des personnes à trouver ont leur nom cité dans le signal sonore. Par ailleurs, 22% des personnes à

		Dev	Test
Indices visuels	Nombre de têtes à l'écran	1 421	1 534
	Nombre de mots dans les textes	13 240	14 764
Indices sonores	Nombre de segments de parole	1 571	1 602
	Nombre de mots transcrits	33 205	33 247
Personnes	Nombre de personnes dont la tête apparaît	216	145
	Nombre de personnes dont le nom apparaît à l'écran	200	141
	Nombre d'anonymes vus à l'écran	177	138
	Nombre de personnes qui parlent	141	122
	Nombre de personnes citées oralement	242	191
	Nombre d'anonymes qui parlent	45	33
	Nombre de personnes à trouver	237	171

TABLE 2 – Données chiffrées sur CORPUS de test à blanc de défi-REPERE

trouver ne sont citées ni oralement ni par écrit. Ainsi, en apprentissage non supervisé, il n'est possible de trouver que 78% des personnes dans le corpus de test. Enfin 56% des personnes apparaissent à l'écran et parlent, 29% des personnes apparaissent à l'écran sans parler et 15% des personnes ne peuvent être repérées que par le signal sonore.

Il est à noter que cette répartition dépend en partie de l'émission traitée. Par exemple, si pour *Entre les lignes*, 66% des personnes ne font qu'apparaître à l'écran (et ne sont donc repérables qu'à travers un mode d'apprentissage supervisé), dans *BFM Story*, seulement 20% des personnes ne font qu'apparaître à l'écran. Dans le même ordre d'idée, pour *BFM Story*, 31% des personnes n'ont que leur nom qui apparaît à l'écran sans être en train de parler ou que leur visage n'apparaisse à l'écran. Dans *Entre les Lignes*, au contraire, aucune personne n'a que son nom qui apparaît à l'écran.

La répartition du temps de parole entre les personnes n'est pas équilibrée comme l'illustre la figure 3. Certaines interviennent longtemps (près de 10 minutes) tandis que d'autres interviennent moins de 20 secondes. Cette situation encourage le développement de solutions pour lesquelles très peu de données sont accessibles.

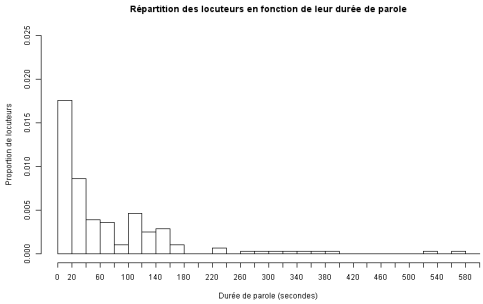

FIGURE 3 – Répartition des locuteurs en fonction de leur temps de parole

En ce qui concerne la vidéo, l'annotation, très coûteuse, n'est faite que sur les images clé et pas sur l'ensemble du corpus. Il est tout de même à noter que 26% des persones n'apparaissent que sur une image tandis que 4% des personnes apparaissent sur plus de 30 images. La Figure 4 illusre le nombre de personnes en fonction du nombre de fois où elles apparaissen à l'écran.

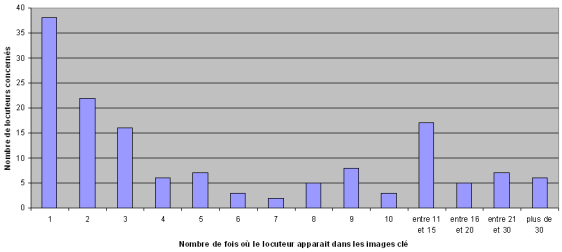

FIGURE 4 – Répartition des locuteurs en fonction de leur temps de parole

Au final, 351 personnes sont présentes entre le corpus de développement et le corpus de test. Seules 57 personnes sont présentes dans les deux.

