الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم الاتصالات

حذف الصدى الصوتي لتحسين دقة أنظمة تعرّف الكلام آلياً

دراسة أُعدت لنيل درجة الماجستير في هندسة الاتصالات من المعهد العالي للعلوم التطبيقية والتكنولوجيا

إعداد

عبير عبدالله

إشراف

د. أميمة الدكاك

د. آصف جعفر

2025

لجنة الحكم

د. علي كاظم المعهد العالي للعلوم التطبيقية والتكنولوجيا

رئيساً ومقرراً

د. هيام خدام جامعة دمشق

مقرراً

د. مجدي مسلم المعهد العالي للعلوم التطبيقية والتكنولوجيا

مقرراً

د. أميمة الدكاك المعهد العالي للعلوم التطبيقية والتكنولوجيا

مقرراً ومشرفاً

تصريح

أنا الموقع أدناه **عبير عبدالله** معدّ أطروحة الماجستير التي تحمل العنوان:

حذف الصدى الصوتي لتحسين دقة أنظمة تعرّف الكلام آلياً

أصرح بأن:

- الأعمال والنتائج المعروضة في هذه الأطروحة هي نتيجة جهودي الشخصية وبتوجيه من المشرف، وأن ما عدا ذلك من معلومات ونتائج قد نُسبت إلى مصادرها ومؤلفيها، وأُشير إلى ذلك في متن النص وفي قائمة المراجع.
- المعطيات والمعلومات المستخدمة في هذه الأطروحة جرى تحصيلها بطرائق سليمة ومشروعة ونُسبت إلى مصادرها في المواضع الملائمة.
- كلّ مكوّن من مكونات هذه الأطروحة (مقطع نصّي، صورة، مخطط، ...) مقتبس من عمل آخر جرى تمييزه بوضوح ونُسب إلى مصدره.
- الأعمال والنتائج المعروضة في هذه الأطروحة لم تُستخدم سابقًا وليست قيد الاستخدام للحصول على أي شهادة أكاديمية أخرى.

التوقيع

دمشق / / 202

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology

المعهد العالي للعلوم التطبيقية والتكنولوجيا مؤسسة حكومية للتعليم العالي أحدثت بموجب المرسوم التشريعي رقم /24/ لعام 1983، وذلك بهدف إعداد كوادر علمية متميزة من مهندسين وباحثين للإسهام الفاعل في عملية التطوير العلمي والتنمية في الجمهورية العربية السورية.

يمنح المعهد العالي درجة الإجازة في الهندسة في الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وعلوم وهندسة المواد وهندسة الطيران. يقبل المعهد العالي لدراسة هذه الاختصاصات شريحة منتقاة من المتفوقين في الشهادة الثانوية من الفرع العلمي. يتيح المعهد العالي أيضاً برامج ماجستير أكاديمي في نظم الاتصالات وفي التحكم والروبوتيك وفي نظم المعطيات الكبيرة ونظم المعلومات ودعم القرار وفي علوم وهندسة المواد وعلوم وهندسة البصريات. وأخيراً، يمنح المعهد العالي درجة الدكتوراه في الاتصالات والمعلوماتية ونظم التحكم والفيزياء التطبيقية. تُحدث في المعهد العالي اختصاصات جديدة بحسب متطلبات سوق العمل وتوجهات البحث والتطوير المحلية والعالمية.

إلى جانب النشاط التعليمي، يمارس المعهد العالي عبر جهود أطره وفعالياته العلمية المختلفة نشاطاً حثيثاً في البحث والتطوير، إذ ينفذ مشاريع ودراسات واستشارات متنوعة لصالح الجهات العامة والخاصة في القطر، كما يتعاون مع جهات خارج القطر في بعض المشاريع البحثية والتطويرية. يسعى المعهد أيضاً، عبر دورات تدريبية نظرية وعملية متاحة للقطاعين العام والخاص وللأفراد، إلى إفادة أوسع فئة من المهتمين من إمكانيات أطره العلمية ومختبراته. واستكمالاً لدوره الرائد في مجال التعليم ونشر العلم، ينشر المعهد العالي كتباً علمية عالية المستوى من نتاج أطره، منها ما هو تدريسي يوافق المناهج في المعهد العالي ويفيد شريحة واسعة من الطلاب الجامعيين عموماً، ومنها ما هو علمي ثقافي. يتيح المعهد العالي بعضاً من منشوراته على موقعه على الشابكة، كما يتيح إمكانية الاطلاع على رسائل الماجستير والدكتوراه المنفذة في المعهد العالي وعلى بعض منشورات طلابه وأطره من المقالات العلمية.

المعهد العالي للعلوم التطبيقية والتكنولوجيا، الجمهورية العربية السورية، دمشق، ص.ب 31983

Higher Institute for Applied Sciences and Technology – HIAST

P. O. Box 31983, Damascus, Syrian Arab Republic

هاتف 00963115123819 - فاكس 00963115140761

بريد إلكتروني contact@hiast.edu.sy

موقع إلكتروني www.hiast.edu.sy

كلمة شكر

أتقدّم بخالص التقدير والامتنان إلى الدكتورة الفاضلة أميمة الدكّاك على إشرافها الكريم ومتابعتها المتواصلة، فقد كان لحرصها الدائم توجيهاتها العلميّة الدقيقة أثر بالغ في إثراء هذا العمل وصقل نتائجه.

كما أتقدّم بخالص الشكر والعرفان للدكتور الفاضل آصف جعفر على دعمه العلميّ وملاحظاته الأكاديمية البنّاءة التي أسهمت في توضيح العديد من الجوانب البحثية وتعزيز الجانب التطبيقي.

ولا يفوتني أن أعبر عن شكري لكلّ من رئيس قسم الاتصالات الدكتور خلدون خرزم والدكتور عمر حمدون على ما قدّماه من دعم علمي ومساندة فعّالة كان لها أثر إيجابي في إتمام هذا البحث.

وفي الختام، أتوجّه بالشكر لكل من قدّم لي يد العون والدعم خلال مراحل هذا العمل، سائلاً الله أن يجزيهم خير الجزاء ويبارك جهودهم.

ملخص البحث

ما تزال مسألة حذف الصدى الصوتي مسألة تحدّ في مجال معالجة الإشارة الكلامية، وقد تناول هذا البحث المشكلة ضمن مجموعة شروط محددة، إذ ركزّ على حذف الصدى الصوتي من قناة أحادية بهدف تحسين دقة أنظمة تعرف الكلام آلياً.

انطلق البحث من مفارقة ASR-Dereverberation والتي تعبّر عن الصعوبة الجوهرية في تحقيق التوازن بين تحسين جودة الإشارة الكلامية من جهة، وضمان رفع دقّة التعرّف على الكلام من جهة أخرى. في المرحلة الأولى من العمل، تم اعتماد نموذج LSTM بوصفه خط الأساس لمعالجة الصدى الصوتي، نظراً لقدرته على تمثيل العلاقات الزمنية في الإشارة الكلامية. أظهر هذا النموذج كفاءة مقبولة في تحسين جودة الإشارة ولكن في المقابل ارتفع معدل الخطأ في الكلمات، وهذا أكدّ أن النهج التقليدي لا يحقق الحل الأمثل لمفارقة ASR–Dereverberation.

ومن هنا سعى البحث إلى تطوير مقاربة أكثر فاعلية لتجاوز هذه المفارقة، من خلال اعتماد نموذج Mamba المبني على نماذج فضاءات الحالة الانتقائية Selective State Space Models–S6 لمعالجة الصدى الصوتي. وقد أظهر هذا النموذج توازناً جيداً بين تحسين جودة الإشارة والحفاظ على دقة التعرّف على الكلام، مما جعله الأساس في بناء النظام المقترح.

وبناءً على نتائج النماذج السابقة، تمّ تصميم نظام متكامل End-to-End يجمع بين نموذج حذف الصدى المعتمد ونموذج HuBERT للتّعرف على الكلام آلياً. تم تدريب النموذج باستخدام تقنية Rank-Stabilized Low Rank Adaptation-rsLORA التي سمحت بتحديث أقل من 1.5% من موسطات نموذج HuBERT.

أظهرت النتائج التجريبية أن النظام المقترح حقّق تحسناً واضحاً في جودة الإشارة الكلامية PESQ، إلى جانب انخفاض ملموس في معدل الخطأ في الكلمات WER على مجموعات معطيات صدى واقعية ومتنوعة، حيث تفوّق النموذج المقترح بوضوح في بيئات الصدى المعقّدة مع تأثير محدود وغير جوهري على الإشارات النظيفة.

يُعدّ هذا النهج خطوة مهمة نحو بناء أنظمة Robust End-to-End ASR قادرة على العمل بكفاءة في ظروف صوتية واقعية، مما يجعلها إطاراً واعداً لتطوير حلول متقدمة للتعرّف على الكلام في اللغات المختلفة، بما فيها اللغة العربية.

الكلمات المفتاحية: حذف الصدى، قناة أحادية، الاستجابة النبضية للغرفة، الضبط الدقيق الموفر للموسطات، نماذج فضاء الحالة الانتقائية.

Abstract

The problem of speech dereverberation remains a significant challenge in the field of speech signal processing. This study addresses the issue under specific controlled conditions, focusing on single-channel dereverberation with the goal of improving the accuracy of Automatic Speech Recognition (ASR) systems. The research is motivated by the ASR–Dereverberation Paradox, which highlights the inherent difficulty in balancing the enhancement of speech signal quality with the preservation of recognition accuracy.

In the initial stage, an LSTM-based model was adopted as a baseline for dereverberation due to its capability to model temporal dependencies in speech signals. While this model achieved a noticeable improvement in perceptual speech quality, it simultaneously increased the Word Error Rate (WER), confirming that traditional recurrent approaches fail to effectively resolve the ASR–Dereverberation paradox.

To overcome this limitation, the study developed a more robust approach by adopting the Mamba architecture, which is based on Selective State Space Models (S6) for dereverberation. The Mamba model demonstrated a better balance between improving speech quality and maintaining ASR accuracy, forming the foundation for the proposed integrated system.

Based on these findings, an End-to-End system was designed by combining the adopted dereverberation model with the HuBERT model for speech recognition. The system was trained using the rsLoRA (Rank-Stabilized Low-Rank Adaptation) technique, which allowed the update of less than 1.5% of HuBERT parameters, ensuring computational efficiency and training stability.

Experimental results showed that the proposed system achieved a clear improvement in speech quality, as measured by PESQ, along with a significant reduction in WER across multiple real and simulated reverberant datasets. The model demonstrated superior performance in complex reverberant environments while maintaining minimal impact on clean signals. These findings indicate that the proposed framework represents a promising step toward developing Robust End-to-End ASR systems capable of operating efficiently in real-world acoustic conditions, offering a scalable foundation for speech recognition in multiple languages, including Arabic.

Keywords: Dereverberation, Single Channel, Room Impulse Response, Parameter-Efficient Fine-Tuning, Selective State Space Models.

المحتويات

الفصل الأول 1

1. تمهيد... 1

1.1- مقدمة عامة..... 1

2.1- مشكلة البحث..... 2

3.1- هدف البحث 2

4.1- أهمية البحث 2

5.1- محددات مسألة البحث 3

6.1- أسئلة البحث 4

7.1- مساهمات البحث..... 4

8.1- مخطط البحث 5

الفصل الثاني..... 6

2. الدراسة المرجعية 6

1.2- مقدمة..... 6

2.2- نظرة عامة موجزة عن نماذج حذف الصدى..... 7

3.2- النماذج الحديثة ذات الصلة 9

4.2- مقارنة بين الدارسات المرجعية 14

5.2- الخلاصة 17

الفصل الثالث 18

3. الدراسة النظرية 18

1.3- مقدمة..... 18

2.3- النمذجة الرياضية لصدى الغرفة..... 19
1.2.3- الاستجابة النبضية للغرفة وأقسامها 19
2.2.3- النموذج الرياضي 20
3.3- تأثير الصدى..... 21
4.3- الشبكات العصبونية العودية 22
1.4.3- شبكات Long-Short Term Memory- LSTM..... 23
5.3- تطور بنية نماذج فضاء الحالة..... 27
1.5.3- مقدمة وهدف النشأة 27
2.5.3- نماذج فضاء الحالة المستمرة State Space Model 28
3.5.3- نماذج فضاء الحالة المتقطعة Discrete State Space Model..... 30
1.3.5.3- آلية تنفيذ نماذج فضاء الحالة المتقطعة بين التمثيل التكراري والتلافيفي 31
2.3.5.3- اختيار وتهيئة مصفوفات نموذج فضاء الحالة في سياق التعلّم العميق..... 34
4.5.3- Structured State Space Sequence Models (S4) 35
5.5.3- Mamba..... 36
6.3- نموذج HuBERT..... 42
1.6.3- منهجية نموذج HuBERT 43
2.6.3- بنية النموذج..... 46
7.3- الضبط الدقيق Fine-Tuning..... 48
1.7.3- فرضية البُعد الجوهري المنخفض..... 49
2.7.3- تصنيف ومقارنة تقنيات الضبط الدقيق الموفِّر للموسطات 50
3.7.3- آلية عمل rsLoRA 52
8.3- الخلاصة 54
الفصل الرابع..... 55

4. تنفيذ واختبار نموذج لحذف الصدى بالاعتماد على بنية LSTM 55
1.4- تجهيز مجموعة المعطيات..... 55
1.1.4- اختيار مجموعات المعطيات 56
2.4- تصميم مجموعة المعطيات..... 62
3.4- البرمجيات والأدوات المستخدمة..... 65
4.4- بنية النموذج 65
1.4.4- مبدأ عمل النموذج 65
2.4.4- بنية النموذج 66
5.4- مرحلة اختيار السمات الطيفية..... 68
6.4- تدريب النموذج 69
7.4- مرحلة الاختبار 70
8.4- الخلاصة 72
الفصل الخامس..... 73
5. تطوير نظام End-to-End لحذف الصدى الصوتي بالاعتماد على بنية Mamba بالتكامل مع ASR 73
1.5- مجموعات المعطيات المستخدمة 73
2.5- مبدأ عمل النموذج وبنيته..... 75
1.2.5- المعالجة في المجال الترددي الخطي 75
2.2.5- المعالجة في مجال Mel 79
3.5- البرمجيات المستخدمة 79
4.5- مرحلة التدريب 80
1.4.5- موسطات النموذج..... 81
2.4.5- موسطات التدريب 81

5.5- مرحلة الاختبار 81
6.5- بنية النموذج المقترح end to end 84
1.6.5- تهيئة النظام المقترح للتدريب 85
2.6.5- مجموعات معطيات التدريب 85
3.6.5- موسطات التدريب 87
4.6.5- مرحلة الاختبار 89
5.6.5- مناقشة النتائج 91
7.5- الخلاصة 92
الفصل السادس 94
6. الخاتمة والآفاق المستقبلية 94
المراجع 96
الملحقات 101

قائمة الأشكال

الشكل 1.2- البنية المعمارية لنموذج TRU-Net.....10

الشكل 2-2- البنية المعمارية لنموذج D^2 Net 11

الشكل 3.2- البنية العامة لنموذج UFormer 11

الشكل 1.3- تمثيل للاستجابة النبضية للغرفة.....20

الشكل 2.3- نشر المخطط الحسابي للشبكات العودية Unfolded Graph 22

الشكل 3.3- بوابة الدخل في LSTM.....24

الشكل 4.3- تعديل ذاكرة الخلية في LSTM.....25

الشكل 5.3-بوابة الحذف في LSTM 25

الشكل 6.3- بوابة الخرج في LSTM.....26

الشكل 7.3- رسم توضيحي لخلية LSTM.....26

الشكل 3 .8- خلية والبوابات LSTM الثلاث.....27

الشكل 9.3- تمثيل مبسط لنظام ديناميكي يوضّح العلاقة بين الدخل والحالة والخرج في نموذج SSM 28

الشكل 10.3– مخطط تمثيلي للنظام الديناميكي في الزمن المستمر 29

الشكل 11.3- تمثيل عملية الانتقال من نموذج النظام الديناميكي في الزمن المستمر إلى التمثيل العودي المتقطع 32

الشكل 12.3- تمثيل عملية الانتقال من نموذج النظام الديناميكي في الزمن المستمر إلى التمثيل التلافيفي المتقطع 32

الشكل 13.3- التمثيل الرياضي لآلية حساب الخرج في نموذج SSM المتقطّع في الصيغة التلافيفية.....33

الشكل 14.3- البنية المعمارية لوحدة Mamba 37

الشكل 15.3- مقارنة بين آلية معالجة الأبعاد في نموذج Mamba والمحولات. 39

الشكل 16.3- البنية الداخلية لوحدة Mamba 40

الشكل 17.3- استخراج وتوليد الأهداف التدريبية 44

الشكل 18.3- المخطط التوضيحي لنموذج HuBERT 45

الشكل 19.3-آلية التنقيح التكراري المعتمدة في نموذج HuBERT 46

الشكل 20.3- وحدة استخلاص السماتHubertFeatureEncoder 47

الشكل 21.3- تقنيات الضبط الدقيق الموفِّر للمعلمات.....50
الشكل 22.3- آلية عمل rsLoRA 52
الشكل 1.4- توزيع الميكروفونات والسماعة في قاعة Octagon بجامعة كوين ماري 58
الشكل 2.4- توزيع الميكروفونات والسماعة في قاعةClassroom.....59
الشكل 3.4- الاستجابة النبضية لإحدى RIR من مجموعة المعطيات AIR 60
الشكل 4.4- مخطط يوضّح أبعاد غرفة التسجيل وإعداداتها في مجموعة معطيات MARDY 61
الشكل 5.4- التوزيع التقريبي لمدة المقاطع الصوتية في LibriSpeech (train-clean-100) 63
الشكل 6.4- تمثيل لإشارة كلامية نظيفة وبعد إضافة الصدى.....64
الشكل 7.4- آلية استخراج الإطارات الهدف من الإشارات الصدى باستخدام نافذة انزلاقية بخطوة واحدة(Stride=1) ...66
الشكل 8.4- بنية النموذج 67
الشكل 9.4- منحني تدريب النموذج.....70
الشكل 1.5- مخطط تمثيلي يوضح بنية النموذج CleanMel.....75
الشكل 2.5- بنية الكتلة عريضة النطاق.....77
الشكل 3.5- بنية الكتلة ضيّقة النطاق 79
الشكل 4.5- مخطط تمثيلي يوضح آلية اختبار أداء النموذج بعد حذف الصدى. 82
الشكل 5-5- مخطط تمثيلي يوضح آلية اختبار أداء النموذج قبل حذف الصدى. 82
الشكل 6.5- مخطط تمثيلي يوضح هيكلية النظام المقترح.....84
الشكل 7.5- منحنيي الخسارة للتدريب والتحقق عبر الدورات التدريبية. 89
الشكل 8.5- مخطط تمثيلي يوضح بنية النظام المقترح–حالة BaseModel.....89
الشكل 9.5- مخطط تمثيلي يوضح بنية النظام المقترح–حالة rsLORA-HuBERT.....90
الشكل 10.5- مخطط يوضح آلية اختبار BaseModel بدون حذف الصدى.....90
الشكل 11.5- مخطط يوضح آلية اختبار rsLORA-Hubert بدون حذف الصدى.....90

قائمة الجداول

الجدول 1.4- مجموعة معطيات LibriSpeech 57
الجدول 2.4- قيم معيار جودة الإشارة الإدراكية (PESQ) مع صدى وبعد حذف الصدى. 71
الجدول 3.4- قيم معيار WER مع صدى وبعد حذف الصدى..... 71
الجدول 1.5- مجموعة معطيات الكلام النظيف المستخدمة في CleanMel 74
الجدول 2.5- مجموعة معطيات RIRs المستخدمة في CleanMel 74
الجدول 3.5- نتائج اختبار نموذج CleanMel وفق معياري PESQ,WER على المجموعة الأولى..... 82
الجدول 4.5- نتائج اختبار نموذج CleanMel وفق معياري PESQ,WER على المجموعة الثانية..... 83
الجدول 5-5- نتائج اختبار نموذج CleanMel وفق معياري PESQ,WER على معطيات نظيفة. 83
الجدول 6.5- توزيع مجموعة معطيات اللغة العربية Common Voice Corpus 14.0 86
الجدول 7.5- نتائج الاختبار للنظام المقترح End-to-End جميع الحالات..... 91
الجدول 8.5- نتائج اختبار النظام المقترح على معطيات نظيفة. 91

جدول الاختصارات

الاختصار	المصطلح
ASR	Automatic Speech Recognition
WER	Word Error Rate
RIR	Room Impulse Response
PESQ	Perceptual Evaluation of Speech Quality
DNMOS	Deep Noise Suppression Mean Opinion Score
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
RT60	Reverberation Time
STOI	Short-Time Objective Intelligibility
IPDs	inter-channel phase difference
GRU	Gated Recurrent Units
FGRU	Frequency Gated Recurrent Units
ML	Machine Learning
DL	Deep Learning
SNR	Signal-to-Noise Ratio
RNNs	Recurrent Neural Networks
LSTM	long-short term memory
FFN	Feed-Forward Network
DDR	Direct-to-Reverberant Ratio

SSMs	State Space Models
NLP	Natural Language Processing
HiPPO	High-Order Polynomial Projection Operator
S4	Structured State Space Sequence Models
DPLR	Diagonal Plus Low-Rank
S6	Selective Structured State Space Models
SiLU	Sigmoid Linear Unit
SSL	Self-Supervised Learning
ZOH	Zero-order hold
HuBERT	Hidden Unit BERT
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked Language Modeling
FFT	Fast Fourier Transform
FC	Fully Connected
LORA	Low-Rank Adaptation
rsLORA	Rank Stabilization LoRA
PEFT	Parameter-Efficient Fine-Tuning
MSE	Mean Squared Error
GAN	Generative Adversarial Network
CTC	Connectionist Temporal Classification

الفصل الأول

تمهيد

1.1- مقدمة عامة

يُعد الصدى (Reverberation) من أبرز التحديات التي تواجه نظم الاتصالات الحديثة وأنظمة تعرّف الكلام آلياً (Automatic Speech Recognition – ASR). ينشأ هذا الصدى نتيجة تداخل الصوت المباشر من المصدر مع انعكاسات متعددة عن جدران الغرفة وسقفها وأرضيتها، مما يجعل الإشارة المسموعة مزيجاً من النسخة الأصلية للكلام ونسخ مضاعفة ومتأخرة عنه زمنياً، هذه الظاهرة تجعل الإشارة المسجّلة أو الملتقطة عبر الميكروفون أقل وضوحاً، إذ تتداخل فيها المعلومات المفيدة مع الصدى الناتج عن الانعكاسات.

وفي حين يمتلك الإنسان قدرة فطرية على انتقاء الصوت الأصلي وتجاهل انعكاسات الغرفة والضجيج المحيط، الأمر الذي يتيح له فهماً واضحاً للكلام حتى في البيئات المعقدة صوتياً. على النقيض من ذلك، تبقى هذه المهمة صعبة للغاية بالنسبة للآلة، حيث تُظهِر أنظمة التعرّف الآلي على الكلام محدودية واضحة في التعامل مع هذه الظاهرة بنفس الكفاءة البشرية، وهو ما ينعكس على انخفاض دقة التعرّف وارتفاع معدل الخطأ في الكلمات Word Error Rate–WER.

يمكن توصيف صدى الغرفة بواسطة الاستجابة النبضية للغرفة Room Impulse Response-RIR وهي بمثابة البصمة الصوتية المميزة لمكان معين مثل الغرف أو القاعات، حيث ترتبط هذه الاستجابة بموقع كل من مصدر الصوت والميكروفون، وتعكس بشكل دقيق مسار انتقال الموجة الصوتية بينهما بما يتضمنه من انعكاسات. وهنا يمكن تعريف المسألة على النحو التالي لدينا إشارة حساس (ميكرفون) تحتوي عدداً غير معروف من الانعكاسات المتراكبة على الصوت الأصلي والمطلوب حذف هذه الانعكاسات، بغرض تحسين دقة أنظمة تعرف الكلام آلياً. ورغم التقدم الكبير الذي تحقق في هذا المجال، لا تزال الأبحاث الحديثة تسعى إلى تطوير حلول أكثر كفاءة وملاءمة للتطبيقات العملية.

تنقسم مسألة حذف الصدى إلى نوعين رئيسيين حسب عدد القنوات، الأولى متعددة القنوات تتميز بوجود مصفوفة ميكرفونات، وفي هذه الحالة يتمّ الاستفادة من المعلومات المكانية spatial information مثل فروق الطور بين القنوات inter-channel phase difference-IPDs إلى جانب المعلومات الطيفية. والثانية أحادية القناة وهي الحالة الأصعب وتتميز بوجود ميكروفون وحيد يتعامل مع إشارة واحدة، وبالتالي تفتقر إلى المعلومات المكانية الموجودة في الأنظمة متعددة القنوات، التي تحقق عادةً أداءً

أعلى بفضل المعلومات المكانية، لكن تكلفتها وتعقيدها أكبر. إنّ الأنظمة الأحادية القناة أقل تكلفة وأكثر تحدياً، مما يجعل مجال البحث فيهما واسعاً ومفتوحاً. سندرس في بحثنا النوع الثاني وهو حالة ميكروفون وحيد، كما سنهتم بحالة منبع صوتي وحيد، أي أنّ الإشارات الصوتية هي إشارات كلامية فقط من متكلم وحيد. الأمر الذي يجعل البحث أكثر تعقيداً كون الإشارة المرغوبة (الكلام المباشر) والصدى مترابطين إحصائياً.

2.1- مشكلة البحث

تتمثل المشكلة البحثية في أن أنظمة تعرف الكلام آلياً (ASR) تُظهر سلوكاً متناقضاً عند التعامل مع معطيات تحتوي على الصدى. فعندما يُدرَّب نظام التعرف على معطيات نظيفة خالية من الصدى، فإن أداءه يتدهور بشكل كبير عند اختبارِه على معطيات تحوي صدى، حيث يرتفع معدل الخطأ في الكلمات (WER). وفي المقابل، فإن إضافة نموذج لإزالة الصدى (Dereverberation Model) قبل أنظمة ASR لا يؤدي بالضرورة إلى تحسين الدقة، بل قد يسبب تشويهات طيفية تقلل من فائدة المعالجة المسبقة وتزداد WER.

أمّا في حالة تدريب أنظمة ASR مباشرة على معطيات تحتوي صدى، فإن الأداء يبقى محدوداً، ويتدهور عند اختباره على معطيات نظيفة خالية من الصدى. وإضافة نموذج إزالة الصدى قبل أنظمة ASR لا يحقق تحسناً جوهرياً على العكس من ذلك ينخفض الأداء وتزداد WER. يُعرف هذا التناقض في الأدبيات باسم مفارقة ASR-Dereverberation، والتي تبرز صعوبة تحقيق توازن بين تحسين جودة الإشارة الكلامية من جهة، وضمان رفع دقة التعرف الآلي على الكلام من جهة أخرى. وانطلاقاً من ذلك، يسعى هذا البحث إلى دراسة هذه المفارقة واقتراح مقاربة منهجية لتجاوزها.

