Université Euro Méditerranéenne Fès
EuroMed University of Fez

Ecole d'Ingénierie Digitale et d'Intelligence Artificielle
(EIDIA)

Rapport

Filière: IA

Semestre : 6

Module : : Synthèse vocale

Thème :

Identification et vérification du locuteur

Encadré par :

Pr. Jamal KHARROUBI

Préparé par :

- Hassan KERROUMI

I Introduction :

L'identification et la vérification du locuteur sont des domaines importants en reconnaissance automatique de la parole. Ce TP vise à explorer l'utilisation des Modèles de Mélange Gaussien (GMM) pour identifier un locuteur à partir de ses caractéristiques vocales.

Objectif :

L'objectif est de développer des modèles IA permettant la reconnaissance et la vérification d'un locuteur à partir d'un enregistrement audio.

II Les outils utilisé:

Bibliothèques Python :

- **Scikit-learn (sklearn)**

Bibliothèque d'apprentissage automatique (machine learning) offrant des outils pour la classification, la régression, le clustering, la réduction de dimension, l'évaluation de modèles, etc.

- **Pandas**

Bibliothèque utilisée pour la manipulation et l'analyse de données, notamment grâce à ses structures comme les **DataFrame** (tableaux étiquetés).

- **NumPy**

Bibliothèque fondamentale pour le calcul scientifique, elle permet de manipuler efficacement des tableaux multidimensionnels et d'effectuer des opérations mathématiques rapides.

- **Matplotlib**

Bibliothèque de visualisation de données, utilisée pour tracer des courbes, histogrammes, spectres, etc. Très utile pour analyser graphiquement les résultats.

- **GMM (Gaussian Mixture Models)**

Modèle statistique utilisé pour modéliser la distribution des données en combinant plusieurs distributions gaussiennes. Très utilisé en reconnaissance de locuteur.

- **MFCC (Mel Frequency Cepstral Coefficients)**

Technique d'extraction de caractéristiques audio qui modélise la perception humaine des sons. Utilisée couramment en traitement de la parole.

- **WebRTC VAD (Voice Activity Detection)**

Algorithme de détection de la parole qui identifie les segments contenant de la voix dans un signal audio. Basé sur WebRTC.

I Présentation de la Base de Données :

Le jeu de données utilisé pour ce TP comprend :

- **23 locuteurs** : 13 femmes et 10 hommes.
- Chaque locuteur a un enregistrement pour l'entrainement et 15 enregistrements pour le test.
- La durée des enregistrements :
    - Pour l'entrainement : 2 min
    - Pour le test :
        - 5 enregistrements de 5 secondes.
        - 5 enregistrements de 10 secondes.
        - 5 enregistrements de 15 secondes.

II Méthodologie

1. Prétraitement des Données et Extraction des Caractéristiques

Avant d'extraire les caractéristiques des signaux audio, il est nécessaire de supprimer les parties de silence pour se concentrer uniquement sur les segments parlés.

- La méthode 1 :
    - Chargement du fichier audio avec `librosa`.
    - Application du **Voice Activity Detection (VAD)** pour supprimer les trames silencieuses à l'aide de WebRTC.

Le **Voice Activity Detection (VAD)** est une technique utilisée pour **détecter les segments de parole dans un signal audio**, qui fonctionne comme suit :

- **Découpage du signal audio en petites trames** (de 30 ms par défaut).
- **Conversion de chaque trame en format PCM 16 bits**, requis par l'algorithme WebRTC VAD.
- Application du VAD sur chaque trame pour savoir **si elle contient de la parole**.
- **Suppression des trames silencieuses**, ne conservant que celles où de la voix a été détectée.

- Extraction des **caractéristiques MFCC** à partir de l'audio nettoyé.
- Sauvegarde des MFCC .

- Méthode 2 :
  - Extraction des caractéristiques MFCC.
  - Calcul de l'énergie.
  - Utilisation d'un modèle GMM pré-entraîné sur un signal audio contenant du silence pour éliminer les trames similaires à ce silence.
  - Élimination des parties contenant du silence des caractéristiques MFCC.
  - Enregistrement des MFCC nettoyés.

2. Modélisation par GMM

Chaque locuteur est modélisé par un ensemble de modèles GMM, chacun étant entraîné sur l'énergie des MFCC pour la premiere méthode et entraîné juste sur les carateristique MFCC pour la deuxième méthode, avec un nombre de composantes gaussiennes variant parmi les valeurs suivantes : 8, 16, 32, 64, 128 et 256.

III Résutats et discussion

Dans ce travail, nous avons réalisé deux types de tests :

1. **L'identification automatique de locuteur**,
2. **La vérification automatique de locuteur**.

