

Signal, Apprentissage et Multimédia
Projet - Prédiction des tours de parole

Eliot Maës

Université Aix-Marseille

2023

Contact : eliot.maes@univ-amu.fr

Objectifs

Objectifs du projet :

- Manipuler différentes modalités et combiner les informations que chaque modalité apporte
- Explorer différents types d'architectures et fusions (early / late / quels features ?)
- Justifier des différents choix effectués lors du projet

Modalités d'évaluation : **code + rapport** de 5 pages par groupes de $\simeq 3$, à remettre pour le **13/01/2024***

Corpus Paco-Cheese

Dataset

- Multimodal corpus (audio + video)
- en français
- 26 dyades de 15-20min

Annotations

- Segmentation de la parole basée sur les silences (**IPUs**)
- Transcription manuelle alignée sur l'audio

Dataset <https://amubox.univ-amu.fr/s/gkfA7rZCWGQFqif>



#	u	i	#	e	@	s	grou	s	O~	t	s	f	re	fp	ll	R	k	s	s	a	d
#	oui	#	sur	le	su	grou		sont	tous	ce	il	mett	ieu	euh	leur	cou	rs	s	sans	dire	es
adjectiv	c		pr	det	pr	noun	auxilia	pron	un	u	o	ent	rs	leur	leur	de	ter	m	pron	nous	te
			on	er	on	ry	pro	o	la	o	o	aux	det	deter	pou	mine	n	pron	nous	un	
S1															S0						
bah après l'année dernière aussi i(l)s aidaient hein les gens + j(t) trouve qu'o																					#
#	oui	#	sur le sur le groupe i(l)s sont tous euh ils mettent leurs euh leurs cours s- sans dire										e est-ce que vous les voulez et tout alors que								#
transition											la solidarité des étudiants entre eux										

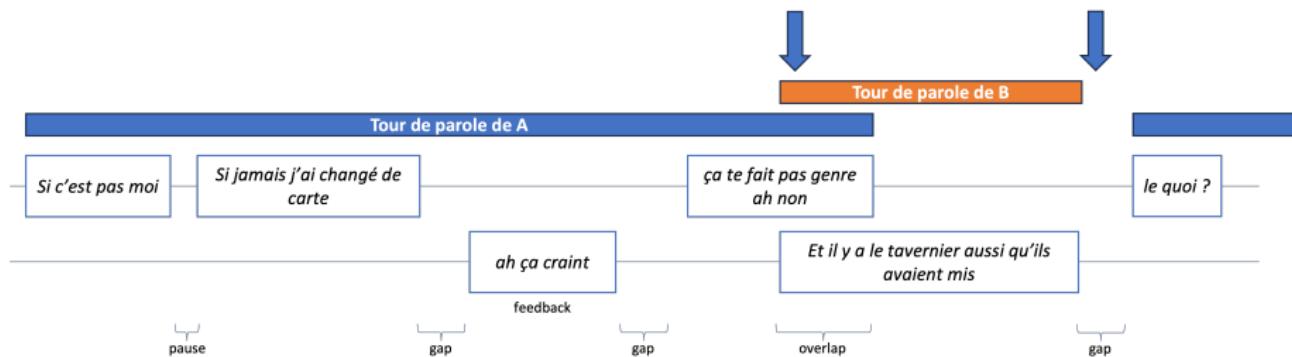
References : Priego-Valverde, Bigi et Amoyal 2020 ; Amoyal, Priego-Valverde et Rauzy 2020

Turn Taking Prediction

Prédiction des tours de parole

Prédire, en conversation, quand un locuteur va laisser la parole à l'autre.

Différents cas de figure peuvent apparaître :



Exemples d'applications : prédiction pour les chatbots du bon moment pour reprendre la parole

