Architectures neuronales bout-en-bout pour la compréhension de la parole

Valentin Pelloin[1] Nathalie Camelin[1] Antoine Laurent[1]
Renato De Mori[2,3] Sylvain Meignier[1]

[1]LIUM, Le Mans Université, 72085 Le Mans, France [2]LIA, Avignon Université, 84029 Avignon, France
[3]Université de McGill, Montréal H3A 0G4, Canada

Introduction

- **Compréhension de la parole** dans le cadre de dialogues homme-machine, en français
- Systèmes **bout-en-bout** ou **cascade** :
  - systèmes neuronaux bout-en-bout (*end-to-end)*
    - ↪ pas de propagation des erreurs ASR
  - systèmes cascade (ASR + NLU)
    - ↪ informations linguistiques (mots) utilisées pour la NLU
    - ↪ le système NLU travaille sur l'ensemble de la séquence de mots (futur compris)
    - ↪ utilisation de données non-supervisées (*transformers* BERT)

→Proposition : **architecture unifiée** (bout-en-bout) **multi-décodeurs** : un décodeur par modalité (mots, SLU mots, SLU valeurs normalisées)

Compréhension automatique de la parole (SLU)

- extraire certaines informations sémantiques contenues dans le signal de parole

Exemple annoté *{mots ; concept ; valeur}*	{est ce qu' il y a ; - ; -} {une piscine ; hôtel-services ; *piscine*} {à ; - ; -} {l' ; lienref-coref ; *singulier*} {hôtel ; objet-bd ; *hotel*}
ASR	est ce qu' il y a une piscine à l' hôtel
SLU_{Mots}	est ce qu' il y a <hôtel-services> une piscine </> à <lienref-coref> l' </> <objet-bd> hôtel </>
SLU_{Norm}	<hôtel-services> piscine </> <lienref-coref> singulier </> <objet-bd> hotel </>

Table 1. Exemple annoté du corpus MEDIA et ses différentes représentations utilisées en sortie des décodeurs.

Architectures à n décodeur(s)

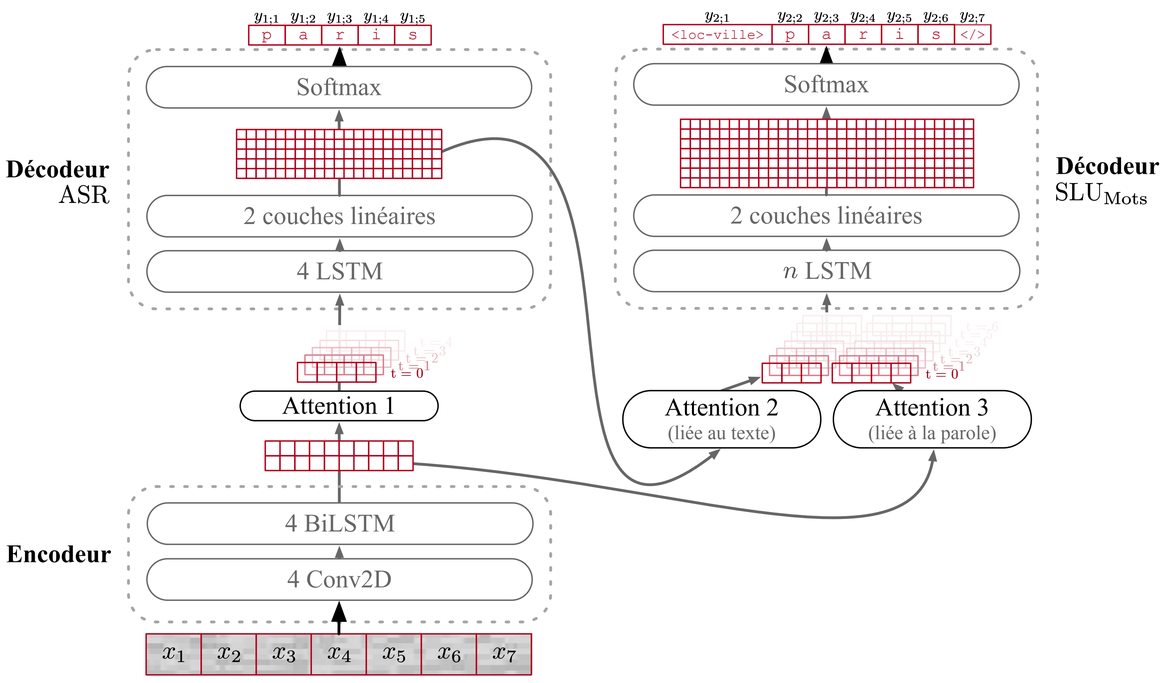

Figure 1. Architecture d'un modèle multi-décodeurs comportant deux décodeurs chaînés : un décodeur ASR et un décodeur SLU_{Mots}.

