PLDAC - Commande vocale

16 mai 2024

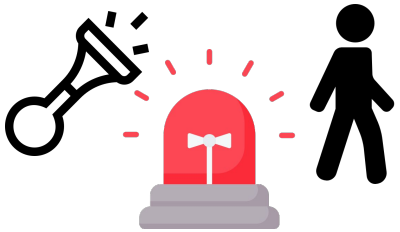
Sarah Eng - Zhile Zhang

Introduction

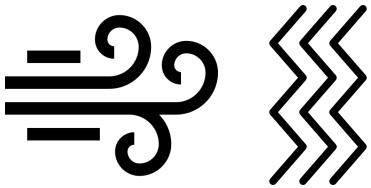
Contexte

Voitures autonomes

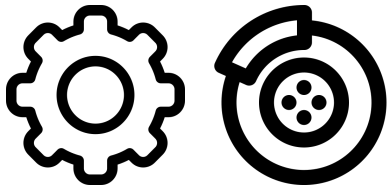
Commandes vocales



Signaux externes



Confort des passagers

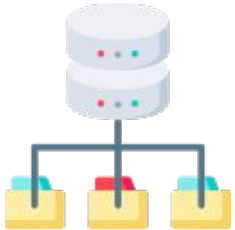


Fonctionnement anormal

Problématique

Quels traitements et quels modèles semblent adapter à la classification de courtes directives audio ?

Données



Google Speech Commands V2

105 829 fichiers audio

35 mots différents

"Yes"	"Zero"	"Bed"	"Backward"
"No"	"One"	"Bird"	"Forward"
"Up"	"Two"	"Cat"	"Follow"
"Down"	"Three"	"Dog"	"Learn"
"Left"	"Four"	"Happy"	"Visual"
"Right"	"Five"	"House"	
"On"	"Six"	"Marvin"	
"Off"	"Seven"	"Sheila"	
"Stop"	"Eight"	"Tree"	
"Go"	"Nine"	"Wow"	

Documentations

Aspects et techniques

Keyword Spotting (KWS)

Non-streaming models

↓

Streaming models

Modèles

Audio Spectrogram Transformer (AST)