3.1- هدف البحث

يهدف هذا البحث إلى تطوير نموذج فعّال لمعالجة الصدى الصوتي Dereverberation بغرض تحسين جودة الإشارة الكلامية، بما يسهم في رفع دقة أنظمة تعرف الكلام آلياً ASR وتقليل معدل الخطأ في الكلمات WER.

4.1- أهمية البحث

يشهد العصر الحالي اعتماداً متزايداً على تطبيقات تعرف الكلام آلياً ASR في مختلف جوانب الحياة اليومية، بدءاً من خدمات المساعدات الصوتية Voice Assistants والكتابة الآلية للاجتماعات Automatic Meeting Transcription، وإنتاج الكتابة الفورية للتسجيلات Automatic Captioning، وتجهيزات المساعدة السمعية وصولاً إلى تطبيقات التعليم عن بُعد.

غير أنّ فعالية هذه التطبيقات ما تزال مقيدة بعدة عوائق جوهرية أبرزها ظاهرة الصدى الصوتي، إذ يؤدي هذا الأخير إلى تشويه الإشارة الكلامية ورفع معدل الخطأ في الكلمات WER، مما يحدّ من موثوقية النماذج الحالية في البيئات الواقعية مثل قاعات المؤتمرات أو القاعات الدراسية أو المكالمات عبر الإنترنت.

ومما يزيد من أهمية هذا البحث، هو أن بصمة الصدى الصوتي (RIR) تختلف جذرياً باختلاف البيئة (أبعاد الغرفة، الأثاث، مواد البناء)، مما يجعل من غير العملي أو الممكن بناء نموذج مخصص لكل سيناريو محتمل. يؤكد هذا الأمر على الحاجة الماسة لمحاولة تطوير نماذج قوية وعمياء قادرة على التكيف مع مختلف الظروف الصوتية دون الحاجة لمعرفة مسبقة بخصائص البيئة المحيطة.

وتزداد أهمية البحث بالنظر إلى أن معظم الدراسات السابقة ركزت على تحسين جودة الكلام من منظور إدراكي دون إيلاء الاهتمام الكافي لمدى انعكاس ذلك على أداء أنظمة ASR، وهو ما يجعل هذه الدراسة هادفة من خلال ربطها المباشر بين مسألة حذف الصدى لتحسين دقة أنظمة الكلام آلياً.

5.1- محددات مسألة البحث

يمكن تلخيص محددات المسألة في عدد من الجوانب:

- معالجة مسألة الصدى (لا تشمل معالجة الضجيج في الخلفية أو مصادر صوتية أخرى).
- قناة وحيدة (ميكروفون وحيد).
- صدى متكلم وحيد.
- مستقل عن اللغة.
- مستقل عن المتكلم.
- بيئة غير محددة.
- التقييم وفق معيار معدل الخطأ في الكلمات WER.
- تبسيط تعقيد النموذج بحيث يتناسب مع الموارد المتاحة.

مع الأخذ بالاعتبار أن تقييم النموذج المقترح في هذا البحث يتمّ بالاعتماد على نموذج تعرّف الكلام آلياً، يعمل باللغة العربية.

6.1- أسئلة البحث

1. ما هو الواقع الراهن لأداء تقنيات إزالة الصدى الصوتي في تعزيز دقة أنظمة وتعرف الكلام آلياً ASR؟
2. ما مدى فعالية النُهج التكاملية من طرف إلى طرف (End-to-End) في تجاوز مفارقة إزالة الصدى وتعرف الكلام آلياً، من خلال مواءمة أهداف تحسين الإشارة مع الحصول على دقة تعرّف جيدة؟
3. ما مدى فعالية نماذج التعلم العميق الحديثة المعتمدة على state space models مقارنة بنماذج التعلم العميق المبنية على الشبكات العصبونية العودية RNN؟

7.1- مساهمات البحث

يقدّم هذا البحث مساهمان نظرية وتطبيقية في مجال معالجة الإشارة الكلامية وإزالة الصدى الصوتي، يمكن تلخيصها في النقاط التالية:

مساهمات نظرية:

- تقديم إطار نظري للتعرف على المفاهيم الأساسية لنماذج فضاء الحالة المتنوعة وآلية الضبط الدقيق للنماذج الضخمة وبعض نماذج التعلم العميق مثل HuBERT, LSTM, Mamba .
- اقتراح مقاربة تكاملية جديدة توازن بين حذف الصدى وتحسين أداء أنظمة تعرف الكلام آلياً، بما يتجاوز المفارقة التقليدية ASR–Dereverberation.

مساهمات تطبيقية:

- تقديم دليل تطبيقي يوضح أنّ العلاقة بين حذف الصدى وتحسين دقة التعرف على الكلام ليست مباشرة من خلال تطبيق نموذج LSTM.

- تطبيق نموذج لحذف الصدى بالاعتماد على بنية Mamba، لتطوير نظام End-to-End فعّال لحذف الصدى من أحادية القناة، وتحسين أداء أنظمة ASR وإثبات تفوقه في بيئات متنوعة.
- كتابة رماز برمجي بلغة Python يتيح تدريب النماذج المستخدمة في هذا البحث مع اختبارها وفق معايير التقييم المعتمدة.

8.1- مخطط البحث

قُسم البحث إلى ستة فصول، يقدّم الفصل الأول الإطار العام لمسألة البحث ويتضمن مقدمة عامة، مشكلة البحث، أهدافه، بالإضافة إلى مساهماته. أما الفصل الثاني فيتناول الدراسة المرجعية المتعلقة بمسألة الصدى الصوتي ويعرض أبرز النماذج ذات الصلة، بينما خُصص الفصل الثالث للإطار النظري الذي يوضح أهم المفاهيم العلمية ونماذج التعلم العميق المستخدمة في هذه البحث، ويركز الفصل الرابع على التنفيذ العملي لنموذج LSTM وتحليل النتائج التي حصلنا عليها، في حين يتناول الفصل الخامس تطوير نظام End-to-End لحذف الصدى الصوتي بالاعتماد على بنية Mamba بالتكامل مع أنظمة تعرف الكلام آلياً ويتضمّن مناقشة شاملة لأبرز النتائج، واستخلاص الاستنتاجات العامة. أمّا الفصل السادس فيقدّم خلاصة البحث والآفاق المستقبلية لتطوير العمل في هذا المجال.

الفصل الثاني

الدراسة المرجعية

يتناول هذا الفصل الأساليب والنماذج التي طوّرها الباحثون لحذف الصدى الصوتي بدءاً من النماذج التقليدية وحتى التعلم العميق، وأثرها على أنظمة تعرف الكلام آلياً، مع التركيز على تحليل المزايا والقيود لكل تقنية للوصول إلى الطريقة الأكثر فعالية لدعم أهداف هذا البحث.

1.2– مقدمة

تُعدّ مشكلة الصدى الصوتي Speech Reverberation من التحديات الجوهرية التي تواجه أنظمة تعرف الكلام آلياً Automatic speech Recognition-ASR إذ تؤدي إلى انخفاض دقة التعرّف وتدهور جودة الكلام المعالج، خاصة في البيئات الواقعية المليئة بالانعكاسات الصوتية المتعددة. وقد ازدادت الحاجة لمعالجة هذه المشكلة مع اتساع استخدام أنظمة التعرف على الكلام في التطبيقات الواقعية مثل التفاعل الصوتي المباشر، والمساعدات الافتراضية، والأنظمة الذكية التي تتطلب مستويات عالية من وضوح الكلام ودقة التعرف. نتيجة لذلك، تطورت جهود البحث في مجال تحسين الكلام Speech Enhancement - SE، وبرزت تقنيات متقدمة تهدف خصيصاً لمعالجة الصدى الصوتي بحيث ينعكس إيجاباً على أداء أنظمة ASR في الظروف الصوتية الصعبة.

وفي إطار هذه الجهود البحثية، انتقينا في هذه الدراسة المرجعية مجموعة من الأوراق البحثية الحديثة، بعد مراجعة عدد كبير من الدراسات، اختيرت بعناية لكونها تمثل الاتجاهات الرئيسية والتطورات الأحدث في مجال حذف الصدى اعتماداً على معايير دقيقة تشمل حداثة الطرح العلمي، وكفاءة النموذج المقترح في معالجة الصدى، وتأثيرها على أداء أنظمة تعرف الكلام آلياً، وبهذا قد نكون أجبنا على السؤال الأول من أسئلة البحث والذي يهدف إلى معرفة الوضع الراهن لأنظمة حذف الصدى الصوتي.

2.2- نظرة عامة موجزة عن نماذج حذف الصدى

تُظهر النماذج التقليدية في مجال صوتيات الغرفة room acoustics تنوعاً منهجياً يمكن تصنيفه على أساسين: أولهما مصدر النموذج، أي إن كان يعتمد على القياسات الفعلية للغرفة Data-driven models أو على المبادئ الفيزيائية الأساسية Physical models؛ وثانيهما طبيعة السلوك الصوتي الذي يفترضه النموذج، سواء كان هندسياً Geometric يقوم على مسارات الأشعة، أم موجياً Wave-based يستند إلى معادلات انتشار الموجات .[1]

تُعد النماذج المعتمدة على المعطيات Data-driven models من أقدم المقاربات التي استُخدمت في مجال صوتيات الغرفة. يتمثل إسهام هذه النماذج في بساطتها وارتباطها المباشر بقياسات الغرفة، وتوصيف الخصائص الزمنية والمكانية للغرفة. ومع ذلك، فإن محدوديتها تظهر في كونها تتطلب عمليات قياس دقيقة ومكلفة، كما أنها تعجز عن التعميم على غرفة جديدة لم تُقَس بعد. لاحقاً ظهرت النماذج Modal Response Models مثل CAPZ[1][2] و[3] OBF[2] و PF[3][4] التي تسعى إلى اختزال سلوك الغرفة إلى مجموعة محددة من الترددات الخاصة بالغرفة. وقد أسهمت هذه النماذج في تبسيط التوصيف الرياضي وتسهيل المحاكاة، لكنها محدودة في قدرتها على تمثيل البيئات الكبيرة والمعقدة.

وتزداد فعالية هذه النماذج عند دمجها مع معلومات فيزيائية مسبقة Physical Prior فإذا كانت هذه المعلومات هندسية Geometric أتاح تطوير نماذج مثل SDM[4] [5] التي تفكك الاستجابة النبضية للغرفة إلى انعكاسات مكانية محددة Spatial Decomposition Method، بينما أدّى إدماج المعلومات الموجية Wave-based إلى نماذج مثل PWD[5] [6] وSHD[6][7] التي اعتمدت تمثيل انتشار الأمواج. تجمع هذه النماذج بين الدقة الفيزيائية والمرونة الحسابية، أما محدوديتها فتكمن في ارتفاع كلفة الحساب وتقييدها بفرضيات تبسيطية حول شكل الغرفة والمواد المكوِّنة لها.

من جهة أخرى، تقوم النماذج الفيزيائية Physical Models على المبادئ الأساسية لانتشار الصوت، فهي لا تعتمد على معطيات القياس المباشرة فقط، بل تحاكي الظواهر الصوتية انطلاقاً من قوانين الفيزياء. ومن أبرزها نموذج تتبع الأشعة RT[7] [8] وتتبع الحزم BT[8] [9] تتميز هذه النماذج ببساطتها النسبية وقدرتها على تمثيل المسارات الصوتية في الغرفة بكفاءة، مما يجعلها

[1] common-acoustical-pole and zero
[2] orthonormal basis function
[3] parallel filter
[4] spatial decomposition model
[5] plane-wave decomposition
[6] spherical-harmonic decomposition
[7] Ray Tracing
[8] Beam Tracing

أساسية في التطبيقات الصوتية الافتراضية .Virtual Acoustics غير أن محدوديتها تكمن في تجاهلها للظواهر الموجية المعقدة مثل التداخل والانعراج Diffraction.

أما النماذج الأكثر دقة فهي تلك المبنية على المعادلات التفاضلية الجزئية PDE[9] Model مثل طريقة العناصر الحدّية BEM[10][10] وطريقة الحجوم المحدودة FVM[11][11]، وتكمن أهميتها في قدرتها على تمثيل السلوك الموجي للصوت بدقة عالية، بما يشمل التداخل والانعراج والانتشار. غير أن هذه الدقة الكبيرة تأتي بكلفة حسابية عالية جداً، إذ تتطلب مثل هذه النماذج موارد ضخمة من حيث الزمن والذاكرة، مما يجعل استخدامها غير عملي في كثير من الأحيان.

وبذلك، يمكن القول إن النماذج الفيزيائية توفر أعلى درجات الدقة النظرية، لكنها تبقى محدودة من الناحية التطبيقية بسبب كلفتها الحسابية العالية وتعقيد تنفيذها مقارنة بالنماذج المعتمدة على المعطيات، فضلاً عن ذلك فهي تعتمد على معرفة الشكل الهندسي للبيئة والمواد المستخدمة بدقة، وهو أمر غير متاح دائماً.

يتضح مما سبق، أن النماذج المعتمدة على المعطيات Data-driven قدّمت حلولاً عملية وبسيطة ذات صلة مباشرة بالقياسات، لكنها تعاني من محدودية التعميم وضعف القدرة على تفسير الآليات الفيزيائية الدقيقة. في المقابل، استطاعت النماذج الفيزيائية Physical models أن تُمثّل الظواهر الصوتية بدرجة عالية من الدقة، بما يشمل التداخل والانعراج، لكنها واجهت تحديات كبيرة من حيث الكلفة الحسابية وصعوبة التطبيق في بيئات متعددة كما أنها أقل مرونة في حال تغيرت ظروف الغرفة بشكل كبير. هذه المفارقة بين البساطة والسرعة من جهة، والدقة والتعقيد من جهة أخرى، شكّلت حافزاً رئيسياً للبحث عن بدائل أكثر توازناً، وهو ما أفسح المجال أمام ظهور مقاربات جديدة قائمة على التعلم العميق. فقد انتقل الاهتمام تدريجياً من النماذج التقليدية إلى مقاربات عميقة معتمدة على Deep, Data-driven Models، الأمر الذي غيّر بصورة جوهرية أحدث ما توصلت إليه الأبحاث في هذا المجال.

انطلقت أولى الجهود البحثية في هذا السياق من معالجة المسألة العكسية Inverse Problems التي تركز على تقدير الموسطات الصوتية Acoustic Parameters Estimation من إشارات الصدى مثل زمن الصدى، وكذلك خصائص مكانية كحجم الغرفة أو المسافة إلى المصدر، كما تناولت بعض الدراسات مهمة تصنيف الغرف بحسب إشارات الصدى.

لم يقتصر دور التعلم العميق على المسألة العكسية، بل امتد إلى مجال تحسين الإشارة الكلامية، وهو ما سنفصله في القسم التالي عند استعراض أحدث نماذج التعلم العميق ذات الصلة التي طرحت في الفترة بين 2021-2025. ولكن قبل ذلك لابدّ من التنويه إلى ظهور بعض التوجهات الحديثة التي تدمج بين التعلم العميق و المبادئ الفيزيائية Physics-informed Deep

[9] Partial Differential Equation
[10] Boundary Element Method
[11] Finite Volume Method

Models عبر إدخال المعلومات الهندسية Geometry-based أو الموجية Wave-based. غير أنّ هذه النماذج تبقى أقرب إلى الحالة غير العمياء Non-Blind، إذ تتطلب معرفة مسبقة بخصائص الغرفة أو استجابتها النبضية، وهو ما يجعلها أقل ملاءمة لمسألة هذا البحث التي تنطلق من فرضية المعالجة العمياء. ومن ثمّ، سيكون التركيز على النماذج العميقة المعتمدة على المعطيات فقط، كونها الأكثر ملاءمة للسيناريو الواقعي المستهدف.

تناولت بعض الأدبيات أيضاً مسألة حذف الصدى في البيئات المتعددة القنوات Multichannel باستخدام التعلم العميق، حيث تُستغل المعلومات المكانية الناتجة عن توزيع الميكروفونات للفصل بين الصوت المباشر والانعكاسات. تعتمد هذه النماذج على مفهوم التنوّع المكاني Spatial Diversity لتقدير مكونات الإشارة القادمة من اتجاهات مختلفة، مما يسمح بعزل الإشارة المباشرة عن الإشارات المرتدة. يمكن دمج المعالجة المكانية والطيفية لتحقيق تحسين متكامل عبر القنوات المتعدد.

في [12] دمجت بين الشبكات العصبونية و تقنيات التشكيل الشعاعي Acoustic Beamforming، حيث تُستخدم الشبكة لتقدير الأقنعة الطيفية Spectral Masks من إشارات القنوات المتعددة، ومن ثمّ اشتقاق مصفوفات الكثافة الطيفية المتقاطعة Cross-Power Spectral Density Matrices – CPSD التي تمثّل الخصائص المكانية بين الميكروفونات. تستخدم هذه المصفوفات لاحقاً لحساب موسطات التشكيل الشعاعي التكيفية Adaptive Beamformer Coefficients، ما يتيح للنظام تعزيز الإشارة المباشرة وقمع الانعكاسات والضجيج بدقة محسّنة.

وفي نموذج SpatialNet [13]أحد أحدث الأمثلة؛ الذي يعتمد على تجميع دوال النقل الصوتي Acoustic Transfer Functions – ATFs ضمن فضاء مكاني–طيفي مشترك، ثم يقوم بدمج المعالجة المكانية والطيفية Spatial and Spectral Processing Fusion بهدف تنفيذ تحسين متكامل للإشارة عبر القنوات المتعددة، مما يتيح للنموذج معالجة الصدى والضجيج والفصل الصوتي في إطار واحد متكامل وفعّال.

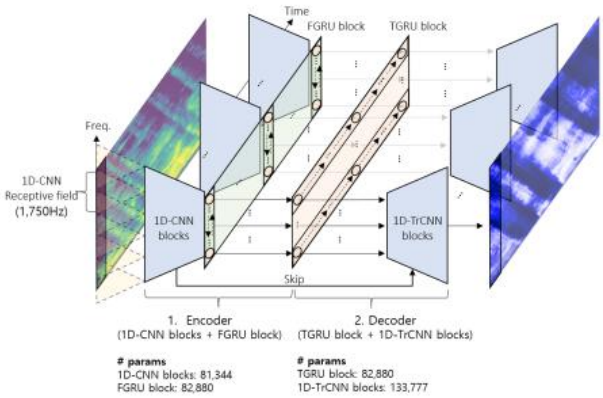
ومع ذلك، فإن هذا المسار البحثي، رغم أهميته، لا يمثل محور البحث الذي يركّز على حذف الصدى في الحالة الأحادية القناة.

3.2– النماذج الحديثة ذات الصلة

طُرح نموذج TRU-Net[12] [14] عام 2021 ليكون حلاً خفيفاً وسببياً لمعالجة الضجيج والصدى معاً. تمثّلت الفكرة الجوهرية في استبدال الشبكات التلافيفية ثنائية البعد 2D Convolution المستخدمة في بنية U-NET التقليدية بشبكات تلافيفة أحادية البعد 1D Convolution على المحور الترددي، مما يجعل النموذج سببياً وقابلاً للتنفيذ في الزمن الحقيقي، بالإضافة إلى

[12] Tiny Recurrent UNet

إدخال وحدات الذاكرة العودية Gated Recurrent Units -GRUs لتحسين فهم النموذج للعلاقات الزمنية بين الإطارات، حيث استخدمت شبكة FGRU للمحور الترددي وشبكة TGRU للمحور الزمني، كما هو مبين في الشكل (1.2).

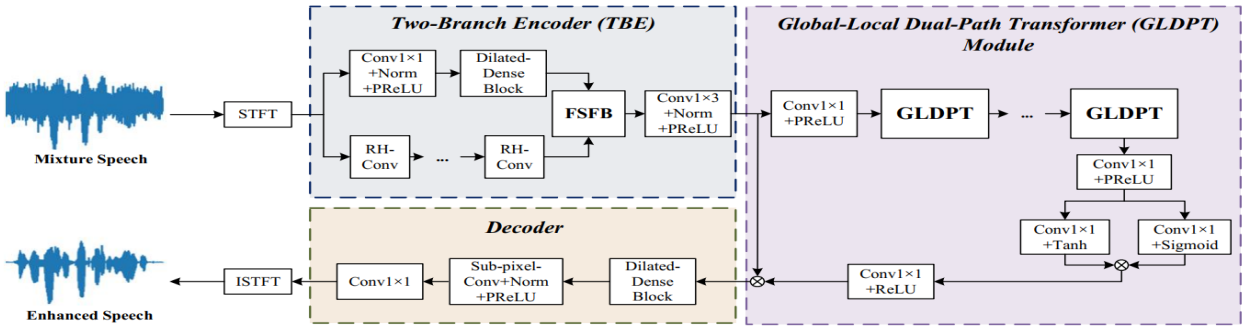


الشكل 1.2- البنية المعمارية لنموذج TRU-Net

دُرّب النموذج على مجموعات التدريب الخاصة بمسابقة DNS Challenge 2020 لتوليد معطيات تركيبية مع أو بدون صدى وعند اختباره في الوضع العائم FP32[13] على مجموعة معطيات WHAMR، حقق النموذج أداءً إدراكياً ملحوظاً بمقاييس PESQ=2.51 و SI-SDR=3.51 dB و STOI=81.22% مع قابلية للعمل في الزمن الحقيقي، بينما حافظ الإصدار المضغوط INT8 على أداء مماثل مع حجم لا يتعدّى 360 KB، ولكن رغم خفة وزن النموذج وسرعته، فقد ركّز على تحسين جودة الكلام إدراكياً دون أن يُختبر على مقياس معدل الخطأ في الكلمات، وهو المعيار الجوهري في تقييم أنظمة تعرف الكلام آلياً.

قُدّم نموذج D² Net[14] [15] عام 2022 كنموذج موحّد لإزالة الصدى والضجيج من قناة أحادية، يعتمد على تكامل مُرمّز ثنائي الفروع Two-Branch Encoder لاستخلاص كل من السمات الشاملة global والمحلية local مع محول ثنائي المسار Dual-Path Transformer لالتقاط الارتباطات الزمنية الطويلة والقصيرة في آن واحد. كما في الشكل (2.2)، حيث تسمح هذه البنية بدمج السياق العام والمحلي معاً بفضل من خلال تكامل الشبكات التلافيفية والمحول ثنائي المسار.

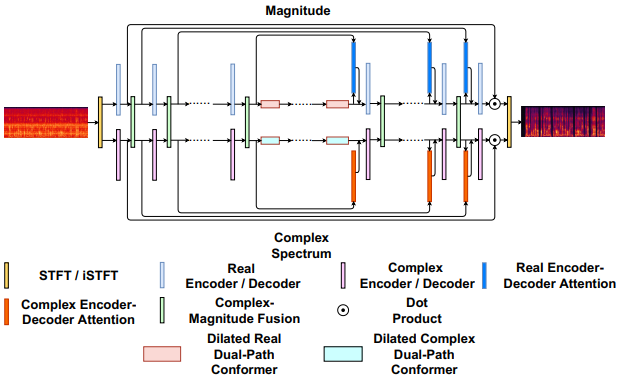
[13] Floating-Point Format
[14] Denoising and Dereverberation Network



الشكل 2.2- البنية المعمارية لنموذج D^2 Net

دُرّب النموذج على معطيات تركيبية مُولّدة من مجموعتي VoiceBank+DEMAND و WHAMR!، وعند اختباره على مجموعة WHAMR، حقق أداءً إدراكياً ملحوظاً بمؤشرات PESQ=2.51، و SI-SNR= 10.25dB، و STOI=95%، ولكن على حساب ارتفاع التعقيد الحسابي. كما أنه لم تُختبر فعاليته على مقياس معدل الخطأ في الكلمات، لتبقى العلاقة بين تحسين جودة الكلام وأداء أنظمة ASR غير محسومة. وهكذا يمكن النظر إلى D^2Net كنموذج يُجسّد التكامل بين السمات الشاملة والمحلية.

طُرح نموذج [16] UFormer عام 2022 بهدف تجاوز القصور الناجم عن معالجة المطال والطور بشكل منفصل، إذ اعتمد على معالجة ثنائية المسار؛ حيث يُمرَّر المطال والطيف العقدي Complex Spectrum عبر مسارين مستقلين، كل منهما مبني على بنية Conformer، ويتضمن كل مسار وحدات انتباه زمنية Time Attention وأخرى ترددية Frequency Attention كما يُظهر الشكل (3.2).



الشكل 3.2- البنية العامة لنموذج UFormer

أجرت الورقة البحثية دراسة تحليلية لتقييم أهمية كل مكون في البنية المعمارية لنموذج UFormer، وخلصت إلى أن إزالة أي مكون من النموذج يؤدي إلى تدهور في جودة استعادة الإشارة الصوتية. وخلصت إلى أنّ النموذج الذي يعتمد على فرع الطيف

العقدي Complex Spectrum Branch فقط يؤدي إلى انخفاض في DNSMOS بنسبة تقارب 3%، في حين أن النموذج الذي يعتمد فقط على فرع المطال Magnitude Branch دون استخدام القيم العقدية يؤدي إلى انخفاض أقل في DNSMOS بنسبة تقارب 7%. تشير هذه النتائج إلى أن النماذج التي تعتمد فقط على تحسين المطال تعاني من فقدان معلومات الطور، وهو ما يؤكد أهمية دمج معلومات المطال والطور معاً لتحقيق استعادة دقيقة وطبيعية للإشارة الكلامية ونتائج أفضل. دُرب النموذج على مجموعات المعطيات DNS Challenge 2021 بالإضافة إلى معطيات أخرى للضجيج مثل MUSAN ومعطيات كلام نظيف مثل LibriTTS.

حقق النموذج عند اختباره على مجموعة معطيات محاكاة تضم صدى وضجيج النتائج التالية:

PESQ=2.4501 عند مستوى ضجيج بين [-5,0] و 2.7472 عند مستوى ضجيج بين [0,5] ،كما حقق أفضل قيمة لمعيار MOS=3.3545 بالمقارنة مع النماذج الأخرى التي كانت State-of-the-art-SOTA حينها.

على الرغم من تقييم هذا النموذج عند مستويات ضجيج مختلفة وفق معايير متنوعة وهذا ما يعكس قوته إلا أنه لم يختبر على مجموعات معطيات حقيقية للصدى والضجيج، فضلاً عن ذلك، لم تُعرض نتائج لمعدل الخطأ في الكلمات، كما أنه أكثر حساسية لفقدان معلومات الطور.

كما قُدم نموذجي MR-SCB[16] , MR-UNet[15] [17]عام 2024 تطبيقاً لمبدأ المعالجة متعددة الدقة Multi-Resolution Processing لإزالة الصدى، يعتمد كلاهما على إطار عمل متعدد الدقة يتكون من عدة فروع، بحيث يتم تقسيم طيف الإشارة الكلامية إلى عدة مقاطع متساوية الطول غير متداخلة، يعمل كل فرع على عدد مختلف من المقاطع بحيث تركز الفروع العالية الدقة على معالجة الصدى على المدى القصير ، والفروع المنخفضة الدقة على معالجة الصدى على المدى الطويل. ويجري تبادل المعلومات عبر وظيفة نقل المعلومات Information Transfer Function-ITF، والتي تضمن تبادل المعلومات بين الفروع بشكل تدريجي حتى الوصول إلى المخرجات النهائية

يعتمد MR-UNet على بنية U-Net داخل كل فرع، بينما يعتمد MR-SCB على تكديس شبكات تلافيفية في كل فرع.