🔻 Identification automatique de locuteur :

Pour ce test, nous avons évalué tous les segments audio de test sur les modèles GMM du même genre (les femmes avec les femmes, les hommes avec les hommes), afin de prédire l'identité du locuteur.
Par exemple, un **segment de test de l'homme H1** est comparé à tous les modèles GMM des hommes (de H1 à H10).
La prédiction est faite en sélectionnant le modèle qui donne **le score de vraisemblance maximal**.

🔻 Vérification automatique de locuteur :

Pour ce test, tous les segments audio de chaque personne sont comparés à **tous les modèles GMM**.
La vérification est considérée comme **valide si le score dépasse un seuil**.
Ce seuil est déterminé comme étant le point où le **taux de fausses acceptations est égal au taux de fausses rejets** (également appelé **Equal Error Rate - EER**).

Les **graphiques ci-dessous** montrent les résultats obtenus pour chaque type de test.

3. Identification automatique de locuteur (Méthode 1)

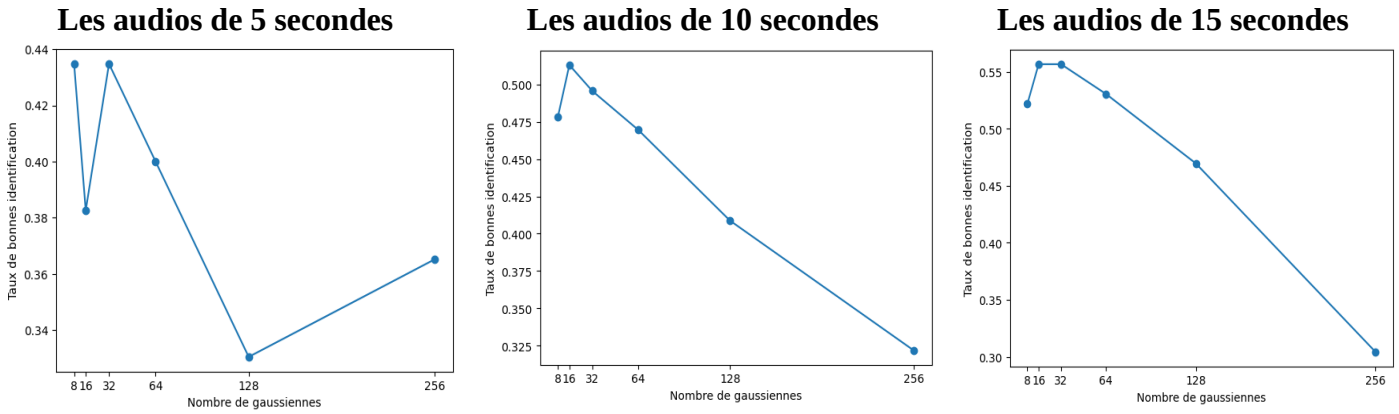

Figure 1 : Taux de bonne identification en fonction du nombre de gaussiennes utilisées pour l'entraînement du modèle.

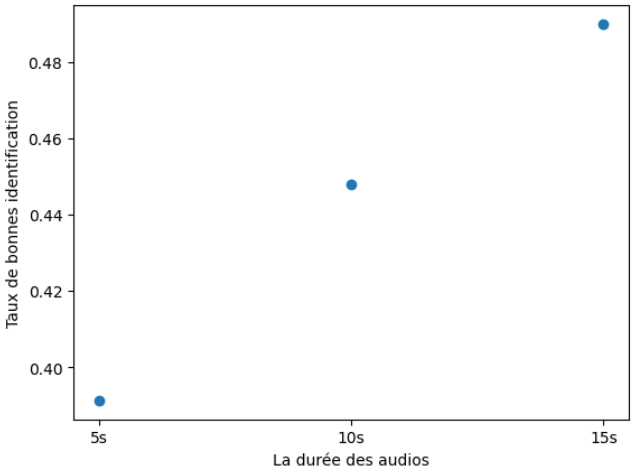

Figure 2 : Taux de bonnes identification en fonction de la durée des audios

Conclusion :

D'après les résultats obtenus, on peut conclure que le nombre optimal de composantes du GMM ayant donné les meilleurs résultats est 32. De plus, le taux de bonne identification augmente avec la durée de l'audio