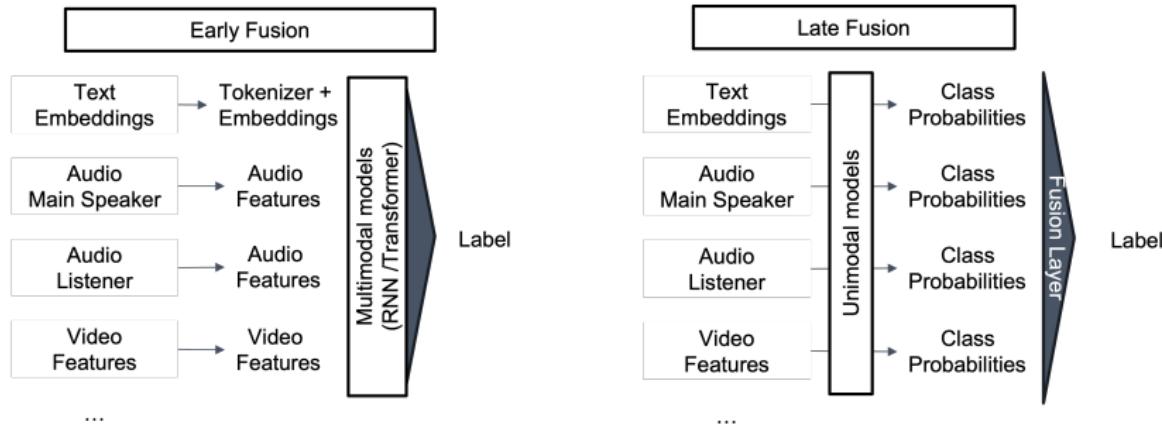
Poser la question de recherche

Pour ce genre de problèmes, différentes questions de recherche se posent :

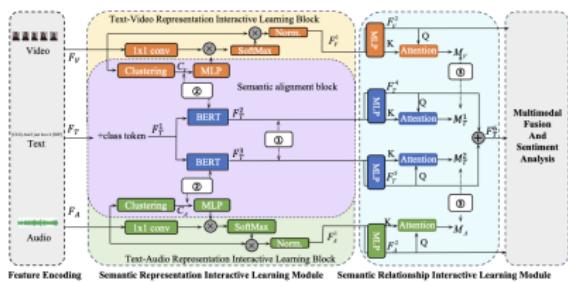
- Prédiction de différents labels (plus ou moins compliquée) : fin de l'IPU, fin d'un tour de parole, locuteur laissant la place pour l'autre ou requérant la parole
- Quelle échelle temporelle pour la prédiction ?
 - En entrée : après la fin de chaque mot | IPU | fenêtre temporelle (par exemple 50ms)
 - En sortie : changement de tour dans une fenêtre à venir (50ms, 500ms, 1s, entre 1s et 2s plus tard...)
- Quels features en entrée ?

text	audio	video
Embeddings (BERT...)	Deep Features (wav2vec2)	Patches
Features extraits (PoS...)	Prosodie, MFCC...	Facial Features (OpenFace...)

Architectures



Architectures



(a) Huang et al (2023)

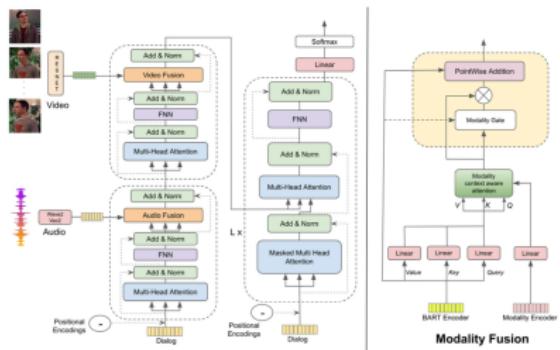


Figure 2: (a) Overall architecture of the MO-Sarcasm model (left side), (b) Architecture of Modality Fusion Network. Here audio and video fusion are done in different layers of the same encoder (right side)

(b) Tomar et al (2023)