دُرب النموذج على مجموعة المعطيات LibriSpeech مع استجابات غرف محاكاة Image Source Model بالإضافة إلى مجموعة المعطيات VOICES.

[15] multi-resolution UNet
[16] Multi-Resolution Stacked Convolutional Blocks

الأداء :أظهر MR-UNet تحسيناً في مقاييس الجودة مثل PESQ و STOI، وحقق خفضًا في WER عند استخدام نظام ASR مدرَّب على معطيات نظيفة. بالمقابل، عند تطبيقه على نظام مدرَّب على معطيات مع صدى، ارتفع WER من 12.49%إلى 15.86%. تفوق النموذج MR-UNet على النموذج MR-SCB ولكن على حساب التعقيد حيث ازدادت موسطات النموذج بما يقارب 13M حيث حقق PESQ= 2.629 و STOI=88.2% عند اختباره على مجموعة معطيات تحاكي أثر الصدى وحقق PESQ=2.223 و STOI=75.2% على مجموعة المعطيات LIBRI-ADHOC40.

غير أن التقييم على أنظمة التعرف الآلي على الكلام كشف عن نتائج متباينة:

- على نظام ASR مدرّب على معطيات نظيفة، ساهم MR-UNet في خفض معدل الخطأ في الكلمات بنسبة 37.2% مقارنةً بأفضل نموذج مرجعي سابق.
- أما على نظام ASR مدرّب على معطيات تحتوي على صدى، فقد كان الأداء سلبياً، إذ ارتفع WER من 12.49% (لإشارة الصدى) إلى 15.86% بعد تطبيق MR-UNet، و16.43% بعد تطبيق .MR-SCB.

يعاني النموذجان من تكلفة حسابية عالية، مما قد يشكل عائقاً عند تطبيقهما على أجهزة ذات موارد محدودة، كما لوحظ أن إزالة الصدى أثرت سلباً على أداء أنظمة ASR المدربة على مجموعات معطيات تحتوي على الصدى.

ومن ثمّ اقتُرح [18] Minimum-Phase And All-Pass Decomposition عام 2024 وهو نموذج ثنائي المسار يعتمد على التحليل الطيفي Cepstral Analysis، لتحليل الطيف إلى مكوّنين: المكوّن ذي الطور الأصغري والمكون ذي التمرير الكامل يحمل معلومات الطور. يعالج كل مسار على حدة باستخدام شبكات عصبونية معقّدة مبنية على Conformer U-Net. تم تدريب هذا النموذج على مجموعة المعطيات REVERB Challenge Corpus و WSJCAM0[17] التي تحوي كلاماً نظيفاً ومجموعة معطيات تحوي الاستجابة النبضية للغرفة اختيرت من OpenSLR26[18] و OpenSLR28[19].

حقق النموذج انخفاضاً في WER؛ إذ ذكرت التجارب أن WER انخفض من 15.72% إلى 7.29% في حالة الميكروفون البعيد، ومن 46.78% إلى 16.22% في حالة الميكروفون القريب. ورغم هذه النتائج المتميزة، يظل للنموذج قيود واضحة، تتعلق بالتكلفة الحسابية العالية وتعقيد تقدير الطور، بالإضافة إلى أنه قد يكون أكثر حساسية في البيئات ذات الانعكاسات المعقدة، مما قد يؤثر على وضوح الكلام في بعض الحالات.

[17] https://catalog.ldc.upenn.edu/LDC95S24
[18] https://openslr.org/26/
[19] https://openslr.org/28/

طُرح نموذج CleanMel [18] عام 2025 كأحد أبرز المحاولات للتوفيق بين تحسين جودة الإشارة الكلامية من جهة، وتحقيق توافق مباشر مع أنظمة ASR من جهة أخرى، حيث اتبع منهجاً جديداً يعتمد على إخراج طيف ميل Mel-spectrogram المحسّن الذي يمكن تغذيته مباشرة لنظام ASR أو استخدامه مع Neural Vocoder لإعادة بناء الإشارة الكلامية. يُبنى النموذج من كتل ضيقة النطاق Narrow-band القائمة على وحدات Mamba لمعالجة كل تردد على حدة، وكتل عريضة النطاق Cross-band لنمذجة الارتباطات بين الترددات. تمّ تطوير وتقييم خمس تشكيلات مختلفة من نموذج CleanMel، بهدف تحليل أثر كل من: نوع المعالجة (online/offline)، واستراتيجية الهدف التدريبي سواء كانت الإسقاط الطباقي Mapping أو توليد قناع Masking، بالإضافة إلى حجم النموذج على جودة الإشارة الصوتية ودقة تعرف الكلام آلياً. أعطى نموذج CleanMel-L-mask أفضل أداء حيث خفض النموذج WER في معطيات CHiME-4 من 15.3% إلى 9.6% كما أعطى PESQ=2.35 بعد أن كانت 1.2، وقد أظهر النموذج قدرته على التعامل بفعالية مع ظروف تحتوي على صدى تركيبي وتسجيلات واقعية فضلاً عن كونه منخفض التعقيد.

4.2– مقارنة بين الدارسات المرجعية

المرجع/ الباحث	النموذج/ العام	المبدأ	نمط السمات	النتائج	الإيجابيات والسلبيات
[14] Choi et al	TRU-Net 2021	يعتمد على بنية U-Net (1D CNN) مع استخدام وحدات تكرارية GRU لمعالجة المحورين الترددي والزمني.	الطيف العقدي للإشارة	DNS-Challenge (صدى) PESQ2 = 2.74 SI-SDR = 14.87 dB STOI = 91.29%. WHAMR(صدى وضجيج معاً) PESQ1 = 2.51 SI-SDR = 3.51 dB STOI = 81.22%	**من الناحية الإيجابية:** نموذج سببي يعمل في الزمن الحقيقي على الأجهزة محدودة الموارد مع حجم صغير جداً، كما أنه قادر على معالجة الصدى والضجيج بشكل متزامن. أما من **الناحية السلبية:** لم يُختبر النموذج على أنظمة ASR كما أنه أظهر انخفاضاً ملحوظاً في الأداء عند البيئات الصوتية شديدة التعقيد كما في مجموعة المعطيات WHAMR
[15] Wang et al	D²Net 2022	بنية مرمز–فك الترميز. مرمز ثنائي المسار (TBE)	الطيف العقدي للإشارة	WHAMR!(صدى فقط) PESQ: 3.68 SI-SNR: 15.64 dB STOI: 99%	**من الناحية الإيجابية:** يعدّ النموذج المطروح قادر على معالجة الضجيج والصدى بشكل متزامن وبفعالية عالية،

		لاستخلاص سمات متعددة الدقة محلية وعامة)، ومحول ثنائي المسار (GLDPT) لنمذجة الارتباطات الزمنية.		WHAMR!(ضجيج وصدى) PESQ: 2.51 SI-SNR: 10.25 dB STOI: 95%	كما حقق أداء متفوقاً على النماذج الأخرى في مهام إزالة الصدى والضجيج أما من **الناحية السلبية**: يتطلب موارد حسابية عالية بسبب تعقيد بنيه وكما أنه لم يتم تقييمه على أنظمة ASR.
[16] Fu et al	UFormer 2022	يعتمد على بنية U-Net التي تستخدم Conformer ثنائي المسار لمعالجة طيف المطال والطيف العقدي بشكل متزامن وعلى التوازي.	يستخدم مسارين متوازيين من السمات: المطال والطيف العقدي	مجموعة المعطيات DNS Challenge 2021: DNSMOS: 3.6032 MOS: 3.3545	**من الناحية الإيجابية** بنية النموذج تسمح بتعزيز جودة الإشارة المستعادة في مجال تحسين الإشارة كما أنه يحافظ على معلومات الطور. أما من **الناحية السلبية**: يتطلب كلفة عالية للموارد الحسابية بالإضافة إلى حساسية معلومات الطور أي أنّ أداء النموذج يتدهور بشكل كبير عند الاعتماد على مسار المطال فقط بالإضافة إلى أنه لم يتم اختباره على أنظمة ASR
[19] Liu et al	Dual-Path Minimum-Phase and All-Pass Decomposition Network 2024	يعتمد على بنية Conformer U-Net ثنائية المسار	Minimum-Phase Cepstrum في المسار الأول. المكون ذي التمرير الكلي في المسار الثاني.	انخفض WER من 15.72% إلى 7.29%في حالة الميكروفون البعيد، ومن 46.78% إلى 16.22% في حالة الميكروفون القريب.(حالة معطيات حقيقية)	يسمح فصل مكونات الطور الأصغري والتمرير الكلي بمعالجة أكثر دقة وفعالية لكل من مطال وطور الإشارة، إلا أن تقدير الطور بدقة في البيئات المعقدة يبقى تحدياً أساسياً يؤثر على جودة الإشارة النهائية. كما ركز النموذج فقط على حذف الصدى.

[17] Zhao et al	نموذجين: (MR-UNet) (MR-SCB) 2024	إطار عمل متعدد الدقة يعالج الإشارة في فروع مختلفة MR-UNet: يستخدم بنية UNet كشبكة فرعية لإزالة الصدى داخل كل فرع. MR-SCB: يستخدم مكدس من الكتل التلافيفية كشبكة فرعية.	الطيف العقدي للإشارة	النموذج (MR-UNet) Libri-adhoc40: PESQ: 2.223 ASR(Reverb&Clean_trained) ، تحسّن 10.89% إلى WERخفّض بنسبة 37.2% مقارنة بأفضل نموذج مرجعي ASR(Reverb&Clean_trained) مقارنة بـ 15.86% إلىWERزيادة 12.49% للإشارة الأصلية المشوهة بالصدى	أثبت إطار العمل متعدد الدقة فعاليته في التعامل مع نطاق واسع من أزمنة الصدى المختلفة كما أنه حسّن بشكل كبير أداء أنظمة التعرف على الكلام التي تم تدريبها على معطيات نظيفة، ولكن طبيعة الإطار المتعدد الدقة الذي يعتمد على تكرار شبكات فرعية معقدة UNetفي كل فرع، تؤدي بطبيعتها إلى زيادة كبيرة في العدد الإجمالي للموسطات والتعقيد الحسابي للنموذج
[18] Shao et al	CleanMel 2025	بنية مكونة من كتل ضيقة النطاق تعالج كل تردد بشكل مستقل عبر الزمن باستخدام وحدات Mamba و كتل عريضة النطاق متداخلة لنمذجة العلاقات بين الترددات المختلفة.	الطيف العقدي للإشارة	حقق النموذج (CleanMel-L-mask)أفضل أداء : على نظام ASR: على مجموعة المعطيات CHiME4 انخفضت WER من 15.3 إلى 9.6 بعد معالجتها عبر النموذج مع العلم أنه في حال المعطيات بدون صدى كان WER=3.1	نجح النموذج في تحسين جودة الكلام المسموعة بشكل ملحوظ وفي نفس الوقت خفض معدل الخطأ في أنظمة التعرف على الكلام بشكل كبير. ولكن لا ينتج النموذج إشارة كلامية مباشرة، وللحصول عليها يجب استخدام شبكة Neural Vocoderمنفصلة، والتي تضيف طبقة أخرى من المعالجة

5.2- الخلاصة

من خلال استعراض وتحليل الدراسات السابقة، تبيّن أن نظم تحسين الإشارة الكلامية ليست بالضرورة مناسبة بشكل مباشر لأنظمة تعرف الكلام آلياً، حيث أنّ الإشارة المحسّنة الناتجة عن نموذج تحسين الكلام قد تكون مناسبة من منظور إدراكيّ سمعيّ، لكنها ليست بالضرورة فعّالة لأنظمة تعرف الكلام آلياً ، وذلك لأن عملية التحسين قد تُحدث تشوهات طيفية غير مقصودة تؤثر على الخصائص الصوتية التي تعتمد عليها نماذج ASR في تحليل وتمييز الكلمات[20] ،كما كشفت الدراسات عن مفارقة جوهرية تتمثّل في أنّ تدريب أنظمة ASR على معطيات نظيفة وأخرى تحتوي صدى لا يؤدي بالضرورة إلى تحسين الدقة، بل إنّ إضافة نموذج حذف الصدى قبل نظام ASR ترفع معدل الخطأ في الكلمات (WER) بدلاً من خفضه، نتيجة لتشوّه الخصائص الطيفية التي يعتمدها النظام في التعرف.

وبالتالي للحصول على نتائج مقبولة يجب على النموذج أن يعمل بصورة تكاملية مع أنظمة تعرف الكلام آلياً بهدف الوصول إلى نتائج أفضل على مستوى الأداء الكلي للنظام وذلك من خلال نهج تدريبي end to end يأخذ في الاعتبار كلاً من جودة الإشارة ومدى قابلية التعرف عليها، بحيث أن يُدرب نموذج تعرف الكلام آلياً على معطيات نظيفة. ومن المتوقع أن يُسهم هذا التكامل في تحسين الأداء الكلي للنظام، لا سيما من حيث تقليل معدل الخطأ في الكلمات، كما أنهّ من المستحسن أن يعالج النموذج الإشارة الكلاميّة بطريقة تأخذ بعين الاعتبار كل من مطال الطيف والطور.

وبناءً على ذلك يظهر نموذج CleanMel كخيار مثالي لهذا النهج، كونه يحقق الشروط السابقة بالإضافة إلى كونه يعتمد على آلية التنبؤ بنموذج فضاء الحالة الانتقائية Selective State Space Model Prediction، والتي تجمع بين السرعة والحجم المقبول، بناءً على هذه المزايا، سيتمّ اعتماد نموذج CleanMel كنموذج أساسي Base Model في هذا البحث، لتحقيق توازن بين تحسين جودة الإشارة وتقليل WER.

الفصل الثالث

الدراسة النظرية

يقدم هذا الفصل دراسة نظرية تُؤطّر العمل ضمن سياق المشروع، حيث يُعرَض فيه أقسام الاستجابة النبضية لغرف التسجيل (RIRs) والنموذج الرياضي للصدى، إلى جانب بيان أثر الصدى على الإشارة الكلامية، كما يتضمن شرح لمبادئ وأساسيات التعلم العميق بشكل عام، مع تركيز خاص على الشبكات العصبونية العودية ونماذج فضاء الحالة State Space Models، بهدف توضيح المفاهيم الأساسية، وخصائص هذه النماذج، وآليات عملها، مع إبراز الفوائد التي تميز كل نموذج وأسباب اختيارها ضمن نطاق هذا المشروع.

1.3– مقدمة

يُعرف تعلّم الآلة Machine Learning-ML بأنه فرع من الذكاء الصنعي يركّز على تمكين الأنظمة الحاسوبية من تحسين أدائها تلقائياً من خلال المعطيات، دون برمجتها بشكل صريح لحل مهمة بعينها حيث يعتمد تعلم الآلة على تصميم خوارزميات قادرة على اكتشاف الأنماط في المعطيات، وبناء نماذج تساعد في اتخاذ القرار، هذه القدرة على التعلم من المعطيات جعلت من تعلم الآلة محوراً هاماً للعديد من التطبيقات المعاصرة، مثل أنظمة التوصية Recommendation System، الكشف عن الاحتيال، والتنبؤ بسلوك المستخدمين وغيرها [21].

في العقود الأخيرة، شهدت الأبحاث تطوراً هائلاً مع بروز التعلّم العميق Deep Learning-DL، وهو فرع متقدّم من تعلم الآلة وأحد الركائز الأساسية في تطوير تقنيات الذكاء الصنعي الحديثة، يتمحور هذا المجال حول تصميم وتدريب الشبكات العصبونية العميقة Deep Neural Networks. يتميز التعلم العميق بقدرته الفائقة على استخلاص السمات التمثيلية Feature Representations من المعطيات تلقائياً عبر طبقات متعددة من المعالجة، على سبيل المثال، تتعلم الطبقات الأولى في الشبكة سمات منخفضة المستوى مثل الخطوط والحواف في الصور، بينما تكتشف الطبقات الأعمق سمات عالية المستوى مثل الأشكال أو الأجسام الكاملة [22] [21].

وانطلاقاً من هذه الفكرة، توسّع استخدام الشبكات العصبونية ليشمل استخلاص المفاهيم التجريدية، متطلباً معالجة متعددة المستويات عبر طبقات عميقة ومتسلسلة من النموذج. فالتعرّف على المفهوم التجريدي لا يعتمد على الخصائص السطحية للإشارة أو النص فقط، بل يحتاج إلى تمثيلات عالية التجريد يمكن للنموذج اكتسابها من خلال تراكم التحويلات غير الخطية في الطبقات العميقة. هذه المعالجة المتسلسلة والعميقة تزيد من تعقيد النموذج، [22] ما يستدعي موارد حسابية أكبر (من

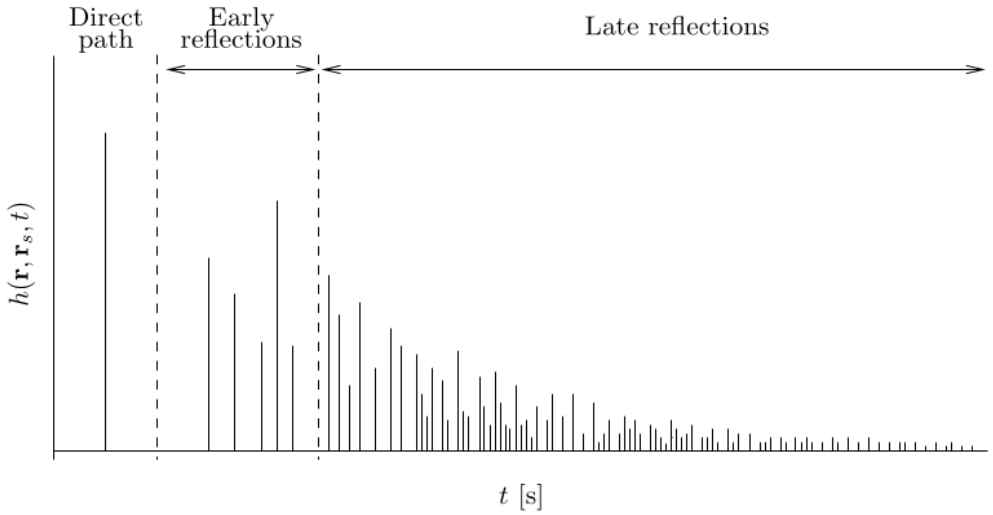
حيث الذاكرة وسرعة المعالجة) إضافةً إلى كميات ضخمة من معطيات التدريب لضمان تعميم النموذج على مختلف الأنماط والسياقات اللغوية. ويُعَدّ استنتاج اللهجة accent مثالاً بارزاً على ذلك، إذ يتطلّب من الشبكة العصبونية التقاط أنماط دقيقة في النطق والإيقاع والترددات الطيفية، ثم ربطها بالسياق اللغوي والثقافي العام بطريقة تجريدية تتجاوز التحليل الصوتي المباشر.

2.3– النمذجة الرياضية لصدى الغرفة

1.2.3– الاستجابة النبضية للغرفة وأقسامها

تُعَدّ الاستجابة النبضية للغرفة RIR المكوّن الأساسي في توصيف صدى الغرفة [23]، وهي إشارة زمنية تُقاس باستثارة المكان بإشارة نبضية قصيرة، ثم تسجيل الاستجابة عند الميكروفون. تُظهر الـ RIRs جميع تفاعلات الصوت مع المكان: المسار المباشر، الانعكاسات المبكرة والانعكاسات المتأخرة كما يبيّن الشكل (1.3). وُتقسم إلى ثلاث مكوّنات رئيسية:

- المسار المباشر Direct Path: وهو الموجة الصوتية التي تنتقل مباشرةً من المنبع إلى الميكروفون دون أي انعكاس، ويُعَدّ المرجع الأساسي في تحديد وضوح الإشارة، وذكرت بعض الأدبيات غياب المسار المباشر في حال عدم وجود خط رؤية مباشر بين المنبع الصوتي والميكروفون.
- الانعكاسات المبكرة Early Reflection: بعد فترة زمنية قصيرة، تصل الأصوات المنعكسة عن سطح واحد أو أكثر (الجدران، الأرضية، الأثاث، إلخ). لا تُدرَك هذه الانعكاسات كصوت منفصل عن الصوت المباشر طالما أن تأخيرها لا يتجاوز تقريباً 80–100ms بالنسبة إلى زمن وصول الصوت المباشر. بل تُدرك على أنها تعزيز للصوت المباشر، ولذلك تُعتبر مفيدة فيما يتعلق بوضوح الكلام Speech Intelligibility ويُشار إلى هذه الظاهرة غالباً باسم تأثير الأسبقية Precedence Effect، وذلك في الغرف الصغيرة حيث تكون الجدران والسقف والأرضية قريبة جداً. لكنها أيضاً تسبب تشويه الطيف بما يُعرف باسم التلوين Colouration.
- الانعكاسات المتأخرة Late Reflection: تبدأ بعد حوالي 100ms من الصوت المباشر وهي السبب الرئيسي في تدهور وضوح الكلام وجودته السمعية وتُدرَك كأصداء منفصلة.



الشكل 1.3-تمثيل للاستجابة النبضية للغرفة

ولقياس طول الصدى عملياً يُستخدم معيار زمن الصدى Reverberation Time-RT60 وهو الزمن اللازم لانخفاض شدة الصوت بمقدار 60dB، حيث تتراوح قيمته عادةً بين 0.5-1Sec في قاعات المحاضرات، وقد تتجاوز 5 Sec في المساحات الكبيرة.

2.2.3– النموذج الرياضي

يُعَدّ صدى الغرفة Reverberation ظاهرة فيزيائية طبيعية تنشأ نتيجة انعكاس الموجات الصوتية عن الأسطح الصلبة داخل البيئات المغلقة مثل الجدران والأرضيات والأسقف. وتؤدي هذه الانعكاسات إلى حدوث تشويه في الإشارة الكلامية، حيث يتراكب الصوت الأصلي مع نسخ أخرى مخمَّدة ومتأخرة زمنياً عنه.

ويمكن توصيف هذه الظاهرة عبر الاستجابة النبضية للغرفة RIR التي تمثل العلاقة بين إشارة المنبع الصوتي والإشارة المستقبلة عند الميكروفون، حيث يُنمذج الصدى رياضياً [24]بالمعادلة (1.3):

$$y(t) = S(t) * h(t) \qquad (1.3)$$

حيث: $S(t)$ تمثل الإشارة الكلامية النظيفة.

$h(t)$ تمثل الاستجابة النبضية للغرفة.

$*$ جداء التلاف.

وعلى الرغم من أنّ حذف الصدى يندرج ضمن تحسين الكلام، إلا أنّ معالجته تُعَدّ أكثر تعقيداً من معالجة الضجيج؛ فبينما يمثل الضجيج إشارة مضافة لا ترتبط بالإشارة الأصلية، فإن الصدى هو ناتج جداء تلافيفي مرتبط بالإشارة نفسها. وبذلك فإن الطرق الفعالة في إزالة الضجيج لا تصلح بالضرورة لإزالة الصدى. وتكمن صعوبة هذه المسألة في كونها مسألة غير عكوسة، حيث لا تتوفر معلومات كافية بشكل مباشر لاسترجاع الإشارة النقية، فضلاً عن أن البيئات الواقعية متغيرة وغير ثابتة، مما يفرض الحاجة إلى حلول قادرة على التكيف مع أنماط مختلفة من الصدى .[25]

3.3- تأثير الصدى

يُعَدّ صدى الغرفة من أبرز العوامل التي تؤثر سلباً على كل من إدراك الكلام البشري وأداء أنظمة تعرف الكلام آلياً [23] .

في الظروف المثالية الخالية من الصدى، تتميز الإشارات الكلامية بوضوح بنيتها الطيفية، لا سيما البواني Formants[20] . كما تحافظ الصوتيمات (Phonemes) على حدود زمنية واضحة ومنفصلة. إلا أنّ وجود الصدى يؤدي إلى تشويه هذه الخصائص من خلال آليتين رئيسيتين:

- طمس البواني Formant Smearing: حيث تتسبب الانعكاسات الصوتية المتعددة في تداخل البواني، مما يؤدي إلى فقدان الدقة في تحديد البواني ويجعلها أقل وضوحاً.
- التداخل الزمني الصوتيمات Phoneme Overlap: تعمل الانعكاسات على ملء الفجوات الزمنية الطبيعية بين الوحدات الصوتية والمقاطع الكلامية، مما يخلق تراكباً زمنياً بينها.

ينتج عن هاتين الظاهرتين انخفاض ملحوظ في وضوح الكلام وجودته السمعية المدرَكة.

على صعيد أنظمة التعرّف الآلي، يتجلى التأثير السلبي للصدى في تدهور مقاييس جودة الإشارة الصوتية، وبشكل خاص نسبة الإشارة إلى الضجيج SNR ونسبة الصوت المباشر إلى الصدى DRR، هذا التدهور يؤدي مباشرةً إلى طمس حدود الكلمات Word Boundary Smearing حيث يصبح من الصعب على النظام تحديد نقاط البداية والنهاية الدقيقة للكلمات وهذا بدوره يزيد معدل الخطأ ويقلل من دقة النظام وفعاليته.

تُبرر هذه الظواهر الحاجة إلى تطوير خوارزميات متقدمة لمعالجة الصدى بما يحقق توازناً بين وضوح الكلام للمستمع البشري ودقة أنظمة التعرّف الآلي. وهنا يبرز السؤال المحوري: **هل جميع التقنيات التي تعمل على تحسين جودة الكلام الإدراكية تساهم بالضرورة في تحسين دقة أنظمة التعرّف الآلي على الكلام؟**

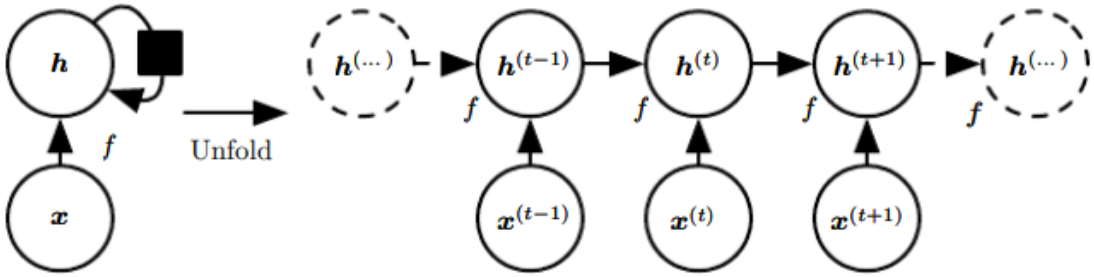
[20] قمم الطيف

أثبتت الدراسات الحديثة أن العلاقة بين جودة الكلام الإدراكية ودقة التعرّف الآلي ليست بالضرورة علاقة خطية أو مباشرة، إذ يمكن لبعض تقنيات تحسين الكلام أن تعزز وضوح الإشارة للمستمع البشري دون أن تحقق تحسناً مقابلاً في أداء أنظمة التعرّف بل على العكس قد تؤدي في بعض الأحيان إلى زيادة معدل الخطأ في الكلمات[26] [20] .

