The People's Democratic Republic of Algeria
Ministry of Higher Education and Scientfiic Research



Ferhat Abbas University of Sétfi-1
Faculty of Sciences
Computer Science Department

Deepfake Audio Detection for Social Engineering Defense

Master Thesis — Computer Science

Major: Cyber Security

Supervised by
Pr. TOUMI Lyazid

Presented by
BOUDRAA Fatima
CHAABNA Manar

Jury Members
BALBAL Samir
SEMCHEDINE Moussa

Academic Year: 2025–2026

Contents

Abstract **v**

Résumé **vi**

ملخص **vii**

General Introduction **1**

1 Background and Fundamentals **6**

Introduction 6

1.1 Cybersecurity and Social Engineering 6

1.1.1 Overview of Cybersecurity 6

1.1.2 Cybersecurity Threats Taxonomy 7

1.1.3 Social Engineering: Definition and Mechanisms 7

1.1.4 Types of Social Engineering Attacks 9

1.1.5 The Rise of AI-Powered Social Engineering 9

1.1.6 Summary 10

1.2 Audio Deepfakes: Definition, History, and Threat Landscape 10

1.2.1 What is a Deepfake? 10

1.2.2 A Brief History of Deepfakes 11

1.2.3 Impact of Audio Deepfakes 11

1.2.4 Why Audio has Become the Primary Attack Vector 12

1.2.5 Summary 13

1.3 Audio Signal Processing Fundamentals 13

1.3.1 The Digital Audio Signal 13

1.3.2 Spectrogram and Mel-Spectrogram 15

1.3.3 Mel-Frequency Cepstral Coefficients 15

1.3.4 Prosodic Features 16

1.3.5 Summary 17

1.4 Chapter Summary 17

2 Related Work **18**
Introduction . . . 18
2.1 Deepfake Audio Generation Methods . . . 18
2.1.1 Text-to-Speech Synthesis . . . 19
2.1.2 Voice Conversion . . . 20
2.1.3 Neural Vocoders . . . 21
2.1.4 Summary . . . 22
2.2 Deepfake Audio Detection Approaches . . . 22
2.2.1 Classical Feature-Based Methods . . . 22
2.2.2 Deep Learning-Based Detection . . . 23
2.2.3 Recent Advances in Deepfake Audio Detection . . . 25
2.2.4 Summary . . . 27
2.3 Datasets and Benchmarks . . . 27
2.3.1 ASVspoof Challenge Datasets . . . 28
2.3.2 WaveFake . . . 29
2.3.3 Open-World and Multilingual Datasets . . . 29
2.3.4 Summary . . . 30
2.4 Evaluation Metrics . . . 30
2.4.1 Equal Error Rate (EER) . . . 31
2.4.2 Area Under the ROC Curve (AUC) . . . 31
2.4.3 Accuracy, Precision, Recall, and F1-Score . . . 32
2.4.4 Comparison and Choice of Metric . . . 33
2.4.5 Summary . . . 34
2.5 Limitations and Research Gaps . . . 34
2.5.1 Cross-System and Cross-Dataset Generalisation . . . 34
2.5.2 Robustness to High-Quality Synthesisers . . . 35
2.5.3 Underutilisation of Prosodic Features . . . 35
2.5.4 Real-Time Deployment Constraints . . . 35
2.5.5 Summary . . . 35
2.6 Chapter Summary . . . 36

3 Proposed Architecture and Methodology **37**
Introduction . . . 37
3.1 Motivation and Design Choices . . . 37
3.2 System Overview . . . 38
3.3 Audio Preprocessing and Data Augmentation . . . 39
3.3.1 Data Augmentation . . . 40
3.4 Multi-Modal Feature Extraction . . . 41

3.4.1 Mel-Frequency Cepstral Coefficients (MFCC) 41
3.4.2 Mel-Spectrogram 42
3.4.3 Fundamental Frequency F0 43
3.4.4 Complementarity of the Three Features 43
3.5 CNN Feature Encoding Branches 44
3.5.1 CNN 2D Branch for MFCC and Mel-Spectrogram 44
3.5.2 CNN 1D Branch for F0 45
3.6 Weighted Multi-Modal Fusion 45
3.7 Transformer Encoder Stack 46
3.7.1 Architecture 46
3.7.2 Global Average Pooling 47
3.8 MLP Classifier Head 48
3.9 Training Strategy 48
3.9.1 Loss Function 48
3.9.2 Optimizer and Learning Rate Schedule 49
3.9.3 Regularization 49
3.9.4 Hyperparameter Summary 50
3.10 Chapter Summary 50

4 Experiments and Results 52
Introduction 52
4.1 Experimental Setup 52
4.1.1 Datasets 52
4.1.2 Hardware and Computational Resources 53
4.1.3 Evaluation Metrics 54
4.1.4 Computational Complexity 55
4.2 Results on ASVspoof 2019 LA 56
4.2.1 Main Results 56
4.2.2 Training Dynamics 57
4.2.3 Comparison with State-of-the-Art Baselines 57
4.3 Results on WaveFake 59
4.3.1 Main Results 59
4.3.2 Comparison with Baselines on WaveFake 59
4.4 Ablation Study 61
4.4.1 ASVspoof 2019 LA 61
4.4.2 WaveFake 63
4.4.3 Summary 64
4.5 Multi-Dataset Generalisation 64

4.5.1 Cross-Dataset Evaluation Results 64
4.5.2 Analysis 64
4.5.3 Summary 65
4.6 Discussion 66
4.6.1 Summary of Main Findings 66
4.6.2 Limitations and future works 66
4.7 Chapter Summary 67

General Conclusion **68**

Bibliography **72**

Abstract

The rapid advancement of speech synthesis has enabled audio deepfakes — synthetic voices virtually indistinguishable from genuine human speech. When exploited through social engineering attacks such as vishing and impersonation, they pose a severe threat to organisational security, as demonstrated by fraud incidents exceeding $25 million in losses.

