Détection audiovisuelle du locuteur actif lors de réunion

Francisco Madrigal[1] Frédéric Lerasle[1,3] Lionel Pibre[2] Isabelle Ferrané[2,3]

[1] LAAS-CNRS, Toulouse, France
[2] IRIT, Université de Toulouse, Toulouse, France
[3] Université Paul Sabatier, Toulouse, France

{jfmadrig, lerasle}@laas.fr, {pibre, ferrane}@irit.fr

Résumé

Caractériser les interactions multi-personnes lors de réunions, par exemple les tours de parole successifs, permet de nombreuses applications concrètes e.g. en multimédia. Lors du déroulement d'une réunion, la détection du locuteur actif est à inférer intuitivement par l'activité vocale. Néanmoins, des informations complémentaires extraites de flux vidéo ou modèles d'interactions humains sont susceptibles de robustifier le processus de détection. Ainsi, ces travaux prototypent puis évaluent une modalité originale de détection de locuteur qui mixtent des percepts audiovisuels et de comportements sociaux inhérents au contexte de réunion. Les percepts visuels sont inférés à l'aide d'un réseau neuronal convolutif qui capture les relations spatio-temporelles de clips vidéo de visages des participants à la réunion. Nous comparons ainsi plusieurs architectures CNN avec deux types de données visuelles en entrée: les images RVB de visages segmentés et les images de flot optique associé. Le modèle de comportement social repose sur la convergence naturelle des directions du regard des participants vers le locuteur courant. Notre détecteur est évalué sur le jeu de données audiovisuelles AMI corpus. Nous montrons que la fusion de ces divers percepts améliorent les performances globales de notre système de détection.

Abstract

Characterizing multi-person interactions in meetings, for example successive speaking tours, is useful for many concrete applications e.g. in multimedia. During the course of a meeting, the detection of the active speaker is intuitively inferred by voice activity. However, additional information extracted from video streams or models of human interactions are likely to strengthen the detection process. Thus, these aspects can create an original modality of active speaker detection mixing audiovisual percepts and social behaviors inherent in the meeting context. Visual percepts are inferred using a Convolutional Neural Network (CNN) that captures spatio-temporal relationships of video clips of participants faces at the meeting. We thus compare several CNN architectures with two types of visual input data: RGB images of segmented faces and their associated optical flow images. The model of social behavior is based on the natural convergence of the directions of gaze of the participants towards the current speaker. Our detector is evaluated on the audiovideo data set AMI corpus. We show that the fusion of these various percepts improves the overall performance of our detection system.

Figure 1: Exemple de détection du locuteur actif utilisant la vidéo *Talking* face mise à disposition par le groupe de travail sur la reconnaissance des visages et des gestes [2].

Keywords

Speaker recognition, Convolutional Networks, Audiovisual modeling, Feature fusion.

1 Introduction

Les réunions sont fréquentes dans notre société. Elles permettent par exemple de coordonner le travail des collaborateurs au sein d'une équipe. Dans tous les cas, la personne d'intérêt est celle qui prend la parole; elle centralise alors l'attention de tous les autres participants. Cette interaction se fait non seulement par la voix, mais aussi par certaines postures notamment faciaux. L'interprétation de ces informations audiovisuelles afin d'identifier le locuteur actif peut être utile dans des scénarios autres : interaction H/M [1], multimédia, etc.

La littérature propose certes des approches sur la détection du locuteur mais sans prise en compte des interactions posturales entre personnes. De plus, elles se focalisent

en général sur l'analyse du canal audio car discriminant. Ainsi, certaines approches portent sur la reconnaissance du locuteur, d'autres sur la segmentation et regroupement en locuteurs qui permettent de partitionner l'audio entrant en segments homogènes i.e. des segments audio contenant au moins un locuteur. Citons enfin les approches qui transforment le signal audio en texte.
Le couplage avec une détection visuelle est louable, en particulier lorsque plusieurs personnes parlent en même temps. Une détection du locuteur via la vision seule est a contrario peu efficace de par les occultations, les expressions faciales, etc.
Il est donc opportun de coupler audio et vidéo comme illustré sur la figure 1. Ici, le mouvement des lèvres est encodé par des caractéristiques spatio-temporelles, généralement estimées via un réseau de neurones convolutionnel (CNN).

Notre détecteur combine ainsi trois percepts : (1) une reconnaissance audio classique, (2) une analyse spatio-temporelle des mouvements faciaux intrinsèques aux visages préalablement segmentés par un détecteur visuel usuel, et (3) une estimation visuelle de leurs orientations (notamment le lacet) et un modèle social qui présuppose la convergence des regards vers le locuteur actif.
A noter que la littérature propose hélas peu de bases de données audio-vidéo publiques adaptée à notre contexte de réunion : interaction multi-personnes, tours de parole successifs, champs de vue large.