4.3- الشبكات العصبونية العودية

لا يبدأ الإنسان تفكيره من الصفر في كل لحظة، بل يبني معرفته الجديدة اعتماداً على ما سبق أن تعلمه، على سبيل المثال عند قراءة نص لغوي يمكن فهم كل كلمة في سياق الجمل السابقة والمحافظة على استمرارية التفكير [27]، هذه القدرة على الاحتفاظ بالمعلومات السابقة وتوظيفها في الفهم اللحظي تُعد سمة أساسية في الذكاء البشري، في المقابل تفتقر الشبكات العصبونية التقليدية إلى هذه الخاصية، فهي تعالج كل مدخل بشكل مستقل دون القدرة على تذكر السياق السابق وهذا بدوره يمثل قيداً جوهرياً خاصة في المهام التي تتطلب فهماً متسلسلاً للمعطيات، مثل تحليل الكلام المستمر.

وهنا جاءت الشبكات العصبونية العودية Recurrent Neural Networks-RNNs لمعالجة هذا القصور، من خلال بنية تحوي حلقات تكرارية تسمح بتمرير المعلومات من خطوة زمنية إلى الخطوة التالية، أي أنها تعيد استخدام الخرج كدخل للمرحلة اللاحقة وتطبق نفس العمليات الحسابية على كل شعاع من سلسلة الدخل ويحسب الخرج في كل مرة بالاعتماد على بعض قيم الخرج السابقة، وهذا ما يُكسبها القدرة على الاستمرارية في معالجة المعلومات. يمكن تبسيط هذه الفكرة بالنظر إلى الشبكة العودية على أنها مجموعة من النسخ المتطابقة لنفس الوحدة العصبونية، تتواصل عبر الزمن من خلال تمرير حالات خفية hidden state من مرحلة إلى أخرى وعند نشر هذه الحلقات يظهر هيكل يشبه السلسلة، كما يُظهر الشكل (2.3) يكشف عن مدى ملاءمتها الطبيعية لمعالجة المعطيات المتسلسلة مثل النصوص، والكلام وغيرها[21] .



الشكل 2.3-نشر المخطط الحسابي للشبكات العودية Unfolded Graph

تعمل الشبكات العودية على سلاسل من أشعة $x^{(t)}$ حيث t يعبر عن الخطوة الزمنية التي تأخذ قيمها بين 1 و τ طول سلسلة الدخل. أهم ما يميز الشبكات العودية ميزة مشاركة الموسطات Parameter Sharing وتعني أنّ مصفوفات الأوزان متماثلة في

بعض أجزاء الشبكة خلال الخطوات الزمنية، و في كل خطوة زمنية يحسب الخرج بتطبيق نفس العمليات الحسابية على قيم الخرج السابقة وبهذه الطريقة يتم مشاركة الموسطات عبر الشبكة وهذا ما أعطى الشبكات العودية القدرة على التعامل مع سلاسل مختلفة الطول والتعرف على المعلومة المتكررة بغض النظر عن موقعها في التسلسل، حيث أنّ كل خرج يعتمد على الدخل الحالي والخرج السابق باستخدام ذات الموسطات، كما تبين المعادلة (2.3):

$$h^{(t)} = f\left(h^{(t-1)}, x^{(t)}, \theta \right) \qquad (2.3)$$

حيث:

h يعبر عن الحالة المخفية وتمثل ملخص للمدخلات السابقة ويتم تحديثه في كل خطوة بناءً على الحالة السابقة والدخل الحالي

f هو تابع يربط حالة النظام في اللحظة t بحالته في اللحظة $t-1$

h^t حالة المخفية للنظام في اللحظة t

x^t الدخل في اللحظة t

θ موسطات التابع f

1.4.3– شبكات Long-Short Term Memory- LSTM

قدّمت شبكات LSTM حلاً ناجحاً لمشكلة تلاشي وتضخم التدرجات Vanishing and Exploding Gradients التي تعاني منها الشبكات العودية التقليدية عند تدريبها، حيث يتم حساب المشتق عبر مسارات ضمن المخطط الحسابي دون أن ينعدم أو يتضخم كما أنه يتميز بكونه نموذجاً واعياً للسياق Context Awareness من خلال بواباته التي تنظم تدفق المعلومات عبر الزمن، حيث يعدّ تذكر المعلومات لفترات طويلة سلوكاً افتراضياً لديها وليس من الصعب تعلمه، وبناء على ذلك، حققت شبكات LSTM نجاحات كبيرة في العديد من التطبيقات، مثل التعرف على الكلام و الترجمة الآلية وغيرها [21].

تتميز شبكات LSTM أيضاً بهيكلية السلسلة نفسها الموجودة في الشبكات العصبونية العودية التقليدية، حيث تتكون من خلايا مكررة زمنياً. لكن الفرق الجوهري يكمن في أن الخلية المكررة في LSTM لها بنية مختلفة، تتألف كل خلية LSTM من ثلاث بوابات [28]:

- بوابة الدخل input: المبينة في الشكل (3.3) تلعب دوراً هاماً في تحديد المعلومات التي ستخزن في ذاكرة الخلية في كل خطوة، حيث يتم تحديث ذاكرة الخلية على مرحلتين، في المرحلة الأولى تحديد المعلومات التي يجب تحديثها في ذاكرة الخلية والثانية انتخاب معلومات جديدة لإضافتها مكان المعلومات الواجب تحديثها. بحيث يتم معرفة المعلومات الواجب تغييرها عن طريق تطبيق تابع sigmoid (سنأتي على شرحه لاحقاً) على الحالة المخفية السابقة والدخل الحالي، كما تُظهر المعادلة (3.3):

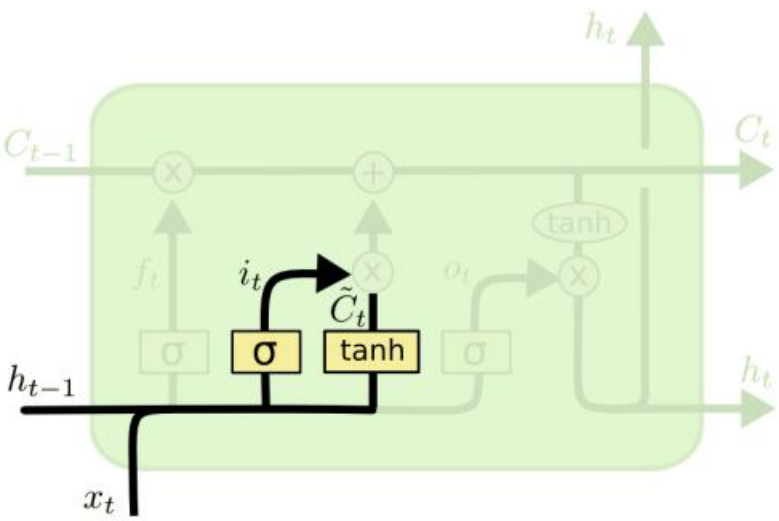
$$i_t = \sigma\left(x_t\, U^i + h_{t-1} W^i + b_i\right) \qquad (3.3)$$

U^i, W^i مصفوفات الأوزان لبوابة الدخل. b_i يمثل الانحياز

تُنتخب المعلومات الجديدة بتمرير h_{t-1}, x_t على تابع tanh فيعطي الخرج ضمن المجال [-1,1]

$$\widetilde{C_t} = \tanh(x_t\, U^c + h_{t-1}\, W^c + b_c) \qquad (4.3)$$

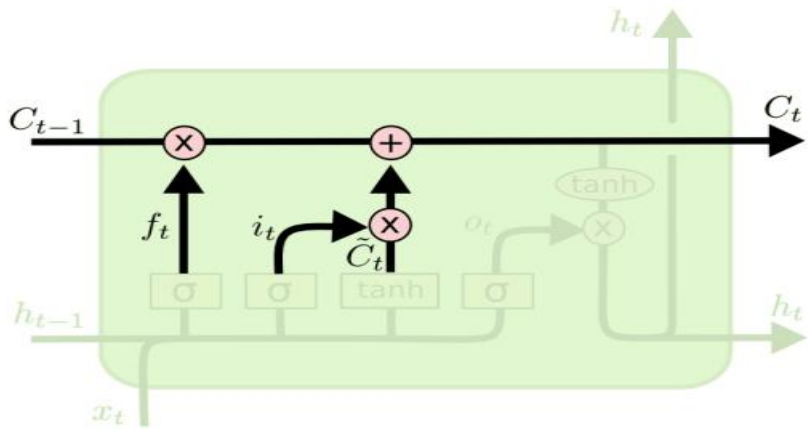
U^c, W^c مصفوفات الاوزان أمّا b_c الانحياز.



الشكل 3.3-بوابة الدخل في LSTM

ترمز $\widetilde{C_t}$ إلى المعلومات المنتخبة لتكون ذاكرة الخلية في اللحظة t، ويتم اختبار كمية معينة من $\widetilde{C_t}$ لتكون الذاكرة الجديدة، كما يُظهر الشكل (4.3). حيث تُحدد الكمية عن طريق البوابة i_t، فتضرب $\widetilde{C_t}$ بـ i_t ويضاف لها مقدار يعبر عن ذاكرة الماضي مضروباً ببوابة الحذف f_t (تُشرح في الفقرة التالية) لنحصل على ذاكرة C_t كما تبيّن المعادلة (5.3):

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \qquad (5.3)$$

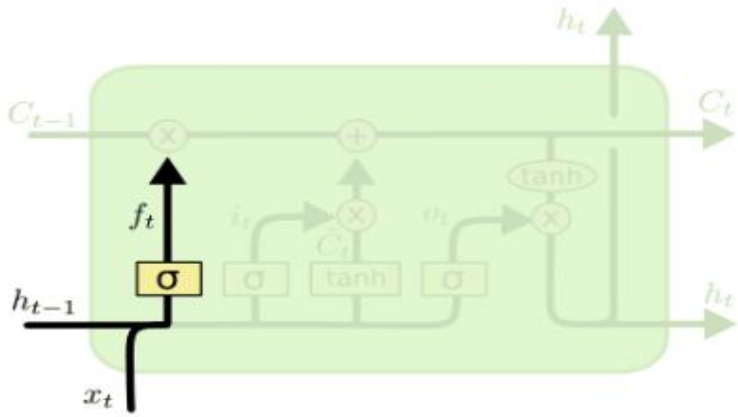


الشكل 4.3- تعديل ذاكرة الخلية في LSTM

- بوابة حذف forget gate: المبينة في الشكل (5.3) تحدد المعلومات التي يجب حذفها من الذاكرة. حيث يُمرر الدخل الحالي مع الحالية المخفية عن الماضي إلى تابع sigmoid، إذا كان الخرج أقرب للصفر يتم حذف تلك المعلومات، وإذا كان أقرب للواحد يتم الاحتفاظ به. تمثل المعادلة (6.3) النموذج الرياضي لبوابة الحذف:

$$f_t = \sigma(x_t U^f + h_{t-1} W^f + b_f) \qquad (6.3)$$

U^f، W^f مصفوفات الأوزان لبوابة الحذف، b_f يُمثل الانحياز.



الشكل 5.3- بوابة الحذف في LSTM

- بوابة الخرج output gate: المبينة في الشكل (6.3) تُمثل الحالة المخفية h_t خرج الخلية في اللحظة t، حيث يتم التحكم بخرج الخلية عن طريق بوابة o_t. يحُسب الخرج بخطوتين: الأولى تقييس ذاكرة الخلية C_t بين -1 و1 باستخدام

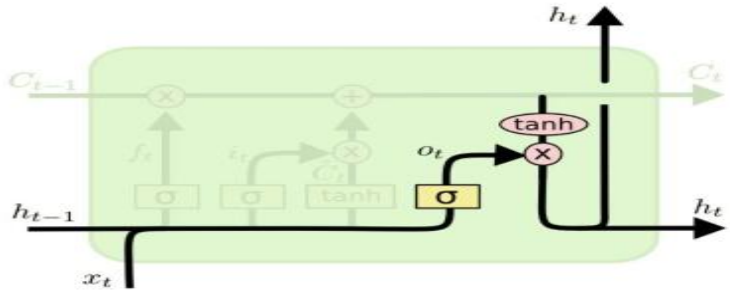
تابع tanh، والخطوة الثانية ضرب ناتج التابع الأخير ببوابة الخرج والتي تحسب بتطبيق تابع sigmoid على الدخل الحالي وعلى الحالة المخفية في اللحظة السابقة كما في المعادلتين (8.3) (7.3):

$$o_t = \sigma(x_t U^o + h_{t-1} W^o + b_o) \qquad (7.3)$$

U^o، W^o مصفوفات الأوزان لبوابة الخرج، b_o يمثل الانحياز.

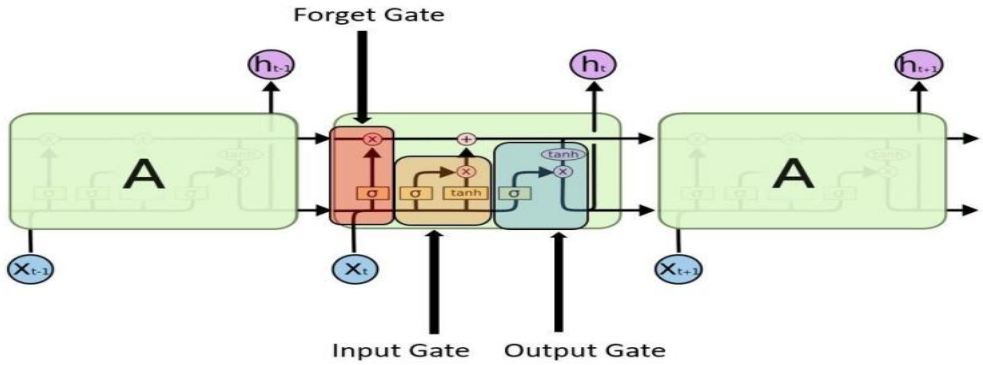
$$h_t = \tanh(C_t) * o_t) \qquad (8.3)$$

تمثل عملية الضرب في المعادلة (8.3) تغيير جزء من معلومات ذاكرة الخلية C_t وفقاً لبوابة الخرج o_t، حيث تحتوي الأخيرة معلومات عن الدخل الحالي، وسيضاف نسبة منه (تتراوح هذه النسبة بين 0 و100%) إلى ذاكرة الخلية، يحدد تابع sigmoid تلك النسبة.



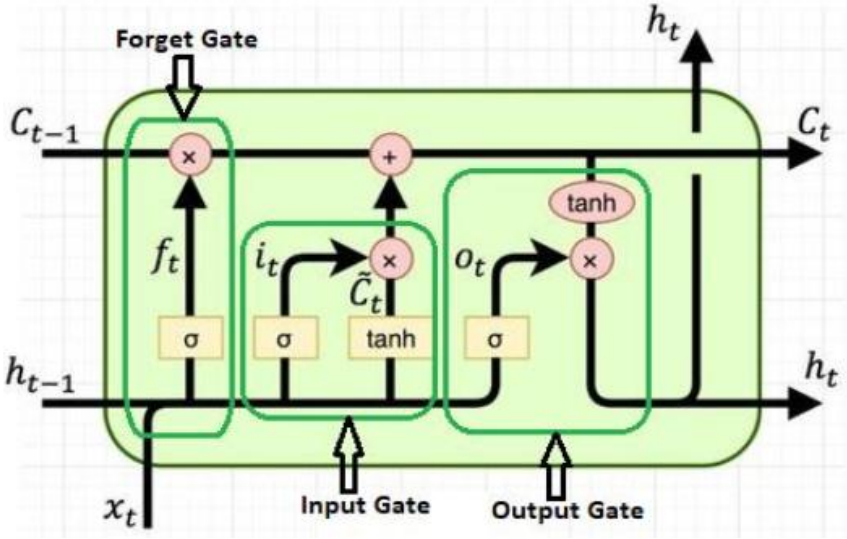
الشكل 6.3- بوابة الخرج في LSTM

يبيّن الشكل (7.3) طريقة ارتباط خلايا LSTM مع بعضها.



الشكل 7.3- رسم توضيحي لخلية LSTM

يبيّن الشكل (8.3) المعادلات الرياضية التي تحدث ضمن شبكات LSTM.



الشكل 8.3- خلية LSTM والبوابات الثلاث

تابع sigmoid:

تكون قيم هذا التابع ضمن المجال [0,1] ويفيد في حذف أو الاحتفاظ بقيمة الخلية، فمثلاً إن كان خرج هذا التابع 0 فهذا يعني أن قيمة الخلية يجب أن تُحذف (تُنسى) forgotten، وإذا كان خرج التابع 1 فيتم الاحتفاظ بالقيمة نفسها، وبهذه الطريقة تتعلم الشبكة أي المعلومات التي يجب الاحتفاظ بها، وأي المعلومات التي لم يعد لها فائدة [29]. تُعطى العلاقة الرياضية للتابع بالمعادلة (9.3):

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (9.3)$$

5.3- تطور بنية نماذج فضاء الحالة

1.5.3- مقدمة وهدف النشأة

تُعد معالجة السلاسل الطويلة واحدة من أكبر التحديات في الذكاء الصنعي الحديث، خاصة في مجالات مثل معالجة اللغة الطبيعية NLP، و الصوتيات، والجينومات (الصبغيات). حققت المحولات Transformerنقلة نوعية في هذه المجالات بفضل آلية الانتباه الذاتي Self-Attention التي تسمح للنموذج بدمج المعلومات عبر سياق طويل ومعقد، من خلال النظر إلى جميع

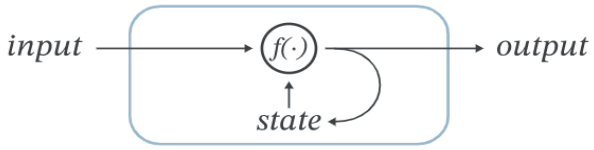
أجزاء سلسلة الدخل في وقت واحد، وهذا بدوره يمنح النموذج القدرة على فهم العلاقات البعيدة المدى ولكن بتكلفة حسابية تتزايد بشكل تربيعي مع طول السلسلة. مما يؤدي إلى استهلاك هائل للذاكرة والوقت عند معالجة السلاسل الطويلة، وهنا برزت الحاجة إلى تصميمات أكثر كفاءة قادرة على الاحتفاظ بالسياق الطويل دون تكلفة حسابية باهظة[30] .

استجابة لهذه التحديات، اتجهت الأبحاث إلى تطوير بنى أكثر كفاءة Efficiency و قادرة على الاحتفاظ بسياقات طويلة وبتعقيد خطي، فظهرت نماذج State Space Models-SSMs كأحد الحلول الواعدة، ومع ذلك فقد عانت هذه النماذج من قيود جوهرية، أهمها الثبات الزمني Time-Invariance الذي يجعلها تعالج جميع المدخلات بالطريقة نفسها بغضّ النظر عن محتواها، وبالتالي تفتقر إلى خاصية إدراك المحتوى Lack of Content-Awareness، هذا القصور جعلها أقل فعالية في المهام المعقدة التي تتطلب التمييز الانتقائي بين المعلومات المهمة وغير المهمة.

في هذا السياق، قدّم الباحثان ألبرت غو (Albert Gu) وتري داو (Tri Dao) عام 2024 نموذج مامبا Mamba وهو جيل جديد من نماذج فضاء الحالة تُعرف بنموذج فضاء الحالة الانتقائي Selective State Space Model يهدف إلى الجمع بين التعقيد الخطي الذي يتيح التعامل مع سلاسل طويلة جداً، والقدرة على الوعي بالمحتوى لانتقاء المعلومات المهمة وتجاهل المعلومات غير الضرورية .[31]

2.5.3– نماذج فضاء الحالة المستمرة State Space Model

تُعَدّ نماذج فضاء الحالة SSM من الأدوات الأساسية في نظرية التحكم، حيث تُستخدم لنمذجة الأنظمة الديناميكية عبر مجموعة من متغيرات الحالة State Variables التي تصف السلوك الداخلي للنظام. تقوم هذه النماذج على تمثيل النظام بوصفه كياناً ديناميكياً يستقبل مدخلات Inputs، ويُنتج مخرجات Outputs، مع الاحتفاظ بحالة داخلية internal state تتغير مع الزمن استجابةً لتلك المدخلات. وتتحدد المخرجات في كل لحظة زمنية للنظام وفقاً لكل من الدخل الحالي والحالة الداخلية في اللحظة نفسها، بينما تتطور الحالة الداخلية للنظام تدريجياً كلما استقبل مدخلات جديدة[32] . كما يُظهر الشكل (9.3) :



الشكل 9.3-تمثيل مبسط لنظام ديناميكي يوضّح العلاقة بين الدخل والحالة والخرج في نموذج SSM

ويُعَبَّر عن معدل تغيّر الحالة عادةً بالعلاقة الرياضية (10.3):

$$\dot{x}(t) = Ax(t) + Bu(t) \qquad (10.3)$$

$u(t)$ دخل النظام في اللحظة t

$x(t)$ تمثل الحالة الداخلية للنظام في اللحظة t

$\dot{x}(t)$ مشتق $x(t)$

كما يُعرف خرج النظام كتابع للزمن وفق العلاقة التالية (11.3):

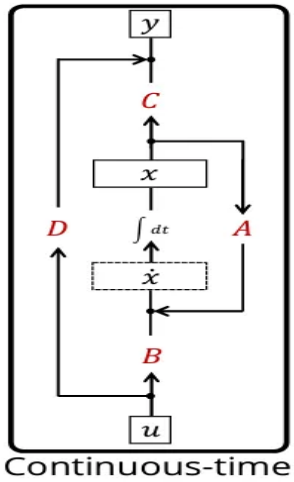
$$y(t) = \mathrm{C}x(t) + \mathrm{D}u(t) \qquad (11.3)$$

تعتمد نماذج SSM على ثلاث متغيرات موضحة في الشكل (10.3) تتغير مع الزمن t :

$x(t) \in R^n$ تمثل متغيرات الحالة وعددها n

$u(t) \in R^m$ تمثل مدخلات الحالة وعددها m

$y(t) \in R^p$ تمثل المخرجات وعددها p



الشكل 10.3- مخطط تمثيلي للنظام الديناميكي في الزمن المستمر

يمكن أيضاً ملاحظة أنّ نماذج SSM تتكوَّن من أربع مصفوفات قابلة للتعلّم هي:

$A^{n\times n}$ تمثل مصفوفة الحالة وتمثل ديناميكية النظام الداخلية.

$B^{n\times m}$ مصفوفة الدخل أو بعبارة أخرى مصفوفة التحكم (control matrix).

$C^{p\times n}$ مصفوفة الخرج، تربط بين الحالة الداخلية ومخرجات النظام.

$D^{p\times m}$ مصفوفة التغذية الأمامية Feedthrough Matrix وتمثل الأثر المباشر للمدخلات على المخرجات دون المرور بالحالة.

وفي التطورات الحديثة لنماذج فضاء الحالة لم تعد مصفوفات نماذج فضاء الحالة A، B، C، D مجرد موسطات ثابتة تصف ديناميكية نظام محدد مسبقاً، بل أصبحت تُعامَل بوصفها موسطات قابلة للتعلّم تُحدَّث باستمرار أثناء عملية التدريب باستخدام خوارزميات التعلّم الآلي بهدف تحقيق أفضل تمثيل ممكن للأنماط الموجودة في معطيات التدريب، بحيث يعبّر النموذج عن العلاقات الديناميكية المعقدة بين الدخل والخرج بطريقة أكثر دقة وتكيّفاً. وفي سياق التعلم العميق، يُترجم هذا المفهوم من خلال الأوزان القابلة للتعلّم في الشبكات العصبونية التي تؤدي دوراً مماثلًا للمصفوفات في توصيف البنية الداخلية للنظام، مما يجعل نماذج فضاء الحالة الحديثة أداةً فعّالة لدمج المبادئ التقليدية لنظرية التحكم مع القدرات التكيفية لأنظمة التعلّم العميق[33] [31].

3.5.3– نماذج فضاء الحالة المتقطعة Discrete State Space Model

تُعتبر نماذج SSM من الأنظمة المستمرة، ولذلك تُعدّ عملية التقطيع Discretization خطوة أساسية لتحويلها من المجال الزمني المستمر إلى المجال الزمني المتقطع بما يتناسب مع متطلبات التطبيقات الحاسوبية.

في المجال الزمني المتقطّع تُحسب قيمة الحالة عند كل خطوة زمنية t استناداً إلى قيمة الحالة في الخطوة السابقة $t-1$، وبذلك يمكن تمثيل نماذج فضاء الحالة في صورتها المتقطّعة بالعلاقات التالية[34] :

$$x_t = \overline{A}x_{t-1} + \overline{B}u_t \qquad (12.3)$$

$$y_t = \overline{C}x_t + \overline{D}u_t \qquad (13.3)$$

يتمّ التقطيع بخطوة زمنية ثابتة Δ حيث $\Delta = t_{n+1} - t_n$

ومن أبرز الدوال (التوابع) المستخدمة لتقطيع نماذج فضاء الحالة في مجال التعلم الآلي[31] :

✓ **Bilinear Transform**

$$\overline{A} = \left(I - \frac{\Delta}{2}A\right)^{-1} \left(I + \frac{\Delta}{2}A\right) \qquad (14.3)$$

$$\overline{B} = \left(I - \frac{\Delta}{2}A\right)^{-1} \Delta B) \qquad (15.3)$$

$$\overline{C} = C \qquad (16.3)$$

✓ **Zero-order hold- ZOH**

$$\overline{A} = e^{\Delta A} \qquad (17.3)$$

$$\overline{B} = (\Delta A)^{-1} \ (\exp(\Delta A) - I)\Delta B \qquad (18.3)$$

$$\overline{C} = C \qquad (19.3)$$

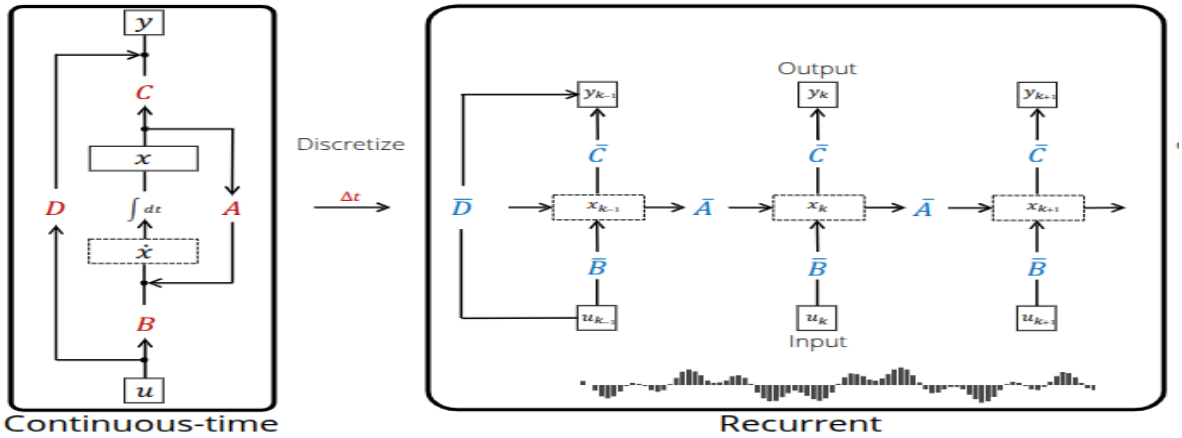
في النماذج التقليدية لفضاء الحالة، تمثل Δ الخطوة الزمنية الثابتة Time Step التي تفصل بين حالتين متتاليتين بعد عملية التقطيع للنموذج المستمر. أي أنها تحدد الفاصل الزمني الذي تُحدَّث فيه الحالة الداخلية. وتُؤثر قيمتها بشكل مباشر في استجابة النظام فكلما كانت كبيرة، كانت الفترة بين التحديثات أطول، مما يجعل استجابة النظام أبطأ ويؤدي إلى تغيّرات كبيرة في حالة النظام عند كل خطوة زمنية. أما عند صِغَر Δ، تصبح استجابة النظام أسرع، مما يمنحه قدرة أعلى على تتبع التغيرات السريعة في الإشارة باستمرار [31] [33].

