Étude de cas

Mehdi Brahmi, Elise Chin, Mathilde Da Cruz, Mohamed Rahmouni

Master IASD - Promo 2021/2022

6 octobre 2022

1 Formulation du problème

2 Analyse de l'existant (AWS)

3 SOTA

4 Solutions proposées

Formulation du problème

Problème

Beaucoup de mes clients déposent leur voitures après les heures d'ouverture, mettent les clés dans notre boîte aux lettres et laissent un message vocal pour nous dire quel est le problème. **Taper le message vocal sur l'ordinateur prend trop de temps** et je veux m'assurer que **mes techniciens ont accès à ces informations au format texte** sur leurs tablettes.

Formalisation du problème

Plusieurs manières d'appréhender le problème.

- "Speech2Keywords" ou Keywords spotting
- "Speech2Text" $\rightarrow$ "Text2Keywords"
- mots clés ? résumé ? analyse de sentiment ? classification ?

Exemple

Bonjour, j'ai eu un gros problème avec ma voiture louée !! Non seulement elle n'était pas propre lorsqu'elle m'a été remise, mais surtout, on m'a demandé de payer des frais de ménage lorsque je l'ai rendue ! C'est inadmissible !! Vos conseillers sont incompétents. En plus de cela, vos voitures sont moches. Je ne reviendrai plus.

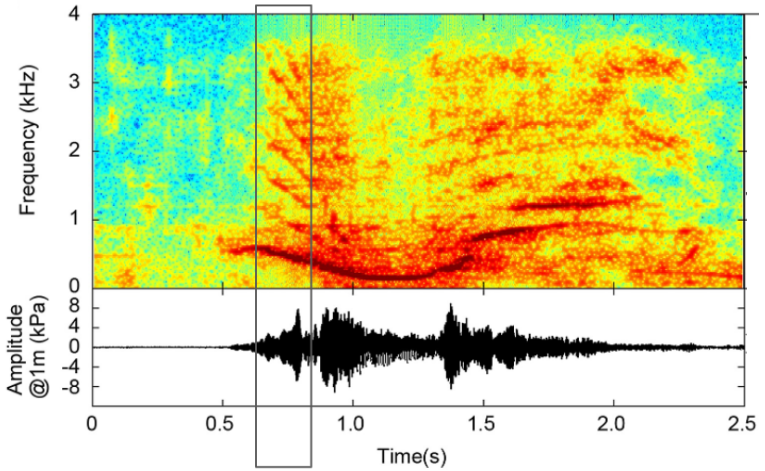
Bonjour, j'ai eu un gros problème avec ma voiture louée !! Non seulement elle n'était pas propre lorsqu'elle m'a été remise, mais surtout, on m'a demandé de payer des frais de ménage lorsque je l'ai rendue ! C'est inadmissible !! Vos conseillers sont incompétents. En plus de cela, vos voitures sont moches. Je ne reviendrai plus.

Solutions AWS

- Speech-to-text : Amazon Transcribe
  `https://aws.amazon.com/fr/transcribe/`
- Keyword extraction : Amazon Comprehend $\longrightarrow$ "Extraction de phrases clés"

Reconnaissance vocale - Définition

Soit X un signal audio d'un enregistrement vocal. On cherche f qui transforme X en une séquence des mots prononcés W.



- X : signal audio ou autres représentations (temporelle,spectrale...)
- Problème intermédiaire : W' = g(X) où W' est une séquence de phonèmes ou syllabes
  Ex : understand = AH N DER S T AE N D

Reconnaissance vocale - Évaluation

Il existe plusieurs métriques d'évaluation de modèle pour la reconnaissance vocale. La plus utilisée est le **Word Error Rate (WER)** .