MatchboxNet

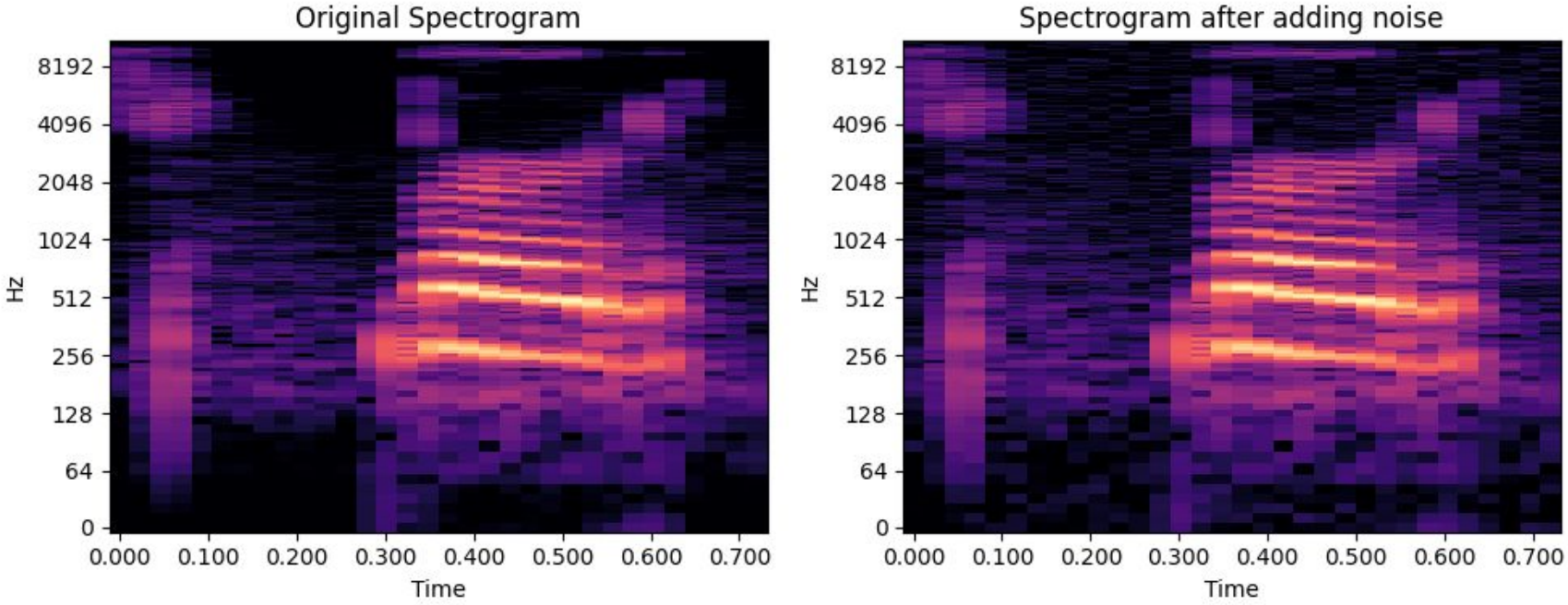
Expériences

Protocole expérimental

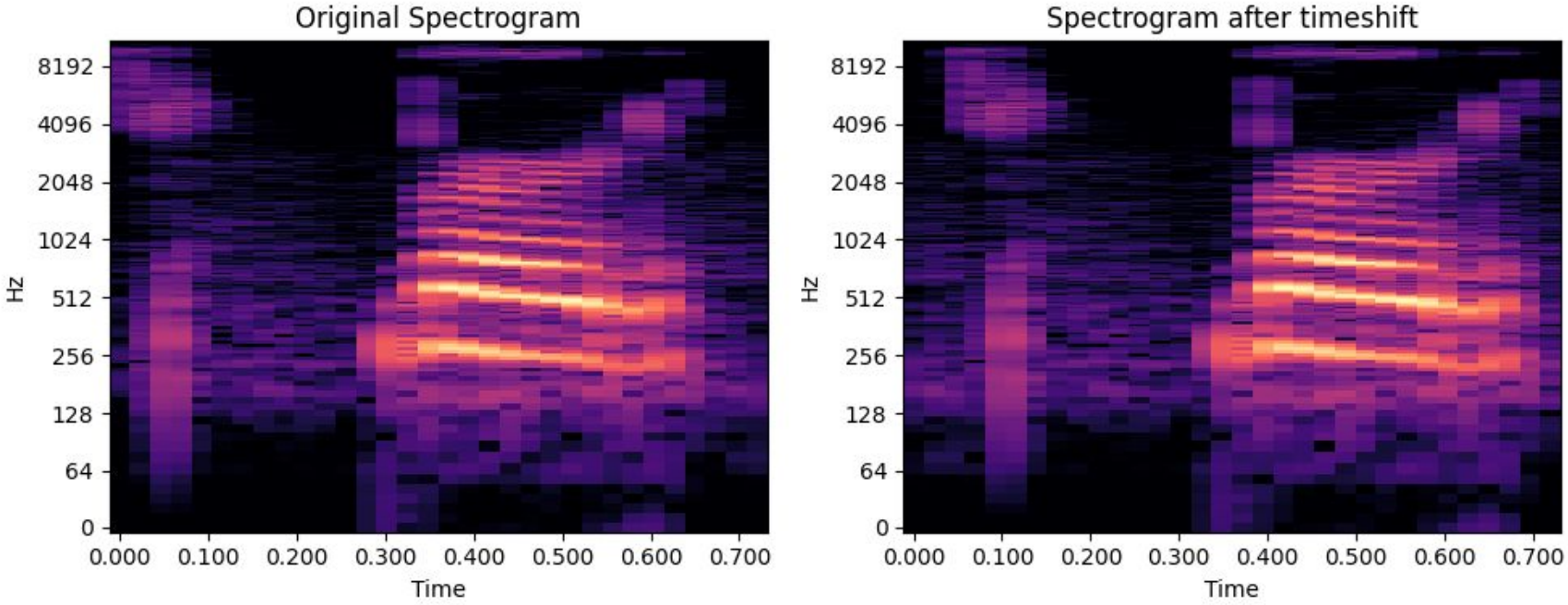
- Collecte des données
- Séparation des ensembles de données
- Prétraitement des données
- Augmentation des données
- Entraînement et validation du modèle
- Evaluation du modèle