1.3.5.3- آلية تنفيذ نماذج فضاء الحالة المتقطعة بين التمثيل التكراري والتلافيفي

تقوم الرؤية العودية Recurrent View لنموذج فضاء الحالة المبيّنة في الشكل (11.3) على تقطيع المصفوفات A، B،C باستخدام طريقة التحويل الثنائي الخطي Bilinear Method، وينتج عن ذلك النموذج المتقطّع كما تبين المعادلتين:

$$x_k = \overline{A}x_{k-1} + \overline{B}u_k \qquad (20.3)$$

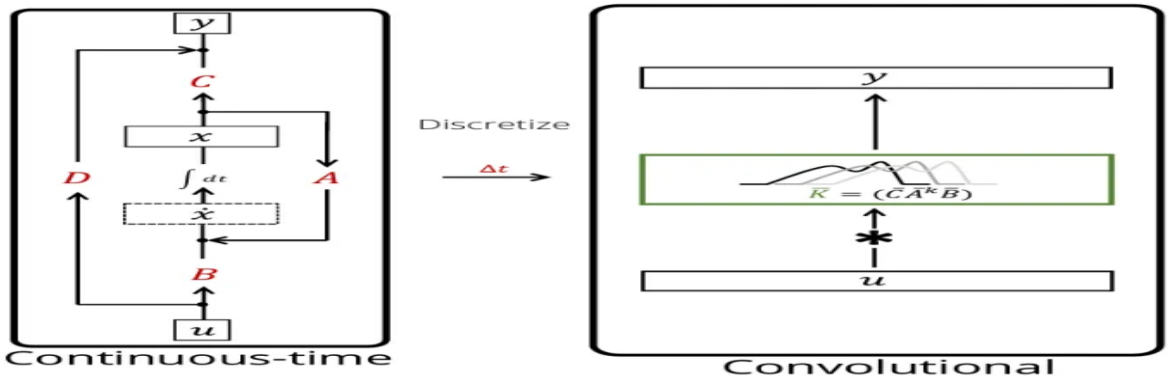
$$y_k = \overline{C}x_k + \overline{D}u_k \qquad (21.3)$$



الشكل 11.3-تمثيل عملية الانتقال من نموذج النظام الديناميكي في الزمن المستمر إلى التمثيل العودي المتقطع

تسمح الصيغة المتقطّعة للنظام بأن يُمثَّل بوصفه علاقة طباق من سلسلة إلى سلسلة Sequence-to-Sequence Mapping، إذ تربط بين سلسلة المدخلات u_k وسلسلة المخرجات y_k كما أن المعادلة الخاصة بتغير بالحالة تُمثل بعلاقة عودية (Recurrence) في المتغير x_k، وهو ما يجعل النموذج يعمل بطريقة مشابهة لبنية الشبكات العصبونية التكرارية، حيث يمكن النظر إلى المتجه $x_k \in R^n$ باعتباره الحالة الخفية Hidden State التي تتحدث عبر مصفوفة انتقال $\bar{A}$.

إلا أنّ الشبكات العودية تُعدّ محدودة الكفاءة من حيث التدريب، نظراً لكونها غير قابلة للمعالجة المتوازية Non-Parallelizable، كما تعاني من مشكلة تلاشي التدرّج Vanishing Gradient التي تُضعف قدرتها على تعلّم الارتباطات الزمنية الطويلة. ولتحقيق كفاءة أعلى في المعالجة المتوازية على وحدات المعالجة الرسومية (GPUs)، يُلجأ إلى تحويل النموذج إلى الصيغة التلافيفية كما يُظهر الشكل (12.3)، مما يزيد من سرعة التدريب ويعزز كفاءة الأداء دون فقدان الخصائص الديناميكية للنظام [34] [33].



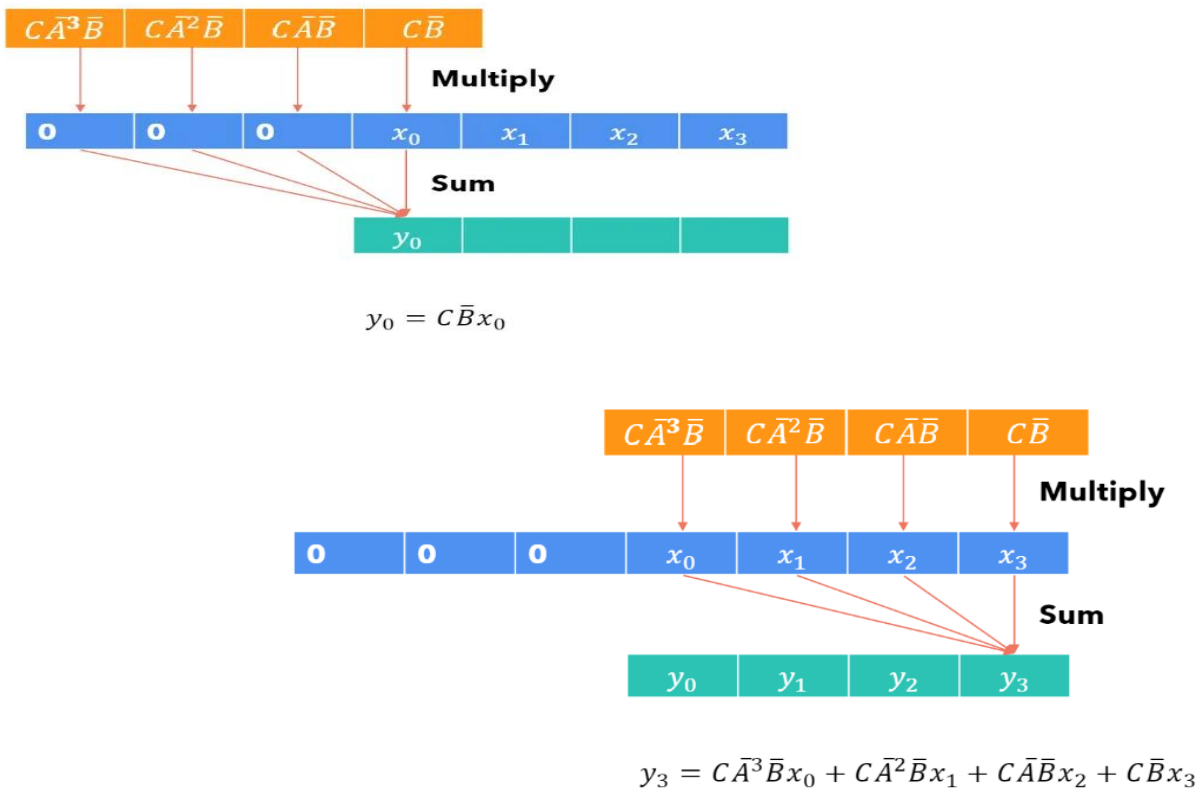
الشكل 12.3-تمثيل عملية الانتقال من نموذج النظام الديناميكي في الزمن المستمر إلى التمثيل التلافيفي المتقطع

إذ يمكن إعادة تمثيل الصيغة العودية للنظام إلى الصيغة التلافيفية، وذلك من خلال تكرار معادلات النموذج المتقطّع على مدى الزمن، كما يُظهر الشكل (13.3) وتصبح معادلات النظام كالتالي [33]:

$$y_k = \overline{CA}^{k}\,\overline{B}u_0 + \overline{CA}^{k-1}\,\overline{B}u_1 + \cdots + \overline{CAB}\,u_{k-1} + \overline{CB}\,u_k \qquad (22.3)$$

$$y = \overline{K} * u \qquad (23.3)$$

K يعبر عن المرشح للشبكات التلافيفية $\overline{K}_L = [\overline{C}\,\overline{B}, \overline{C}\,\overline{A}\,\overline{B}, \ldots\ldots, \overline{CA}^{\,L-1}\overline{B}]$: $\overline{k} \in R^L$ و $x_{-1} = 0$



الشكل 13.3- التمثيل الرياضي لآلية حساب الخرج في نموذج SSM المتقطّع في الصيغة التلافيفية

بعد تحويل الموسطات من الصيغة $(\Delta,\ A,\ B,\ C)$ إلى $(\overline{A}, \overline{B}, \overline{C})$ يمكن تنفيذ النموذج بطريقتين مختلفتين من حيث آلية العمل، مع الأخذ بالاعتبار أنّ المصفوفات $\overline{A}, \overline{B}, C$ ثابتة زمنياً Time-Invariant وتتمثل هاتان الطريقتان إمّا على شكل علاقة عودية خطية Linear Recurrence أو على شكل علاقة تلافيفية عامة Global Convolution.

ومن الشائع، اعتماد النموذج على الوضع التلافيفي Convolutional Mode أثناء مرحلة التدريب لما يتميز به من كفاءة حسابية عالية وقدرته على المعالجة المتوازية، إذ يتيح للنموذج معالجة سلسلة الدخل بالكامل دفعة واحدة مما يسهم في تسريع عملية التعلم وتحسين استقرار التدريب. أما في مرحلة الاستدلال التراجعي الذاتي Autoregressive Inference، فيتم الانتقال إلى الوضع العودي Recurrent Mode حيث تُقدَّم المدخلات للنموذج بصورة متتابعة خطوة زمنية واحدة في آن One Timestep at a Time، يتعامل مع كل عنصر على حدة، مستعيناً بالحالة السابقة لتوليد الخرج الحالية. مما يتيح تنفيذاً فعّالاً يتوافق مع طبيعة النماذج الزمنية التي تعتمد على العلاقات المتتابعة في معالجة السلاسل الزمنية[33] .

2.3.5.3- اختيار وتهيئة مصفوفات نموذج فضاء الحالة في سياق التعلّم العميق

عند توظيف نماذج فضاء الحالة ضمن إطار التعلّم العميق تعتبر المصفوفات $\overline{A}, \overline{B}, \overline{C}$ موسطات قابلة للتعلّم Learnable Parameters يتم تحديثها أثناء عملية التدريب، غالباً ما يتمّ إلغاء المصفوفة D واستبدالها بـ وصلة تخطّ Skip Connection بهدف تبسيط البنية وتقليل التعقيد الحسابي للنموذج.

تُعد بنية المصفوفة $\overline{A}$ من العناصر المحورية في بناء نماذج فضاء الحالة المخصصة لنمذجة السلاسل الزمنية Sequence Modeling إذ تتحكم في آلية انتقال المعلومات من الحالة السابقة إلى الحالة الحالية. كما أن كيفية تعريف المصفوفة $\overline{A}$ وطريقة تهيئتها الابتدائية من العناصر التي تميز بين البنى المختلفة لنماذج فضاء الحالة المطروحة في الأدبيات البحثية. وقد أظهرت الدراسات التجريبية أن تهيئة المصفوفة $\overline{A}$ بشكل عشوائي تؤدي إلى نتائج ضعيفة، في حين أن تهيئتها اعتماداً على مصفوفة High-Order Polynomial Projection Operator- HiPPO الموضحة في العلاقة الرياضية (24.3) تحقق أداءً متفوقاً:

$$A_{nk} = -\begin{cases} (2n+1)^{\frac{1}{2}}(2k+1)^{\frac{1}{2}} & if\ n > k \\ n+1 & if\ n = 0 \\ 0 & if\ \ n < k \end{cases} \qquad (24.3)$$

إذ سجّلت النماذج التي استخدمت هذه التهيئة تحسناً كبيراً في الدقة تتراوح بين 60% و98 % في اختبار MNIST التسلسلي Sequential MNIST Benchmark الذي يتطلب من النموذج معالجة صورة ثنائية البعد كسلسلة من البكسلات، مما يؤكد فعالية هذه المقاربة في تعزيز قدرة النموذج على تمثيل الأنماط الزمنية المعقّدة بدقة أعلى [34] [33].

على الرغم من المزايا الكبيرة التي تقدمها نماذج SSMs من حيث القدرة على معالجة السلاسل الطويلة بكلفة خطية إلا أنها تواجه تحديات أساسية أبرزها الثبوتية بالزمن Time-Invariance ،تؤدي هذه القيود إلى ضعف في قدرة النموذج على إدراك محتوى المدخلات وتجعله غير قادر على انتقاء المعلومات المهمة وتجاهل المعلومات غير الضرورية وتتجلى هذه المشكلة بوضوح في مهام مثل النسخ الانتقائي Selective Copying، والتي تتطلب قدرة النموذج على تحديد أجزاء محددة من السلسلة وإعادة

ترتيبها أو إخراجها كما هي، بينما يتجاهل باقي السلسلة؛ كذلك، في مهمة رؤوس الاستقراء Induction Heads التي تتطلب التقاط نمط أو علاقة في موضع سابق من السلسلة، ثم تطبيقه لاحقاً لاستخلاص الخرج الصحيح في الموضع المناسب باختصار، نقص إدراك المحتوى في نماذج SSM التقليدية يمنعها من حل المهام التي تتطلب انتقاء المعلومات بشكل ذكي وديناميكي [31].

4.5.3– Structured State Space Sequence Models (S4)

تعدّ نماذج S4 من النماذج المتقدمة التي طُوِّرت لزيادة كفاءة نماذج فضاء الحالة SSMs. وتتمثل الفكرة الرئيسة لهذا النموذج في فرض بنية رياضية مهيكلة Structured Form على المصفوفات الداخلة في تكوين النموذج، ولا سيما مصفوفة الحالة $\overline{A}$.

يستخدم نموذج S4 مجموعة من المصفوفات القابلة للتعلّم Learnable Matrices لالتقاط الارتباطات الزمنية بعيدة المدى Long-Range Dependencies ضمن السلاسل، وتُشكّل هذه المصفوفات المكوّن الأساسي في الصياغة الرياضية للنموذج ضمن إطار فضاء الحالة State-Space Formulation، حيث تجري أمثَلة قيمها أثناء التدريب Optimization لتكييف النموذج مع خصائص المعطيات وطبيعتها.

تُعد مصفوفة الانتقال $\overline{A}$ المكوّن المحوري في نموذج S4 إذ تتحكم في تطور الحالات الخفية Hidden States عبر الزمن، وقد صُممت هذه المصفوفة بطريقة مهيكلة Structured تهدف إلى تحقيق تنفيذ حسابي فعّال مع القدرة على الاحتفاظ بالمعلومات الطويلة المدى Long-Range Memory.

تُهيّأ المصفوفة مبدئياً باستخدام مصفوفة HiPPO ، ثمّ يُعاد حساب موسطاتها Reparameterized على شكل مصفوفة قطرية مضاف إليها تصحيح منخفض الرتبة DPLR – Diagonal Plus Low-Rank، وفق المعادلة الرياضية (24.3):

$$A = \Lambda - PP^{*}\, u \qquad (25.3)$$

حيث: Λ مصفوفة قطرية تحوي عناصر عقدية قابلة للتعلّم Learnable Complex-Valued Entries.

PP^{*} تصحيح منخفض الرتبة Low-Rank Correction حيث P شعاع قابل للتعلّم Learnable Vector.

يوفّر هذا التمثيل ميزة إجراء العمليات التلافيفية بسرعة عالية Fast Convolution من خلال الاستفادة من هوية وودبري Woodbury Identity وتحويل فورييه، وهذا بدوره يقلل التعقيد الحسابي من $O(N^2)$ إلى $O(NlogN)$. [33]

Mamba –5.5.3

1.5.5.3– مساهمات نموذج Mamba في تطوير نماذج فضاء الحالة

جاء تطوير Mamba بوصفه استجابة مباشرة للقيود التي واجهتها نماذج SSMs التقليدية، إذ يهدف إلى الجمع بين ميزة الكلفة الخطية التي توفرها نماذج SSM، وبين القدرة على معالجة المعلومات بطريقة انتقائية تعتمد على محتوى المدخلات، وقد حقق ذلك بمساهمتين أساسيتين وهما [31]:

- نموذج فضاء الحالة الانتقائي Selective State Space Model الذي يمنح النموذج القدرة على التركيز الانتقائي على أجزاء محددة من المدخلات السابقة أو تجاهلها وفقاً لدرجة أهميتها الراهنة بالنسبة للدخل الحالي، من خلال استخدام موسطات ديناميكية معتمدة على الدخل مثل المصفوفتين B، C، والخطوة الزمنية Δ. وتُعدّ هذه الآلية بمثابة بديل متطور لآلية الانتباه Attention Mechanism المستخدمة في نماذج المحوّلات، التي تقوم بتعديل أوزان الانتباه Attention Weights بما يتيح تعزيز أو تقليل تأثير كل علّام سابق Token تبعاً لمدى ارتباطه بالسياق الزمني الحالي.
- خوارزمية المسح المتوازي المدركة للعتاد Hardware-Aware Parallel Scan Algorithm، وهي خوارزمية تهدف إلى تحسين إدارة وحدة معالجة الرسوميات (GPU) للعمليات الحسابية الخاصة بالنموذج ضمن هرمية الذاكرة Memory Hierarchy، بما يضمن زيادة سرعة التنفيذ ورفع الكفاءة الحسابية، وتُسهم هذه البنية في تحقيق تكامل فعّال بين التصميم الرياضي للنموذج ومتطلبات العتاد الحاسوبي، مما يجعل نموذج Mamba قادراً على الجمع بين القدرة التكيفية في تمثيل المعلومات والكفاءة العالية في المعالجة المتوازية.

2.5.5.3– نماذج فضاء الحالة الانتقائية Selective State Space Models (S6)

سعياً لتجاوز القيود التي تعاني منها نماذج فضاء الحالة التقليدية SSMs ولا سيما عجزها عن تعديل تركيزها بشكل ديناميكي على أجزاء محددة من المدخلات السابقة أو إهمالها وفقاً لأهميتها اللحظية، قدّم Dao وGu فئة جديدة من نماذج فضاء الحالة تُعرف باسم نماذج فضاء الحالة (S6) الانتقائية، ويعود السبب في تسميتها S6 لأنها Structured State Space Sequence Model With Selective Scan والتي تعتمد على آلية المسح الانتقائي Selective Scan بوصفها مكوّناً أساسياً في بنيتها.

كما ذكرنا سابقاً تستند نماذج فضاء الحالة التقليدية إلى خاصية الثبوتية الزمنية الخطية Linear Time-Invariance–LTI، أي أنّ الموسطات التي تتحكم في تحديث الحالة الخفية (Hidden State) تبقى ثابتة لجميع المدخلات وعبر مختلف الخطوات الزمنية. غير أنّ هذا الثبات الزمني يُقيد قدرتها على الاستدلال المعتمد على السياق Context-Dependent Reasoning والذي يستدعي إعطاء أولوية انتقائية للمعلومات السابقة استناداً إلى أهميتها بالنسبة للدخل الحالي.

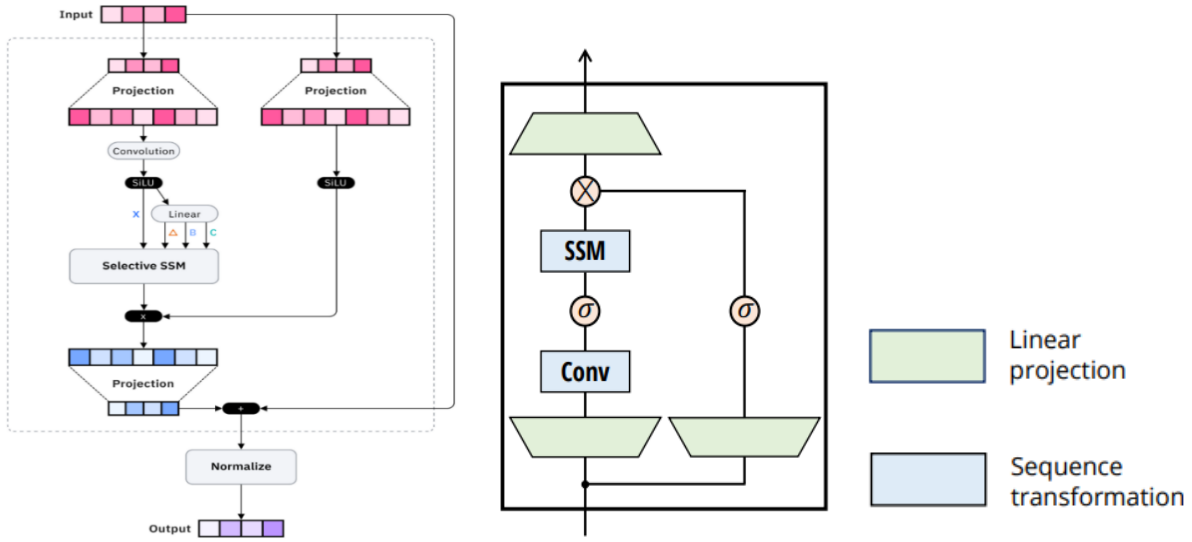
ولمعالجة هذا القيد، قدّم الباحثان Dao وGu مفهوم التمثيل المعتمد على الدخل Input-Dependent Parameterization بحيث تُحسب الموسطات الأساسية للنموذج مثل المصفوفات B، C، والخطوة الزمنية Δ، بشكل ديناميكي وكتابع للدخل الحالي. وبهذا الأسلوب، يتم كسر خاصية الثبوتية الزمنية، ويعطي النموذج مرونةً أعلى في تمثيل السياق الزمني المتغير وقدرةً أكبر على التكيّف مع طبيعة المعطيات وتسلسلها عبر الزمن. وبهذه الطريقة يضيف النموذج خاصية إدراك المحتوى التي كانت مفقودة في SSM التقليدي.

3.5.5.3– البنية والأبعاد في نموذج Mamba

يتكوّن نموذج Mamba من تكديس عدة وحدات متماثلة تُعرف باسم وحدات Mamba Blocks وتشكل بنية كل وحدة العنصر الأساسي في نجاح النموذج. فيما يلي وصف لتدفق المعطيات داخل وحدة Mamba الواحدة المبينة في الشكل (14.3) خطوة بخطوة[31] :

الدخل هو سلسلة من العلاّمات sequence of tokens يُعبَّر عنه بأبعاد (B,L,D)، حيث يشير B إلى حجم الدفعة (Batch Size)، L تمثل طول سلسلة الدخل Sequence Length، بينما تعبر D عن البُعد التمثيلي لكل علّام Embedding Dimension، والذي يُعرَّف d_model أو عدد القنوات Channel.

تُقسَّم المعطيات إلى مسارين متوازيين: الأول يُعرف بالمسار الرئيسي Main Path ويتضمن وحدة S6، أما الثاني فهو مسار البوابة (Gating Path).



الشكل 14.3- البنية المعمارية لوحدة Mamba

- **المعالجة في المسار الرئيسي:**

- تبدأ المعالجة في المسار الرئيسي بمرحلة الإسقاط الخطي [31]، حيث يمر الدخل عبر طبقة إسقاط خطي Linear Projection Layer تعمل على توسيع الأبعاد Dimension Expansion بمقدار ثابت يُعرف بمعامل التوسيع، ويُحدَّد عادةً بالقيمة $E = 2$ بحيث تتحول البعد من D إلى $E * D$. مع العلم أنّ هذه الطبقة ذاتها في كلا المسارين. بُعد دخل هذه الطبقة (B,L,D)، أمّا بُعد الخرج فيكون (B,L,E×D).

بعد مرحلة الإسقاط الخطي، تمر المعطيات في المسار الرئيسي بسلسلة من العمليات تهدف إلى تعزيز تمثيل الأنماط الزمنية.

- تُطبَّق طبقة تلافيفية أحادية البعد 1D Convolution Layer على طول السلسلة الزمنية L، تعمل هذه الطبقة على كل علّام داخل السلسلة بشكل مستقل، دون أن تخلط بين العينات المختلفة ضمن نفس الدفعة B أو بين السمات التمثيلية D لكل علّام. تُعد هذه الطبقة تلافيفية سببية Causal Convolution، أي أنّ كل خرج عند لحظة زمنية معينة يعتمد فقط على الرموز السابقة أو الحالية، دون النظر إلى الرموز المستقبلية. ولتحقيق هذه الخاصية، يُضاف حشو Padding إلى يسار السلسلة بمقدار k−1، حيث يُمثّل k حجم النواة Kernel Size. بُعد دخل هذه الطبقة (B,L, E×D)، وبُعد الخرج يبقى كما هو (B,L,E×D).
- يمرر الخرج إلى تابع التفعيل غير الخطي Sigmoid Linear Unit- SiLU لإدخال اللاخطية وتحسين القدرة التمثيلية للنموذج.

ومن ثم ترسل مخرجات الطبقة السابقة ذات البعد (B,L, E×D) إلى ثلاث طبقات إسقاط خطية متوازية مستقلة Parallel Linear Projection Layers وتعطي على خرجها الموسطات التالية:

$$dt: (B,\ L,\ dt_{rank}) \qquad (26.3)$$
$$B_t: (B,\ L, d_{state}) \qquad (27.3)$$
$$C_t: (B,\ L, d_{state}) \qquad (28.3)$$

حيث:

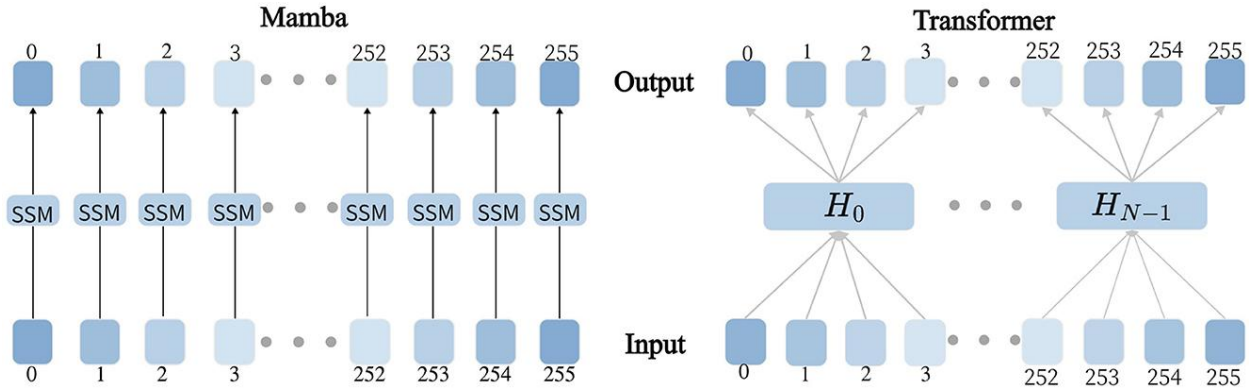
$d_state = N$ تمثل بعد متجه الحالة الخفية h_t، أمّا dt_rank فتمثل البعد الوسيط Intermediate Dimension المستخدم لتمثيل الخطوة الزمنية، ويكون عادةً $dt_rank << ED$ لتحقيق توازن بين الكفاءة الحسابية والقدرة التمثيلية، ويُحدّد تجريبياً بالعلاقة:

$$dt_{rank} = ceil(d_model / 16) \qquad (29.3)$$

إذ يتم تمرير الخرج الناتج عن الطبقة التلافيفة السابقة وتابع التفعيل عبر طبقة خطية أولى تُستخدم لاستخلاص تمثيل مضغوط يعكس المعلومات الزمنية الأساسية لكل علّام في السلسلة وتقوم بتقليل الأبعاد مؤقتاً من ED إلى بعد وسيط dt_{rank}، ثم يُعاد تمرير التمثيل المضغوط عبر طبقة خطية ثانية تعيد توسيعه إلى البعد الأصلي E×D، كما يُستخدم تابع التفعيل Softplus لضمان قيم موجبة ومستقرة للخطوة الزمنية. طريقة حساب الخطوة الزمنية يقلل من التكلفة الحسابية ويجعل النموذج أكثر استقرار عند تعلم قيمة الخطوة.

