Analyse des Confusions dans la Classification des Sentiments Vocaux avec IA

AROUN Jeevya, NGAUV Nicolas, THEZENAS Anissa
Université Sorbonne-Nouvelle, INaLCO,
8 avenue de Saint-Mandé 75012 Paris FRANCE, 2 rue de Lille 75007 Paris FRANCE
jeevya.aroun@sorbonne-nouvelle.fr, ngauv.nicolas@gmail.com, thezenasanissa@gmail.com

Abstract

À travers différents modèles comme les mémoires à long terme (LSTM), les réseaux neuronaux convolutifs (CNN), et des algorithmes les machines à vecteurs de support (SVM), l'IA a démontré son efficacité dans la reconnaissance des émotions vocales, mais des confusions significatives persistent entre certaines classes émotionnelles. Cet article présente une comparaison des résultats obtenus avec ces différentes techniques, et une analyse approfondie des confusions interclasses à partir des matrices de confusion obtenues. Nous montrons que les overlaps acoustiques, en particulier entre les émotions proches comme sad et calm ou happy et fearful, sont des sources majeures d'erreurs. Une discussion sur les limitations des modèles utilisés face à ces overlaps est proposée, accompagnée de suggestions pour améliorer la classification.

Keywords: Analyse Vocale, Analyse des Sentiments, Intelligence Artificielle

1. Introduction

La reconnaissance des émotions vocales (RAV) est essentielle dans des domaines variés, allant des interactions homme-machine au suivi de la santé mentale. Bien que les LSTM et les CNN aient amélioré les performances de classification, des erreurs persistent, notamment des confusions entre émotions acoustiquement similaires.

Ce travail se concentre sur l'analyse des confusions interclasses en utilisant un modèle LSTM et un modèle CNN appliqués à des caractéristiques acoustiques standardisées (MFCC) et une comparaison avec des résultats obtenus également via l'utilisation de SVM. Contrairement aux études précédentes (notamment celle de l'article choisi (Singh and Nagrath, 2022)), nous examinons en détail les causes des erreurs pour mieux comprendre les limites des modèles actuels et proposer des axes d'amélioration.

2. Méthodologie

2.1. Données

Nous avons utilisé les datasets suivants : Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018c), composé de deux datasets (Speech pour la voix parlée (Livingstone and Russo, 2018a) et Song pour la voix chantée (Livingstone and Russo, 2018a)).

Le RAVDESS contient 24 acteurs professionnels (12 femmes, 12 hommes), vocalisant deux phrases lexicalement identiques avec un accent nord-américain neutre. Les émotions exprimées dans les chansons incluent les expressions de calme, de joie, de tristesse, de colère et de peur. Chaque émotion est produite à deux niveaux d'intensité émotionnelle (normal et fort), avec une expression neutre supplémentaire.

Conformément à l'article choisi (Singh and Nagrath, 2022), nous avons réduit nos étiquettes de 7 à 5 pour réduire la complexité de notre taux d'apprentissage et ainsi améliorer la précision. Ainsi, les expérimentations utilisent un jeu de données contenant 5 classes d'émotions : angry, calm, fearful, happy et sad. On utilise un split Train/Test de respectivement 80% et 20% du corpus. Les coefficients MFCC sont extraits comme caractéristiques audio principales.

Le dépôt git avec tout notre code ainsi que les différents graphes et matrices de confusion mais aussi la documentation, se trouve ici (Aroun et al., 2024).

2.2. Extraction des caractéristiques MFCC

L'extraction des caractéristiques audio dans cette étude repose sur les coefficients cepstraux des fréquences de Mel (MFCC), qui sont calculés à l'aide de la bibliothèque librosa. L'objectif des MFCC est de capturer les caractéristiques acoustiques les plus significatives d'un signal sonore tout en filtrant les variations superficielles ou bruitées. Contrairement aux spectres bruts ou aux transformées de Fourier, les MFCC intègrent une modélisation perceptive grâce à l'échelle de Mel. Cette échelle reflète la non-linéarité de la perception des fréquences par les humains : nous sommes plus sensibles aux basses fréquences qu'aux hautes

fréquences.

