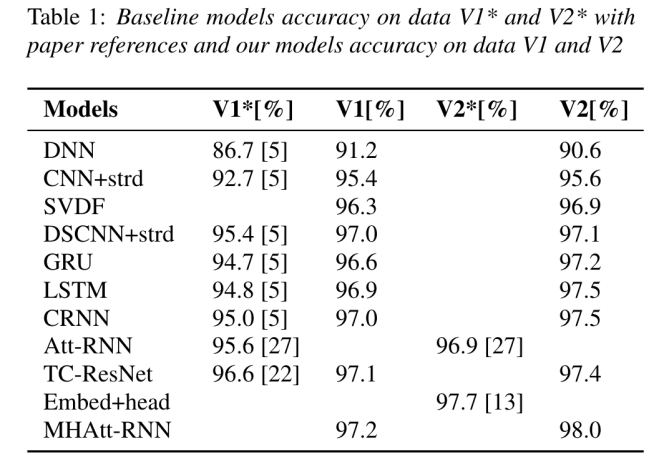
## Streaming keyword spotting on mobile devices

**Prétraitement des déonnees:**

1. Segmentation de l'audio : Le signal audio d'entrée d'une durée de 1 seconde est divisé en cadres (frames) superposés d'une durée de 40 millisecondes, avec un chevauchement de 20 millisecondes.
2. Extraction des caractéristiques：Chaque cadre est traité par un extracteur de caractéristiques basé sur les MFCC (Mel Frequency Cepstral Coefficients), qui est capable de capter les propriétés essentielles du signal audio pour la reconnaissance de la parole.
3. Augmentation des données :
   1. Le décalage temporel dans une plage de -100 ms à 100 ms
   2. le rééchantillonnage du signal avec un facteur de rééchantillonnage dans la plage de 0,85 à 1,15 .
   3. Ajout de bruit de fond et le masquage de fréquence/temps basé sur SpecAugment.

**Evaluation:**

****

Pour les données Speech Commands V1 et V2, plusieurs modèles de reconnaissance de mots-clés (KWS) ont été évalués, y compris DNN, CNN, RNN (LSTM et GRU), CRNN, DSCNN, MHAtt-RNN, etc.

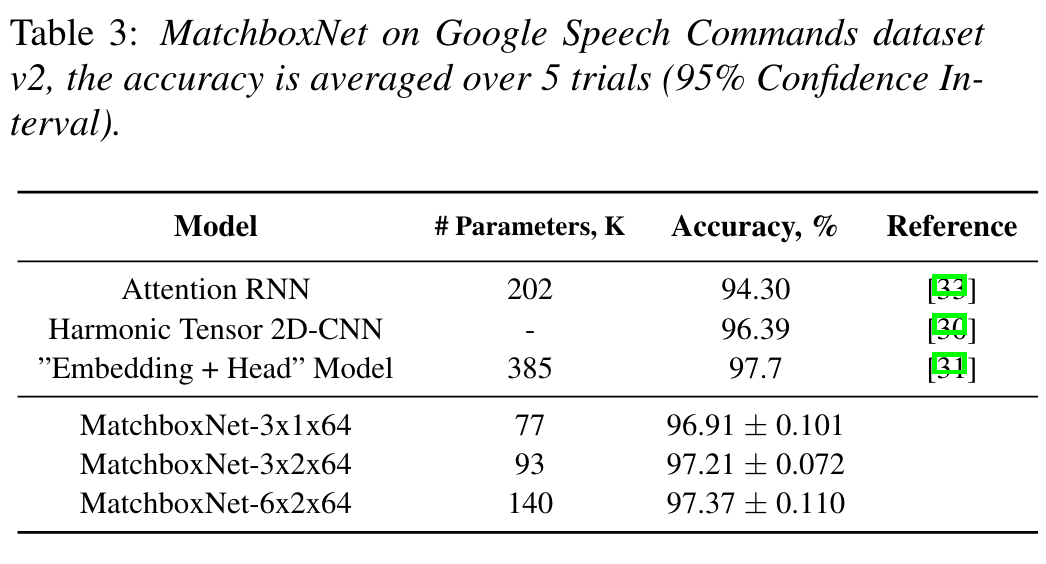
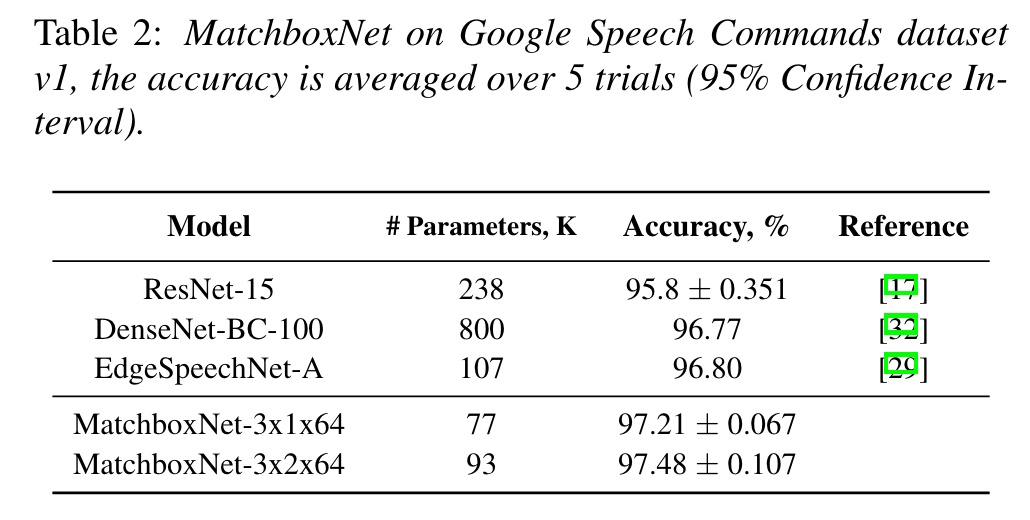
Sur l'ensemble de données V2, le modèle MHAtt-RNN a montré les meilleures performances parmi les différents modèles, atteignant un taux de précision de 98,0 %, ce qui représente une amélioration significative par rapport aux autres modèles et aux recherches précédentes. Par exemple, par rapport au modèle Embed+head, MHAtt-RNN a réduit le taux d'erreur de classification d'environ 10 %.