Forts de ces constats, nos contributions sont donc les suivantes :

1. Un détecteur visuel de locuteur actif, basé sur des CNN 3D, et adapté à notre contexte de réunion. Ces réseaux prennent en entrée des clips vidéos RVB et leur flot optique intrinsèque pour caractériser les mouvements faciaux inhérent au visage préalablement détecté.

2. Un modèle de comportement social des participants reposant sur la topologie spatiale (position relative des participants) de ceux-ci dans la salle et leurs orientations de visages inférées dans le flux vidéo.

3. Une stratégie de fusion combinant ces deux percepts et une détection audio du locuteur.

4. Des évaluations exhaustives sur des jeux de données publiques adaptés à notre application. La fusion, en désambiguïsant certaines situations, apporte une réelle plus-value.

L'article est structuré comme suit. La section 2 présente un état de l'art et la section 3 décrit notre approche. Les évaluations associées sont présentées en section 4. La section 5 décrit enfin les conclusions et perspectives.

2 Etat de l'art associé

La communication entre personnes est assurée classiquement par la voix et les gestes (notamment faciaux). L'analyse des signaux audio, surtout, et vidéo associés permet la détection automatique du locuteur actif lors des tours de parole entre personnes.

Audio - La littérature propose de nombreux travaux sur la reconnaissance automatique du locuteur [2]. Une stratégie repose sur la segmentation et regroupement en locuteurs. On partitionne ainsi le flux audio en segments de parole/non-parole, puis on associe un locuteur à chaque segment de parole. Bonastre et al. dans [3] proposent une méthode de segmentation et regroupement en locuteurs basée sur la modélisation par clé binaire (BK), qui transforme l'audio en un vecteur caractéristique représentant le locuteur dans l'espace binaire. Puis, le regroupement en locuteurs est effectué par un algorithme de regroupement aggloméré itératif qui forme des segments du même locuteur. Patino et al. dans [4] bonifient cette approche en utilisant un regroupement spectral. Un défi majeur est de reconnaître la même personne quelle que soit l'intensité de la voix du locuteur e.g. lors de chuchotement ou bruit de fond. Vestman et al. dans [2] réalisent une taxonomie approfondie des différentes caractéristiques qui répondent à ces problèmes et proposent une fonction de variation temporelle du son. Avec l'essor des réseaux CNN, certains travaux [5] déroulent cette approche de bout en bout pour la segmentation et regroupement en locuteur. Les architectures récurrentes sont basées sur des architectures utilisant des couches récurrentes à mémoire court et long terme (LSTM)[6, 7] car elles captent les variations de la voix du locuteur. Les approches requièrent hélas, pour la plupart, des cartes GPU puissantes. Dans [8], les auteurs proposent une approche pour la segmentation et regroupement en locuteurs de bout en bout au niveau de l'énoncé (*utterance*). Cette méthode propose une nouvelle architecture appelée "thinResNet", qui intègre une couche GhostVLAD permettant d'agréger les caractéristiques dans le temps. Ce réseau est entraîné et évalué sur la base de données VoxCeleb [9]. Il s'agit d'une base de données audiovisuelles de courts clips d'interviews extraits de YouTube. Cette base de données est constituée de plus de 2000 heures d'enregistrement et plus de 7000 personnes. Xie et al. dans [8] exhibent de bonnes performances avec ce réseau compact.
Néanmoins, ces approches restent sensibles à des artefacts sonores e.g. les murmures, les bruits de fond ou le chevauchement des sons. Une alternative est donc de considérer aussi le signal vidéo.

Vidéo - Zhou et al. dans [10] synthétisent les travaux menés sur la détection visuelle du locuteur jusqu'en 2014 et les jeux de jeux de données publiques associés. Les méthodes sont catégorisées comme suit :

1. Basées image brute. Les pixels sont ici transformés en caractéristiques via des méthodes telles que l'Analyse en Composantes Principales [11, 12].

2. Basées mouvement. Les caractéristiques capturent le mouvement observé pendant la conversation. Citons par exemple le flot optique [12].

3. Basées géométrie. Ces caractéristiques capturent les informations géométriques d'une bouche en mouvement. Korshunov et al. dans [13] infèrent ce mouvement par la distance inter-image entre points détectés sur la bouche.