Augmentations des données

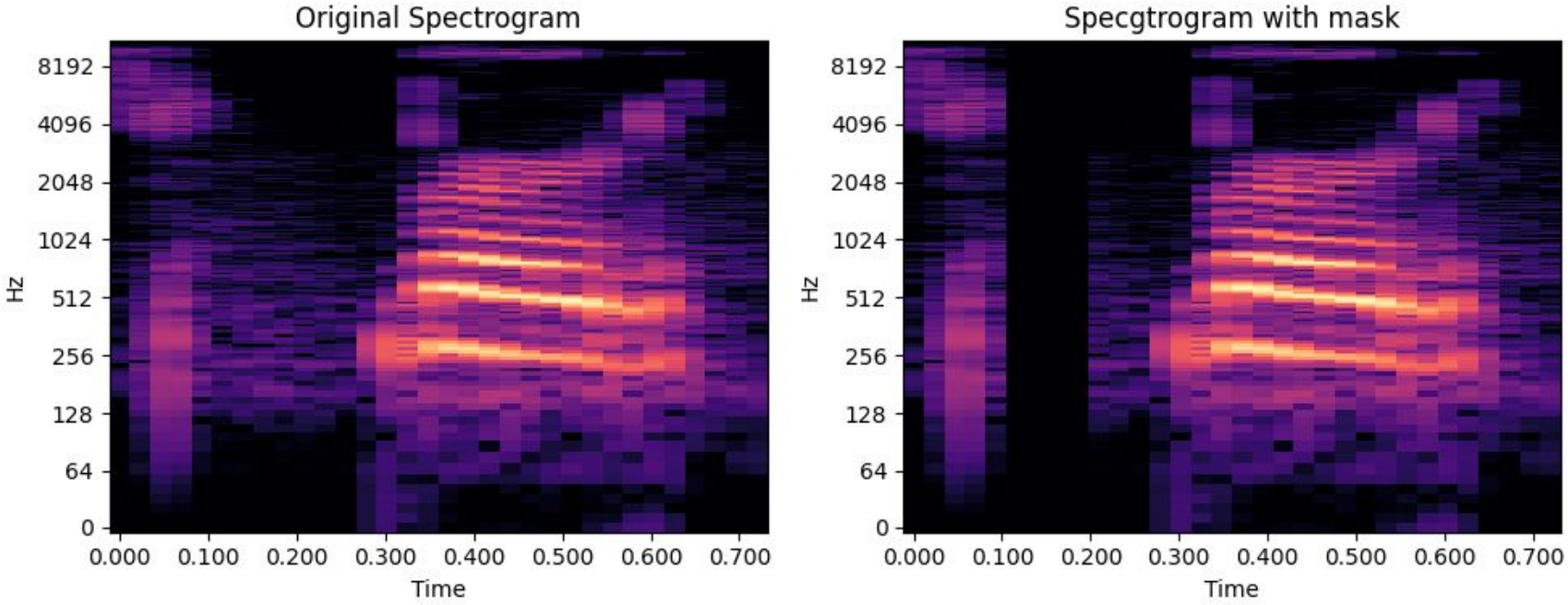
Bruit



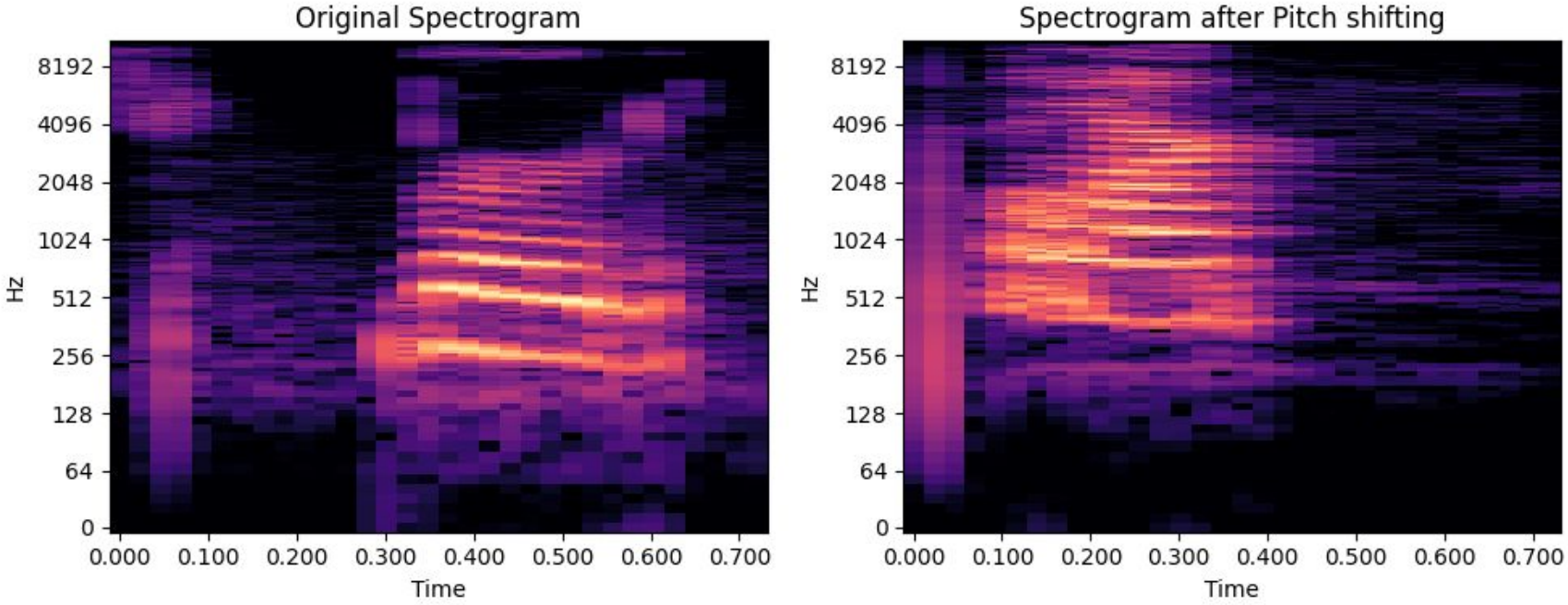