4. Vérification automatique de locuteur (Méthode 2)

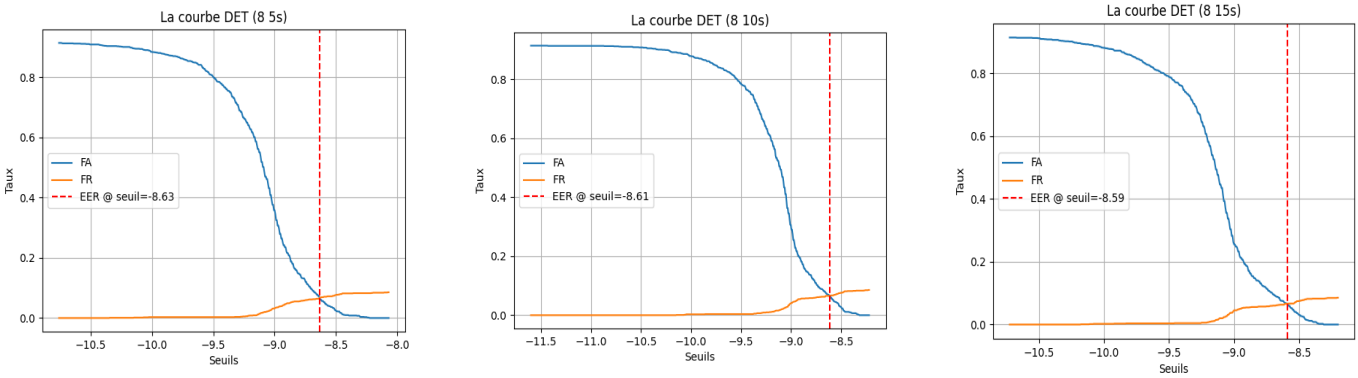

Figure 3 : Courbe DET du modèle GMM avec 8 composantes pour des durées audio de 5s, 10s et 15s.

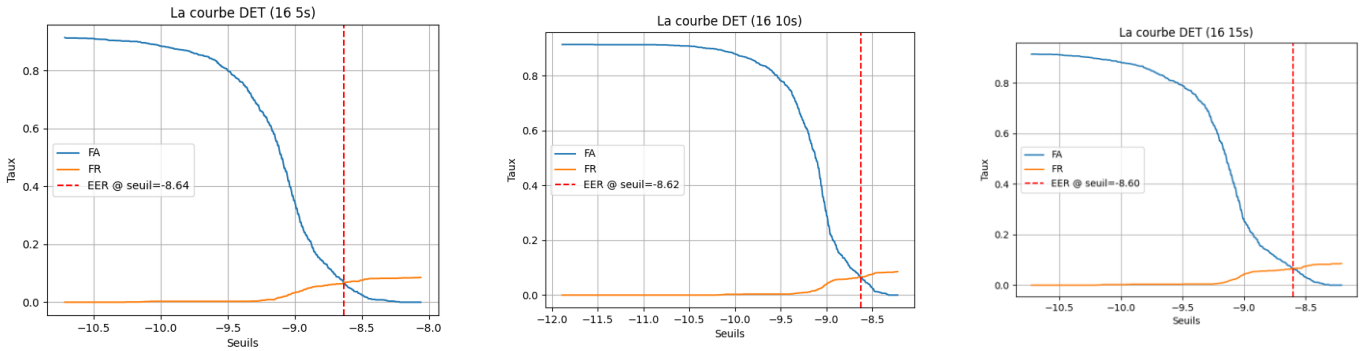

Figure 4 : Courbe DET du modèle GMM avec 16 composantes pour des durées audio de 5s, 10s et 15s.

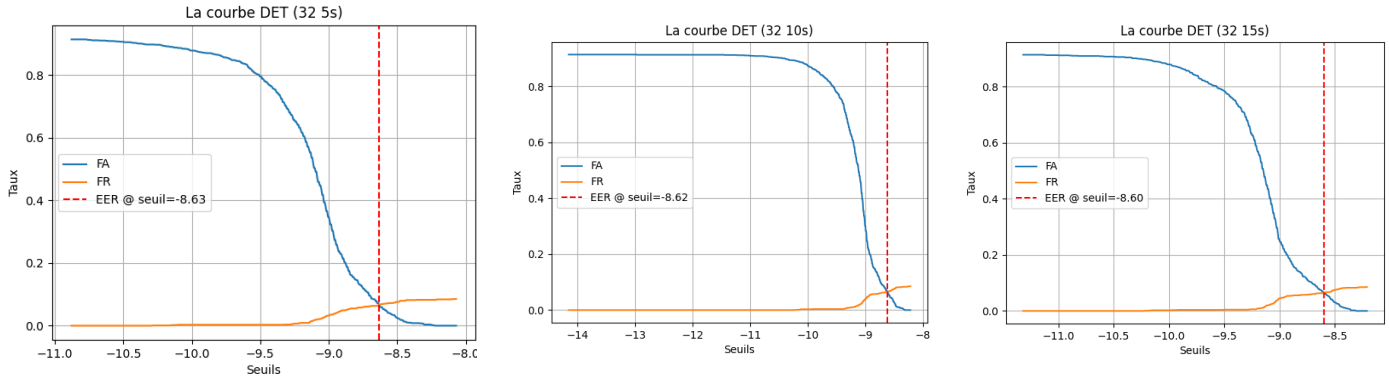

Figure 5 : Courbe DET du modèle GMM avec 32 composantes pour des durées audio de 5s, 10s et 15s.