Audio-Vidéo - Ces approches sont a priori plus robustes au bruit de fond et aux variations d'intensité de la voix. Récemment, l'apprentissage profond a permis des avancées dans la détection du locuteur actif audiovisuel en capturant les relations temporelles des informations visuelles et acoustiques. L'utilisation de réseaux de neurones récurrents (RNN) est explorée dans [14] pour extraire des caractéristiques vidéo via un réseau CNN 2D. Puis, à l'instar de [13], ils entraînent une couche de mémoire à court terme pour chaque entité. Les sorties des deux couches sont concaténées et le résultat est transmis à une dernière couche LSTM. On parle ici de fusion précoce. Pedritis et al. dans [15] comparent fusion précoce et tardive pour la reconnaissance des locuteurs. Afouras et al. dans [16] explorent l'impact de diverses fonctions de perte pour l'apprentissage d'un réseau de neurones pour la lecture labiale avec des données audiovisuelles.
Dans la littérature, l'application des couches LSTM a largement été étudiée car elles captent bien les informations spatio-temporelles d'un locuteur. Mentionnons aussi les réseaux convolutionnels 3D i.e. C3D [17] mais dans un cadre applicatif autre, par exemple la reconnaissance d'actions. Il existe des variantes e.g. le réseau neuronal résiduel 3D (ResNet3D) [18]. Tous ces réseaux, étudiés ici pour la détection de locuteur, visent à capturer des caractéristiques 3D i.e. apparence et mouvement simultanément.
Notons enfin que ces réseaux 2D ou 3D exploitent usuellement des images brutes mais l'ajout de caractéristiques manuelles (dites *hand crafted features*) est souvent concluante; citons le canal profondeur [19] ou le flot optique [12, 19, 18] privilégié ici.
Ces modalités audio-vidéo sont hélas souvent étudiées en environnement contrôlé contrôlé i.e. une seule personne regardant en fronto-parallèle la caméra. Ce scénario n'est pas compatible pour notre contexte de réunion induisant des tours de parole successifs et non contrôlés entre plusieurs personnes. Leurs interactions sociales, captées par la vidéo, est alors très informative quant au locuteur courant; le regard des participants se focalise naturellement vers l'orateur principal. Liu et al. dans [20] estime ainsi l'orientation du regard avec un CNN différentiel focalisé sur les yeux ce qui requiert une bonne résolution image du visage. Ce pré-requis est incompatible avec notre contexte de réunion (champs de vue large de la scène pour observer tous les participants).
Une alternative, privilégiée ici, pré-suppose que l'orientation du regard est assimilée à celle de la tête. Citons ici la modalité *opensource* HyperFace [21] basée sur le réseau ResNet101. Celle-ci infère simultanément détection de visage, estimation de la position de la tête, localisation de repères et reconnaissance du genre. Cette modalité est plébiscitée pour ses performances. La figure 2 en montre une illustration; l'orientation tête est représentée par les axes en rouge, vert et bleu.

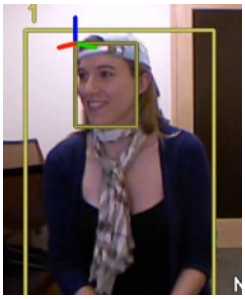

Figure 2: Exemple d'inférence par HyperFace sur la base de données ICT-3DHP [22]. Les axes R, V, B définissent l'orientation de la tête en termes de roulis, lacet et tangage respectivement.

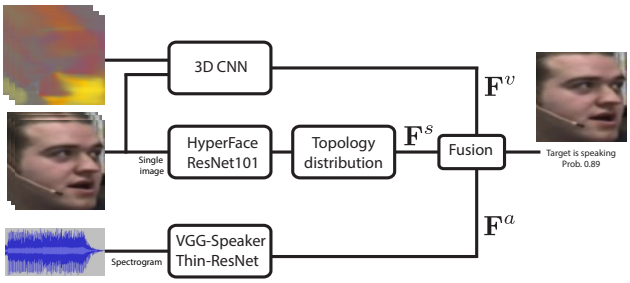

Figure 3: Synoptique de notre approche. De gauche à droite : entrées, modules de détection indépendants, fusion des percepts associés $\mathbf{F}^*$.

3 Détection multimodale du locuteur

Pour rappel, notre détection audio du locuteur est combiné avec : (1) des caractéristiques extraites des clips vidéo, et (2) des informations contextuelles provenant de l'interaction sociale des personnes. La figure 3 montre ainsi le synoptique de notre approche. Nous privilégions le détecteur audio de Xie et al. [8] pour ses performances sur la base de données VoxCeleb. Concernant la vidéo, nous évaluons différents réseaux 3D qui extraient des informations spatio-temporelles sur des clips. Pour l'orientation de la tête, nous privilégions HyperFace [21] comme évoqué précédemment.