Décalage temporel (Time Shift)



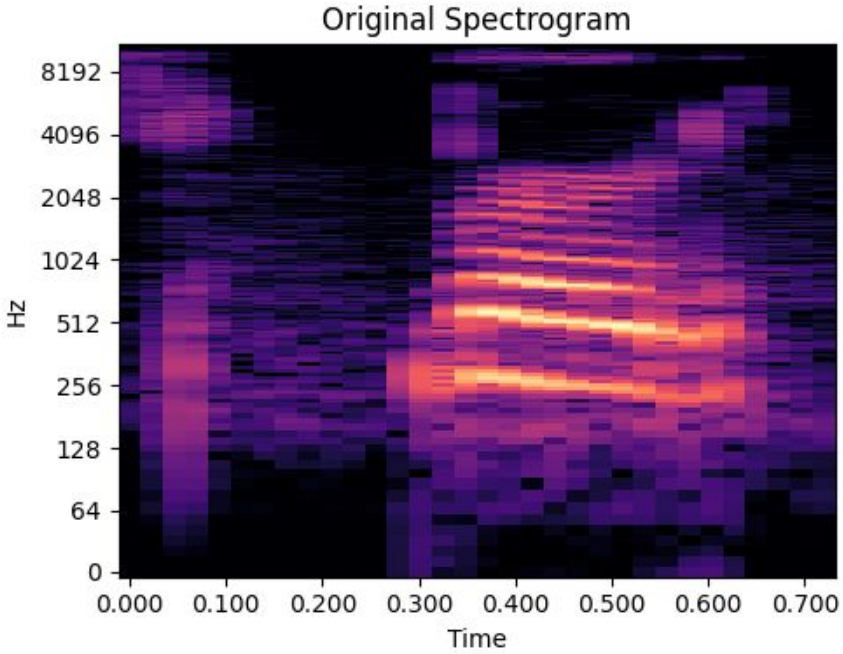
Masque (Time Mask et Frequency Mask)