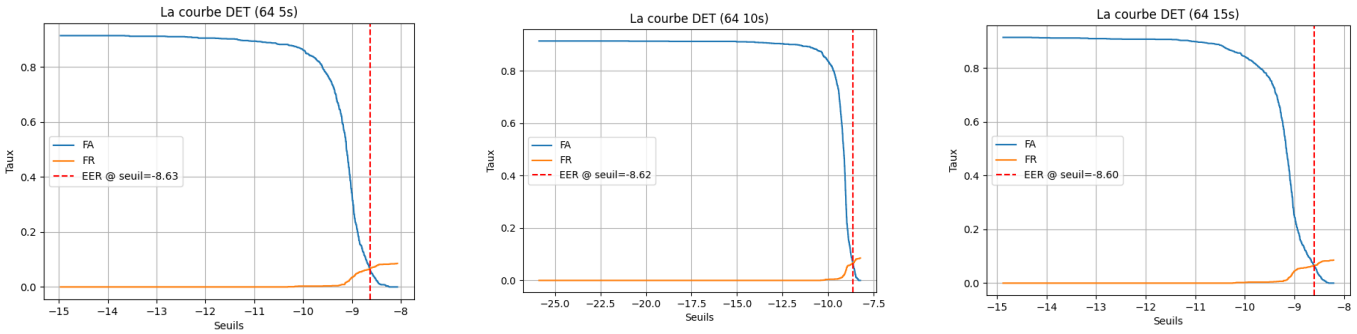

Figure 6 : Courbe DET du modèle GMM avec 64 composantes pour des durées audio de 5s, 10s et 15s.

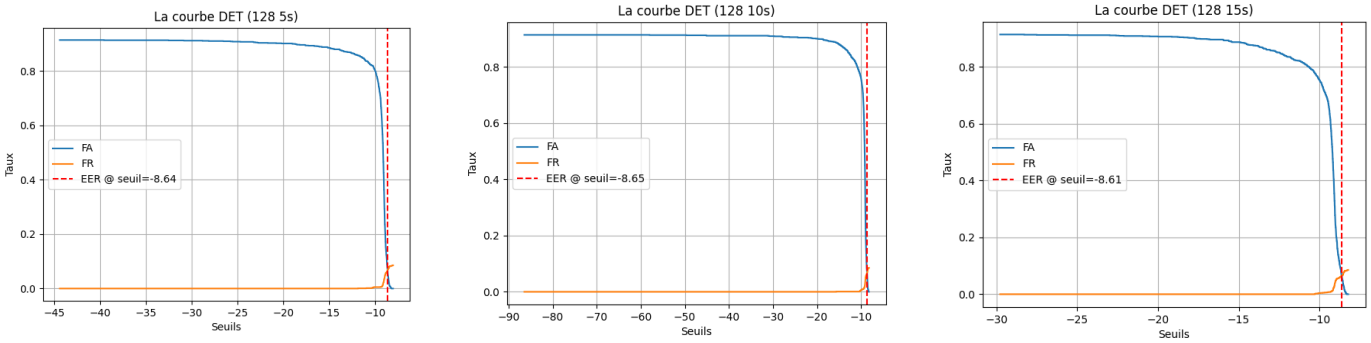

Figure 7 : Courbe DET du modèle GMM avec 128 composantes pour des durées audio de 5s, 10s et 15s.

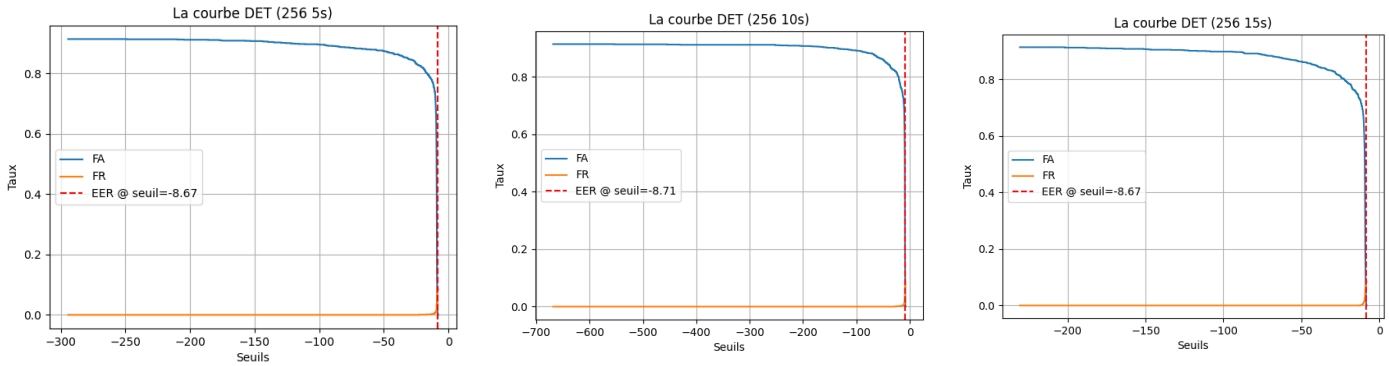

Figure 8 : Courbe DET du modèle GMM avec 256 composantes pour des durées audio de 5s, 10s et 15s.