يتم تطبيق نموذج S6 بشكل مستقل على كل قناة من قنوات التمثيل Embedding Channels، أي أن بُعد التضمين (التمثيلي) لكل علّام Embedding Dimension مساوياً لبُعد الدخل Input Dimension. وبناءً على ذلك، يتم إنشاء نموذج فضاء حالة منفصل لكل بُعد من أبعاد التمثيل، وهو ما يتيح للنموذج التعامل مع كل قناة بصورة مستقلة مع الحفاظ على البنية العامة الموحدة، أبعاد خرج هذه الطبقة (B,L, E×D).

مع العلم أنه في نماذج المحوّلات يجري تحويل التمثيلات Embeddings إلى عدة أبعاد فرعية للرؤوس Head Dimensions تُعالَج بشكل متوازٍ ثم تُدمَج Concatenated لاحقاً للحصول على التمثيل النهائي. في المقابل، يحتفظ نموذج Mamba بتماثل الأبعاد ويُطبِّق نموذج فضاء الحالة مباشرة على كل بعد من دون عملية تقسيم أو دمج لاحقة، مما يُسهم في تقليل التكلفة الحسابية وتحقيق كفاءة في المعالجة المتوازية .[35] كما يبيّن الشكل (15.3):



الشكل 15.3- مقارنة بين آلية معالجة الأبعاد في نموذج Mamba والمحولات.

على الرغم من امتلاك نماذج S6 القدرة الكاملة على معالجة الأنماط العامة والارتباطات طويلة المدى، إلّا أنّ إدراج الطبقة التلافيفية يُعدّ خياراً تصميمياً مدروساً يهدف إلى معالجة جوانب محددة لا تستطيع S6 التعامل معها بكفاءة. حيث تتيح الطبقة التلافيفية للنموذج القدرة على استخلاص الأنماط المحلية Local Pattern Extraction داخل السلسلة الزمنية في حين تتولى وحدة S6 تمثيل السياق العام طويل المدى، مما يُكمّل تمثيل النموذج للسياق الكلي وبالتالي تُمثّل جزءاً من آلية نمذجة متممة Complementary Modeling Mechanism، كما تُساعد على تحسين استقرار عملية التدريب، إذ

تُوفّر حسابات تدرّج قوية، إضافة إلى ذلك أكدت الدراسات التجريبية على الأهمية العملية لهذه الطبقة، إذ أظهرت أن إزالة الطبقة التلافيفية تؤدي إلى انخفاض ملحوظ في أداء النموذج، مما يجعلها عنصراً أساسياً لضمان تحقيق الأداء الأمثل من حيث الدقة والاستقرار.

- **المعالجة في مسار البوابة:**

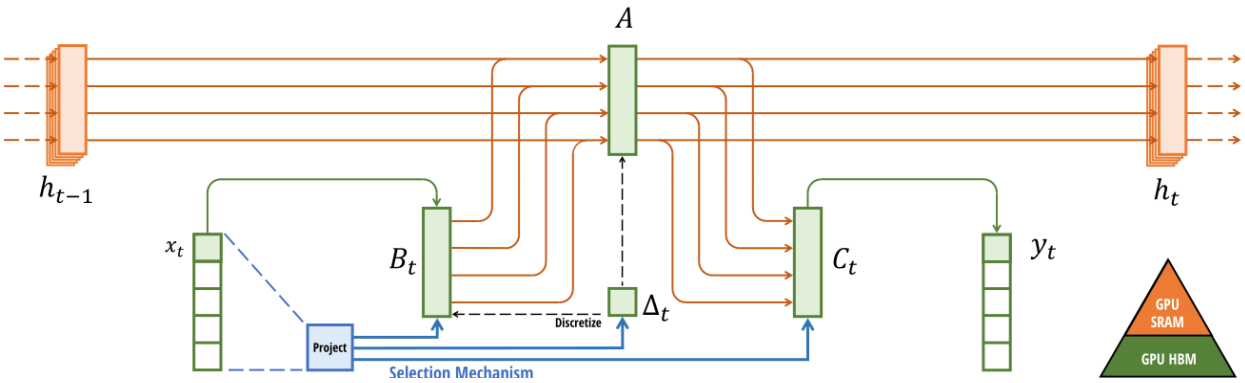
يعمل مسار البوابة بالتوازي مع المسار الرئيسي داخل بنية وحدة Mamba Block ويستخدم للتحكم بمقدار المعلومات التي يُسمح بتمريرها من مخرجات نموذج فضاء الحالة الانتقائي S6 إلى المخرجات النهائية للوحدة، حيث يُقرّر هذا المسار مدى أهمية التحديث المستمد من الحالة الزمنية لكل رمز في السلسلة.

يتكون من طبقة خطية لتوسيع الأبعاد ويصبح الخرج (B,L, E×D) ثم يُطبق دالة تفعيل غير خطية SiLU.

بعد المعالجة في كلا المسارين، يتم دمج الخرج لكل منها وذلك عن طريق عملية ضرب عنصر بعنصر Element-wise Multiplication ثم تطبق طبقة خطية لتكون الأبعاد النهائية للوحدة (B,L, D) [31].

4.5.5.3– موسطات نماذج فضاء الحالة الانتقائية S6

يتمّ جعل كل من المصفوفات B_t، C_t، والخطوة الزمنية Δ_t،كتابع للدخل الحالي x_t ، بحيث تتغير هذه القيم تبعاً لكل علّام (token) في سلسلة الدخل.



الشكل 16.3- البنية الداخلية لوحدة Mamba

ويتحقق ذلك من خلال تمرير شعاع التمثيل Vector Embedding للدخل x_t عبر طبقات الإسقاط الخطية SSM Input Projection Layer في نموذج S6 وحينها ينتج لدينا كل من B_t، C_t، والخطوة الزمنية Δ_t. ولكل منها تأثيره، حيث:

- تتحكم قيمة الخطوة الزمنية Δ_t في مقدار تأثير الدخل الحالي x_t على ذاكرة النموذج وسياقه السابق، بمعنى آخر في مقدار التغيير الذي يحدث بين الحالة الخفية السابقة h_{t-1} والحالة الجديدة h_t حيث تدخل في حساب المصفوفتين A،B_t.

كلما كانت قيمة Δ_t أكبر، زادت درجة التحديث في الحالة، وتسارعت عملية نسيان المعلومات القديمة المخزنة في الذاكرة وزاد التركيز على الدخل الحالي. أما عندما تكون Δ_t صغيرة، فإن التحديث يكون طفيفاً وهذا يعني احتفاظ أكبر بالمعلومات السابقة وتركيز أقل على الدخل الحالي، وعند قيم صغيرة جداً قد لا يكون للمدخل الحالي أي تأثير يُذكر على الحالة الخفية. وبالتالي الخطوة الزمنية تعمل كآلية تحديث انتقائي.

- تغيرات المصفوفة B_t تحدّد الطريقة التي يؤثر بها الدخل الحالي في تحديث الحالة الخفية، أي الكيفية التي تُعدَّل بها مكونات الحالة الخفية للنموذج استجابةً للمعلومة الجديدة الواردة عند هذه اللحظة الزمنية.
- بينما تتحكم تغيرات المصفوفة C_t بكيفية ترجمة معلومات السياق المخزنة داخل الحالة إلى تأثير مباشر على مخرجات النموذج y_t.

يتم حساب المصفوفتين B_t و C_t لكل علّام Token ضمن السلسلة الزمنية ثم تُطبَّق كلٌّ منهما بشكل منفصل على كل قناة Channel أثناء عملية العودية داخل نموذج فضاء الحالة SSM recurrence.

- أما بالنسبة لمصفوفة الانتقال A، تُعَدّ نماذج S6 من النماذج التي بُنيت على نماذج فضاء الحالة المهيكلة Structured State Space Models–S4، والتي تعتمد على مصفوفة قطرية Diagonal Matrix وذلك لأن تنفيذ حسابات نماذج فضاء الحالة بكفاءة عالية يتطلب فرض بنية محددة على المصفوفة A، وتُعد البنية القطرية Diagonal Structure الشكل الأكثر شيوعاً لتحقيق هذا الغرض، نظراً لقدرتها على تبسيط العمليات الحسابية وتقليل التعقيد الزمني، بالإضافة إلى كونها ذات قيم عقدية جزأها الحقيقي سالب لضمان استقرار التدريب بينما يسمح الجزء التخيلي بوجود سلوك اهتزازي Oscillatory Behavior مفيد في معالجة السلاسل الزمنية. في نموذج Mamba، لا يتم تعلم A مباشرة، بل يتمّ تعلم $\log A$ ثمّ تُقطع المصفوفة $A \in R^{d_{state} \times d_{state}}$ عند كل $\Delta_t \in R^{ED}$ وفق ضرب عنصر بعنصر element wise، كما تبيّن المعادلات الرياضية التالية:

$$\Delta = [\delta_0, \dots, \delta_i, \ \dots, \delta_{ED-1}] \qquad (30.3)$$

$$\mathrm{A} = [A_0, \dots, A_i, \ \dots, A_{ED-1}] \qquad (31.3)$$

وبالتالي:

$$\Delta\mathrm{A} = [\delta_0 A_0, \dots, \delta_i A_i, \ \dots, \delta_{ED-1} A_{ED-1}] \qquad (32.3)$$

تبيّن العلاقة (31.3) وجود مصفوفة A مستقلة لكل قناة من قنوات التمثيل Embedding Channels أي ED مصفوفة وبالتالي كل عنصر A_i يتحكم في عنصر واحد من الحالة. وتحدد هذه المصفوفة في انتقال المعلومات عبر الزمن وتتحكم في مدى احتفاظ النموذج بالحالة السابقة أو نسيانها في كل قناة من قنوات التمثيل.

– المصفوفة D تمثّل وصلة تخطي Skip Connection ضمن بنية نموذج فضاء الحالة، حيث تتيح مرور الدخل مباشرة إلى الخرج دون المرور بعمليات الحالة الداخلية. تكون قيمها قابلة للتعلّم.

وبذلك، تُمكّن النماذج فضاء الحالة الانتقائية من تحكم ديناميكي في تدفق المعلومات عبر الزمن، بحيث تتفاعل استجابة النموذج بشكل مرن مع أهمية كل دخل لحظياً، مما يمكّنها من التكيّف مع السياق اللحظي للمدخلات وتحقيق توازن دقيق بين الاحتفاظ بالمعلومات طويلة المدى والاستجابة الفورية للتغيرات في السلسلة الزمنية.

كما تُضيف آلية الانتقاء Selection Mechanism ديناميكية تعتمد على الدخل Input-Dependent Dynamics، مما يستلزم استخدام خوارزمية مدركة للعتاد Hardware-Aware Algorithm قادرة على تنظيم تموضع الحالات الموسعة expanded states ضمن ذاكرة وحدة المعالجة الرسومية GPU Memory Hierarchy لضمان سرعة وكفاءة التنفيذ والمعالجة المتوازية**[31]** **[35]**.

6.3– نموذج HuBERT

شهد العقد الأخير ثورة في مجال معالجة اللغة الطبيعية Natural Language Processing-NLP، مدفوعة بنجاح نماذج التدريب المسبق Pre-training القائمة على بنية المحوّلات Transformers ، وقد أدت هذه النماذج إلى تحقيق إنجازات غير مسبوقة في مهام متعددة، حيث تعتمد على مرحلتين متكاملتين: مرحلة التدريب المسبق حيث يتعلم النموذج تمثيلات عامة وغنية من معطيات ضخمة وغير موسومة Unlabeled Data، تليها مرحلة التدريب اللاحق Fine-tuning التي يُعاد فيها ضبط النموذج باستخدام معطيات موسومة خاصة بمهمة معينة، مثل الترجمة الآلية أو التصنيف وغيرها [36].

وبموازاة نجاح هذه النماذج في مجال النصوص، سعى الباحثون في مجال معالجة الكلام إلى استلهام هذا التوجه لمواجهة أحد أبرز التحديات، وهو الاعتماد الكبير على المعطيات الموسومة Labeled Data، التي يتطلب إعدادها وقتاً وجهداً وتكلفة باهظة.

من هذا المنطلق، برز التعلم ذاتي الإشراف Self-Supervised Learning-SSL كحل واعد، حيث يُمكِّن النماذج من تعلم تمثيلات مفيدة من الإشارات الصوتية الخام Raw Audio Signals مباشرةً دون الحاجة إلى نصوص مقابلة. وقد ظهرت محاولات رائدة في هذا السياق، مثل نماذج wav2vec التي اعتمدت على التعلم التبايني Learning Contrastive بهدف تدريب النموذج مسبقاً على تمييّز بين المقطع الصوتي الصحيح لسياق معين Positive Sample من بين عدد كبير من العينات

الخاطئة Negative Samples. وعلى الرغم من نجاحها فقد فرضت هذه المقاربة تحديات تتعلق بتعقيد تابع الخسارة، وأهمية اختيار العينات السلبية بعناية لضمان فعالية التدريب[36] .

في خضم هذا التطور، ظهر نموذج Hidden Unit BERT-HuBERT ليقدم منظوراً جديداً ومبتكراً، استُلهمت فكرته الجوهرية من النموذج الذي أحدث ثورة في مجال معالجة اللغات الطبيعية، وهو Bidirectional Encoder Representations from Transformers-BERT.

يرتكز BERT على المرمز Encoder في بنية المحوّلات، ويُدرّب في إطار التعلم الذاتي الإشراف SSL عبر نمذجة اللغة المقنّعة Masked Language Modeling-MLM، حيث تُحجب نسبة من الكلمات ويُدرَّب النموذج على التنبؤ بها من سياقها [37].

وهنا يكمن التحدي الجوهري الذي عالجه HuBERT في إمكانية تكييف استراتيجية BERT لتناسب الإشارة الكلامية في طبيعتها المختلفة جذرياً عن النصوص؛ فالنصوص تتألف من وحدات رمزية منفصلة (حروف، كلمات) ذات حدود واضحة وقاموس محدد، بينما تتسم الإشارة الكلامية بكونها مستمرة، ليس لها قاموس محدد، ولا تحتوي على فواصل طبيعية بين الوحدات الصوتية (الفونيمات)، كما أن هذه الوحدات تتفاوت في طولها وتتداخل فيما بينها [36].

تتجلى الفكرة المحورية في نموذج HuBERT، الذي أدخل مفهوم الوحدات الخفية Hidden Units المستخرجة مسبقاً عبر التجميع غير الخاضع للإشراف Unsupervised Clustering، بحيث تعمل هذه الوحدات كوسم تقريبي Pseudo-labels للنموذج. وبذلك يصبح بإمكان HuBERT محاكاة فلسفة نمذجة اللغة المقنّعة Masked Language Modeling-MLM على الإشارة الكلامية: إذ يُجبر النموذج على التنبؤ بالوحدة الخفية الصحيحة لجزء مقنّع من الصوت اعتماداً على سياقه، مما يمكّنه من تعلم تمثيلات صوتية غنية وقابلة للاستخدام في مهام متقدمة، خاصة بعد المرور بمرحلة التدريب اللاحق Fine-tuning على معطيات موسومة مرتبطة بالتطبيق المستهدف مثل التعرف التلقائي على الكلام ASR [38].

1.6.3– منهجية نموذج HuBERT

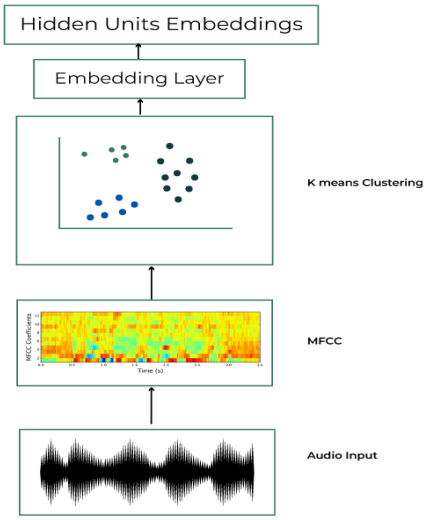
يمكن تقسيم منهجية عمل نموذج HuBERT إلى مرحلتين أساسيتين متمايزتين [38]: مرحلة غير خاضعة للإشراف لاكتشاف وتوليد الوحدات الخفية Hidden Units، تليها مرحلة تدريب ذاتي الإشراف تهدف إلى تعليم النموذج التنبؤ بهذه الوحدات.

1.1.6.3- المرحلة الأولى: اكتشاف وتوليد الوحدات الخفية

في البداية يقوم HuBERT بإنشاء أهدافه التدريبية الخاصة به عبر عملية عنقدة Clustering تتم لمرة واحدة قبل بدء التدريب (Offline Pre-Training Targets) كما يوضح الشكل (17.3). تبدأ هذه المرحلة باستخلاص موسطات MFCC من

الإشارة الصوتية الخام، ثم تطبيق خوارزمية K-Means Clustering بهدف تصنيف الإطارات المتشابهة معاً ضمن عدد محدد مسبقاً من المجموعات، والذي تم تحديده بـ K=100 في التجارب الأولية. في الخطوة النهائية، وبعد تحديد مراكز المجموعات، يُخصص لكل إطار رقم معرّف فريد وفقاً لأقرب مركز مجموعة، وهو ما يُعرف بـ الوحدة الخفية (Hidden Unit).

وبهذه الطريقة، تتحول الموجة الصوتية المستمرة $X = [x_1, x_2, x_3, \dots, x_T]$ حيث T تمثل عدد الإطارات، إلى سلسلة من الوحدات المنفصلة $Z = [z_1, z_2, z_3, \dots, z_T]$ التي تعمل بمثابة أهداف تدريبية (Pseudo-Labels) يتعلم النموذج التنبؤ بها في المرحلة التالية، وبذلك يتم تحويل الموجة المستمرة إلى سلسلة من الوحدات المنفصلة التي يمكن للنموذج التعامل معها.



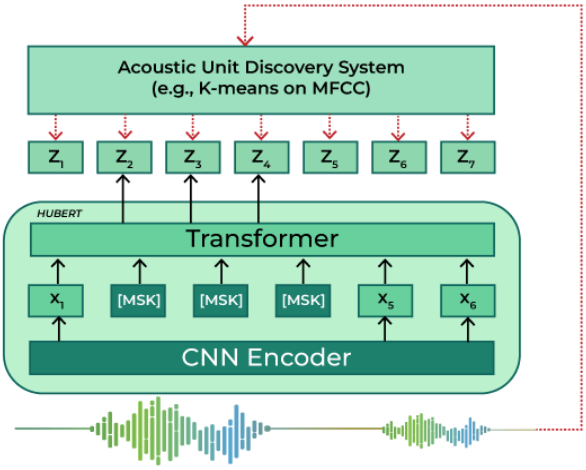
الشكل 17.3- استخراج وتوليد الأهداف التدريبية

2.1.6.3- المرحلة الثانية: تدريب النموذج عبر التنبؤ المقنّع

بعد تحديد الوحدات الخفية الأولية تبدأ مرحلة التدريب الذاتي الإشراف، وفي هذه المرحلة، يُدخل إلى النموذج الإشارة الصوتية الخام حيث تتم معالجتها أولاً عبر مُرمّز تلافيفي (CNN Encoder) لاستخراج السمات، ثم تُمرَّر هذه السمات إلى مُرمّز المحوّل (Transformer Encoder) الذي يتولى مهمة النمذجة السياقية.

كما أشرنا سابقاً يعتمد HuBERT على مبدأ التقنيع (Masking)، التي تطبق على المخرجات الناتجة عن المرمز التلافيفي كما هو مبيّن في الشكل (18.3)، واستناداً إلى منهجية BERT في التقنيع SpanBERT، يتم اختيار نسبة 8% من الإطارات عشوائياً كنقاط بداية، ومن كل نقطة بداية يقنع امتداد زمني متصل بطول 10 إطارات متجاورة وهذا يعني أن النموذج لا يُخفي إطارات متناثرة، وإنما يخفي مقاطع صوتية قصيرة ومتصلة، على نحو يحاكي آلية نمذجة اللغة المقنّعة (MLM) في نموذج BERT

مع تكييفها لتلائم المعطيات الصوتية. وبذلك، فإن نسبة الإطارات المقنّعة الفعلية تكون أكبر من 8%، وهو ما يزيد من تعقيد المهمة ويدفع النموذج للاعتماد على السياق المتاح لاستعادة الأجزاء المحجوبة.



الشكل 18.3- المخطط التوضيحي لنموذج HuBERT

بعد تطبيق التقنيع، تنتقل مهمة التنبؤ إلى المحوّل الذي يسعى إلى استنتاج الوحدة الصوتية الخفية الصحيحة لكل إطار محجوب بالاعتماد على السياق المستمد من الإطارات غير المقنّعة. ولتحقيق ذلك، تمر المخرجات عبر طبقة الإسقاط (Projection Layer) التي تُعيد تمثيلها في فضاء الوحدات الخفية، ثم يُحسب احتمال انتماء كل إطار مقنّع إلى وحدة معينة باستخدام مقياس تشابه جيب التمام (Cosine Similarity) بين التمثيل الناتج ومتجهات الوحدات المستهدفة.

تُقاس قدرة النموذج على التنبؤ الصحيح من خلال دالة خسارة الإنتروبية المتعارضة (Cross-Entropy Loss)، والتي تُطبَّق حصراً على المواقع المقنّعة، مما يجبر النموذج على تركيز عملية التعلّم على الأجزاء المفقودة. وتُعرَّف هذه الخسارة بالصيغة التالية:

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f\left(z_t \middle| \tilde{X}, t\right) \qquad (33.3)$$

حيث: M مجموعة الإطارات المقنعة

z_t الوحدة الخفية المستهدفة

$\tilde{X}$ النسخة المقنّعة من الإشارة الصوتية بعد إخفاء بعض الإطارات.

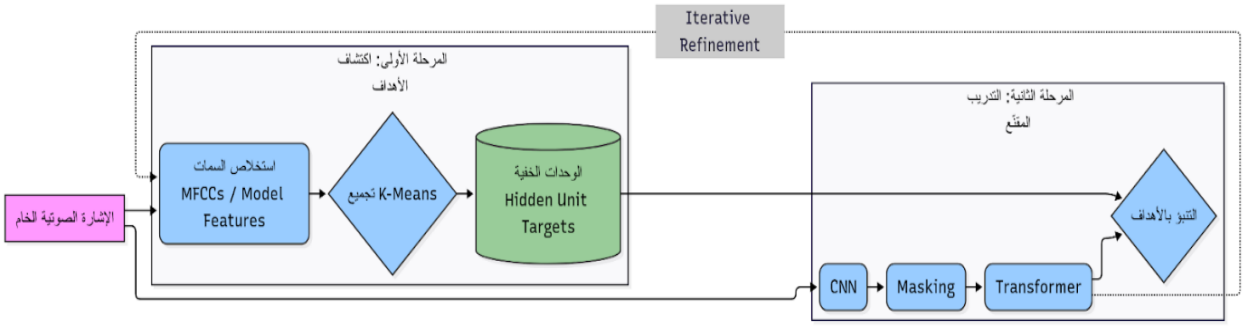
3.1.6.6- التنقيح التكراري Iterative Refinement

يعتمد نموذج HuBERT على آلية التنقيح التكراري [38] الموضحة في الشكل (19.3)، والتي تهدف إلى تحسين جودة الوحدات الخفية المستخدمة كأهداف تدريبية. ففي مرحلة التدريب الأولي، يُدرّب النموذج اعتماداً على وحدات خفية مستخرجة من

موسطات MFCC بعد تجميعها بواسطة خوارزمية K-Means بعدد مجموعات قدره 100، وبعد إتمام هذا التدريب، تُستخدم المخرجات من الطبقة السادسة في مرمز المحول في نموذج HuBERT-BASE لتوليد تمثيلات أكثر دقة، ثم يُعاد تطبيق التجميع باستخدام K=500 لتشكيل قاموس وحدات جديد يُستخدم في التكرار الثاني.

أمّا في حالة النماذج الأكبر مثل HuBERT-LARGE و HuBERT-X-LARGE، فلا يُكتفى بنتائج التكرار الثاني، بل يُستفاد من المخرجات العميقة للطبقة التاسعة في مرمز المحول في نموذج BASE، حيث يُجرى عليها التجميع باستخدام K=1000 لتوليد قاموس وحدات أكثر ثراءً ودقة، وهو ما يُعتمد في التكرار الثالث [38] [39] .

تبرز أهمية التنقيح التكراري في تمكين النموذج من تحسين أهدافه التدريبية ذاتياً، إذ يستفيد من مخرجاته الداخلية لإعادة تشكيل القاموس الصوتي وتوليد أهداف أعلى جودة. وقد أدّى هذا النهج إلى تحقيق قفزة نوعية في جودة التمثيلات الصوتية، مما جعل HuBERT أحد أبرز النماذج في مهام تعرف الكلام آلياً.



الشكل 19.3- آلية التنقيح التكراري المعتمدة في نموذج HuBERT

2.6.3– بنية النموذج

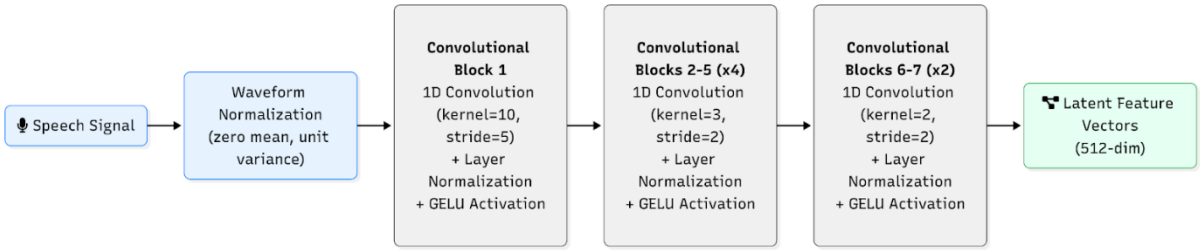
تتألف بنية نموذج HuBERT من مراحل متسلسلة تبدأ بتمرير الإشارة الصوتية الخام، الممثلة بمصفوفة أحادية البعد وبمعدل أخذ عينات قدره 16KHz إلى وحدة استخلاص السمات (HubertFeatureEncoder) يليها مرمز المحوّل (Transformer Encoder) ثم طبقة إسقاط (Projection Layer)، سنبيّن فيما يلي بنية كل وحدة والهدف منها [36] [38].

