Expérimentations et résultats

Pour évaluer la faisabilité de l'ajoute de l'idée de correspondance entre la vidéo et l'audio dans *liveness-control*, on a refait l'expérimentation de «3D Convolutional Neural Networks for Cross Audio-Video Matching Recognition». Cette expérimentation a utilisé la base *Lip Reading in the Wild*, qui contient 500 mots différents. Selon l'article, on a fait les pré-traitements ci-dessous avant de faire l'entraînement et le test:

1) Convertir toutes les vidéos en 30 fps;
2) Retarder 0.3s en l'axe d'audio pour chaque vidéo qui a le numéro pair ;
3) Segmenter toutes les vidéos en 4 parties, on utilise seulement la deuxième partie qui contient en général le mot clé;
4) Extraire l'audio des vidéos;
5) Extraire les caractéristiques MFEC;
6) Extraire les régions de bouche de chaque vidéo;
7) Redimensionner les images de bouche;
8) Extraire les caractéristiques de vidéos.

Expérimentation avec un mot «ABOUT»

Dans cette expérimentation, on a mis le nombre de «epoch» en 1, «batch-size» en 16 et pour chaque «epoch», on a effectué 62 fois d'entraînement. La variation d'erreur peut être représenté par la figure [1]:

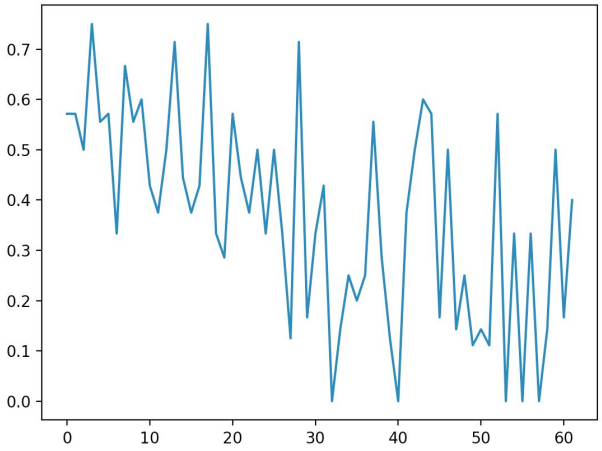

Figure 1 Variation de *ERR*

On peut aussi regarder les résultats de précision par la figure [2]:

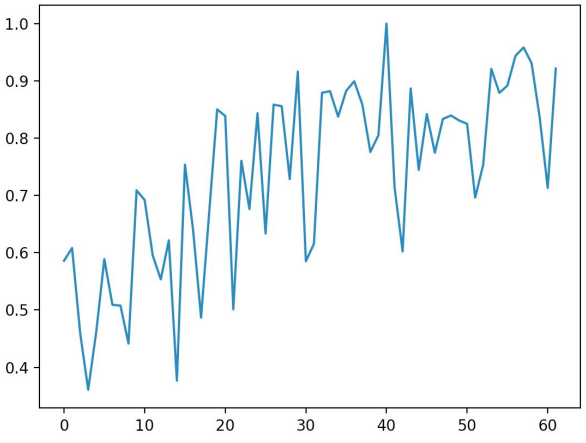

Figure 2 Variation de *AP*

La variation de AUC est similaire au AP, elle est comme la figure [3]:

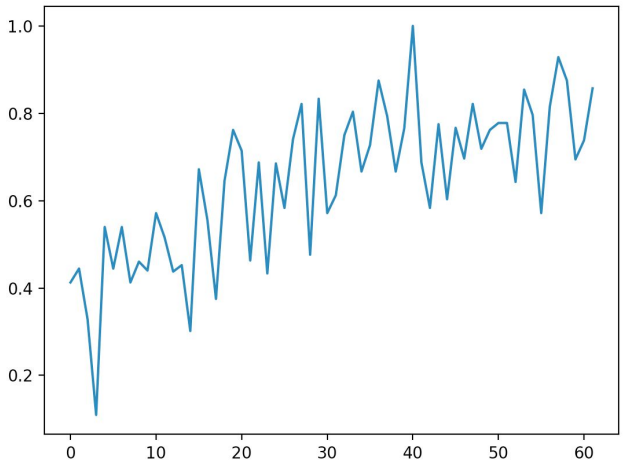

Figure 3 Variation de *AUC*

Expérimentation avec 2 mots «ABOUT» et «ABUSE»

Comme l'expérimentation utilisant un mot, maintenant on a utilisé deux mots. Dans ce cas, on a juste modifié la taille de «batch-size» en 32, et les tailles de la base d'apprentissage et de la base de tests deviennent 2 000 et 100 respectivement. La variation de *ERR*, *AP*, *AUC* sont comme les figures [4], [5], [6] respectivement.