- **Le taux d'égale erreur pour chaque modèle et selon la durée de l'audio :**

Nombre de guassiennes	5 secondes	10 secondes	15 secondes
8	0.07	0.06	0.07
16	0.07	0.07	0.07
32	0.07	0.07	0.07
64	0.06	0.07	0.07
128	0.06	0.06	0.06
256	0.07	0.07	0.07

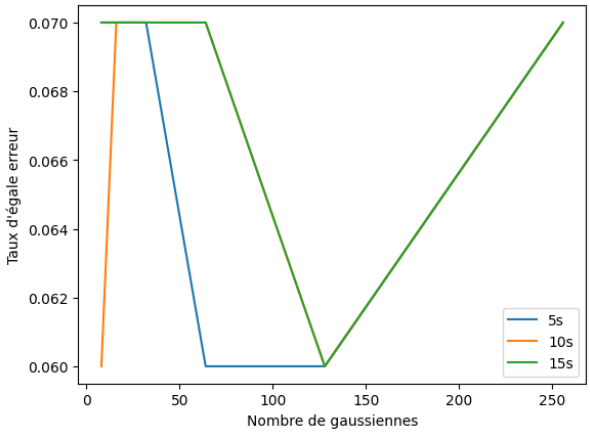

Figure 9 : Visualisation des taux d'erreur pour les trois types d'audio (5s, 10s, 15s).

Conclusion :

D'après les résultats obtenus, on observe que le **taux d'erreur égale (EER)** reste relativement **stable**, avec des valeurs comprises entre **0.06 et 0.07**.

Le **meilleur compromis** entre complexité du modèle et performance est obtenu avec **32 ou 64 gaussiennes**, en particulier pour des enregistrements de **10 à 15 secondes**. Cela montre qu'un modèle trop simple (8 gaussiennes) ou trop complexe (128 ou 256) **n'améliore pas forcément les performances**, et qu'un bon **équilibre est crucial**.

5. Identification automatique de locuteur (Méthode 2)

Les audios de 5 secondes **Les audios de 10 secondes** **Les audios de 15 secondes**

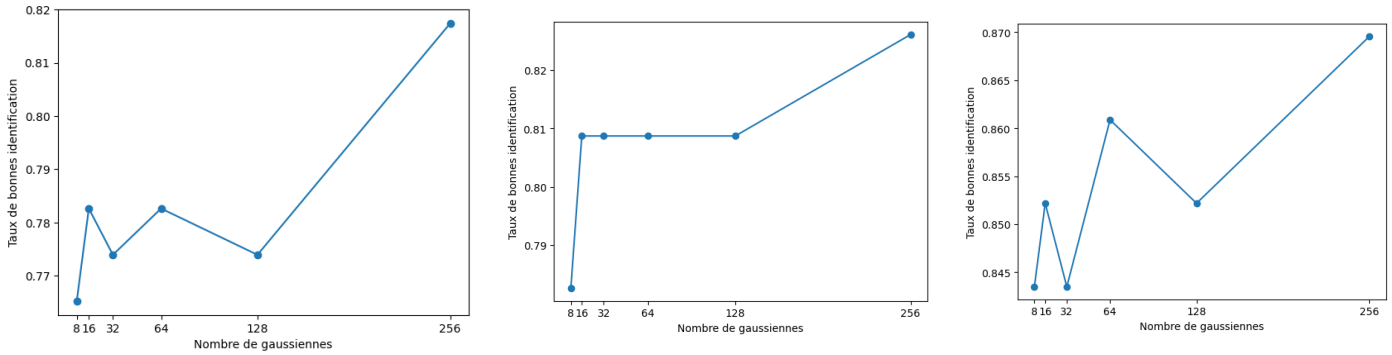

Figure 10 : Taux de bonne identification en fonction du nombre de gaussiennes utilisées pour l'entraînement du modèle.

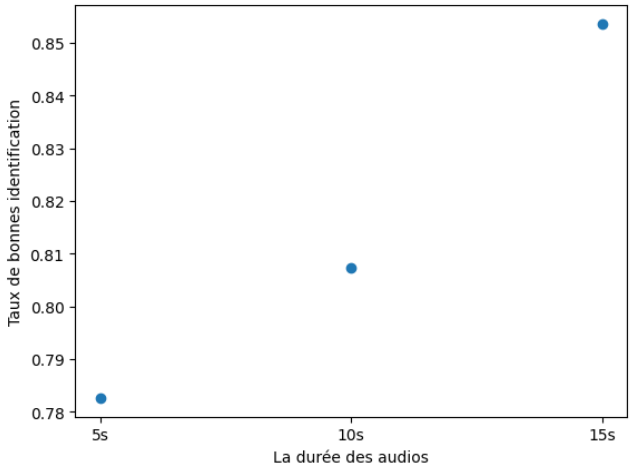

Figure 11 : Taux de bonnes identification en fonction de la durée des audios

Conclusion :

D'après les résultats obtenus, on peut conclure que le nombre optimal de composantes du GMM ayant donné les meilleurs résultats est 256 avec un pourcentage de 86.95%. De plus, le taux de bonne identification augmente avec la durée de l'audio comme dans la premiere méthode.