3.1 Caractéristiques audio

Nous utilisons l'implémentation proposée dans [8], appelée VGG-Speaker-Recognition framework [3]. Les spectrogrammes sont calculés à partir de clips audio de 2,5 secondes, sur 256 fréquences. Le spectrogramme est normalisé en soustrayant la moyenne et en divisant par

[3] https://github.com/WeidiXie/VGG-Speaker-Recognition

l'écart type. Le réseau "thinResNet" est entraîné avec l'optimiseur Adam et un taux d'apprentissage initial de $1-e3$. Dans notre cas, l'évaluation est faite avec un audio à une fréquence d'échantillonnage de 40kHz. Ce réseau produit un vecteur de caractéristiques audio f^a qui est comparé à une banque de vecteurs de caractéristiques (f_i^a), préalablement calculés pour chaque participant i. Ainsi, on obtient un vecteur $\mathbf{F}^a$ qui indique la probabilité que l'audio enregistré appartienne à une personne.

3.2 Caractéristiques visuelles

Paramètres - Nous considérons quatre architectures de réseau : un CNN en 2D et trois variantes CNN 3D qui s'appuient sur la concaténation de 16 images consécutives en clip vidéo sans chevauchement. Le clip d'entrée a une taille de $3 \times L \times H \times W$, où H et W sont la hauteur et la largeur de l'image dans les 3 canaux RVB et $L = 16$ est la taille du clip.

Réseau ResNet2D - Nous privilégions ici le réseau ResNet50; il sera utilisé comme *benchmark*. Pour cette architecture 2D, les informations spatio-temporelles peuvent être considérées en traitant les trames L comme les canaux d'une même image. Par conséquent, la dimension d'entrée de ResNet2D est de $3 \times L \times H \times W$, formant un tenseur tridimensionnel. Dans ce cas, la taille de l'image est fixée à $H = 224$ et $W = 224$. Dans ce réseau, la convolution est appliquée sur la dimension spatiale et la première couche réduit les informations temporelles en cartes de caractéristiques 2D. Les couches suivantes ne tiennent donc pas compte de l'aspect temporel.

Réseau C3D - Contrairement aux réseaux 2D, le CNN 3D peut traiter l'information spatio-temporelle à toutes les couches. Ce réseau [17] a cinq couches de convolution, suivies d'un *pooling*, deux couches denses puis finalement un softmax. La taille d'entrée est de $3 \times 16 \times 112 \times 112$. En comparaison avec ResNet2D, nous avons ici un tenseur à quatre dimensions. Nous pouvons observer que l'architecture a une structure classique, sauf que les convolutions sont appliquées en 3D, ce qui encapsule la temporalité. Ainsi, les couches de convolution créent des caractéristiques en 3D où la partie initiale se concentre sur les caractéristiques spatiales des premières images et le reste considère le mouvement image saillant [17].

Réseau ResNet3D - Cette architecture reprend l'idée des blocs résiduels ResNet mais en utilisant une convolution 3D et non 2D. Cela permet de préserver et de propager l'information temporelle à travers les couches du réseau. L'entrée du réseau est un tenseur 4D comme C3D mais la taille de l'image est la même que celle de ResNet2D : $3 \times 16 \times 224 \times 224$. Nous évaluons deux versions de ResNet3D, resp. avec 18 et 34 blocs.

Tous les réseaux précités sont entraînés avec des imagettes RVB de visages préalablement segmentés. A l'instar de [12, 19, 18], et comme évoqué, nous utilisons en complément le flot optique (FO) intrinsèque à chaque clip vidéo. La figure 4 illustre les images utilisées en entrée des réseaux. Nous privilégions ici une architecture Bi-CNN avec deux réseaux appris sur une information visuelle, resp. pour les images RVB et pour les images de FO. Nous tronquons la dernière couche de chaque réseau afin de concaténer les deux caractéristiques visuelles. Enfin, nous formons deux couches entièrement connectées et une fonction softmax.

La sortie de ces réseaux indique la probabilité que le participant i soit locuteur ou non. Si l'entrée est un vecteur clip, alors le réseau donne un vecteur visuel $\mathbf{F}^v$ qui représente la probabilité locuteur pour chaque participant.

Figure 4: Exemple d'images utilisées pour l'apprentissage des CNN. Image brute RVB (gauche), image de l'amplitude du flot optique (droite).