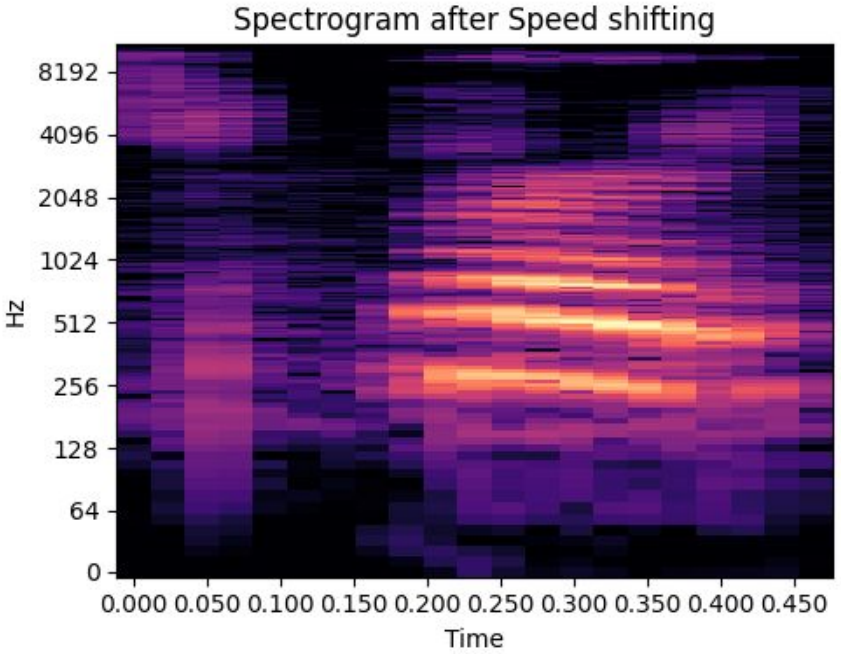


Changement de hauteur (Pitch Shift)



Changement de vitesse (Speed Change)





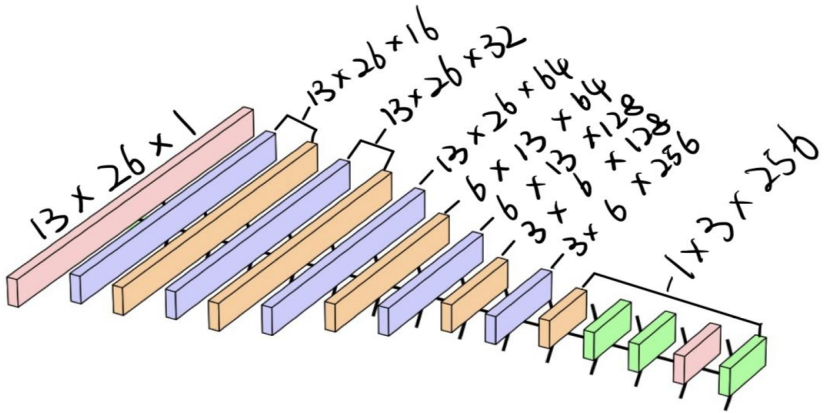
MLP :

- Première couche: La couche entrée, qui reçoit des données avec 338 caractéristiques et les traite à travers 500 neurones.
  - avec fonction d'activation ReLU

- Deuxième couche: La couche de sortie, qui mappe les sorties de la première couche vers 35 nœuds de sortie.
  - avec fonction d'activation Softmax