6. Vérification automatique de locuteur (Méthode 2)

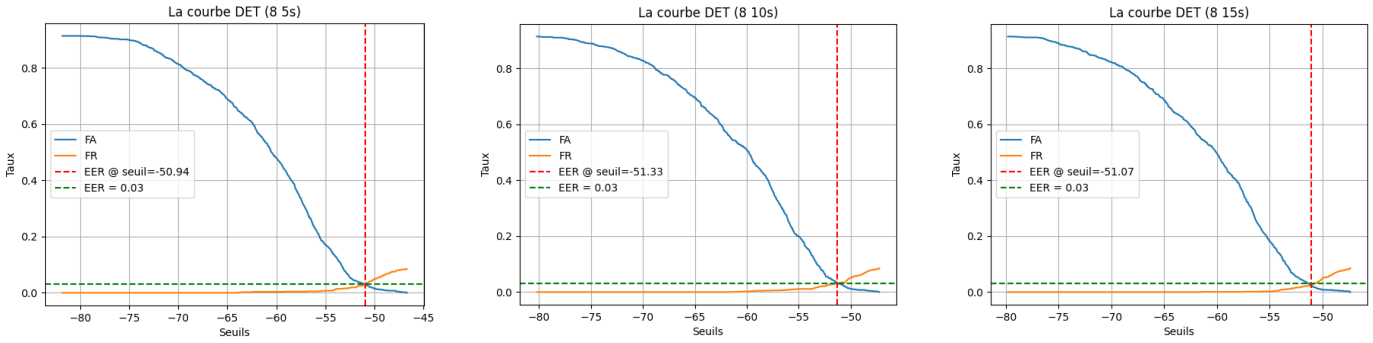

Figure 12 : Courbe DET du modèle GMM avec 8 composantes pour des durées audio de 5s, 10s et 15s.

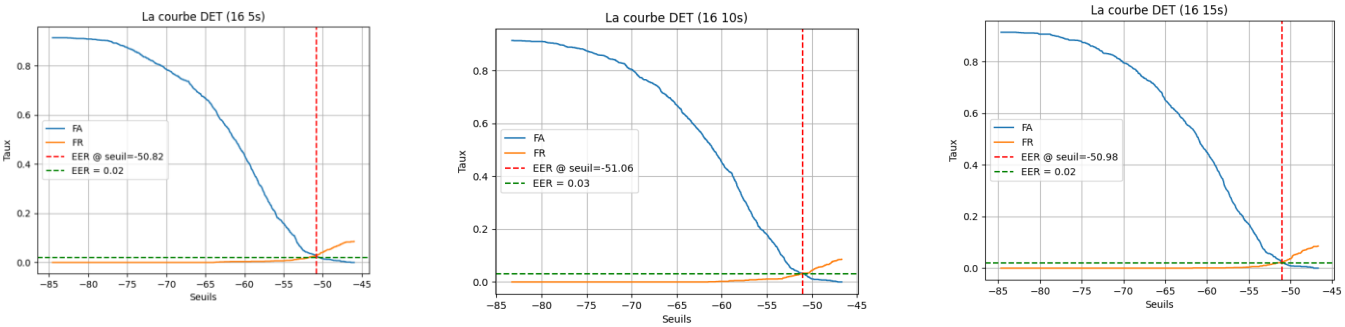

Figure 13 : Courbe DET du modèle GMM avec 16 composantes pour des durées audio de 5s, 10s et 15s.