1.2.6.3- وحدة استخلاص السمات HubertFeatureEncoder

هي مكدس stack مكون من سبع طبقات من الشبكات العصبونية التلافيفية أحادية البعد (1-D Convolutional Neural Network layers) صُممت وفق الترتيب التالي:

طبقة تلافيفية أولية بحجم نواة (kernel size) يساوي 10 وخطوة (stride) تساوي 5، تليها أربع طبقات تلافيفية متتالية بحجم نواة 3 وخطوة 2 لكل منها، وأخيراً طبقتان تلافيفيتان بحجم نواة 2 وخطوة 2.

بعد كل طبقة تلافيفية، يوجد طبقة استنظام (Layer Normalization) لضمان استقرار عملية التدريب، تليها دالة التنشيط غير الخطية (GELU) لإضافة اللاخطية على السمات المستخلصة. كما يبيّن الشكل (20.3):



الشكل 20.3- وحدة استخلاص السمات HubertFeatureEncoder

يعدّ استخدام مكدس تلافيفي ممارسة قياسية وفعالة في معالجة الكلام الحديثة، حيث تؤدي هذه البنية وظيفتين أساسيتين متمثلتين في استخلاص السمات عبر المرشحات التلافيفية، وتقليل الأبعاد الزمنية (Temporal Downsampling) للإشارة الكلامية تدريجياً ويتحقق ذلك عبر الخطوات التي جرى توزيعها وفق الطبقات كما يلي: (2, 2, 2, 2, 2, 2, 5) إذ إن الإشارة الصوتية الخام، عند تردد تقطيع قدره 16KHz ، تكون غير مناسبة للمعالجة المباشرة بواسطة طبقات المحولات اللاحقة التي تكون تكلفتها الحسابية متناسبة تربيعياً مع طول التسلسل، وبالتالي هذا الإجراء يقلل من العبء الحسابي. وبناء على ذلك يتم تقليص طول التسلسل الزمني بمقدار 320 مرة، ونتيجة لذلك، يتم تحويل تردد الدخل البالغ 16KHz إلى تردد خرج 50Hz.

وبالتالي يُنتج المرمّز متجه سمات واحد كل 20ms، بحيث يُمثل كل متجه خلاصة نافذة زمنية طولها 25ms من الصوت الخام، بتداخل 5ms بين النوافذ الصوتية المتعاقبة. وفي نهاية هذه المرحلة، تتحول الموجة الصوتية أحادية البعد إلى تسلسل من متجهات السمات (Feature Vectors) عالية الأبعاد، يبلغ بُعد كل منها 512، لتصبح جاهزة للمعالجة اللاحقة داخل مرمز المحول.

2.2.6.3- مرمز المحوّل Transformer Encoder

يتألف مرمز المحوّل في نموذج HuBERT-BASE من اثنتي عشرة طبقة محوّل بينما يتألف في HuBERT-LARGE، الذي سنعتمده في دراستنا، من أربعٍ وعشرين طبقة محوّل. وكل طبقة هي كتلة Transformer Block متكاملة. وتجدر الإشارة إلى أنّ جميع الأبعاد التي سنذكرها في هذا السياق خاصة بنموذج LARGE دون غيره.

يبدأ الدخل إلى مرمز المحوّل من متجهات السمات الناتجة عن وحدة استخلاص السمات والتي يبلغ بُعدها 512، قبل إدخالها إلى الطبقات، تمر هذه المتجهات عبر طبقة إسقاط خطي (Linear Projection) لرفع أبعادها إلى 1024، وهو البعد المعتمد

في جميع طبقات المرمز، كما يُضاف إلى هذه المتجهات تضمين موضعي (Positional Embedding) يمكّن النموذج من ترميز الترتيب الزمني للإطارات، وذلك لتعويض افتقار بنية المحوّل إلى البنية التتابعية الصريحة، وبذلك يصبح تسلسل السمات جاهزًا للدخول في عملية النمذجة السياقية العميقة.

أما داخل كل طبقة محوّل، فيُطبق أولاً الانتباه الذاتي متعدد الرؤوس Multi-Head Self Attention، حيث يُقسَّم الدخل البالغ بُعده 1024 إلى ستة عشر رأس انتباه مستقل (16 Attention Heads)، بحيث يعمل كل رأس على فضاء فرعي ببعد 64.

يقوم كل رأس انتباه بإنشاء ثلاث مصفوفات: الاستعلام (Query – Q)، والمفتاح (Key – K)، والقيمة (Value – V). بعد ذلك تُحسب مصفوفة العلاقات Attention Scores بين كل إطار والإطارات الأخرى، ما يمكّن النموذج من استحضار معلومات من المواقع القريبة والبعيدة في التسلسل على حد سواء. بعد ذلك، تُدمج نواتج الرؤوس وتُسقط خطيًا لتعود إلى نفس البعد الأصلي (1024).

بعد تجميع معلومات السياق عبر الانتباه الذاتي، يتم تمرير كل متجه بشكل مستقل عبر شبكة تغذية أمامية Feed-Forward Network-FFN تتألف من طبقتين خطيتين متتاليتين وتُطبق بينهما دالة التنشيط (GELU) لإضافة اللاخطية اللازمة، بهدف توسيع البعد الوسيط إلى 4096 ثم إسقاطه مجدداً إلى 1024.

بعد المرور عبر أربع وعشرين طبقة من هذا النمط، ينتج تسلسل من التمثيلات السياقية عالية المستوى، حيث يحتفظ كل إطار صوتي ببعد 1024 لكن معزز بمعلومة سياقية مستمدة من كامل التسلسل الصوتي. هذه التمثيلات تشكّل الخرج النهائي لمرمز المحوّل، وتُرسل لاحقاً إلى طبقة الإسقاط النهائية Projection Layer التي تتولى مواءمتها مع فضاء الوحدات الخفية المستهدفة، تمهيداً لحساب دالة الخسارة.

7.3– الضبط الدقيق Fine-Tuning

مع كل ما أظهرته نماذج التعلم الذاتي الإشراف المدربة مسبقاً مثل HuBERT من قدرات مميزة في استخراج تمثيلات صوتية غنية وقابلة للتخصيص، وإثباتها قدرة فائقة على التعميم في مهام متعددة، تظل الاستفادة الكاملة من إمكانياتها مرتبطة بقدرتها على التكيف مع متطلبات مهام تطبيقية محددة. يبرز سؤال جوهري: كيف يمكن استثمار هذه الإمكانات الهائلة بصورة عملية وفعّالة في المهام التطبيقية؟ تكمن الإجابة في مرحلة الضبط الدقيق Fine-Tuning، التي تمكّن من تكييف النموذج المدرب مسبقاً مع متطلبات مهمة محددة مثل التعرف الآلي على الكلام.

لكن مع تضخم أحجام النماذج الحديثة لتشمل مئات الملايين أو حتى المليارات من الموسطات، أصبحت عملية الضبط الدقيق الكامل لجميع الأوزان مكلفة إلى حد كبير، سواء من حيث استهلاك الموارد الحاسوبية وذاكرة وحدة معالجة الرسومات GPU Memory، أو من حيث الحاجة إلى تخزين نسخة جديدة من النموذج لكل مهمة مستقلة. هذه الصعوبات لا تقتصر على مجال الكلام فقط، بل تعكس التحديات العامة التي تواجه النماذج التأسيسية (Foundation Models) في مختلف تطبيقات الذكاء الصنعي.

ومن هنا نشأت الحاجة إلى مقاربات أكثر كفاءة، تُعنى بتقليص عدد الموسطات الواجب تدريبها مع الحفاظ على جودة الأداء، وهو ما عُرف لاحقاً باسم تقنيات الضبط الدقيق الموفّر للموسطات Parameter-Efficient Fine-Tuning–PEFT والذي يرتكز على مبدأ تجميد (freezing) الغالبية العظمى من أوزان النموذج المدرب مسبقاً، و التركيز على تدريب جزء صغير جداً من الموسطات الإضافية أو الموجودة مسبقاً، وقد أثبتت هذه الطرق قدرتها على حل العقبات التي تواجه الباحثين أمن حيث متطلبات الذاكرة الرسومية وموارد الحوسبة و زمن التدريب مع المحافظة على أداء مماثل — بل وفي بعض الحالات متفوق — على أداء الضبط الدقيق الكامل [40].

1.7.3– فرضية البُعد الجوهري المنخفض

قبل البدء في شرح الأساس الرياضي من الضروري التوقف عند الإطار المفاهيمي الذي تستند تقنيات الضبط الدقيق، والمتمثل في فرضية البُعد الجوهري المنخفض low intrinsic Dimensionality، وقد طُرحت هذه الفرضية في السنوات الأخيرة في محاولة للإجابة عن تساؤل بحثي، وهو:

> كيف يمكن لنماذج تحتوي على مئات الملايين أو المليارات من الموسطات أن تُضبط بدقة على مهام جديدة باستخدام مجموعات معطيات صغيرة نسبياً وباعتماد خوارزميات انحدار متدرج بسيطة؟

في هذا السياق، اقترح Aghajanya تحليل عملية الضبط الدقيق finetune من خلال مفهوم البُعد الجوهري، والذي يُعرّف بأنه الحد الأدنى لعدد الموسطات اللازم تعديلها للوصول إلى أداء جيد في المهمة المستهدفة، وذلك ضمن فضاء الموسطات الكامل للنموذج، وتشير الفرضية إلى أن التعديلات التي يُجريها النموذج خلال الضبط الدقيق لا تنتشر عشوائياً في جميع أبعاد الفضاء عالي الرتبة، بل تتركز داخل فضاء فرعي منخفض الأبعاد مضمن فيه، بمعنى آخر، يمكن تكييف النموذج بفعالية من خلال تحديث عدد محدود جداً من الموسطات [41].

وقدمت الورقة البحثية أدلة تجريبية قوية لدعم هذه الفرضية، حيث أظهرت أن البُعد الجوهري للنماذج المدربة مسبقاً منخفض للغاية. فعلى سبيل المثال، تمكن نموذج RoBERTa-Large من تحقيق نحو 90٪ من أداء الضبط الكامل على مهمة تصنيف

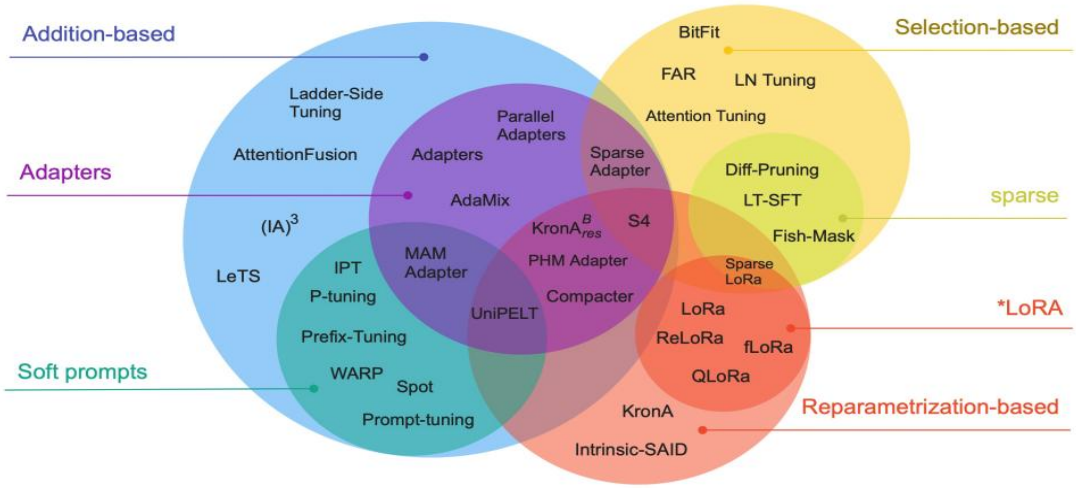
باستخدام 200 موسطاً فقط. كما بينت الدراسة أن عملية التدريب المسبق لا تمنح النموذج تمثيلات لغوية عامة فحسب، بل تُعيد هيكلة بنيته الداخلية بشكل يجعل عملية التكيّف اللاحق أكثر فعالية وأقل تطلباً من حيث عدد المعلمات المتغيرة[41] .

وبالإضافة إلى ذلك، أوضحت الورقة أن النماذج الأكبر حجماً تتمتع ببُعد جوهري أقل، أي أن كلما ازداد عدد موسطات النموذج، قلّ عدد الموسطات الفعلية المطلوبة لتحقيق أداء عالٍ في مهمة جديدة. هذا الاكتشاف يقدّم تفسيراً علمياً لفعالية النماذج الضخمة في التكيّف مع مجموعة واسعة من المهام، ويعزّز من قيمة الأساليب التي تستغل البُعد المنخفض في عملية التحديث بحيث تقتصر على فضاء فرعي منخفض الأبعاد، بدلاً من تعديل كل الموسطات، مما يتيح تعلماً أكثر كفاءة بأقل عدد ممكن من المعلمات القابلة للتدريب [41].

وبناءً على ما سبق، تُعد فرضية البُعد الجوهري المنخفض الركيزة النظرية الأهم التي تستند إليها معظم تقنيات الضبط الدقيق الموفّر للمعلمات، وعلى الأخص الأساليب القائمة على إعادة التمثيل.

2.7.3– تصنيف ومقارنة تقنيات الضبط الدقيق الموفِّر للموسطات

تندرج تقنيات الضبط الدقيق الموفِّر للموسطات (PEFT) تحت ثلاث فئات رئيسية[42] كما هو مبيّن في الشكل (21.3)، تختلف فيما بينها من حيث الأسلوب المتبع في تقليل عدد الموسطات القابلة للتدريب، وطبيعة تدخلها في بنية النموذج الأصلي، وكفاءتها من حيث الأداء والزمن والموارد.



الشكل 21.3-تقنيات الضبط الدقيق الموفِّر للمعلمات

1.2.7.3- الأساليب القائمة على الإضافة Addition-Based Methods

تقوم هذه الفئة على مبدأ تجميد جميع أوزان النموذج المدرب مسبقاً وإضافة مكونات صغيرة قابلة للتدريب، من أبرز أمثلتها المكيّفات (Adapters) حيث يتمّ إدخال شبكات عصبونية صغيرة متصلة بالكامل (small fully-

connected networks) بعد طبقات المحولات. يقتصر التدريب في هذه الحالة على هذه المكونات الجديدة فقط، دون المساس بأوزان النموذج الأصلي.

تمتاز هذه الأساليب بمرونتها العالية وكفاءتها في استهلاك الذاكرة والتخزين مقارنة بالضبط الدقيق الكامل، إلا أن إدخال طبقات إضافية يزيد من عمق النموذج وهذا بدوره يتطلب معالجة تسلسلية إضافية وبالتالي يؤدي إلى زمن استدلال أطول.

2.2.7.3- الأساليب القائمة على الاختيار Selection-Based Methods

على النقيض من الفئة السابقة، لا تضيف هذه الأساليب أي مكونات جديدة إلى النموذج، بل تعتمد على تحديد مجموعة فرعية صغيرة جدًا من الموسطات الموجودة مسبقاً وتحديثها فقط، مع إبقاء الغالبية العظمى من الأوزان مجمّدة. المثال الأبرز على ذلك هو BitFit، والذي يكتفي بتحديث موسطات التحيز (Bias Terms) فقط.

تتميز هذه الأساليب بعدم إضافة أي تكلفة زمنية أثناء الاستدلال وبحجم تحديثات بالغ الصغر، إلا أن قدرتها على التكيف محدودة، حيث يتراجع أداؤها مع النماذج الكبيرة التي تتجاوز المليار موسطاً بشكل ملحوظ مقارنةً بالضبط الدقيق الكامل، فضلًا عن عدم قابليتها للتطبيق في بعض النماذج الحديثة مثل LLaMA التي تفتقر أصلاً إلى موسطات تحيز.

3.2.7.3- الأساليب القائمة على إعادة التمثيل Reparameterization-Based Methods

تستند هذه الفئة إلى فرضية البُعد الجوهري المنخفض التي تفترض أن التحديثات المطلوبة لأوزان النموذج تقع فعلياً في فضاء فرعي منخفض الرتبة. ومن أبرز أمثلتها Low-Rank Adaptation-LoRA [43] حيث تتم إضافة مصفوفتين منخفضتي الرتبة بجانب الأوزان الأصلية ليتم تدريبهما فقط خلال مرحلة الضبط الدقيق. أثبتت LoRA قدرتها على تقليل عدد الموسطات القابلة للتدريب بمعدل قد يصل إلى عشرة آلاف مرة، وخفض استهلاك الذاكرة الرسومية بما يقارب ثلاثة أضعاف مقارنة بالضبط الدقيق الكامل، مع المحافظة على أداء مماثل أو أفضل.

غير أن LoRA التقليدية تعاني من بطء التعلم وانهيار التدرجات عند الرتب المرتفعة مما يحد من إمكانية الاستفادة القصوى من النموذج، وهنا جاء تطوير Rank Stabilization LoRA-rsLoRA كحل مباشر لهذه المعضلة، بقدرتها على تحقيق استقرار أكبر في عملية التدريب، وإتاحة المجال للاستفادة من الرتب العالية لتحسين الأداء، مع المحافظة على نفس كفاءة LoRA في الاستدلال وعدم فرض أي تكلفة إضافية في زمن التنفيذ، ولهذه الأسباب وقع الاختيار على rsLoRA كإطار معتمد في مرحلة الضبط الدقيق ضمن هذا البحث[44] .

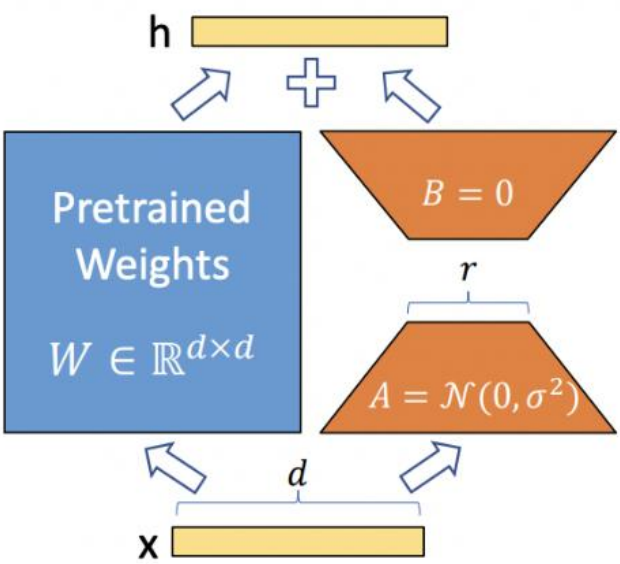
3.7.3– آلية عمل rsLoRA

تعتمد تقنية rsLoRA على آلية عمل تتكون من خطوتين أساسيتين: التجميد والحقن[44] .

أولاً: كما هو الحال في LoRA تعتمد rsLoRA على تجميد مصفوفة الأوزان الأصلية W_0، هذا يعني أن الأوزان التي المتضمنة المعرفة العامة المكتسبة أثناء مرحلة التدريب المسبق تظل ثابتة ولا يتم تحديثها أثناء عملية الضبط الدقيق.

ثانياً: يتم حقن (inject) مصفوفات قابلة للتدريب ذات رتبة منخفضة low-rank matrices في بنية المحول من النموذج، وتحديداً في مصفوفات الإسقاط الخاصة بآلية الانتباه الذاتي self-attention، وهي الوحيدة التي يتم تدريبها لتعلم التحديثات الخاصة بالمهمة الجديدة.

كما يوضح الشكل (22.3) تعمل rsLoRA بشكل متوازٍ مع مصفوفة الوزن الأصلية، بدلاً من أن يمر الخرج من خلال طبقة جديدة، يتم حساب مسارين في نفس الوقت: المسار الأصلي عبر مصفوفة الوزن المجمدة W_0 والمسار الجديد عبر rsLoRA يتم بعد ذلك جمع مخرجات المسارين.



الشكل 22.3- آلية عمل rsLoRA

هذا الاختيار المعماري هو حجر الزاوية الذي يمنح rsLoRA ميزتها الأكثر أهمية: القدرة على دمج الأوزان مرة أخرى في النموذج الأساسي بعد انتهاء التدريب بحيث لا تشكل أي تكلفة زمنية أثناء الاستدلال.

1.3.7.3- الأساس الرياضي لـrsLoRA

كما ذكرنا سابقاً تقوم تقنية [44] rsLoRA على فرضية أساسية مفادها أن التغيير في مصفوفة الأوزان ΔW أثناء عملية الضبط لمهمة جديدة تمتلك رتبة جوهرية منخفضة (low intrinsic rank). وبالتالي من منظور الجبر الخطي يعني أنه يمكن تمثيل المصفوفة ΔW من خلال حاصل ضرب مصفوفتين أصغر حجماً بكثير B و A. يُعرف هذا الإجراء باسم تحليل الرتبة المنخفضة (low-rank decomposition)، ويتم التعبير عنه رياضياً كالتالي:

$$\Delta W \approx BA \qquad (34.3)$$

حيث أنّ: مصفوفة الوزن الأصلية $W_0 \in R^{d\times k}$ وبالتالي $A \in R^{r\times k}$ $B \in R^{d\times r}$ و $r \ll \min(d,k)$ تمثل الرتبة. أثناء التدريب، يتم تعديل عملية التمرير الأمامي forward pass للطبقة المستهدفة بحيث إذا كان x هو متجه الدخل للطبقة، فإن متجه الخرج h يتم حسابه وفق المعادلة الرياضية التالية:

$$h = W_0\, x + \frac{\alpha}{\sqrt{r}}\, \Delta W\, x = \left(W_0 + \frac{\alpha}{\sqrt{r}}\, BA \right) x \qquad (35.3)$$

α معامل قياس للتحكم بشدة التحديث الذي تُضيفه rsLoRA إلى الأوزان الأصلية.

تتمثل الإضافة المحورية لـ rsLoRA في تعديل معامل القياس التقليدي المستخدم في LoRA، حيث يتم استبدال $\frac{\alpha}{r}$ بـ $\frac{\alpha}{\sqrt{r}}$ وذلك بهدف تفادي تلاشي المشتق الناتج عن استخدام رتب عالية، وهو ما يتيح لـ rsLoRA الاستفادة من مرونة أكبر دون التضحية باستقرار التدريب.

2.3.7.3- التهيئة والدمج

لضمان أن يبدأ التدريب من الحالة المستقرة للنموذج المدرب مسبقاً وتجنب أي مشاكل في بداية عملية التعلم يتم تهيئة مصفوفة A باستخدام قيم عشوائية من توزيع غاوسي (Gaussian distribution)، بينما يتم تهيئة مصفوفة B بالأصفار [44].

وهذا يضمن أنه في بداية التدريب (عند الزمن t=0)، يكون ΔW=BA=0 نتيجة لذلك، يكون خرج الطبقة المعدلة مطابقاً تمامًا لخرج الطبقة الأصلية، وهذا يسمح للنموذج بالبدء من نقطة معروفة، ويتم تعلم التكييف بشكل تدريجي ومستقر مع تقدم التدريب.

بعد اكتمال التدريب، يجري دمج التحديثات مرة واحدة داخل الوزن الأصلي وفق المعادلة الرياضية التالية:

$$\acute{W} = W_0 + \frac{\alpha}{\sqrt{r}}\, \Delta W \qquad (36.3)$$

ويُستخدم الوزن الجديد $\acute{W}$ خلال مرحلة الاستدلال عبر المعادلة المبسطة:

$$h = \acute{W}\, x \qquad (37.3)$$

كخلاصة بعد الدمج يصبح لدى النموذج نفس البنية المعمارية وعدد العمليات الحسابية تماماً مثل النموذج الأصلي المدرب مسبقاً ولا توجد وحدات إضافية لمعالجتها، ولا توجد مسارات متوازية لحسابها أثناء الاستدلال. هذا يعني أن rsLoRA تحقق كفاءة هائلة في التدريب دون أي تكلفة على سرعة الاستدلال [44].

8.3– الخلاصة

بعد استكمال التمهيد النظري لأهم المفاهيم والنماذج المرتبطة بمعالجة الإشارة الكلامية، ينتقل البحث فيما يلي إلى القسم التطبيقي ضمن الفصلين الرابع والخامس، حيث يجري تنفيذ النماذج المدروسة واختبار أدائها عملياً في معالجة الصدى وتحسين دقّة أنظمة تعرّف الكلام آلياً.

الفصل الرابع

تنفيذ واختبار نموذج لحذف الصدى بالاعتماد على بنية LSTM

يستعرض هذا الفصل منهجية إعداد مجموعة المعطيات التي تم تدريب النموذج عليها، وآلية اختيار السمات الطيفية المناسبة، بالإضافة إلى عرض بنية النموذج المقترح والبرمجيات المستخدمة في تنفيذ عملية التدريب، وصولًا إلى استعراض نتائج الاختبار وتحليلها.

1.4- تجهيز مجموعة المعطيات

تلعب معطيات التدريب دوراً محورياً في نجاح أي نموذج من نماذج الذكاء الصنعي، حيث تنقسم المعطيات المستخدمة في عملية التعلم إلى ثلاثة أنواع رئيسية:

- معطيات التدريب (Training Dataset): وهي المجموعة الأساسية من الأمثلة التي تحتوي على أزواج من الدخل والخرج المقابل، ويُستخدمها النموذج لتعديل أوزانه وموسطاته أثناء عملية التعلم. تشكل هذه المعطيات عادة ما بين 70%إلى 80% من إجمالي المعطيات.
- معطيات التحقق (Validation Set): تُستخدم هذه المعطيات لمراقبة أداء النموذج أثناء التدريب والكشف عن حالات التلبيق الزائد(Overfitting)، وذلك باختباره على أمثلة لم تُستخدم في عملية التعلم. تشكل هذه المعطيات ما بين 10% إلى 15% من مجموع المعطيات الكلي.
- معطيات الاختبار (Testing Set): وهي مجموعة مستقلة تماماً تُستخدم بعد انتهاء التدريب لتقييم قدرة النموذج على التعامل مع معطيات جديدة لم يرها من قبل. تمثل عادة ما يقارب 10% إلى 15% من المعطيات الكلية.

وفي إطار هذا العمل، تحتاج الشبكة العصبونية إلى مجموعة معطيات صوتية ذات خصائص محددة، إذ تتطلب عملية التدريب توفر عدد كبير من الملفات الصوتية يصل إلى نحو ثلاثين ساعة من التسجيلات، بما يتيح للنموذج استخلاص السمات الصوتية الدقيقة. كما ينبغي أن تتضمن ملفات صوتية قبل إضافة الصدى وبعده، بحيث يتمكن النموذج من التعلم على الأزواج الصوتية المكوّنة من الإشارة النظيفة والإشارة المشوبة بالصدى. ومن المهم أن تشمل على تسجيلات متنوعة من متحدثين ذكور وإناث لضمان حيادية النموذج وقدرته على التعميم عبر مختلف أنماط الصوت البشري، كما

يجب أن تقتصر هذه التسجيلات على الكلام البشري فقط دون أي أصوات خارجية مثل الموسيقى أو الضجيج البيئي، بهدف تركيز عملية التعلم على مكونات الإشارة الكلامية المرتبطة بظاهرة الصدى.