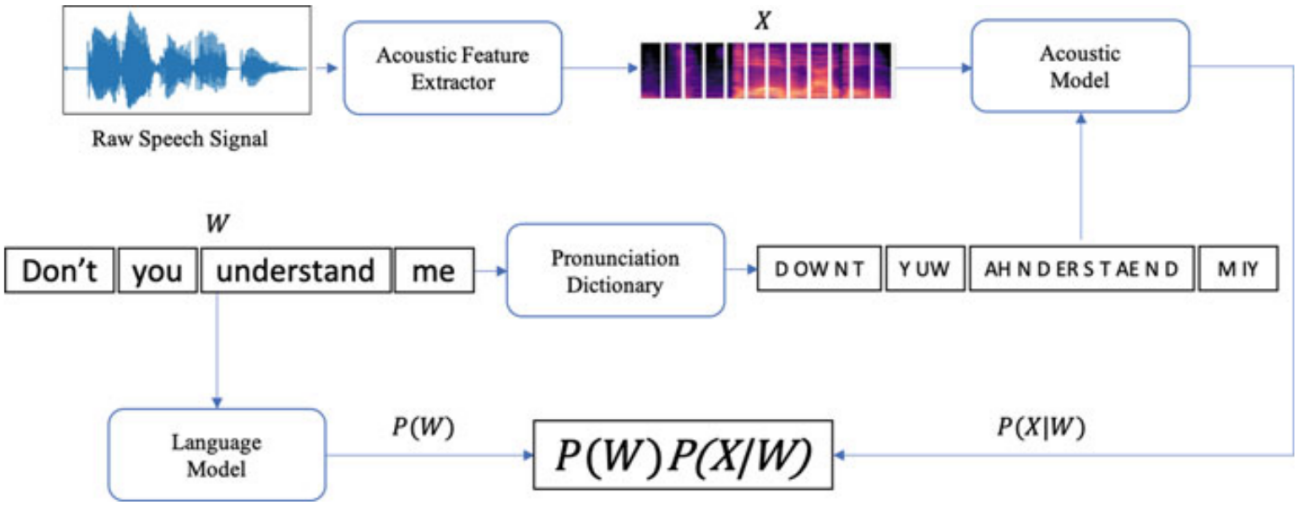
$$\text{WER} = \frac{S+D+I}{N} \text{ où } \begin{cases} S = \text{Nombre de mots remplacés} \\ D = \text{Nombre de mots supprimés} \\ I = \text{Nombre de mots ajoutés} \\ N = \text{Nombre de mots prononcés} \end{cases}$$

Exemples :
(S) "I surf small waves" $\rightarrow$ "I surf **all** waves"
(D) "I surf small waves" $\rightarrow$ "I surf waves"
(I) "I surf waves" $\rightarrow$ "I surf **small** waves"

Approche statistique

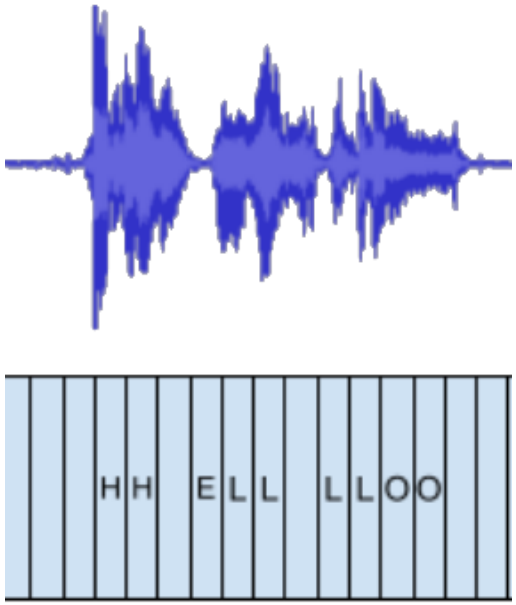
Trouver la séquence de phonèmes/syllabe W la plus probable, étant donné une séquence audio X