Le calcul des MFCC requière plusieurs étapes. Premièrement, Le signal audio est découpé en segments courts, appelés fenêtres, pour capturer les propriétés temporelles locales. Ensuite, Une transformée de Fourier est appliquée à chaque fenêtre pour convertir le signal temporel en un spectre de fréquences. Celui-ci est projetée sur une échelle de Mel à l'aide d'une banque de filtres triangulaires qui simule la sensibilité de l'oreille humaine aux différentes bandes de fréquences. Les amplitudes des bandes fréquentielles sont converties en leur logarithme. Ce procédé aide à comprimer le signal et reflète la perception humaine de l'intensité. Puis, une DCT (Transformée de Cosinus Discrète) est appliquée aux coefficients issus de l'échelle de Mel pour obtenir un ensemble de coefficients cepstraux. Cette étape réduit la corrélation entre les coefficients et produit un vecteur.

Dans cette étude, nous utilisons une fréquence d'échantillonnage de 22 050 Hz, un nombre de coefficients MFCC de 40, et une longueur normalisée de 300 timesteps.

2.3. Modèles et Algorithmes

Nous avons utilisés 3 modèles différents pour cette étude : LSTM, CNN et SVM afin de pouvoir comparer les performances.

2.3.1. Modèle LSTM

Le premier modèle testé est LSTM (Long Short-Term Memory). Il s'agit d'une architecture de réseau neuronal récurrent (RNN) spécialement conçue pour apprendre et exploiter les dépendances à long terme dans des séquences de données.

Pour notre expérience, nous utilisons la configuration suivante : une couche LSTM à 128 unités avec un retour de séquence ; une seconde couche LSTM à 64 unités ; une couche dense à 64 neurones avec activation ReLU ; une couche de sortie dense avec activation softmax pour produire des probabilités sur les classes d'émotions ; une régularisation est effectuée via un taux de dropout de 30 % pour réduire le surapprentissage.

2.3.2. Modèle CNN

On utilise un modèle de réseau de neurones convolutifs (ou réseau de neurones à convolution) pour détecter des motifs dans des matrices comme les MFCC, à travers des couches convolutionnelles, denses, d'activation et de pooling avec une couche de sortie dont la fonction d'activation est softmax. Cela permet de construire un modèle capable de classifier les données avec efficacité.

Ici, on optimise avec Adam et on a essayé de jouer avec la régularisation L2, le Dropout et un apprentissage plus lent et progressif (avec un learning rate scheduler) pour essayer de stabiliser l'entraînement, d'éviter le sur-apprentissage et d'améliorer la capacité de généralisation du modèle.

On évalue et visualise avec les Accuracy du Train et du Test, les Loss du Train et du Test, et des graphiques et une matrice de confusion des émotions.

2.3.3. Algorithme SVM

L'algorithme SVM (machines à vecteurs de support) est choisi pour sa capacité à gérer efficacement des espaces de grande dimension et pour sa robustesse même avec un ensemble d'entraînement relativement restreint. Le processus commence par l'extraction des coefficients MFCC qui condensent le signal audio en un ensemble de caractéristiques représentant les aspects acoustiques essentiels. Ces coefficients sont ensuite normalisés pour minimiser les biais et améliorer la performance du modèle.

Le modèle SVM est configuré avec divers noyaux linéaire, polynomial, et RBF pour évaluer et sélectionner l'option qui maximise la précision de classification sur notre jeu de données spécifique. L'entraînement du modèle vise à trouver un hyperplan qui sépare au mieux les différentes classes d'émotions, en maximisant la marge entre elles. Ce processus est critique pour assurer que le modèle soit non seulement précis mais aussi capable de généraliser à de nouvelles données audio non vues.

3. Résultats et Analyses

3.1. Performances Globales

3.1.1. Modèle LSTM

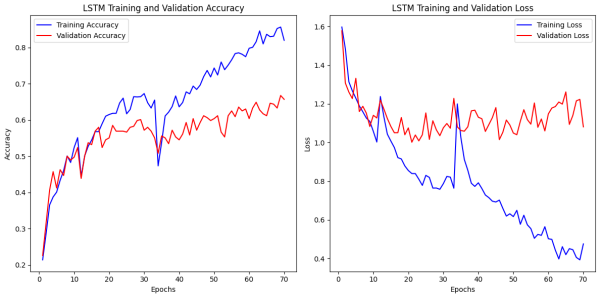

Figure 1: Graphique de l'historique d'entrainement et de validation pour le modèle LSTM

Le modèle LSTM nous permet d'obtenir **66% d'accuracy** sur le test set.