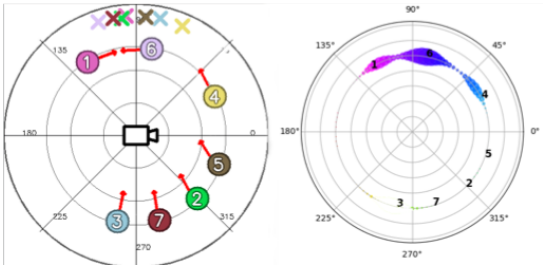

Figure 5: Gauche : Exemple de l'espace topologique d'une réunion (un cercle = un participant). La position angulaire d'un participant est la position du point de vue de la caméra i.e. le côté gauche est à $0°$ et l'extrémité droite à $360°$. Droite : distribution de probabilité des locuteurs.

3.3 Caractéristique sociale

La figure 6 illustre un scénario type de réunion avec plusieurs locuteurs potentiels. Ici, quatre caméras sont placées au centre de la table et observent chacune un participant. Nous obtenons une vue de la salle à $360°$ en concaténant les quatre images des caméras, Cf. figure 7. La configuration spatiale connue de la caméra est projetée dans un espace topologique similaire à celui de la figure 5-gauche. Dans cet exemple, sept participants sont représentés sous forme de cercles et leur position correspond à la position relative du participant par rapport à la caméra dans la vue de $360°$ i.e. la bordure gauche (resp. droite) représente $0°$ (resp. $360°$). Les flèches symbolisent les orientations de tête des participants et les croix sont la projection de

l'orientation estimée. Ces projections permettent alors le calcul d'une distribution de probabilité représentant la détection du locuteur eu égard aux regards des participants.
Pour estimer l'orientation de la tête, nous utilisons la modalité HyperFace [21] entraînée sur la base de données AFLW [23]. HyperFace calcule la pose de la tête d'un participant i comme suit : $\theta_i^{HF} = \{\theta_i^x, \theta_i^y, \theta_i^z\}$, où chaque angle se situe dans la plage $\theta_i^* = [-1, 1]$. Supposons que c_i est la position centrale sur l'image de la $i^{ème}$ personne en considérant uniquement l'axe horizontal. On projette alors sur l'espace topologique par :

$$\mathbf{C}_i = 2\pi * \frac{c_i}{I_w},$$

où I_w est la largeur de l'image 360°. Puis nous la retournons selon l'orientation de la pose estimée. Seul l'angle θ_i^z est alors considéré car il modélise les rotations gauche-droite de la tête. Le regard projeté est donc défini par :

$$\mathbf{G}_i = \mathbf{C}_i + \pi * (1 - 2 * \theta_i^z),$$

On peut observer que la rotation se fait proportionnellement à la pose de la tête. Si $\theta_i^z = 0$, alors la cible se trouve face à la caméra. Cela signifie que le regard est à la position opposée de $\mathbf{C}_i$. Même si $\theta_i^z = [-1, 1]$, la valeur réelle ne dépassera jamais $[-0,5, 0, 5]$ en raison des contraintes physiques du mouvement de la tête. Nous doublons alors la valeur de θ_i^z permettant de couvrir toutes les orientations possibles. Nous transformons le point de vue $\mathbf{G}_i$ en probabilité en utilisant la distribution de von Mises :

$$f_i(x|\mathbf{G}_i, \kappa) = \frac{\exp^{\kappa \cos(x - \mathbf{G})}}{2 * \pi \mathbf{I}_0(\kappa)},$$

où $\mathbf{I}_0(\kappa)$ est la fonction de Bessel d'ordre 0 et $\kappa = 10$ mesure la concentration. La distribution finale est alors calculée comme la moyenne de la distribution de chaque participant. Après normalisation, on obtient une probabilité similaire à celle de la figure 5-droite. Nous observons que la distribution est concentrée principalement sur les personnes 1, 4 et 6; ce dernier est le locuteur le plus probable. En extrayant la probabilité exacte de chaque participant, nous obtenons le vecteur $\mathbf{F}^s$.

3.4 Fusion des caractéristiques

Les étapes précédentes permettent, pour chacune des caractéristiques, de calculer des vecteurs de probabilité de locuteur pour chaque participant. Une stratégie de fusion naïve consiste à faire la moyenne de ces vecteurs. Cependant, la défaillance d'une modalité induit l'échec systématique du système. L'alternative est de fusionner les probabilités via une heuristique :

$$\mathbf{F}_i = \begin{cases} (\mathbf{F}_i^a + \mathbf{F}_i^v + \mathbf{F}_i^s)/3, & \text{if } \forall \mathbf{F}_i^* > 0.5 \text{ or} \\ & \forall \mathbf{F}_i^* < 0.5 \\ (\mathbf{F}_i^\alpha + \mathbf{F}_i^\beta)/2, & \text{if } \forall \mathbf{F}_i^\alpha > 0.5 \text{ and } \mathbf{F}_i^\beta > 0.5| \\ & \alpha = \{a, v, s\}, \beta = \{a, v, s\} \\ & \text{and } \alpha \neq \beta \end{cases}$$

Figure 6: Exemple de scénario/réunion du corpus AMI [24]. Quatre caméras, une caméra focalisée sur chaque participant.

