Le benchmark MEDIA revisité : données, outils et évaluation dans un contexte d'apprentissage profond

Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, Salima Mdhaffar, Nathalie Camelin, Sahar Ghannay, Bassam Jabaian, Yannick Estève

Un peu de contexte...

La **compréhension de la parole** fait référence aux tâches de traitement automatique du langage naturel liées à l'extraction d'informations sémantiques depuis le **signal de la parole**. Le jeu de données français MEDIA traite une tâche visant la compréhension de la parole par une **annotation sémantique riche et complexe**.

Le jeu de données MEDIA

→ Création lors du projet Technolangues, en 2002
→ Distribution académique gratuite par ELRA, depuis 2005

" Dialogues Humain-Machine pour réservation d'hôtel par appel téléphonique grâce à la méthode Wizard-of-Oz "

1258 dialogues -250 locuteurs

	Nb. Échantillons	Nb. Dialogues	Nb. Heures	Durée moyenne Tour de Parole
train	13.7k	727	16h56m	4.69s
dev	1.4k	79	01h40m	4.77s
test	3.8k	208	04h47m	4.89s
???	4.0k	244	05h35m	5.30s

<concept [valeur] mots-support >

"Je <tâche [réservation] voudrais réserver >
une <chambre-type [double] chambre double > ...

Relax
77 concepts

...pour <temps [12h00] midi >."

Full
159 concepts

...pour <temps-**début** [12h00] midi >."

Métriques d'évaluation

- **Concept Error Rate** : erreurs de concepts
- **Concept-Value Error Rate** : erreurs de tuples concept et valeur

$$\frac{\textit{Suppressions + Insertions + Substitutions}}{\textit{Nb. d'éléments (Concept +/- Valeur) dans la référence}}$$

r-CVER
rules-based CVER

u-CVER
unnormalized CVER

Référence : Je <tâche [réservation] voudrais réserver >...
Prédiction : Je <tâche voudrai réserver >...

→ Concept : tâche
→ Valeur normalisée : réservation

→ Concept : tâche
→ Valeur non-normalisée : *voudrai* réserver

Problèmes rencontrés

Normalisation de la Valeur pour r-CVER

- Règles établies par des humains en fonction des corpus de train et dev
- Utilisation insuffisante des règles grammaticales françaises
- **5.7%** de r-CVER sur la référence du corpus de test

Correction des annotations

- Mauvaises interprétations des concepts sémantiques par les annotateurs
- Problèmes de segmentation du signal audio

Données non-utilisées

→ Intégration d'un nouveau corpus de test à la distribution de MEDIA par ELRA

Corrections apportées

1. Normalisation de la transcription
   - ➢ Case des noms propres
   - ➢ Caractères spéciaux
   - ➢ Orthographe
   - ➢ ...
2. Annotations sémantiques
3. Informations supplémentaires
   - ➢ Indication du canal audio (Droite / Gauche) du locuteur
   - ➢ Correction de son identifiant
4. Utilisation des nouvelles données
   - ➢ Création d'un corpus nommé *test2*

Recette SpeechBrain

Préparation des données

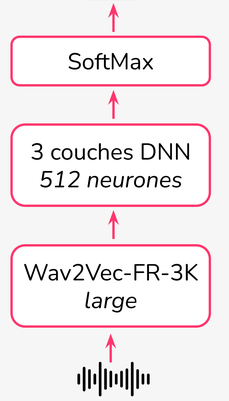
- Retrait de certains caractères spéciaux
- Retrait des traits d'union entre les nombres
- Rattachement de l'apostrophe au mot le précédent
- Respect de la plus stricte des segmentations audio
- Mise en majuscules de tous les caractères

... Conversion au format SpeechBrain

	Nb. Heures	Durée moyenne Tour de Parole
train	10h52m	2.85s
dev	01h13m	3.23s
test	03h01m	2.88s
test2	03h16m	2.94s

Architecture neuronale et Apprentissage

→ media-base : Wav2Vec-FR-3K **LeBenchmark**
→ media comvoice : Wav2Vec-FR-3K LeBenchmark fine-tuned sur 450h de parole **CommonVoice FR**



- Activation : LeakyReLU
- Optimizer : AdaDelta
- Initialisation : Aléatoire

- Optimizer : Adam
- Initialisation : Pré-apprentissage

Résultats préliminaires

	Model	test		
		CER	u-CVER	r-CVER
Relax	**media-base**	21.8	34.1	29.4
Relax	**media-comvoice**	16.3	27.7	23.7

	Model	test2			
		ChER	CER	u-CVER	r-CVER
Full	**media-comvoice**	6.7	21.1	30.9	-
Relax	**media-comvoice**	6.4	16.4	27.1	21.0

➢ Le taux d'erreur u-CVER est plus **fiable** mais plus **stricte** que le r-CVER de 4 à 5 points

➢ **media-comvoice** donne de largement meilleurs résultats que media-base

➢ la recette est opérationnelle, elle nécessite simplement quelques réglages de la part de l'utilisateur pour atteindre l'état-de-l'art

➢ le second corpus de test est **cohérent** avec les résultats du jeu de données originel