3.1.2. Modèle CNN

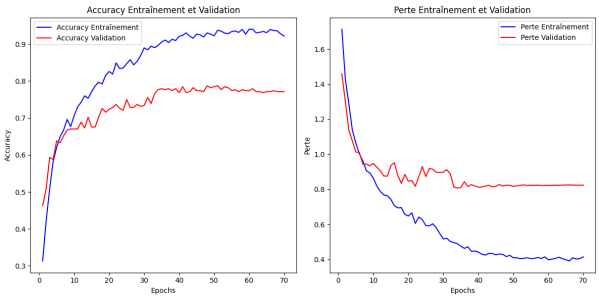

Figure 2: Graphique de l'historique d'entrainement et de validation pour le modèle CNN

- Training Accuracy : 98,20%.
- Validation Accuracy : 77,13%.
- Training Loss : 0,3042.
- Validation Loss : 0,8233.
- Perte stable à partir de l'epoch 50 : le modèle n'apprend plus ensuite.

3.1.3. Algorithme SVM

Les performances du modèle SVM sont résumées dans le rapport de classification, où le modèle atteint une précision globale de 69.27%. Les émotions "happy" et "sad" sont identifiées avec les meilleures précisions de 0.81 et 0.80 respectivement, indiquant une forte aptitude du modèle à reconnaître ces états émotionnels. Toutefois, l'émotion "angry" présente des défis significatifs avec une précision de seulement 0.64, ce qui peut refléter des chevauchements acoustiques avec des émotions similaires comme "fearful".

Le score F1, qui balance précision et rappel, est relativement uniforme à travers les émotions, avec "happy" montrant le meilleur équilibre à 0.77. Ces scores F1 indiquent que le modèle est capable de minimiser les faux positifs et les faux négatifs pour ces émotions. Cependant, les difficultés avec l'émotion "angry" soulignent la nécessité d'une amélioration de la classification pour cette émotion spécifique, potentiellement en ajustant les caractéristiques ou en expérimentant avec d'autres configurations de noyaux SVM.

En conclusion, bien que le SVM montre une bonne performance générale, les résultats indiquent des domaines pour des améliorations futures, notamment en optimisant le traitement des émotions difficiles à distinguer. Ces ajustements pourraient inclure des expérimentations avec différentes techniques de prétraitement des données ou l'intégration de méthodes ensemblistes pour renforcer la distinction entre les classes d'émotions proches.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.81      0.72        31
           1       0.81      0.74      0.77        34
           2       0.80      0.67      0.73        42
           3       0.64      0.57      0.61        40
           4       0.62      0.71      0.66        45

    accuracy                           0.69       192
   macro avg       0.70      0.70      0.70       192
weighted avg       0.70      0.69      0.69       192