Modèles

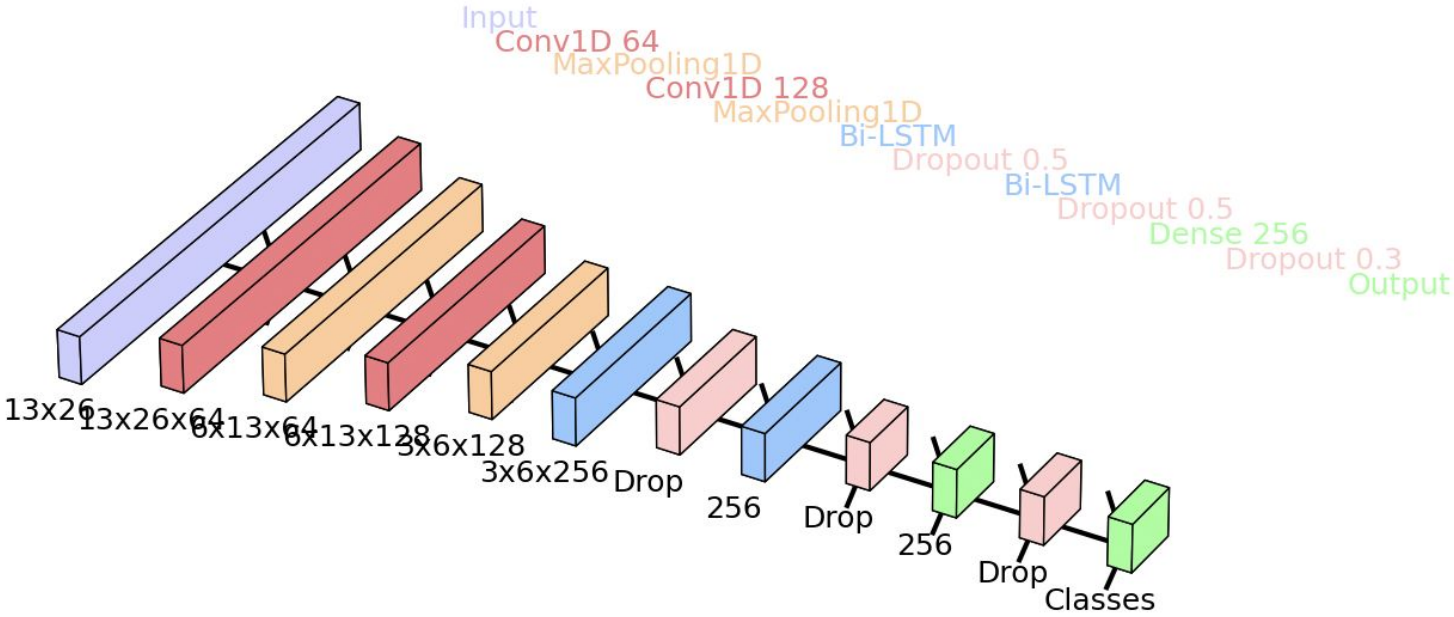
CNN:



- Input
- Conv1 16, ReLu
- MaxPool
- Conv2 32, ReLu
- MaxPool
- Conv3 64, ReLu
- MaxPool
- Conv4 128, ReLu
- MaxPool
- Conv5 256, ReLu
- MaxPool
- Flatten
- Dense 512
- Dropout 0.5
- Output

Modèles

Bi-LSTM:



Résultats

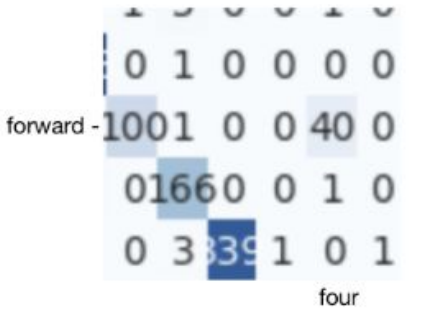
Table

	CNN		Bi-LSTM		MLP	
	accuracy	**time**	**accuracy**	**time**	**accuracy**	**time**
Sans Augmentation	88.98%	34m	73.39%	22m	72.67%	2m
Avec Bruit	89.30%	62m	75.46%	44m	73.28%	3m
Avec TimeShift	90.04%	61m	75.41%	43m	74.88%	3m
Avec Mask	89.33%	72m	74.38%	44m	72.68%	3m
Avec PitchShift	89.12%	66m	75.05%	44m	73.54%	3m
Avec SpeedChange	89.16%	65m	74.63%	44m	71.95%	3m
Avec Mask et TimeShift	89.66%	88m	75.22%	75m	74.88%	3m
Avec Mask et PitchShift	89.87%	92m	76.25%	75m	72.98%	3m
Avec Bruit et SpeedChange	89.50%	82m	75.57%	66m	74.37%	3m
Avec All	89.59%	168m	77.92%	107m	75.62%	5m