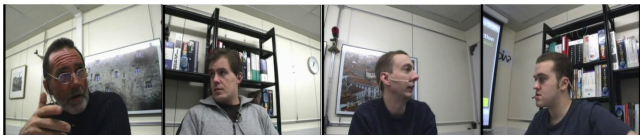

Figure 7: Exemple d'image 360° générée par la concaténation des quatre caméras chacune enregistrant un seul participant.

4 Evaluations

Jeu de données - Nous évaluons notre détecteur de locuteur sur le corpus publique AMI (pour *Augmented Multiparty Interaction*) [24]. Il est composé de 100h de séquences pré-enregistrées en contexte de réunion. Nos évaluations se bornent au sous-scénario IDAIP qui agrège 38 réunions pré-enregistrées et mettant en jeu pour chacune quatre participants, Cf. figure 6. La scène est observée par quatre caméras, chacune étant focalisée à un participant. La figure 7 illustre les quatre champs de vue des caméras. Le corpus AMI corpus dispose de la vérité terrain (GT) pour le canal audio seulement, donc rien sur la vidéo.

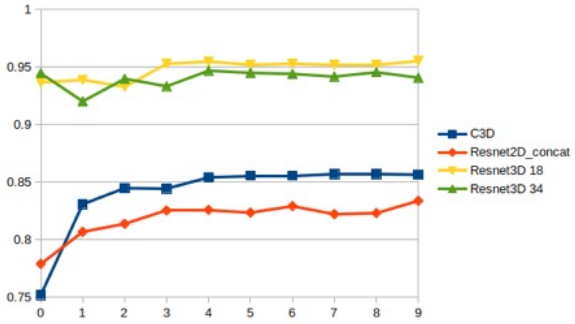

Figure 8: Apprentissage de notre détecteur de locuteur : justesse vs. nombre d'époques.

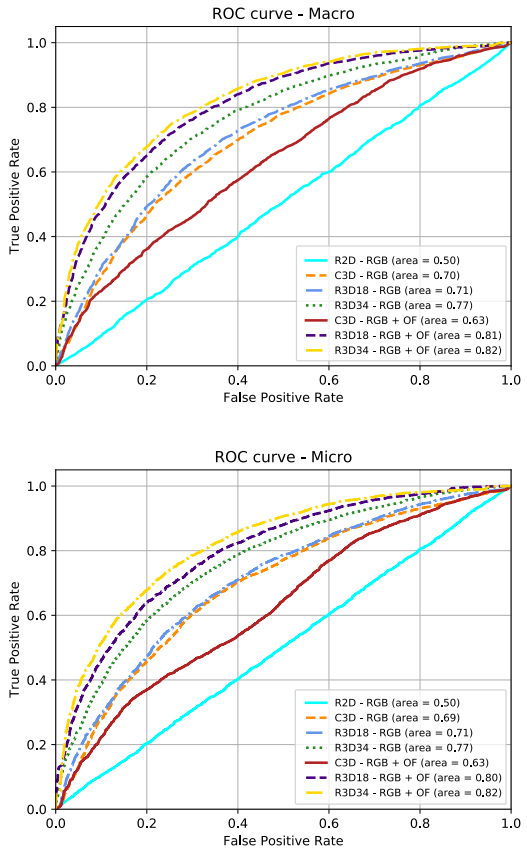

Figure 9: Critères macro- et micro- justesse, courbes ROC sur le *fold* CV2.

Ce corpus permet tout d'abord l'apprentissage et l'évaluation des performances des réseaux CNN visuels en s'appuyant sur une méthode de validation croisée. Les séquences sont divisées en 5 *folds*, 4 (resp. 1) *folds* sont dévolus à l'apprentissage (resp. aux tests). Ces *folds* sont dénommés *CVi*, i indexant le *fold* utilisé pour le test. Ainsi, CV1 exploite le *fold* 1 pour tester et les autres pour apprendre, idem pour CV2 à CV5.