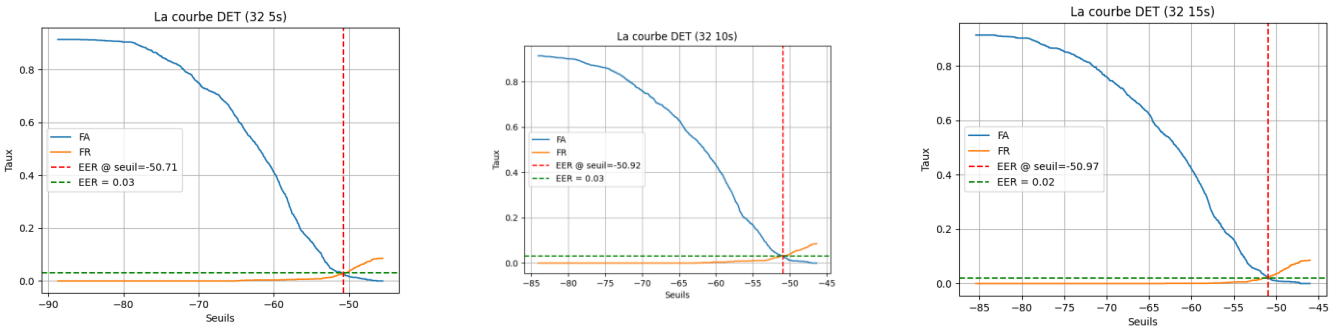

Figure 14 : Courbe DET du modèle GMM avec 32 composantes pour des durées audio de 5s, 10s et 15s.

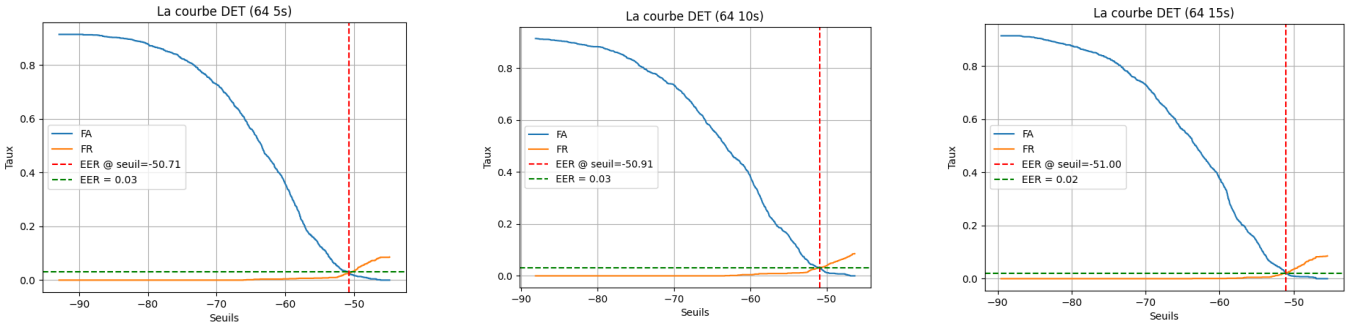

Figure 15 : Courbe DET du modèle GMM avec 64 composantes pour des durées audio de 5s, 10s et 15s.

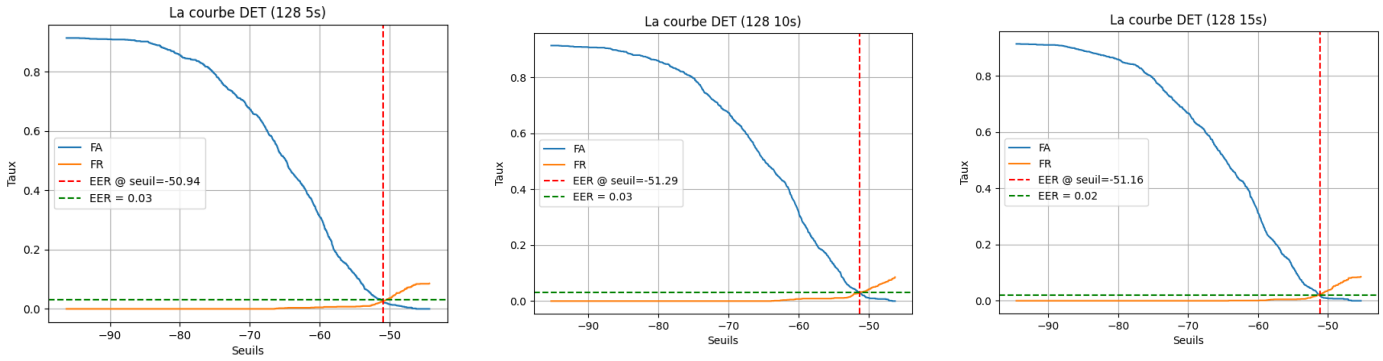

Figure 16 : Courbe DET du modèle GMM avec 128 composantes pour des durées audio de 5s, 10s et 15s.