Figure – Exemples d’architectures de fusion basées sur des Transformers

Architectures

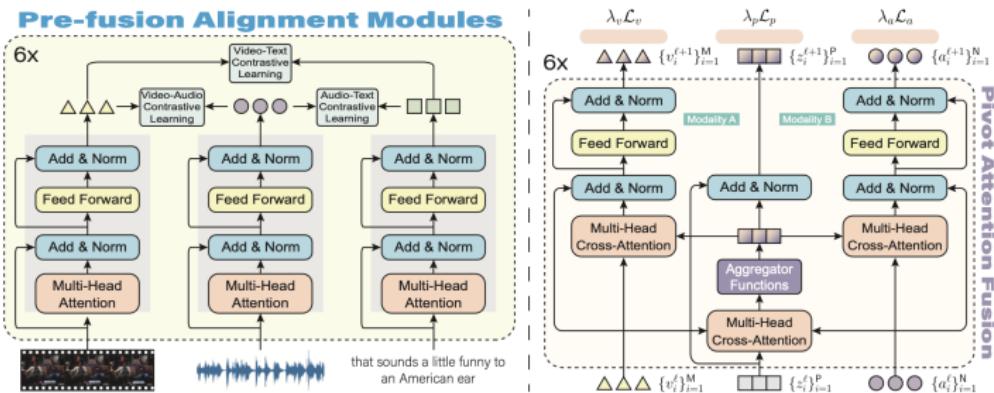


Figure – Exemples d’architectures de fusion basées sur des Transformers (Zhang et al (2023))

Télécharger les données



Télécharger les données en CLI :

- Télécharger l'archive complète :

```
curl https://amubox.univ-amu.fr/s/  
gkfA7rZCWGQFqif/download -output archive.zip
```

- Télécharger un fichier en fonction de son chemin :

```
...download?path=%2Fvideo%2F&files=audio.zip
```

- Télécharger plusieurs fichiers :

```
...files=%5B%22audio.zip%22%2C%22transcr.zip%22%5D
```

Organisation des données

Sur AMUBox :

- README avec l'organisation des dossiers
- Audio (brut) et videos (brut / OpenFace)
- par dyade : un fichier CSV par IPU / mots avec les labels

ipu_id	speaker	start	stop	text	is_main Speaker	turn_at_start	turn_after	yield_at_end	request_at_start
420	PO	643.47	645.26	ça complète ton truc de droit c'est bien	True	False	True	True	False
421	MH	645.44	647.55	ah bah ouais du coup ça me permet aussi genre enfin	True	True	False	False	True
422	MH	648.10	649.89	plutôt l'an dernier parce que on avait	True	False	False	False	False
423	MH	650.11	650.73	justement	True	False	False	False	False
424	MH	651.62	654.16	des des cours en pénal là c'est au à ce semestre on en a plus	True	False	False	False	False
speaker	ipu_id			text_ipu	start_words	stop_words	text_words	request_after_word	is_ipu_end
MH	207	on y pense pas souvent qu'i y a deux genre versants à ouais		328.40	329.00	versants		False	False
MH	207	on y pense pas souvent qu'i y a deux genre versants à ouais		329.00	329.28	à		False	False
MH	207	on y pense pas souvent qu'i y a deux genre versants à ouais		329.28	329.60	ouais		False	True
PO	208	et ouais ça rejoint au final ouais ouais bien sûr		328.27	328.42	et		False	False
PO	208	et ouais ça rejoint au final ouais ouais bien sûr		328.42	328.53	ouais		False	False

Pour la classification, attention aux % des classes et à la manière / quantité de données chargées (le chargement dynamique ralentit le calcul mais peut aider à ne pas surcharger la RAM)

ELAN format in Python

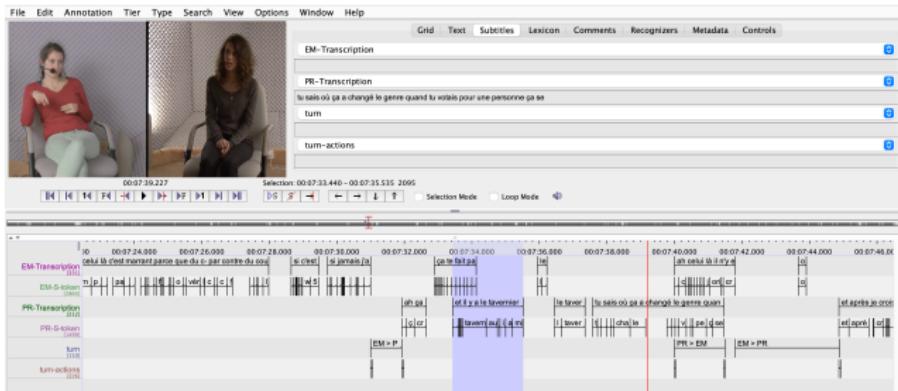


Figure – Utiliser ELAN pour aller regarder dans les données

```
from speach import elan
eaf = elan.read_eaf(file)
dial = pd.DataFrame(eaf.to_csv_rows(),
    columns=[ 'tier', '?', 'start', 'stop', 'duration', 'text' ])
```