Les réseaux CNN sont appris à partir des boites englobantes (imagettes RGB) de visages préalablement segmentés et du flot optique (FO) inhérent au flux vidéo associé. Chaque visage est segmenté via le détecteur de visage ResNetSSD (CAFFE) tandis que le flot optique est inféré par l'algorithme de Farneback; ces deux outils sont impléméntés sous OpenCV. On constitue alors des clips vidéo de 16 imagettes; ceux-ci sont alors labellisés en locuteur/non locuteur grâce à la vérité terrain audio. Les imagettes sont alors ré-échantillonnées en 224×224 pixels, résolution requise pour le réseau ResNet. Pour le réseau C3D, les imagettes sont ré-échantillonnées en 11×112 pixels. Chaque séquence durant plusieurs heures, nous nous limitons à des échantillons de 100 clips par personne et par classe. Chaque classe est, au final, représentée par 15000 clips.

Apprentissage - Les réseaux CNN visuels sont entraînés avec des *batchs* de 20 clips, un optimiseur Adam, enfin un taux d'apprentissage initialisé à 0.003, à l'instar de [17], et divisé par 10 tous les quatre époques. L'apprentissage est stoppé après 10 époques.

Evaluations - Nous comparons tout d'abord les performances des architectures CNN. La figure 8 illustre le processus d'apprentissage en évaluant le critère précision (axe vertical) vs. le nombre d'époques (axe horizontal). Cette évaluation préliminaire est réalisée en validation croisée à partir de CV2. Tous les batchs modulo le $n^\circ 2$ sont donc exploités pour l'apprentissage; ce batch étant représentatif de l'ensemble des séquences. On note que le réseau 2D ResNet, qui concatène le clip vidéo en une image, est le moins performant. A contrario, les réseaux 3D atteignent des performances proches de 95%.
La figure 9 illustre les performances des divers réseaux sur le jeu de tests du lot $n^\circ 2$. Nous privilégions les courbes ROC pour métrique d'évaluation afin de caractériser les performances de classification. Pour rappel, plus le taux de vrai positif est élevé, plus les performances sont excellentes. Ces évaluations confirment le bon comportement des réseaux CNN 3D, notamment de réseaux profonds tels que ResNet3D (blocs de 18 ou 34). L'ajout du flot optique ne bonifie pas le réseau C3D contrairement au réseau ResNet3D dont les performances associées augmentent de 5%. Notons enfin que les performances sur critères micro- ou macro-justesse sont similaires car le nombre d'échantillons par classe est plutôt équilibré.
Les tableaux 1 et 2 illustrent les résultats en considérant tous les lots. La métrique privilégiée ici est l'aire sous la courbe ROC; elle quantifie les performances en terme de classification binaire. On note qur le flot optique induit des gains mais seulement pour certains lots. A contrario, les performances avec RestNet3D sont améliorées quelque soit le lot, pour certains de 10%.
Les évaluations précédentes sont réalisées sur les échantillons de test donc pré-enregistrés. En complément, nous évaluons sur des séquences *live* donc image par image à la volée. Pour chaque flux vidéo, nous avons choisi aléatoirement de limiter notre analyse aux images acquises sur l'intervalle de temps $[3, 18]$ min.
Les résultats sont synthétisés dans le tableau 3, chaque colonne indique la caractéristique considérée pour la classification e.g la colonne 2 montre les résultats avec le seul percept audio pour détecter le locuteur.
On observe que le percept audio est logiquement le plus discriminant des trois pour notre détecteur. Les deux percepts visuels impactent moindrement, mais de façon similaire, les performances. La dernière colonne est relative à la fusion des trois percepts. Les gains observés sont notables. Comme attendu, la vidéo apporte des informations complémentaires que le canal audio seul ne peut capturer.

Une vidéo illustrant le comportement qualitatif de notre

Table 1: Aire sous la courbe (AUC) du critère macro-justesse pour tous les *folds* du corpus AMI pour les réseaux CNN 3D.

	C3D		ResNet3D-18		ResNet3D-34	
Fold	RVB	RVB-FO	RVB	RVB-FO	RVB	RVB-FO
CV1	0.5	0.55	0.7	0.75	0.7	0.78
CV2	0.7	0.63	0.71	0.81	0.77	0.82
CV3	0.65	0.5	0.79	0.82	0.79	0.85
CV4	0.5	0.77	0.76	0.85	0.78	0.84
CV5	0.67	0.5	0.68	0.76	0.68	0.76
Mean	0.60	0.59	0.73	0.80	0.74	0.81

Table 2: Aire sous la courbe (AUC) du critère micro-justesse pour tous les *folds* sur le corpus AMI pour les réseaux CNN 3D.