$$W^* = \text{argmax } P(W|X) = \text{argmax } P(X|W)P(W)$$



Alignement de séquences

Alignement de séquences audio / texte



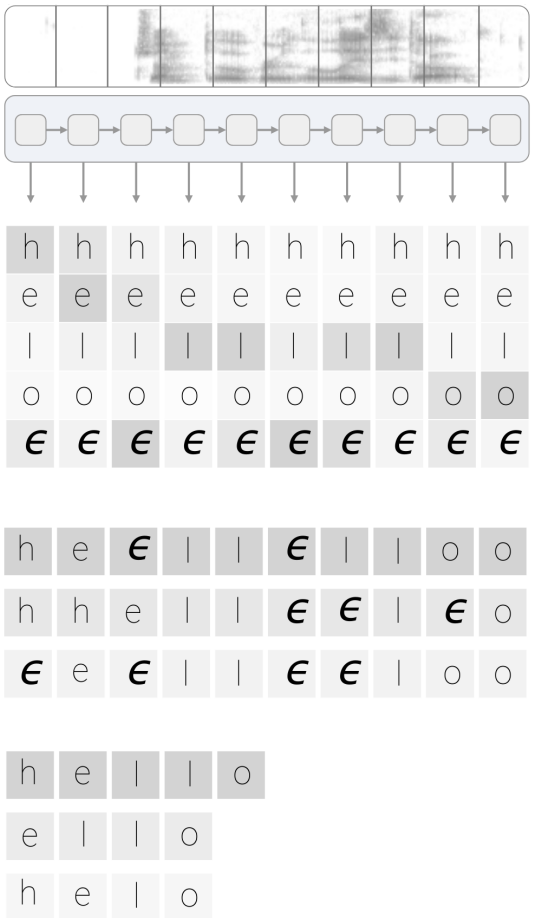
Comment aligner une séquence audio (fenêtre du signal) avec les données textuelles (phonèmes) ?

Solution naïve : regrouper les phonèmes consécutifs

Input Acoustic Features X:	x_1	x_2	x_3	x_4	x_5	x_6
Naive Alignment:	c	c	a	a	a	t
Output Y:	c		a			t

Problèmes : périodes de silence, termes avec caractères répétitifs...

Classification temporelle connectionniste (CTC)