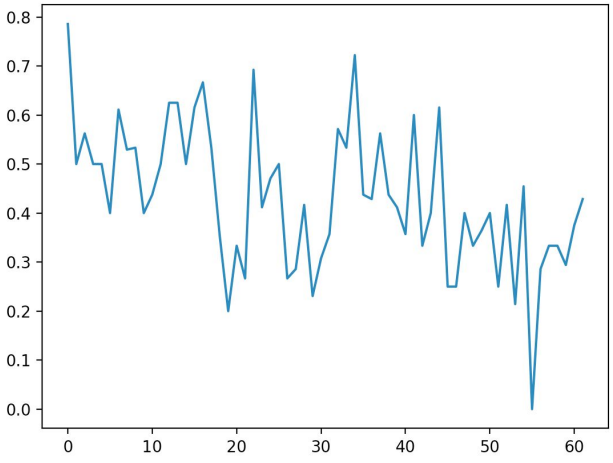

Figure 4 Variation de *ERR* avec deux mots

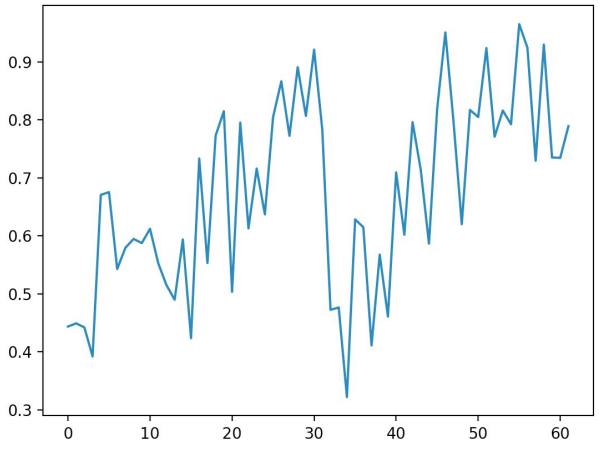

Figure 5 Variation de *AP* avec deux mots

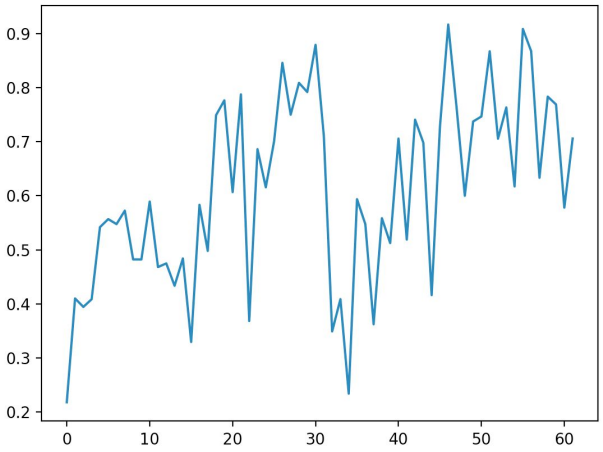

Figure 6 Variation de *AUC* avec deux mots

Résultats

Les figures de variation nous donnent le sentiment indicatif, nous pouvons voir qu'il existe une tendance à la baisse pour *ERR* et la tendance à la hausse pour *AP* et *AUC*, ça correspond aux règles générales de l'apprentissage profond. Mais on peut aussi voir que les variations sont très instables, surtout pour le test avec deux mot: il existe une exception entre la 30ème et 40ème d'itération. (Après une étude de Log, je pense qu'elle est causé par la mauvaise distribution d'échantillon positive et négative). Les résultats sont résumés dans le tableau ci-dessous.

Nombre de mots	EER	AP	AUC
1	0.16	0.84	0.76
2	0.44	0.67	0.61

Pendant les expérimentations, on a aussi trouvé quelques problèmes:

1) Selon notre idée, on voudrait faire la deuxième test avec trois mots. Mais pendant l'extraction de caractéristiques de MFEC, on a trouvé que la base ACROSS ne peut pas être utilisée pour former le tableau MFEC (pour les deux autres base, il n'y pas de problèmes). Pour trouver la raison, j'ai parcourt la caractéristique de chaque audio, et j'ai trouvé qu'il existe une ligne qui n'est pas 15 (en général, la première dimension de caractéristique de MFEC est 15). A cause de ce problème, on ne peut pas former le tableau MFEC pour la base ACROSS et ici, on doit faire attention à ce type de problème, parce que c'est très normal pour un audio qui est un peu court (pour le contenu) et ne peut pas produire 15 caractéristiques de MFEC.
2) Le deuxième problème est l'utilisation de détection de bouch, dans certains cas, on ne peut pas toujours détecter les bouches pour les vidéos, et les caractéristique de vidéos ne sont pas correct car il manque quelques images de bouche. Pour l'instant, nous avons complété les images manuellement, mais ça nous donne aussi le risque de perdre la précision.
3) Le troisième problème est le couplage de vidéos, maintenant on utilise toujours la parti 0.3s - 0.6s, mais dans certains cas, cette partie ne contient pas le mot clé, et cette situation peut influencer la précision si on l'étiquette comme exemple positif.