	C3D		ResNet3D-18		ResNet3D-34	
Fold	RGB	RVB-FO	RVB	RVB-OF	RVB	RGB-FO
CV1	0.48	0.55	0.69	0.75	0.7	0.78
CV2	0.69	0.63	0.71	0.8	0.77	0.82
CV3	0.64	0.5	0.76	0.82	0.79	0.85
CV4	0.5	0.73	0.76	0.84	0.77	0.84
CV5	0.64	0.5	0.68	0.75	0.68	0.76
Mean	0.59	0.59	0.72	0.79	0.74	0.81

Table 3: Evaluation de notre système sur les séquences du corpus AMI.

	Audio Caract.	Social Caract. ResNet3D-34	Visuel Caract.	Fusion Proba.
Macro	0.79	0.66	0.7	0.84
Micro	0.78	0.68	0.67	0.84

système sur une séquence AMI est accessible via le lien URL : https://drive.google.com/file/d/1b7n0NHVu342BBP2DHbS-hV7N2fIJy5w3/view?usp=sharing

5 Conclusion et perspectives

Ces travaux portent sur la détection des locuteurs successifs dans un contexte de réunion à partir de trois percepts combinés. Le premier percept est relatif à l'audio; il permet de détecter les orateurs à partir de leur activité vocale. Le second percept porte ainsi sur la détection visuelle de locuteur à partir de détection préalable du visage. Nous privilégions ici un réseau convolutionnel 3D qui prend en entrée un clip vidéo de 16 images successives RVB mais aussi le flot optique intrinsèque à ce clip. Il est alors montré que l'ajout de cette caractéristique manuelle (*hand crafted*) induit des gains.
Le dernier percept fusionné est original; il s'appuie sur le comportement social de participants en réunion. Il vise à inférer, dans le flux vidéo, leurs orientations de visages sachant que ceux-ci sont majoritairement dirigés vers l'orateur courant. La fusion probabiliste de ces trois percepts est concluante et montre des gains dans ce contexte applicatif.
Les travaux futurs consistent à évaluer des stratégies autres de fusion tardive, e.g. basée théorie de l'évidence, afin de pallier encore plus efficacement les erreurs de l'une ou l'autre des modalités.

Remerciements

Ces travaux menés au LAAS-CNRS sont financés par BPI France dans le cadre du projet LinTO associé à l'appel à projets PIA3.

References

[1] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2018, pp. 74–79.

[2] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Communication*, vol. 99, pp. 62 – 79, 2018.

[3] J. Bonastre, X. Anguera, G. Sierra, and P. Bousquet, "Speaker modeling using local binary decisions," in *Interspeech*, 2011.

[4] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first DIHARD challenge," in *Conf. of the International Speech Communication Association, September 2-6, 2018, Hyderabad, India*, Hyderabad, INDIA, 09 2018.

[5] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *2017 IEEE International Con-*

ference on Acoustics, Speech and Signal Processing (ICASSP), March 2017, pp. 4945–4949.

[6] A. Sarkar, S. Dasgupta, S. K. Naskar, and S. Bandyopadhyay, "Says who? deep learning models for joint speech recognition, segmentation and diarization," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5229–5233.

[7] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5239–5243.

[8] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[9] A. Nagrani, S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild." *Computer Speech Language*, 2019.

[10] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590 – 605, 2014.

[11] X. Hong, H. Yao, Y. Wan, and R. Chen, "A pca based visual dct feature extraction method for lip-reading," in *2006 Int. Conf. on Intelligent Information Hiding and Multimedia*, Dec 2006, pp. 321–326.

[12] N. Le and J.-M. Odobez, "Learning multimodal temporal representation for dubbing detection in broadcast media," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 202–206.

[13] P. Korshunov, M. Halstead, D. Castan, M. Graciarena, M. McLaren, B. Burns, A. Lawson, and S. Marcel, "Tampered speaker inconsistency detection with phonetically aware audio-visual features," in *Proceedings of the International Conference on Machine Learning*, 2019.

[14] F. Tao and C. Busso, "End-to-end audiovisual speech activity detection with bimodal recurrent neural models," *Speech Communication*, vol. 113, 2019.

[15] D. C. M. P. S. Petridis, J. Shen, "Visual-only recognition of normal, whispered and silent speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[16] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE Int. Conf. on Computer Vision (ICCV)*, December 2015.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Computer Vision and Pattern Recognition*, 2018.

[19] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] G. Liu, Y. Yu, K. A. Funes Mora, and J. Odobez, "A differential approach for gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[21] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, Jan 2019.

[22] T. Baltrušaitis, P. Robinson, and L. P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *Conf. on Computer Vision and Pattern Recognition*, June 2012, pp. 2610–2617.

[23] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 2144–2151.

[24] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 2, pp. 181–190, 2007.