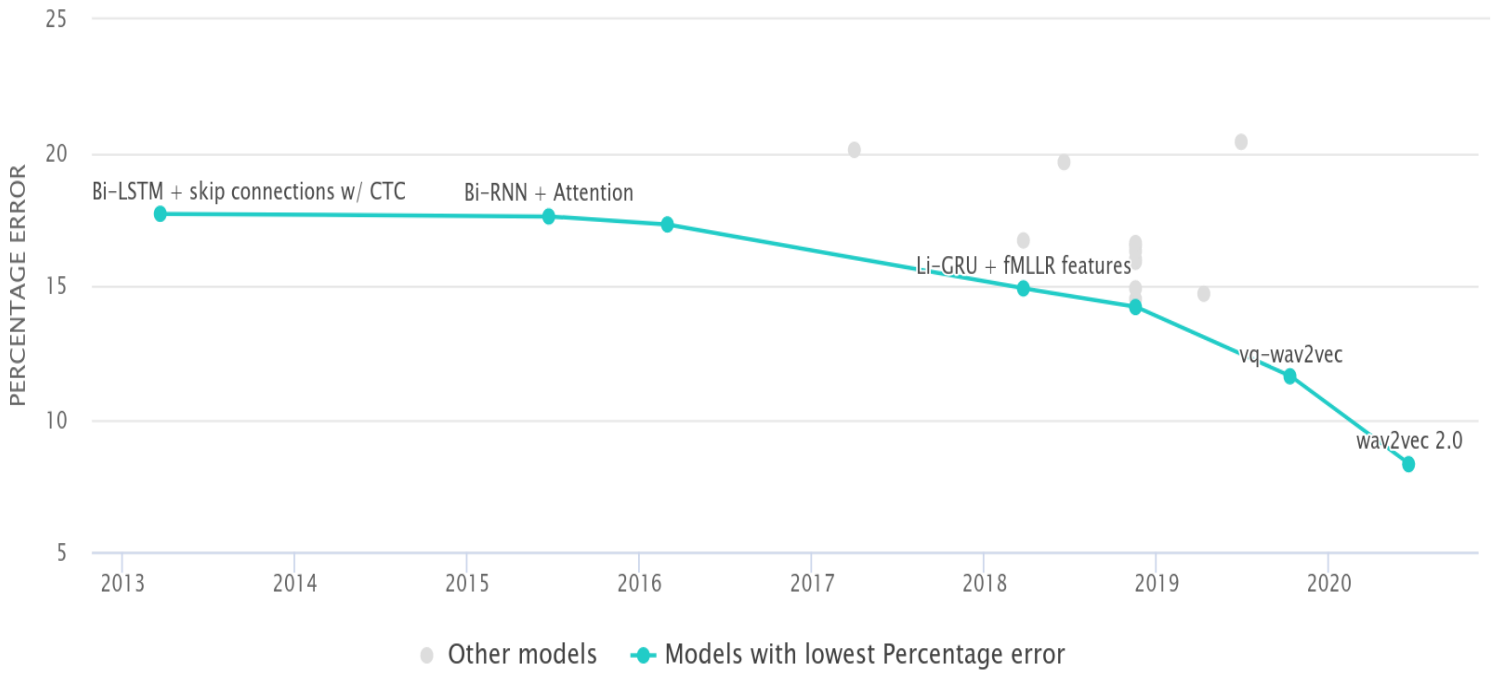


- Séquence à prédire : "hello"
- Le réseau donne $p_t(w|X)$ une distribution sur le vocabulaire de sortie pour la fenêtre t du signal audio.
- Calcul de la probabilité pour chaque séquence possible (produit des probabilités des caractères de la séquence)
- Fusion des lettres répétitives et retrait de ϵ

  $\rightarrow$ Association "many-to-one"