## MatchboxNet: 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition

**Prétraitement des déonnees:**

1. Conversion en MFCC : Conversion du signal audio en 64 coefficients cepstraux de fréquence de mélodie (MFCC), en utilisant des fenêtres de 25 millisecondes avec un chevauchement de 10 millisecondes entre chaque fenêtre.
2. Padding symétrique : Application d’un padding symétrique sur la dimension temporelle pour atteindre une longueur fixe de 128 vecteurs de caractéristiques par échantillon.
3. Augmentation des données :
4. Application de perturbations de décalage temporel (T = [-5, 5] millisecondes) et de bruit blanc (intensité entre [-90, -46] dB).
5. Utilisation de SpecAugment pour les masques temporels et fréquentiels.
6. Utilisation de SpecCutout pour ajouter des masques rectangulaires.
7. Entraînement avec bruit de fond : Intégration d’échantillons de bruit de fond pendant l'entraînement pour augmenter la robustesse du modèle au bruit.

**Evaluation:**

****

Sur le jeu de données v1, MatchboxNet-3x1x64 et MatchboxNet-3x2x64 ont atteint une précision de pointe, avec respectivement 97,21% (±0,067) et 97,48% (±0,107), tout en ayant nettement moins de paramètres que d'autres modèles de précision similaire.

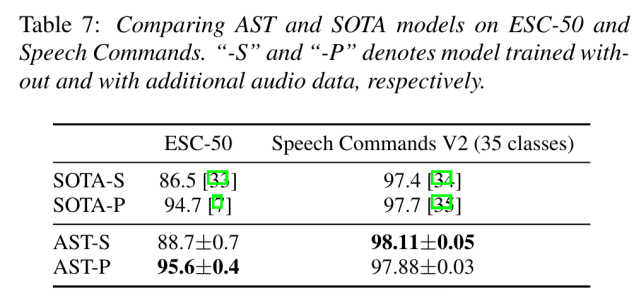
Sur le jeu de données v2, la précision des modèles MatchboxNet était également proche du niveau le plus avancé, avec des précisions de 96,91% (±0,101), 97,21% (±0,072) et 97,37% (±0,110) pour MatchboxNet-3x1x64, MatchboxNet-3x2x64 et MatchboxNet-6x2x64 respectivement.

## AST: Audio Spectrogram Transformer

**Prétraitement des déonnees:**

1. Extraction de caractéristiques : Conversion de l'audio d'entrée en caractéristiques de filtre log Mel (fbank), générant un spectrogramme de 128×100t.
2. Découpage et incorporation des patchs : Découpage du spectrogramme en une séquence de blocs 16×16 (avec un chevauchement de 6), et conversion de chaque bloc en une incrustation 1D de taille 768 à l'aide d'une couche de projection linéaire.
3. Incrustation de position : Pour permettre au modèle de comprendre l'ordre des patchs d'entrée, chaque incrustation de patch est ajoutée à une incrustation de position apprenable.
4. Adaptation du modèle pré-entraîné : Commencer par un modèle pré-entraîné sur ImageNet pour l'initialisation du modèle AST.

**Evaluation:**



D'après les résultats, nous pouvons voir que le modèle AST surpasse les performances des meilleurs modèles existants sur le jeu de données. Cela est dû au fait qu'il repose uniquement sur le mécanisme d'attention, sans utiliser de réseaux neuronaux convolutifs, permettant au modèle de se concentrer sur les parties les plus importantes pour la décision de classification. De plus, l'apprentissage par transfert pré-entraîné sur ImageNet permet au modèle de bénéficier de paramètres de meilleure qualité.