4 Premiers résultats

4.1 Métrique

La métrique d'évaluation pour la tâche principale est le Estimated Global Error Rate (EGER). Elle se fonde sur la comparaison des noms de personnes présents dans les références et les sorties systèmes.
Pour chaque image annotée de la référence, i, la liste des personnes présentes et/ou parlant à l'instant t_i est constituée pour la référence d'une part et pour la soumission d'autre part. Ces deux listes sont comparées en associant les personnes une à une. Chaque personne ne peut être associée au plus qu'une fois. Ceci permet de caractériser les listes fournies selon les cas suivants :
– Sont considérées comme correctes une association entre deux personnes nommées ou une association entre deux personnes anonymes.
– Une confusion, C, est une association entre deux personnes avec des noms différents ou entre un nommé et un anonyme.
– Une fausse alarme, FA, est comptabilisée pour chaque personne de l'hypothèse non associée.
– Un oubli,M, est considéré pour chaque personne de la référence non associée à une personne de la soumission.
Un coût est associé à chaque type d'erreur selon la gravité de l'erreur. Ainsi, une confusion a un coût de 0,5 tandis qu'une fausse alarme ou un oubli ont un coût de 1. Ainsi pour les N images à analyser, EGER se définit de la manière suivante :

$$EGER = \frac{\sum_{i=0}^{i=N} 0.5 * C_i + FA_i + M_i}{\sum_{i=0}^{i=N} P_i} \quad (1)$$

Où P_i est le nombre de personnes à trouver à l'image ou à l'instant i.

Les premiers résultats présentés s'appuient sur cette métrique. Dans le cadre du test à blanc d'autres métriques accompagnent cette mesure globale afin de définir où les systèmes se sont trompés : est-ce dans l'OCR, dans la transcription ou le répérage des entités nommées ? Nous ne pourrons pas développer l'ensemble des résultats obtenus lors de la campagne dans ce papier. Nous ous focalisons sur la comparaison des résultats obtenus pour la tâche principale.

4.2 Variation de performance sur la tâce principale

Trois consortiums ont participé au test à blanc. Ils ont proposé plusieurs systèmes pour répondre aux tâches. En apprentissage supervisé, l'EGER total varie de 43.4% à 64.7% selon le système. Il est à noter que l'EGER calculé sur les personnes repérées à partir du signal sonore présente des meilleurs résultats que l'EGER calculé sur les personnes repérées à l'aide des images ($EGER_{MeilleurAudio} = 20.9\%$ vs $EGER_{MeilleurVideo} = 51.8\%$). La même tendance est observée en apprentissage non-supervisé : les performaces

5 Conclusion et perspectives

Le Défi REPERE vise à encourager le développement de solutions pour la reconnaissance multimodale de personnes. Le corpus final comportera 60 heures de vidéo avec des annotations précises concernant les indices visuels et sonores permettant de savoir qui parle, qui apparait à l'écran, de qui l'on parle et quels noms s'affichent. Les premiers résultats obtenus lors du test à blanc montrent qu'il existe un potentiel réel de progression. Les personnes ne sont parfois présentes que quelques secondes à l'écran ou dans le signal sonore. Ce défi permet également de s'interroger sur les possibilités de fusions des indices idiosyncratiques et ouvrent de nombreuses perspectives. Comment améliorer la détection de têtes ? Comment fusionner les informations pertinentes ? Telles sont quelques questions auxquelles les prochaines campagnes de 2013 et 2014 tenteront de répondre.

Références

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (2000). Transcriber : development and use of a tool for assisting speech corpora production. *In Speech Communication special issue on Speech Annotation and Corpus Tools*, volume 33.

GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K., BONASTRE, J.-F. et GRAVIER, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. *In European Conference on Speech Communication and Technology*, pages 1149–1152.

ORTEGA-GARCIA, J., FIERREZ, J., ALONSO-FERNANDEZ, F., GALBALLY, J., FREIRE, M., GONZALEZ-RODRIGUEZ, J., GARCIA-MATEO, C., ALBA-CASTRO, J., GONZALEZ-AGULLA, E., OTERO-MURAS, E. *et al.* (2010). The multiscenario multienvironment biosecure multimodal database (bmdb). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1097–1111.

SMEATON, A., OVER, P. et KRAAIJ, W. (2006). Evaluation campaigns and trecvid. pages 321–330.