- encodeur-décodeur·s avec mécanisme·s d'attention
- **chaînage entre les décodeurs** :
  - le décodeur D_N dispose en entrée des sorties des décodeurs D_0 à D_{N-1}
  - le décodeur D_N dispose également de la sortie de l'encodeur
  - concaténation des vecteurs de contexte des mécanismes d'attention
- apprentissage en utilisant une *loss* pondérée pour chaque décodeur
- **entrées** : MelFBanks
- **sorties** : caractères + tags de concepts

Protocole expérimental

- corpus de **transcriptions ASR** :
  - 414 heures d'apprentissage : MEDIA, PortMEDIA, DECODA, EPAC, ESTER1, ESTER2, ETAPE, QUAERO, REPERE, et émissions TV
- corpus de **dialogues homme-machine** (woz) :
  - **MEDIA** : réservation de chambres d'hôtel (23 heures)
  - **PortMEDIA** : informations touristiques lors du festival d'Avignon (12 heures)

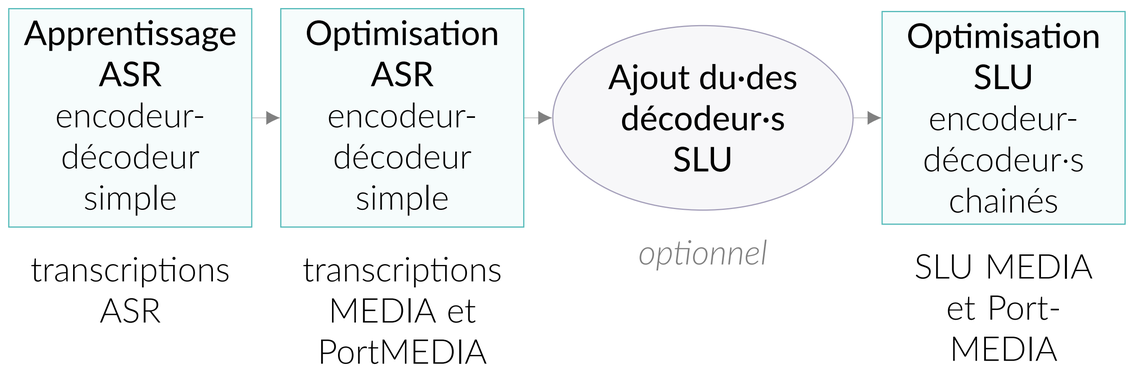

Figure 2. Processus d'apprentissage des modèles (ASR et SLU)

- **décodage en faisceau (*beam search*)**
  - chaque décodeur possède son *beam search*
  - les hypothèses du décodeur D_N sont agrégées pour le décodeur D_{N+1}
- en mode SLU_{Norm} : **traduction directe** vers les valeurs normalisées
  - aucune règle : pas d'expertise humaine requise (en dehors de l'annotation)

Résultats

	Simple-decodeur	décodage en *beam* %CER	%CVER
(a)	SLU_{Mots}	$13,6$	$18,5$
(b)	SLU_{Norm}	$15,4$	$21,6$

Table 2. Résultats sur le corpus MEDIA Test du modèle (*optimisé* sur MEDIA, **avec modèle de langage**) simple décodeur.

#		Multi-décodeurs		décodage en *beam* %CER	%CVER
1	(c)	SLU_{Mots}		$16,74$	$21,51$
2	(d)	ASR	SLU_{Mots}	$16,58$	$22,70$
	(e)	ASR	SLU_{Norm}	$18,98$	$26,00$
3	(f)	ASR	SLU_{Mots} SLU_{Norm}	$16,40/22,28^*$	$22,31/27,77^*$

Table 3. Résultats sur le corpus MEDIA Test de l'architecture multi-décodeurs (uniquement *optimisé* sur MEDIA + PortMEDIA, **sans modèle de langage**).

Conclusions et perspectives

- Architecture de compréhension de la parole **bout-en-bout, sans règles pour l'extraction des valeurs**
- Nouvelle architecture hybride bout-en-bout & cascade
- Nombreuses perspectives d'amélioration de l'architecture multi-décodeurs
  - intégration de modèles de langage
  - optimisation des modèles sur MEDIA et améliorations du *beam search*
  - intégration d'informations supplémentaires (parties de mots, POS, BERT, ...)

Références

[1] V. Pelloin, N. Camelin, A. Laurent, R. De Mori, A. Caubrière, Y. Estève, and S. Meignier. End2End Acoustic to Semantic Transduction. In *ICASSP 2021*, Toronto, ON, Canada, June 2021.

Projet ANR AISSPER (contrat ANR-19-CE23-0004-01)

valentin.pelloin@univ-lemans.fr