

الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbas



Master's Thesis

To obtain the diploma of Master's Degree

Field of Study: **Computer Science**

Specialization: **Intelligence Artificielle et Science de Données (IASD)**

Theme

A Comprehensive Study of Deep Learning Approaches for Speech-to-Text Recognition

Presented by
BELMANA Soufyane

Defended on: **28 June, 2026**
In front of the jury composed of:

Dr. BEZZAOUCHA Fatima Souad
Dr. CHIKH Asma
Pr. AMAR BENSABER Djamel
Dr. ABDELHAK Soumia

President
Examiner
Supervisor
Incubator Representative

Academic Year: 2025/2026

Abstract

Automatic Speech Recognition (ASR) has undergone a remarkable transformation over the past decade, driven primarily by advances in deep learning. This thesis presents a comprehensive study of deep learning approaches for speech-to-text recognition, tracing the evolution from traditional hybrid systems based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) to modern end-to-end architectures powered by Transformers and self-supervised learning.

We systematically analyze the major deep learning architectures that have been applied to ASR, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, attention mechanisms, and Transformer-based models such as the Conformer. We examine the three main end-to-end paradigms—Connectionist Temporal Classification (CTC), attention-based encoder-decoder models, and RNN-Transducers—along with hybrid approaches that combine their strengths.

Furthermore, we investigate state-of-the-art systems including OpenAI’s Whisper, Meta’s wav2vec 2.0, and Google’s Conformer, analyzing their architectures, training methodologies, and performance characteristics. Our comparative analysis across standard benchmarks demonstrates that modern systems have achieved Word Error Rates (WER) as low as 1.5% on LibriSpeech, surpassing estimated human-level performance. We conclude by discussing remaining challenges including robustness in adverse conditions, multilingual coverage, computational efficiency, and fairness, as well as promising future research directions.

Keywords— Speech Recognition, Deep Learning, Automatic Speech Recognition (ASR), Transformer, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), LSTM, Attention Mechanism, Connectionist Temporal Classification (CTC), wav2vec 2.0, Whisper, Conformer, Self-Supervised Learning, End-to-End Models

Résumé

La reconnaissance automatique de la parole (RAP) a connu une transformation remarquable au cours de la dernière décennie, principalement grâce aux avancées de l'apprentissage profond. Cette thèse présente une étude approfondie des approches d'apprentissage profond pour la reconnaissance vocale (conversion de la parole en texte), retraçant l'évolution depuis les systèmes hybrides traditionnels basés sur les modèles de Markov cachés (HMM) et les modèles de mélange gaussien (GMM) jusqu'aux architectures modernes de bout en bout alimentées par les Transformers et l'apprentissage auto-supervisé.

Nous analysons systématiquement les principales architectures d'apprentissage profond appliquées à la RAP, notamment les réseaux de neurones profonds (DNN), les réseaux de neurones convolutifs (CNN), les réseaux de neurones récurrents (RNN) avec leurs variantes LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit), les mécanismes d'attention, ainsi que les modèles basés sur les Transformers tels que le Conformer. Nous examinons les trois principaux paradigmes de bout en bout : la Classification Temporelle Connexionniste (CTC), les modèles encodeur-décodeur à base d'attention, et les RNN-Transducers, ainsi que les approches hybrides combinant leurs forces.

En outre, nous étudions les systèmes à l'état de l'art, notamment Whisper d'OpenAI, wav2vec 2.0 de Meta et le Conformer de Google, en analysant leurs architectures, leurs méthodologies d'entraînement et leurs caractéristiques de performance. Notre analyse comparative sur des benchmarks standards démontre que les systèmes modernes ont atteint des taux d'erreur de mots (WER) aussi bas que 1,5% sur LibriSpeech, dépassant les performances estimées au niveau humain. Nous concluons en discutant les défis restants, notamment la robustesse dans des conditions défavorables, la couverture multilingue, l'efficacité computationnelle et l'équité, ainsi que les directions de recherche futures prometteuses.

Mots-clés : Reconnaissance vocale, Apprentissage profond, Reconnaissance automatique de la parole (RAP), Transformer, Réseaux de neurones convolutifs (CNN), Réseaux de neurones récurrents (RNN), LSTM, Mécanisme d'attention, Classification Temporelle Connexionniste (CTC), wav2vec 2.0, Whisper, Conformer, Apprentissage auto-supervisé, Modèles de bout en bout

ملخص

شهد التعرف التلقائي على الكلام تحولاً ملحوظاً خلال العقد الأخير، مدفوعاً بشكل رئيسي بالتقدم في مجال التعلم العميق. تقدم هذه الأطروحة دراسة شاملة لمقاربات التعلم العميق في التعرف على الكلام وتحويله إلى نص، متتبعاً التطور من الأنظمة الهجينة التقليدية المبنية على نماذج ماركوف المخفية ونماذج المزيج الغاوسي إلى البنى الحديثة من طرف إلى طرف المدعومة بالمحولات والتعلم الذاتي الإشراف. نقوم بتحليل منهجي لأهم بنى التعلم العميق المطبقة في التعرف على الكلام، بما في ذلك الشبكات العصبية العميقة والشبكات العصبية التلافيفية والشبكات العصبية المتكررة بمتغيراتها مثل الذاكرة الطويلة قصيرة المدى ووحدة البوابة المتكررة، وآليات الانتباه، والنماذج المبنية على المحولات مثل الكونفورمر. ندرس ثلاثة نماذج رئيسية من طرف إلى طرف: التصنيف الزمني الاتصالي، ونماذج المشفر-فاك الشفرة المبنية على الانتباه، ومحولات الشبكات العصبية المتكررة، بالإضافة إلى المقاربات الهجينة التي تجمع بين نقاط قوتها. علاوة على ذلك، نستقصي أحدث الأنظمة بما في ذلك ويسبر من أوبن أي آي، وواف تو فيك 0.2 من ميتا، والكونفورمر من غوغل، محللين بنياتها ومنهجيات تدريبها وخصائص أدائها. يُظهر تحليلنا المقارن عبر المعايير القياسية أن الأنظمة الحديثة حققت معدلات خطأ في الكلمات منخفضة تصل إلى 1.5% على مجموعة ليبري سبيتش، متجاوزة الأداء البشري المقدر. نختم بمناقشة التحديات المتبقية بما في ذلك المتانة في الظروف الصعبة، والتغطية متعددة اللغات، والكفاءة الحسابية، والعدالة، فضلاً عن اتجاهات البحث المستقبلية الواعدة.

الكلمات المفتاحية: التعرف على الكلام، التعلم العميق، التعرف التلقائي على الكلام، المحولات، الشبكات العصبية التلافيفية، الشبكات العصبية المتكررة، الذاكرة الطويلة قصيرة المدى، آلية الانتباه، التصنيف الزمني الاتصالي، Conformer، Whisper، wav2vec 2.0، التعلم الذاتي الإشراف، نماذج من طرف إلى طرف.