Accuracy Score: 0.6927083333333334
```

Figure 3: Performances globales pour SVM

3.2. Analyse des Confusions Interclasses

Les lignes représentent les labels réels, et les colonnes représentent les labels prédits. Les valeurs diagonales indiquent les prédictions correctes, et les autres cellules montrent les erreurs de classification.

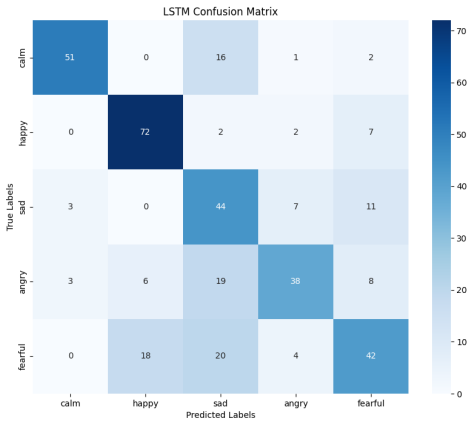

Figure 4: Matrice de confusion pour le modèle LSTM

Avec le modèle LSTM, on observe sur la figure 5 une précision plus élevée pour les émotions telles que 'happy' et 'calm', avec respectivement 72 et 51 prédictions correctes. Les exemples de la classe "fearful" sont souvent prédits comme "sad" ou "happy". Les exemples de la classe "calm" sont souvent faussement prédits comme "sad", et les exemples de la classe "angry" comme "sad". Ce qui pourrait indiquer une similarité dans les caractéristiques MFCC entre ces émotions.

La matrice de confusion de l'algorithme SVM montre que le modèle identifie relativement bien certaines émotions tout en confondant d'autres. Il a correctement classifié "angry" 25 fois mais l'a confondu avec "calm" et "sad". La classe "calm" a été bien reconnu 25 fois, mais souvent confondu avec "sad". La classe "fearful" a été

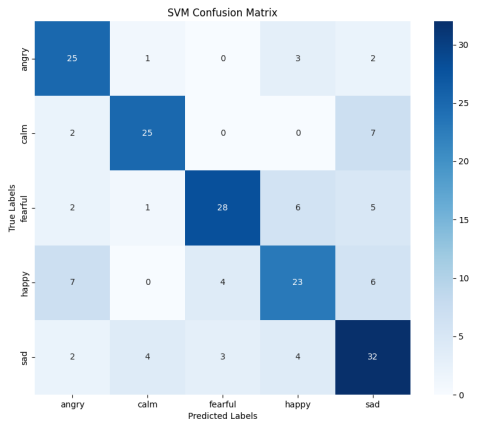

Figure 5: Matrice de confusion pour SVM

précisément identifié 28 fois, bien que confondu avec “happy” et “sad”. L’émotion “happy” a été reconnue 23 fois, mais également mal interprétée comme “angry” et “fearful”. Enfin, “sad” a montré la meilleure reconnaissance avec 32 classifications correctes, malgré quelques confusions avec “calm” et “happy”.

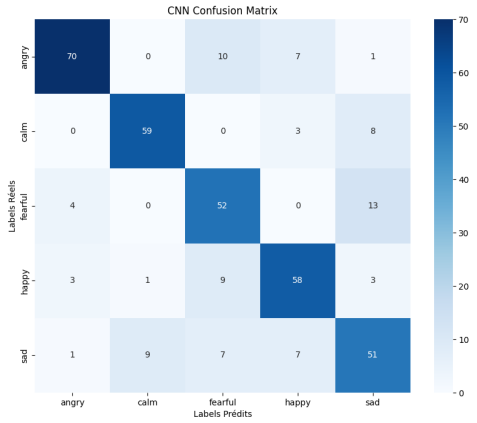

Figure 6: Matrice de confusion pour le modèle CNN

Voici l’analyse par classe en prenant pour illustration la matrice de confusion du modèle CNN.
Pour la classe "angry", le modèle la confond avec des émotions comme "fearful" ou "happy", ce qui peut s’expliquer par des similitudes acoustiques entre des émotions intenses comme la colère et le bonheur ou encore la peur (le corpus étant composé de voix parlées et chantées par des acteurs, et donc plus préparées que spontanées).

À noter que la confusion entre "angry" et "fearful" provient de l’apport de l’apprentissage avec les voix chantées, car lors des tests avec les voix parlées seulement, nous n’avions pas ce résultat.

Pour la classe "calm", le modèle la confond parfois avec des émotions négatives comme "sad". Compréhensible, car une voix calme peut être difficile à distinguer d’une voix légèrement triste ou posée dans le contexte audio (les spectres audio peuvent être similaires).

Pour les petites confusions entre "calm" et "happy", toutes les expressions de la joie ne sont pas forcément associées à des hauteurs vocales élevées, et si les variations de pitch dans "happy" ne sont pas marquées (par exemple, une expression de bonheur tranquille ou réservé), elles peuvent être perçues comme similaires à une voix calme.

Pour la classe "fearful", on remarque qu’il y a une confusion notable entre la peur ("fearful") et la tristesse ("sad"), souvent liée à des signaux audio de faible intensité ou tonalité similaire : certaines caractéristiques acoustiques associées à la peur (comme un ton bas ou tremblant) peuvent ressembler à celles d’une tristesse prononcée.

Pour la confusion entre la peur ("fearful") et la colère ("angry"), elle vient certainement de l’apport des voix chantées (car lors de tests avec seulement les voix parlées, nous n’avions pas ce résultat) : la prosodie lors du chant ou de la voix parlée est différente, et lorsque qu’on entraine et évalue le modlèle sur un corpus composé de ces 2 types d’expression orale, il y a forcément plus de confusion.

Pour la classe "happy", les confusions avec "fearful" peuvent indiquer une confusion dans les fréquences audio plus aiguës. On peut aussi penser à la variation rapide du pitch : les variations fréquentes ou abruptes du ton sont communes dans les émotions comme la joie, la peur ou la colère. Par exemple, une voix joyeuse peut avoir des montées rapides, tout comme une voix apeurée ou colérique.

On peut aussi se dire que cela peut être dû à la diversité des expressions vocales du bonheur, qui peuvent parfois être interprétées comme des émotions intenses (ce que sont la peur et la colère) ou plus mélancoliques (ce qui peut expliquer la confusion avec "sad").

Pour la classe "sad", il existe des confusions significatives avec "calm". Cette confusion est fréquente dans les modèles de classification audio, car des émotions comme la tristesse et le calme partagent souvent des tonalités douces et des rythmes lents. Pour "sad" et "fearful", ces deux émotions sont souvent exprimées avec des tonalités graves ou basses dans la voix, ce qui peut expliquer la confusion.