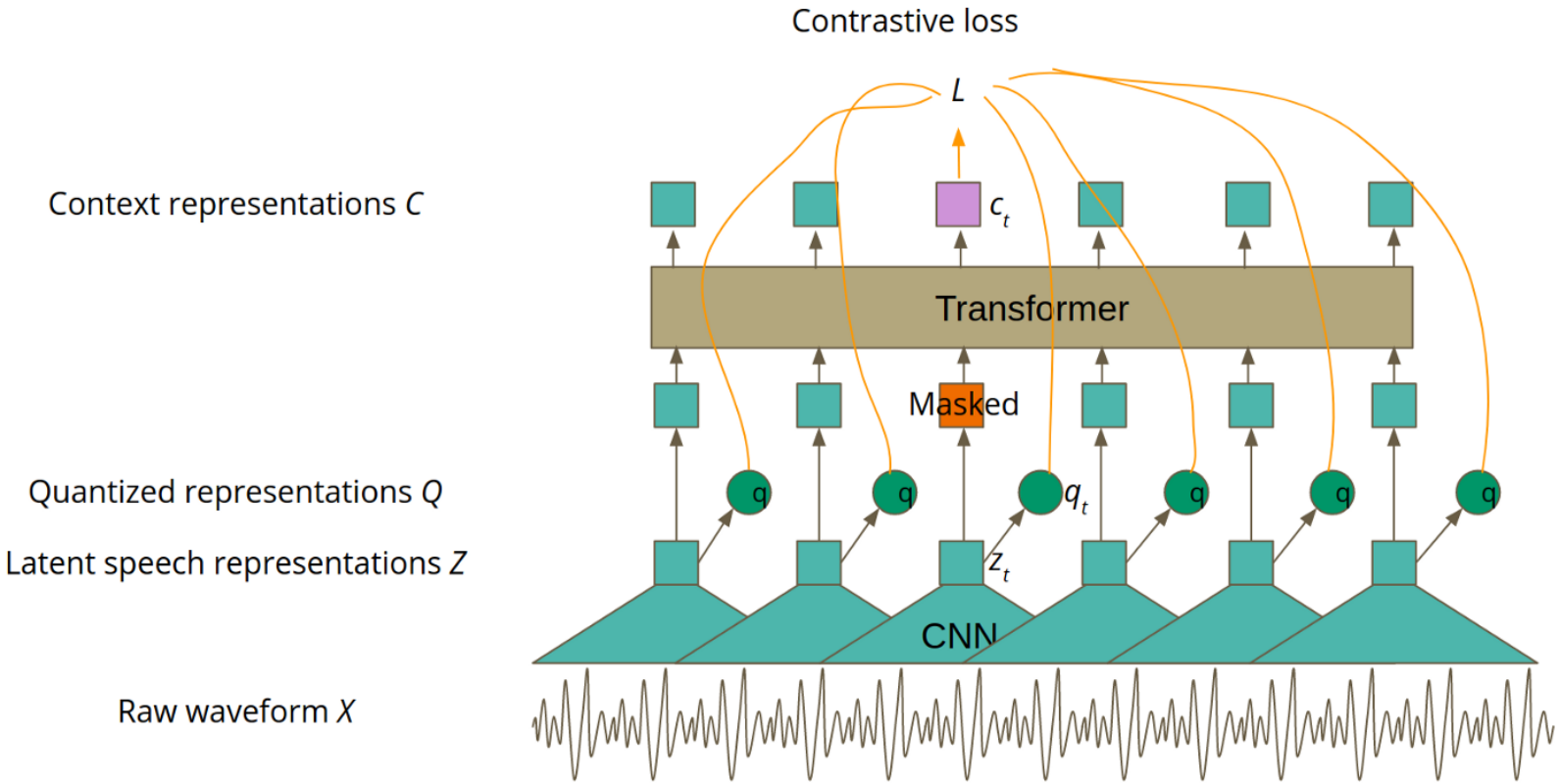
Apprentissage profond

TIMIT : corpus de discours transcrits phonétiquement et lexicalement par des des anglophones américains (326 voix d'hommes et 136 voix de femmes)



Source : https ://paperswithcode.com/

Wav2Vec 2.0



Wav2Vec 2.0 : Framework for Self-Supervised Learning of Speech Representations : Alexei Baevski,Henry Zhou, Abdelrahman Mohamed and Michael Auli (2020)

Extraction de mots clés - Approches

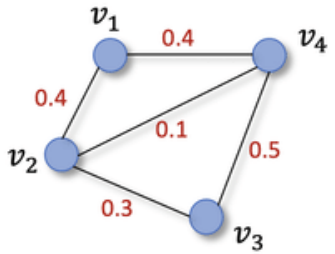
- Statistiques : utiliser statistiques en tant que score
- Graphes : convertir document en un graphe de co-occurrence où les noeuds représentent les mots et les arêtes la relation entre deux mots dans une fenêtre de contexte
- Embeddings

Approches statistiques

- Fréquence des mots, co-occurrence...
- YAKE (Yet Another Keyword Extractor, 2018)
  1. Prétraitement du texte
  2. Extraction de features :
     - Casing
     - Term position
     - Term frequency normalisation
     - Term relatedness to context
     - Term different sentence
  3. Calcul du score pour chaque terme
  4. Génération de n-gram et calcul du score
  5. Classement des mots-clés suivant leur score

Approches basées sur les graphes (1/2)

- **TextRank** (2004), utilise PageRank :
  1. Tokenization et identification de mots candidats (noms et adjectifs)
  2. Construction du graphe : noeud = mot, arête = co-occurrence des deux mots dans une fenêtre de taille N (fixé à $N = 2$). Le graphe est **non-orienté et non pondéré.**
  3. Application de l'algorithme de PageRank pour classifier chaque noeud selon leur score
- SingleRank (2008) : poids sur les arêtes représentant le nombre de fois où deux mots apparaissent dans une même fenêtre
- PositionRank (2017) : ajout de la position des mots dans le texte