This thesis proposes a hybrid multi-modal detection architecture that simultaneously exploits three complementary acoustic features: Mel-Frequency Cepstral Coefficients (MFCC), Mel-Spectrogram, and Fundamental Frequency (F0), each processed by a dedicated convolutional branch. The resulting embeddings are combined through a learnable weighted fusion layer, refined by a two-layer Transformer encoder with multi-head self-attention, and classified by a four-layer MLP head.

Evaluated on ASVspoof 2019 LA and WaveFake, the system achieves an EER of **2.03%** and **0.0452%** respectively, outperforming RawNet2 (5.13%) and LCNN (5.06%), and remaining competitive with AASIST (0.83%) at 7.8× lower computational cost (1.233 GFLOPs vs. 9.618 GFLOPs). A mixed training strategy (70% ASVspoof + 30% WaveFake) reduces the cross-dataset generalisation gap from 46.41% to 1.11% EER on WaveFake.

Keywords: deepfake audio detection, social engineering, multi-modal feature fusion, MFCC, Mel-Spectrogram, fundamental frequency, Transformer encoder, ASVspoof, WaveFake, generalisation.

Résumé

Les progrès rapides de la synthèse vocale ont permis la création de *deepfakes* audio — des voix synthétiques quasiment impossibles à distinguer de la parole humaine authentique. Exploitées dans le cadre d'attaques d'ingénierie sociale telles que le *vishing* et l'usurpation d'identité, elles représentent une menace grave pour la sécurité des organisations, comme l'illustrent des incidents réels ayant entraîné des pertes dépassant 25 millions de dollars.

Ce mémoire propose une architecture de détection hybride et multimodale exploitant simultanément trois représentations acoustiques complémentaires : les Coefficients Cepstraux à Échelle de Mel (MFCC), le Mel-Spectrogramme et la Fréquence Fondamentale (F0), chacune traitée par une branche convolutionnelle dédiée. Les embeddings résultants sont combinés par une couche de fusion pondérée apprise, raffinés par un encodeur Transformer à deux couches avec auto-attention multi-têtes, puis classifiés par un MLP à quatre couches.

Évalué sur ASVspoof 2019 LA et WaveFake, le système atteint un EER de **2,03%** et **0,0452%** respectivement, surpassant RawNet2 (5,13%) et LCNN (5,06%), et restant compétitif avec AASIST (0,83%) à un coût de calcul 7,8× inférieur (1,233 GFLOPs contre 9,618 GFLOPs). Une stratégie d'entraînement mixte (70% ASVspoof + 30% WaveFake) réduit le fossé de généralisation inter-corpus de 46,41% à 1,11% d'EER sur WaveFake.

Mots-clés : détection audio deepfake, ingénierie sociale, fusion multimodale de caractéristiques, MFCC, Mel-Spectrogramme, fréquence fondamentale, encodeur Transformer, ASVspoof, WaveFake, généralisation.

ملخص

أتاح التقدم السريع في تقنيات تركيب الكلام إنشاء تسجيلات صوتية مزيفة، يكاد يستحيل تمييزها عن الكلام البشري الحقيقي. وعندما تُوظَّف في هجمات الهندسة الاجتماعية كانتحال الهوية الصوتية، فإنها تُمثّل تهديدًا بالغ الخطورة لأمن المؤسسات، كما تجلّى في حوادث احتيال موثَّقة تجاوزت خسائرها 25 مليون دولار.

تقترح هذه الأطروحة معمارية كشف هجينة متعددة الأنماط تستغل ثلاث تمثيلات صوتية متكاملة: معاملات السيبستروم (MFCC)، ومطياف ميل (Mel-Spectrogram)، والتردد الأساسي (F0)، تُعالَج كل منها بفرع تلافيفي مخصص. تُدمَج التمثيلات عبر طبقة دمج موزونة قابلة للتعلم، ثم تُكثَّف بمشفِّر Transformer من طبقتين بالانتباه الذاتي متعدد الرؤوس، وتُصنَّف بطبقة MLP متعددة الطبقات.

على مجموعة ASVspoof 2019 LA حقَّق النظام معدل خطأ متساوٍ (EER) بلغ **2.03%**، متفوقًا على RawNet2 (5.13%) وLCNN (5.06%)، ومنافسًا لـAASIST (0.83%) بتكلفة حسابية أقل بـ7.8 أضعاف (1.233 مقابل 9.618 GFLOPs). وعلى WaveFake بلغ EER **0.0452%**. وتُقلِّص استراتيجية التدريب المختلط (ASVspoof + 30% WaveFake %70) فجوة التعميم من 46.41% إلى 1.11% على WaveFake.

الكلمات المفتاحية: كشف التزوير الصوتي، الهندسة الاجتماعية، دمج الخصائص متعدد الأنماط، MFCC، مطياف ميل، التردد الأساسي، مشفِّر Transformer، ASVspoof، WaveFake، التعميم.