Les confusions entre "sad" et "happy" peuvent sembler contre-intuitives étant donné que ces deux émotions ont une valence émotionnelle opposée (négative pour "sad", positive pour "happy")... Cependant, elles peuvent survenir dans un modèle de classification audio à cause de l'overlap qu'il peut parfois y avoir dans les spectres acoustiques : les descripteurs acoustiques utilisés par les modèles, comme les MFCC, se concentrent principalement sur les propriétés spectrales (fréquence et énergie). Ils peuvent ne pas capturer les différences subtiles de valence émotionnelle.

4. Discussion

Les meilleurs résultats sont obtenus avec le modèle CNN. On remarque que c'était également le cas dans l'article choisi (Singh and Nagrath, 2022). Ceci peut s'expliquer par plusieurs facteurs liés à la nature des données, aux architectures des modèles, et à leur capacité à extraire des caractéristiques pertinentes.

Les CNN sont plus adaptés aux données structurées comme les MFCC car ils sont spécifiquement conçus pour capturer des motifs locaux dans des représentations bidimensionnelles où les dimensions représentent les coefficients et les cadres temporels (ici les coefficients cepstraux en fréquence mél (MFCC)). On peut préciser qu'ils sont robustes aux petites variations locales dans les données, comme les variations d'accent ou de prononciation dans les émotions vocales.

Les LSTM sont conçus pour les données séquentielles et sont efficaces pour capturer les dépendances à long terme dans les séries temporelles. Cependant, Les LSTM traitent les séquences de manière itérative, ce qui peut être plus lent et moins efficace pour les données comme les MFCC, où des motifs locaux sont souvent plus importants que les dépendances à long terme. De plus, si les séquences MFCC sont relativement courtes et peu variées, les LSTM peuvent être plus susceptibles de sur-apprendre les données d'entraînement.

Les SVM sont limitées pour les données complexes : elles sont efficaces pour des données de faible dimension et des frontières de décision simples. Les MFCC nécessitent une vectorisation ou une transformation en entrée plate pour les SVM (ce qui peut entraîner une perte d'information structurelle) et elles peuvent être moins performantes pour les grands ensembles de données (faible scalabilité) ou les tâches nécessitant des modèles très expressifs, comme la classification des émotions vocales.

4.1. Proposition d'Amélioration

On peut aujouter des caractéristiques supplémentaires (compléter les MFCC avec des spectrogrammes ou des indicateurs prosodiques comme le rythme et l'amplitude), mais on peut également penser différemment et privilégier des approches multimodales (combiner les données audio avec des informations visuelles ou contextuelles), ou encore essayer d'utiliser des architectures plus avancées (explorer des modèles attentionnels ou des transformers pour capturer des nuances émotionnelles complexes).

5. Conclusion

Cet article propose une analyse ciblée des confusions interclasses dans la classification des émotions vocales. Bien que les CNN offrent de bonnes performances globales (et meilleures que celles offertent par les LSTM ou utilisant des algorithmes SVM), leurs limites face aux overlaps acoustiques montrent la nécessité de méthodes complémentaires pour améliorer la reconnaissance des émotions proches. Les travaux futurs pourront inclure l'utilisation de caractéristiques supplémentaires, des données augmentées, des approches multimodales ou encore des architectures plus poussées pour tenter de dépasser ces limitations.

6. Bibliographie

M. Alhlffee. 2020. Mfcc-based feature extraction model for long time period emotion speech using cnn. *Revue d'Intelligence Artificielle*, 34(2):117–123.

Jeevya Aroun, Nicolas Ngauv, and Anissa Thezenas. 2024. Audio sentiment analysis. Depository available at https://github.com/JeevArn/AudioSentimentAnalysis/tree/main.

Steven R. Livingstone and Frank A. Russo. 2018a. The ryerson audio-visual database of emotional speech and song (ravdess). Dataset available at https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.

Steven R. Livingstone and Frank A. Russo. 2018b. The ryerson audio-visual database of emotional speech and song (ravdess). Dataset available at https://www.kaggle.com/uwrfkaggler/ravdess-emotional-song-audio.

Steven R. Livingstone and Frank A. Russo. 2018c. The ryerson audio-visual database of emotional

speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13(5):e0196391.

Praveen Singh and Preeti Nagrath. 2022. Vocal analysis and sentiment discernment using ai. *ASPG*.