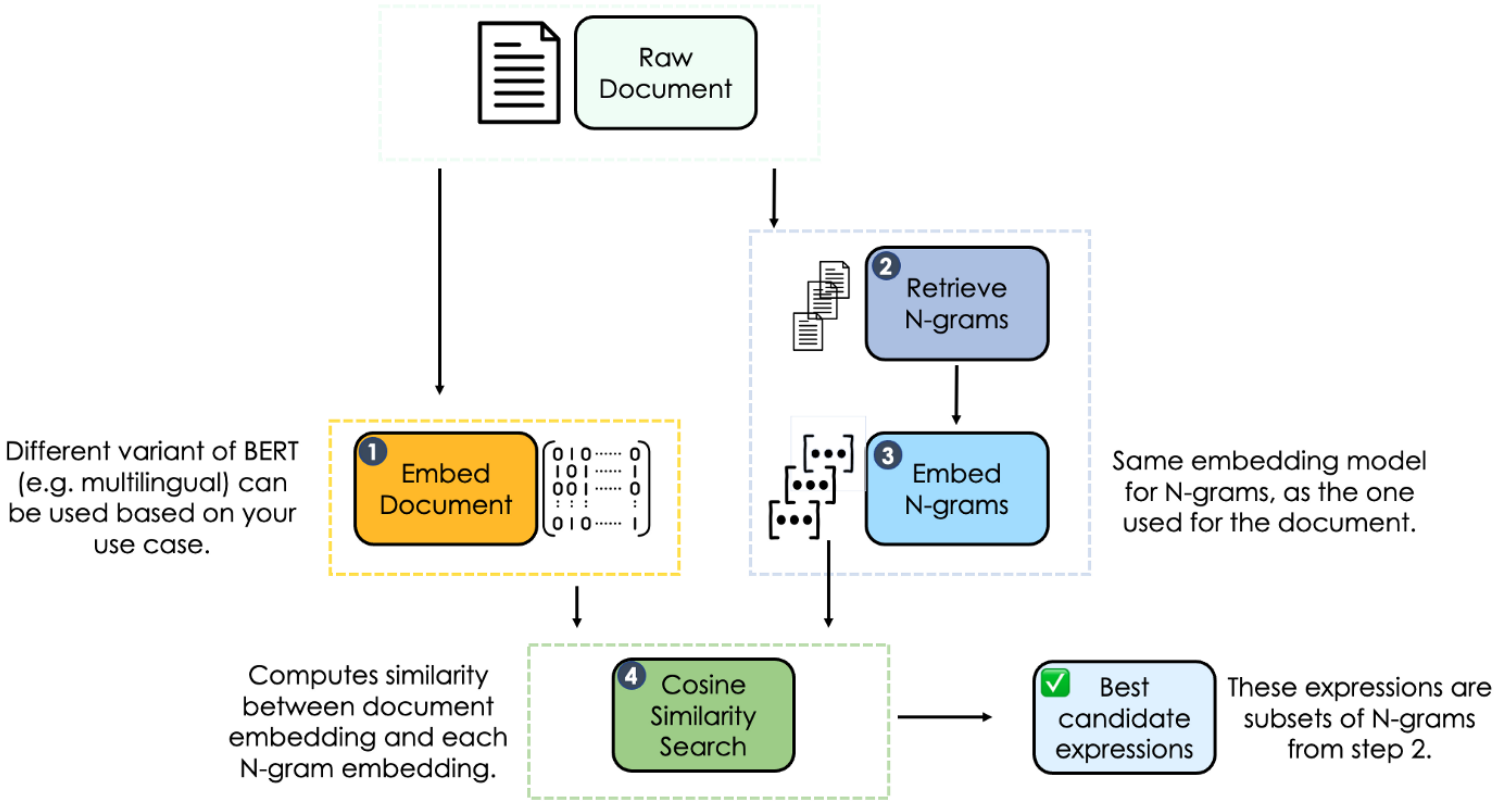
	v_1	v_2	v_3	v_4
v_1	0	0.4	0	0.4
v_2	0.4	0	0.3	0.1
v_3	0	0.3	0	0.5
v_4	0.4	0.1	0.5	0

Approches basées sur les graphes (2/2)

RAKE

Approches basées sur les embeddings (1/2)

KeyBERT (2020)