TABLEAU 1 – Tableau des scores et temps(minute) pour les modèles CNN, Bi-LSTM et MLP

Matrices de confusion

Exemples de Mal classé :



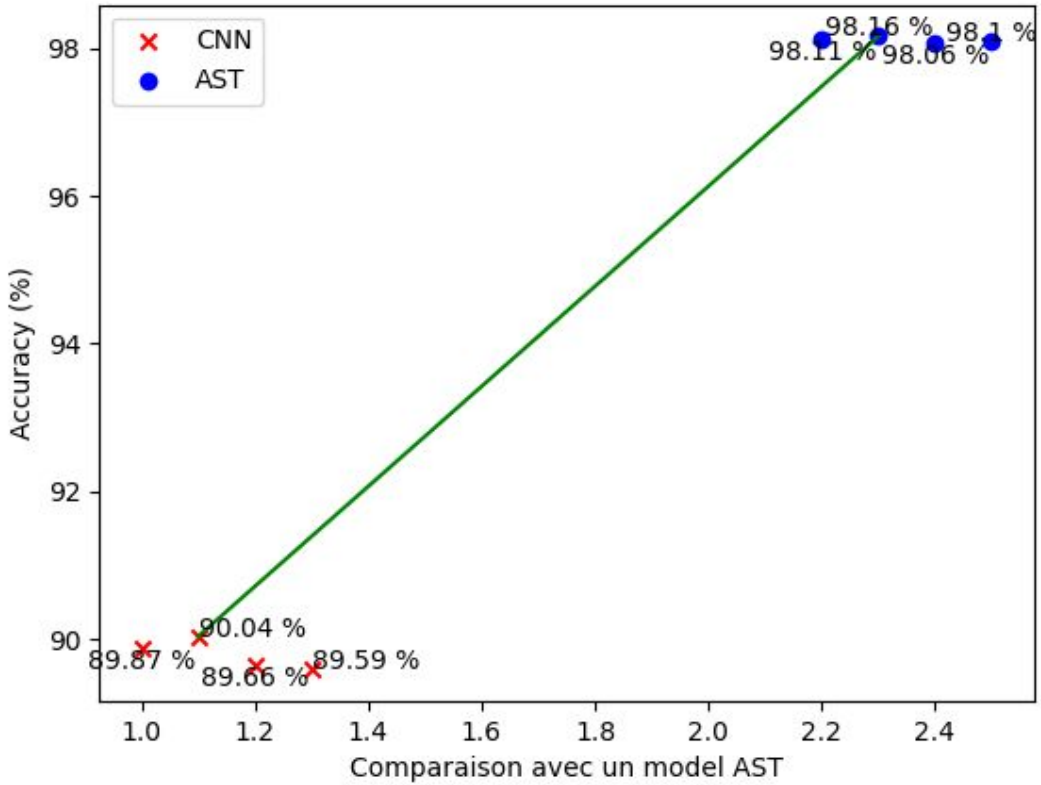
(a) forward-four-noise

(b) no-go-timeshift

(c) tree-three-speed

Comparable

Nous comparons notre modèle CNN avec un modèle Attention-based



Conclusion

Classification de courtes directives audio

- un problème classique d'apprentissage automatique
- les solutions sont utilisées par de nombreux objets connectés pour permettre une interaction avec les utilisateurs

Références

Références

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST : Audio Spectrogram Transformer. https://arxiv.org/abs/2104.01778

Majumdar, S., & Ginsburg, B. (2020). MatchboxNet : 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition. https://doi.org/10.21437/interspeech.2020-1058

Rybakov, O., Kononenko, N., Subrahmanya, N., Visontai, M., & Laurenzo, S. (2020). Streaming Keyword Spotting on Mobile Devices. https://doi.org/10.21437/interspeech.2020-1003

Zhang, Y., Suda, N., Lai, L., & Chandra, V. (2018). Hello Edge : Keyword Spotting on Microcontrollers. https://arxiv.org/abs/1711.0712817

Merci de votre attention