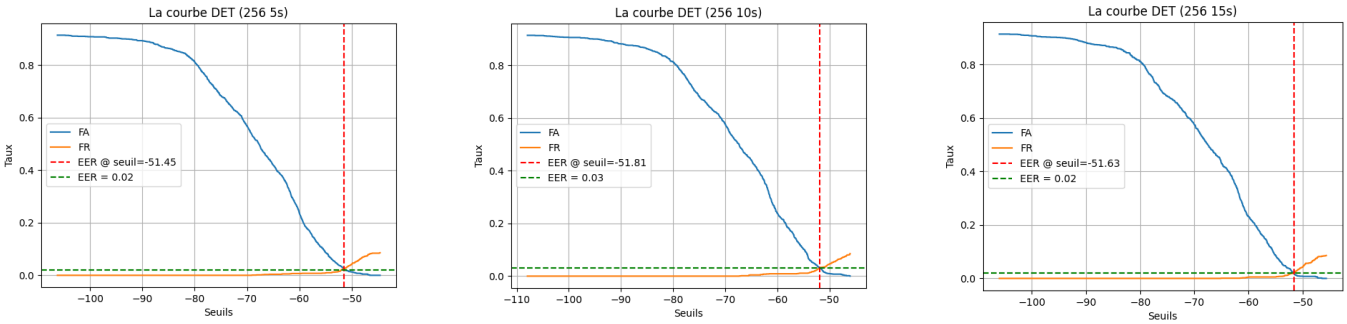

Figure 17 : Courbe DET du modèle GMM avec 256 composantes pour des durées audio de 5s, 10s et 15s.

- **Le taux d'égale erreur pour chaque modèle et selon la durée de l'audio :**

Nombre de guassiennes	5 secondes	10 secondes	15 secondes
8	0.03	0.03	0.03
16	0.02	0.03	0.02
32	0.03	0.03	0.02
64	0.03	0.03	0.02
128	0.03	0.03	0.02
256	0.02	0.03	0.02

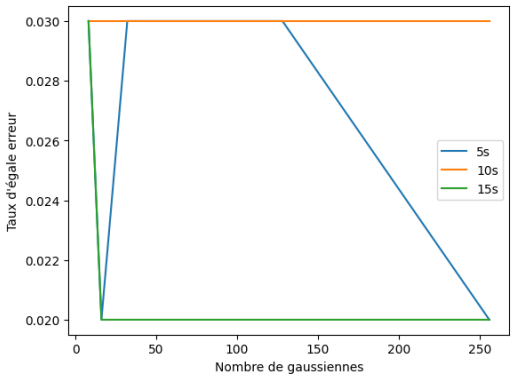

Figure 20 : Visualisation des taux d'erreur pour les trois types d'audio (5s, 10s, 15s).

Conclusion :

D’après les résultats obtenus, on observe que le **taux d’erreur égale (EER)** reste relativement **stable**, avec pour les audios de 5 et 15 secondes entre 0.02 et 0.03 et stable pour les audios de 10 secondes.

7. Analyse des résultats :

Les résultats obtenus, avec une précision maximale d’environ 52 %, s’expliquent en grande partie par la qualité des données utilisées, mais aussi par la nature des caractéristiques extraites. Dans la première méthode, les modèles GMM ont été entraînés uniquement sur l’énergie du signal, qui représente un nombre réduit d’informations (une seule valeur par trame).

En revanche, dans la deuxième méthode, les modèles ont été entraînés directement sur les coefficients MFCC, sans passer par le calcul de l’énergie. Ces caractéristiques offrent une représentation bien plus riche (typiquement 13 coefficients par trame), ce qui permet aux modèles d’identifier plus facilement les locuteurs. Cela se traduit par une nette amélioration du taux de bonne identification, passant de 52 % à 86,95 %.

De plus, l’augmentation de la durée des enregistrements audio a également un impact positif sur les performances. Dans la première méthode, le taux de bonne identification est passé de 44 % pour des audios de 5 secondes à 56 % pour ceux de 15 secondes. Dans la deuxième méthode, ce taux est passé de 78 % à 86,95 %. Cela confirme que plus l’audio est long, plus il contient d’informations utiles à l’identification du locuteur, ce qui améliore la performance des modèles.

Pour la verification de locuteur les resulats montre la premier methode a donnees des par ce que il y a des changes brusque au niveau de seuil optimal

IV Conclusion:

Ce travail a permis d'explorer en profondeur l’utilisation des **Modèles de Mélange Gaussien (GMM)** dans le cadre de l’identification et de la vérification du locuteur. À travers l’extraction des caractéristiques **MFCC**, la suppression du silence via des techniques de **Voice Activity Detection (VAD)**, et l’apprentissage supervisé de modèles GMM, nous avons pu évaluer les performances du système selon différents paramètres.

Les résultats expérimentaux montrent que :

- Le **nombre optimal de composantes gaussiennes** pour le GMM est **32 et 256**, offrant le meilleur selon le nombre de caractéristiques avec laquelle est le modèle est entrainé.

- L’**augmentation de la durée de l’audio** (de 5s à 15s) **améliore significativement le taux de bonne identification**, confirmant l’importance de la quantité d’informations vocales pour la robustesse du système.
- Le **taux d’erreur égale (EER)** reste relativement stable.

En résumé, les résultats obtenus démontrent l’efficacité des GMM pour la reconnaissance vocale, à condition d’utiliser des **caractéristiques bien traitées** et des **modèles de complexité adaptée**.