Source : https ://towardsdatascience.com/semantic-keywords-and-keyphrases-extraction-with-keybert-999234cab7f

Approches basées sur les embeddings (2/2)

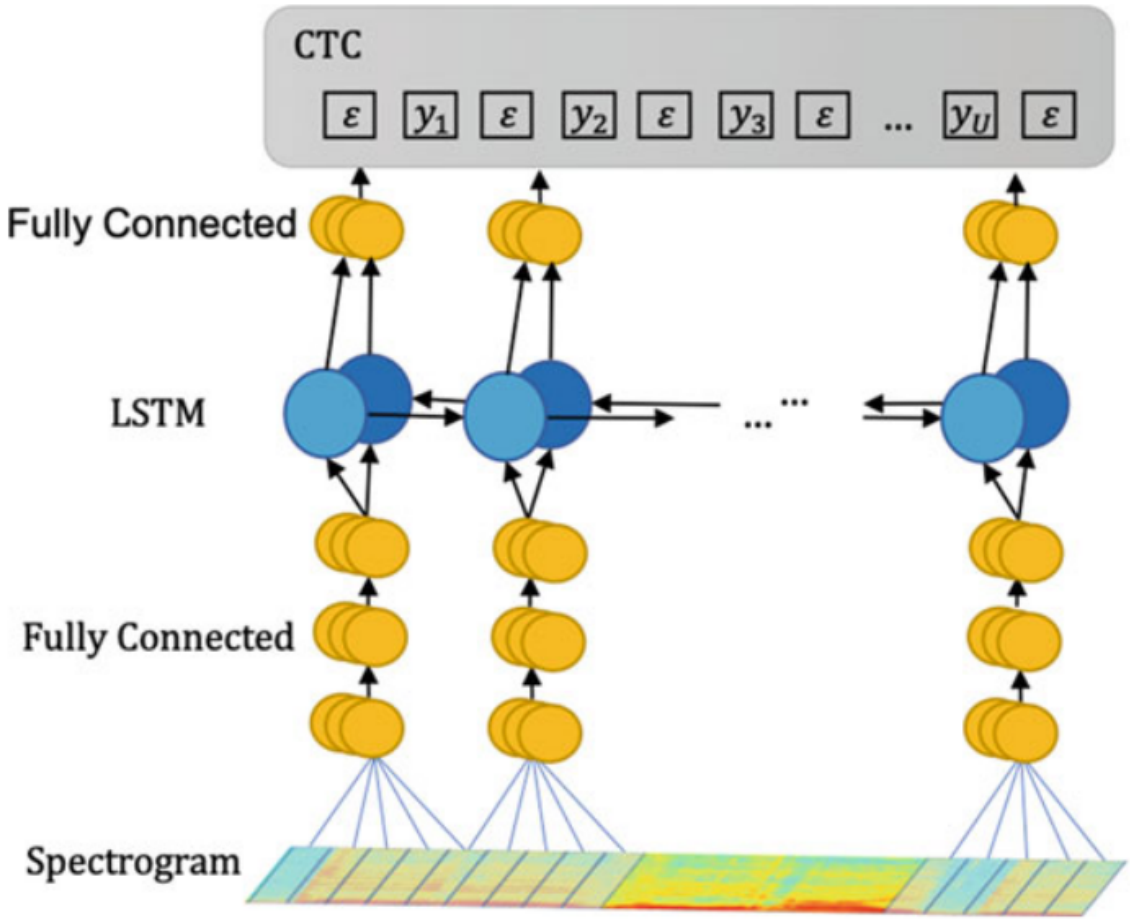
PhraseFormer

Proposition de solution

- Récupération et pré-traitement des données audio
- Transformer le signal audio en texte en utilisant Wav2vec2.0
- Nettoyage et pré-traitement du texte généré (POS tagging, NER)
- Extraction des phrases/mots clés avec Phraseformer ou KeyBERT

Merci pour votre attention

Annexe I : Bi-RNN + CTC



Annexe II : Bi RNN + Attention