يُعَدّ مجال حذف الصدى الصوتي (Dereverberation) من المجالات التي تعاني من ندرة في مجموعات المعطيات المخصّصة للتدريب، إذ تفتقر الأبحاث إلى مجموعات معطيات عامة وشاملة تجمع بين الإشارات النظيفة ونظيراتها المتأثرة بالصدى في بيئات متنوعة. لذلك، يُلجأ إلى استخدام مجموعة معطيات تحتوي على كلام نظيف بالتوازي مع مجموعات معطيات تتضمن الاستجابات النبضية للغرف RIRs .

غير أن الحصول على استجابة نبضية حقيقية لكل غرفة محتملة يُعَدّ أمراً غير عملي، نظراً لما يتطلبه من موارد ضخمة في التسجيل والمعالجة والتخزين، بالإضافة إلى تضخم كبير في حجم مجموعات المعطيات الناتجة. ولهذا السبب، يتم إعادة استخدام مجموعة محدودة من الـ RIRs عبر عدد كبير من الملفات الصوتية المختلفة، وهو إجراء ضروري للحفاظ على توازن واقعي بين تنوّع البيئات الصوتية وكفاءة التدريب الحسابي.

كما أن هذا التكرار ليس مجرد حل تقني لتقليل الموارد فحسب، بل هو أيضاً عنصر أساسي في تحسين جودة التعلم؛ إذ يمكّن النموذج من فهم خصائص الصدى الناتج عن الغرفة، بقطع النظر عن محتوى الجملة المنطوقة. فحين يُطبَّق الـ RIR ذاته على عينات كلامية متعددة، يتعلّم النموذج الأنماط الطيفية والزمنية المشتركة التي تُميز استجابة الغرفة، مما يساعده على تمييز الصدى كظاهرة صوتية فيزيائية مستقلة عن المحتوى اللغوي. وبذلك، يصبح النموذج أكثر قدرة على التعميم والاستقرار أثناء الاختبار، ويستطيع التعامل مع بيئات صوتية جديدة بكفاءة أعلى.

1.1.4– اختيار مجموعات المعطيات

1.1.1.4- مجموعة معطيات الكلام

في هذا العمل تم الاعتماد على مجموعة معطيات LibriSpeech التي تُعَدّ من أشهر مجموعات المعطيات المجانية، يبلغ حجمها حوالي 60GB وقد تم تطويرها في جامعة Johns Hopkins وتضم ما يقارب 1000ساعة من الكلام المقروء باللغة الإنكليزية، مأخوذة من مجموعة كتب صوتية ضمن مشروع LibriVox لتأليف الكتب الصوتية،بتردد تقطيع 16kHz . وتمتاز هذه المجموعة بتوزيع متوازن من حيث المفردات والمتكلمين والجنس (ذكور وإناث)، مما يجعلها مناسبة لتدريب نماذج تتسم بالاستقلال عن المتحدث وتحليل خصائص الإشارة، وقد تم تصنيفها إلى مجموعات فرعية بناءً على الحجم وجودة التسجيل كما يظهر الجدول التالي:

subset	hours	Minutes per speakers	Female speakers	Male speakers	total speakers
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7 3	30	564	602	1166

الجدول 1.4- مجموعة معطيات LibriSpeech

استخدمت هذه الدراسة الجزء "train-clean-100" من مجموعة المعطيات للتدريب، والذي يتضمن حوالي 100 ساعة من التسجيلات الصوتية النظيفة موزعة على 251 متحدث بمتوسط 25 دقيقة لكل متحدث. وقد تم اختيار هذا الجزء تحديداً لعدة أسباب؛ أولها أنه يمثل معطيات نظيفة وموثوقة خالية من الضجيج أو التشويش، وهو ما يجعلها مثالية لتوليد ملفات صوتية وثانياً، يوفّر هذا الجزء تنوعاً كافياً في الأصوات البشرية مع حجم معطيات معتدل، مما يوازن بين جودة التدريب وكفاءة الموارد الحسابية.

2.1.1.4- مجموعات معطيات الاستجابة النبضية

تمّ اعتماد مجموعات المعطيات التالية:

Great Hall, Octagon, Classroom, MARDY, Aachen Impulse Response -AIR[21] حيث تتضمن الاستجابة النبضية لغرف متنوعة ومختلفة، بعض منها يحاكي غرفاً كبيرة وأخرى صغيرة وأسطح عاكسة.

- Great Hall, Octagon, Classroom[22]

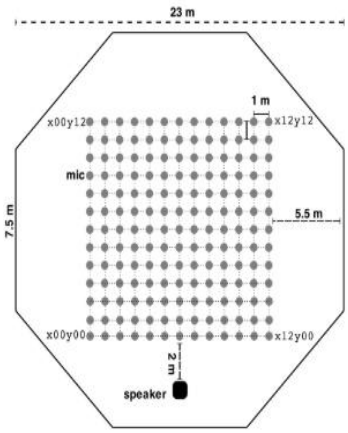
هي مجموعات معطيات صادرة عن مخبر Centre for Digital Music (C4DM) في جامعة Queen Mary والمتاحة عبر موقع Isophonics، والتي تم توثيقها في بحث Stewart & Sandler (ICASSP 2010) تحت عنوان Database of Omnidirectional and B-Format Impulse Responses [45]، تم تسجيل هذه المعطيات عام 2008، وذلك بالاعتماد على مكبر صوت من نوع Genelec 8250A وميكروفونات من نوع:

[21] https://openslr.org/20/
[22] http://isophonics.org/content/room-impulse-response-data-set

- ميكروفون omnidirectional DPA 4006 يلتقط الصوت من جميع الاتجاهات بانتظام إلى قناة واحدة دون تمييز الجهة التي يرد منها الصوت.
- ميكروفون B-format Soundfield SPS422B يمتلك 4 قنوات Sound Field يلتقط الموجات الصوتية في المحاور الثلاثة (X,Y,Z) بالإضافة إلى شدة الصوت W وبالتالي يستخدم لالتقاط وتمثيل المجال الصوتي المكاني Spatial Sound Field في الاتجاهات المختلفة.

توفّر جميع التسجيلات بصيغة WAV بتردد تقطيع 96kHz، وتمثل قياسات لثلاث بيئات معمارية رئيسية داخل حرم جامعة Queen Mary في Mile End Campus، تم اختيارها لتغطية نطاق واسع من خصائص الصدى الصوتي:

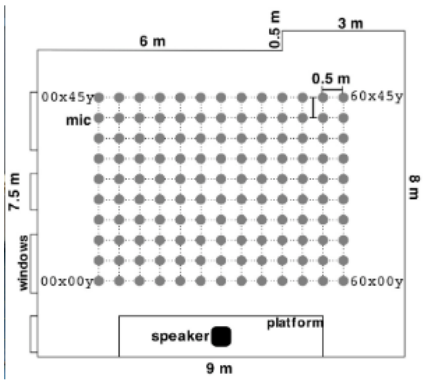
- قاعة Great Hall قاعة متعددة الاستخدامات تتسع لحوالي 800 مقعد، تبلغ أبعادها نحو 16 ×23 متر، مع سقف مرتفع. تم وضع 169 ميكروفون موزعة في أرضية الغرفة على أبعاد 12 ×12 متر، تُسجِّل هذه البيئة استجابات ذات زمن صدى طويل وانعكاسات متعددة، ما يجعلها نموذجاً مثالياً لمحاكاة البيئات الواسعة كقاعات العروض والمحاضرات الكبرى، وتتضمن 169 ملف WAV.
- قاعة Octagon مُبينة في الشكل (1.4) وهي مبنى فيكتوري بُني عام 1888 كان يُستخدم مكتبة ثم تحوّل إلى قاعة مؤتمرات، وتتكوّن من ثمانية جدران بطول 7.5 متر لكل منها مع قبة يبلغ ارتفاعها 21 متر .تمتاز هذه القاعة بصدى متوسط إلى طويل وبتوزيع متعدّد الاتجاهات للانعكسات، مما يجعلها بيئة تمثيلية متوسطة التعقيد الطيفي والزماني. وتتضمن 169 ملف WAV. وتتضمن 169 ملف WAV.



الشكل 1.4- توزيع الميكروفونات والسماعة في قاعة Octagon بجامعة كوين ماري

- قاعة Classroom مُبينة في الشكل (2.4) غرفة تدريس في كلية الهندسة الإلكترونية وعلوم الحاسوب، أبعادها تقريباً 3.5×9×7.5متر، تتميّز بأسطح عاكسة كالأرضية البلاستيكية والجدران المطلية، مع وجود عناصر ماصّة جزئياً. يقدّم

هذا المكان استجابات بصدى قصير نسبياً، مما يجعله مناسباً لمحاكاة البيئات اليومية الصغيرة مثل المكاتب أو الصفوف الدراسية. وتتضمن 130 ملف WAV.



الشكل 2.4- توزيع الميكروفونات والسماعة في قاعة Classroom

وقد اختيرت هذه البيئات الثلاث لأنها تُغطي نطاقاً متدرجاً من زمن الصدى RT-60 الطويل في القاعات الواسعة إلى القصير في الغرف الصغيرة، وبذلك تمكّن النموذج من التعلّم على بيئات صوتية متنوعة وتعميم قدراته على سيناريوهات مختلفة في عملية إزالة الصدى الصوتي.

تجدر الإشارة إلى أنّ مجموعة المعطيات الأصلية تتضمّن نوعين من التسجيلات تم الحصول عليهما باستخدام نوعين مختلفين من الميكروفونات، وذلك بهدف توفير استجابات نبضية تلائم تطبيقات صوتية متعددة:

- تسجيلات (Omnidirectional) أُجريت بواسطة ميكروفون DPA 4006.
- تسجيلات بصيغة B-format تحوي معلومات مكانية، أُجريت بواسطة ميكروفون Soundfield SPS422B.

وقد تم في هذه الدراسة اختيار واستخدام ملفات الـ Omnidirectional فقط، التي تمثّل الاستجابة النبضية من قناة أحادية، بما يتوافق مع مسألة البحث.

- <u>Aachen Impulse Response -AIR</u>

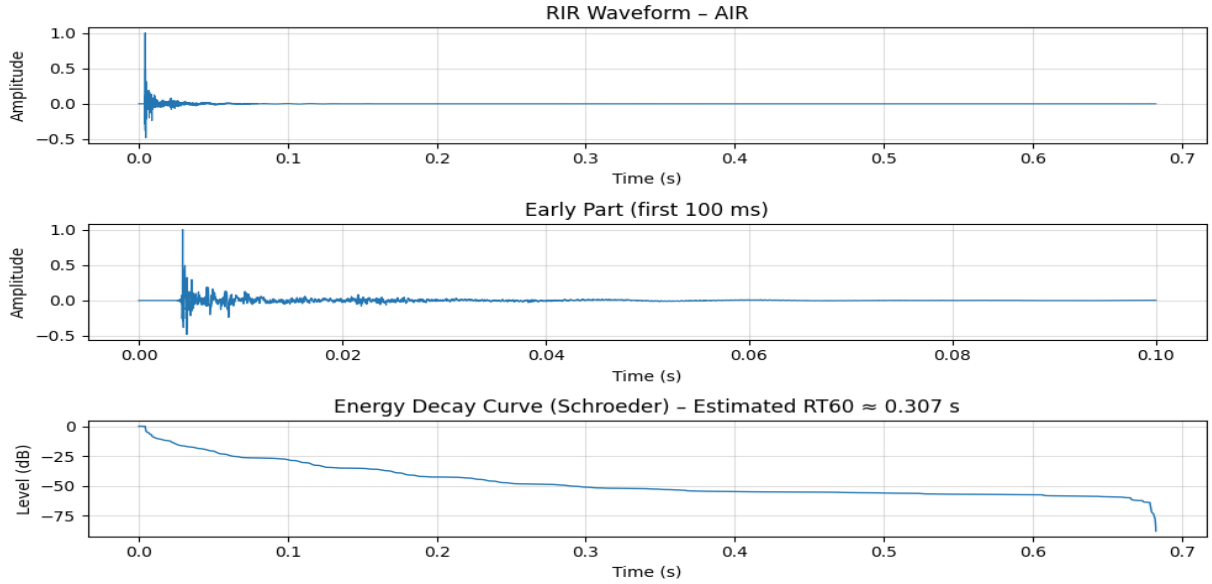
تعدّ إحدى مجموعات المعطيات المعتمدة في دراسة أداء الخوارزميات الصوتية ضمن بيئات الصدى، إذ تتيح مجموعة واسعة من الاستجابات النبضية في غرف ذات خصائص صوتية متنوعة. نُشرت النسخة الأولى من هذه المجموعة عام 2009 في معهد Institute of Communication Systems (IKS) بجامعة RWTH Aachen في ألمانيا، وكان الهدف الأساسي منها

دعم الدراسات المتعلقة بمعالجة الإشارة في البيئات ذات الصدى، مع تركيز خاص على تطبيقات المساعدات السمعية Hearing Aids وتحليل السلوك السمعي البشري في ظروف صوتية مختلفة. [47] [46]

تضم مجموعة AIR استجابات نبضية لغرف مختلفة بتردد تقطيع 48 kHz بدءاً من البيئات منخفضة الارتداد وحتى القاعات الكبيرة ذات زمن صدى مرتفع، وتشمل أربع غرف رئيسية هي:

- Studio Booth غرفة صغيرة منخفضة الصدى، أبعادها 3.0 × 1.8 × 2.2 m وزمن الصدى RT60 ≈ 0.12 Sec ذات جدران عازلة وألواح امتصاصية خاصة.
- Office Room بيئة مكتبية متوسطة الصدى، أبعادها 5.0 × 6.4 × 2.9 m وزمن الصدى RT60 ≈ 0.43 Sec تحتوي على أثاث وأسقف عاكسة.
- Meeting Room قاعة اجتماعات متوسطة الحجم ، أبعادها 8.0 × 5.0 × 3.1 m وزمن الصدى RT60 ≈ 0.23 Sec تُظهر انعكاسات واضحة ناتجة عن الطاولة والجدران.
- Lecture Room قاعة محاضرات واسعة أبعادها 8.0 × 5.0 × 3.1 m و RT60 ≈ 0.78 Sec تُعد نموذجاً لبيئة ذات صدى قوي وزمن صدى طويل.

يمثل الشكل (3.4) إحدى الاستجابات النبضية من غرفة Lecture Room مجموعة المعطيات AIR.



الشكل 3.4- الاستجابة النبضية لإحدى RIR من مجموعة المعطيات AIR

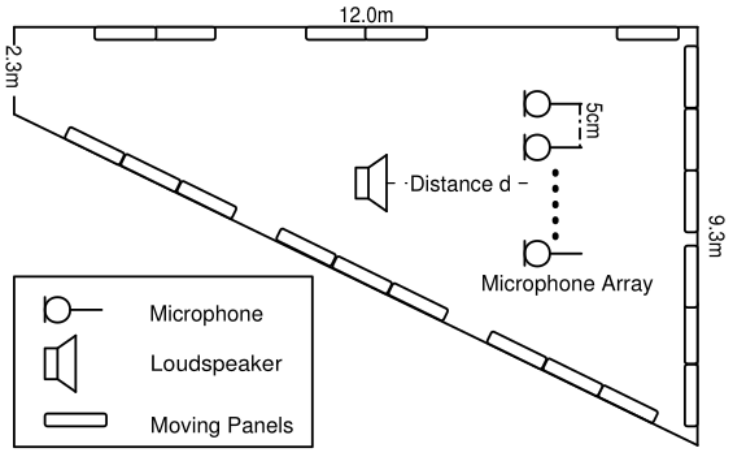
تمّ الحصول على هذه الملفات باستخدام ميكروفونين من نوع omnidirectional Beyerdynamic MM1 في مواقع متعددة داخل الغرف السابقة، عند مسافات مختلفة بين المصدر والميكروفون (0.5 – 10.2) متر لتوفير تنوّع مكاني وواقعي.

الإصدار المستخدم في هذه الدراسة V1.4 وتتضمن 344 ملف WAV.

- MARDY

تُعدّ Multichannel Acoustic Reverberation Database at York- MARDY [48] من أوائل مجموعات المعطيات البحثية التي أُنشئت لتقييم أداء خوارزميات حذف الصدى الصوتي ضمن بيئات واقعية متعددة القنوات. تم تطوير هذه المجموعة في جامعة يورك University of York,UK، وتحديداً في مركز أبحاث الموسيقى Music Research Centre، بالتعاون مع Imperial College London وTechnische Universiteit Eindhoven، وقد أُنشئت بهدف توفير معطيات حقيقية تمكّن الباحثين من اختبار خوارزميات المعالجة الصوتية على قياسات دقيقة.

أُجريت التسجيلات داخل غرفة صوتية متغيّرة الخصائص (Varechoic Room) في الشكل (4.4) ضمن منشأة Trevor Jones Recording Facility، وهي غرفة يمكن تغيير خصائصها الصوتية من خلال ألواح عاكسة وممتصّة متحرّكة تسمح بتبديل خصائص الانعكاس الصوتي للجدران، مما أتاح جمع معطيات تحت ظروف انعكاسية (Reflective) وأخرى ماصّة Absorbent. تمّ استخدام مكبر صوت من نوع Genelec 1029A كمصدر صوتي، ومصفوفة ميكروفونات خطية Linear Microphone Array مكوّنة من ثماني ميكروفونات من نوعAKG C417، متباعدة بمسافة 5سم بين كل عنصر، موضوعة على ارتفاع 1متر عن الأرض.



الشكل 4.4- مخطط يوضّح أبعاد غرفة التسجيل وإعداداتها في مجموعة معطيات MARDY

تم تسجيل الاستجابات النبضية مع تغيير المسافة بين المصدر ومصفوفة الميكروفونات (1m، 2m، 3m، 4m) في كل من حالتي الألواح العاكسة والماصّة. زمن الصدى (RT60) يتراوح بين 0.29s في الحالة الماصّة و 0.45s في الحالة العاكسة، حيث يكون الصدى واضحاً بما يكفي ليؤثّر على دقة تعرف الكلام آلياً، لكنه غير مرتفع إلى درجة تُحدث تشويهاً سمعياً ملحوظاً في الإشارة الكلامية. وتتضمن 72 استجابة نبضية بصيغة ملف WAV.

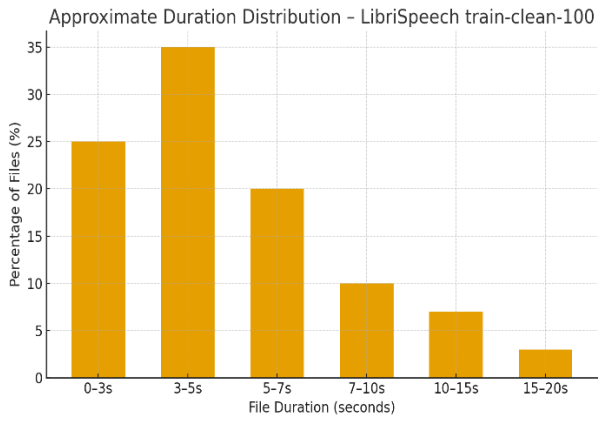
تجدر الإشارة إلى أنه وجود عدة ميكرفونات أثناء تسجيل RIRs لا يعني أنّ المسألة أصبحت متعددة القناة أو تتعارض مع محددات مسألتنا أحادية القناة، إذ يمكن اعتبار كل استجابة نبضية مسجَّلة من ميكروفون مختلف استجابة مستقلة أحادية القناة وتمثل انتقال الصوت من المصدر إلى موقع محدّد داخل الغرفة. وعليه، يمكن معالجة كل قناة على حدة لتوليد معطيات أحادية القناة متنوّعة طيفياً وزمنياً، وهو نهج معتمد وموثّق في الأدبيات العلمية الخاصة بخوارزميات إزالة الصدى الصوتي في قناة أحادية فعلى سبيل المثال، تُعدّ مجموعة المعطيات!WHAMR [49] من أبرز المجموعات التي طُوِّرت لتقييم أداء أنظمة فصل الكلام وتحسينه (Speech Separation and Enhancement) في بيئات تحتوي على الضجيج والصدى ضمن إطار أحادي القناة. وهي المجموعة امتداد مباشر لمجموعتي المعطيات WSJ0-2mixو!WHAM، حيث أُضيف إليها مكوّن الصدى (Reverberation) لتوفير محاكاة واقعية للظروف الصوتية في الغرف المغلقة، وقد تمّ اتباع النهج ذاته في توليد هذه المجموعة من خلال اعتبار الاستجابات النبضية أحادية القناة التي تمثّل مواقع ميكروفونات مختلفة داخل الغرفة، وذلك لتوليد معطيات تدريب متنوّعة تُحاكي ظروف الصدى الواقعية وتُستخدم في تقييم أداء النماذج الصوتية في بيئات متعددة.

2.4- تصميم مجموعة المعطيات

كما ذكرنا سابقاً، يفتقر مجال حذف الصدى إلى مجموعات معطيات جاهزة ومتكاملة مناسبة للتدريب، مما استدعى تصميم مجموعة معطيات مخصّصة، كما هو متّبع في العديد من الدراسات.

- تمّت معالجة مجموعة المعطيات LibriSpeech نظراً لتفاوت أطوال ملفاتها الصوتية، إذ تتراوح مدة الملف بين 3 ثواني و20 ثانية تقريباً، في حين تمتدّ مدتها الإجمالية إلى نحو 100ساعة. ولضمان توحيد أطوال الملفات الصوتية، وتسهيل معالجتها عبر النموذج المقترح، تمّ تحديد طول يبلغ 5 ثوان لكل ملف صوتي. حيث تُستبعد الملفات الأقصر من هذا الحدّ، بينما تُقسَّم الملفات الأطول إلى عدة مقاطع متتالية بطول 5 ثواني لكل منها، بحيث تُغطّي كامل التسجيل الأصلي دون تداخل زمني. وقد تمّ اعتماد هذا الإجراء نظراً إلى أن المدة الكلية لمجموعة المعطيات كبيرة بما يكفي (100 ساعة)، الأمر الذي يسمح بالتضحية بجزء محدود منها، حيث أنّ 25% فقط من عدد الملفات الصوتية أقل من 5 ثواني، كما يُظهر الشكل (5.4)، وإنّ إضافة أصفار للملفات الأقصر يؤثر سلباً على تعلم الشبكة ويزيد من

التعقيد الحسابي. لذلك كان خيار توحيد الأطوال مع استبعاد المقاطع الأقصر هو الأنسب من الناحيتين الحسابية والتعلّمية للشبكة.



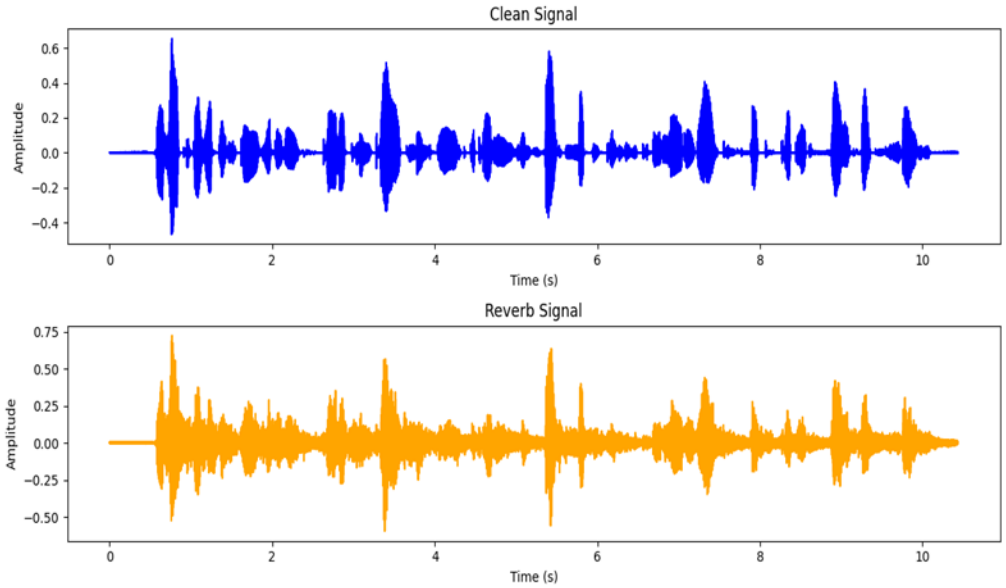
الشكل 5.4- التوزيع التقريبي لمدة المقاطع الصوتية في (LibriSpeech (train-clean-100

بعد القيام بهذه الخطوة أصبحت مدة مجموعة المعطيات حوالي 80 ساعة. ونظراً لأن تدريب الشبكة العصبونية على كامل هذه المعطيات يتطلّب موارد حسابية كبيرة من حيث الذاكرة وزمن المعالجة، وإدراكاً لما أشرنا إليه سابقاً بأن تدريب الشبكات العصبونية يتطلّب عادةً ما يقارب 30ساعة لتحقيق أداء مستقر، فقد تمّ اختيار نحو 35 ساعة من مجموعة المعطيات الكلية للتدريب الفعلي، في حين خُصِّصت 10ساعات للتحقق Validation set. تمّ توزيع العينات بحيث تغطي نطاقاً واسعاً من المتحدثين لضمان التنوع الصوتي، مع الحفاظ على استقلال المتحدثين بين المجموعتين، بمعدل أخذ العينات 16KHz.

- نظراً لأن كل استجابة نبضية (RIR) تمثل قياساً مختلفاً لانتقال الصوت من المصدر إلى موقع محدّد داخل الغرفة، فقد اعتُبرت كل RIR مسجّلة من ميكروفون مختلف استجابة مستقلة تعبّر عن خصائص زمنية وطيفية مميّزة للموقع المكاني الذي سُجّلت منه. وعليه، تمّ تقسيم الاستجابات النبضية داخل كل غرفة بنسبة 80% لمعطيات التدريب، و 10% للتحقق، و 10% للاختبار، مع الحفاظ على استقلالية مواقع الميكروفونات بين المجموعات.بحيث تُستخدم مواقع الميكروفونات المختلفة داخل الغرف المتنوعة لتوليد استجابات متنوعة. يهدف هذا الإجراء إلى إثراء معطيات التدريب بخصائص الانتشار الصوتي المتنوّعة ضمن البيئة الفيزيائية، مع الحفاظ على استقلال المجموعات الثلاث من حيث مواقع التسجيل داخل الغرفة، مما يمكّن النموذج من التعلّم على أنماط صدى متمايزة.
- توليد مجموعة المعطيات المشوبة بالصدى بالاعتماد على مكتبة SoundFile لقراءة ومعالجة الملفات الصوتية بصيغة WAV، وذلك من خلال تنفيذ جداء التلاف Convolution بين الإشارة الكلامية النظيفة والاستجابة النبضية وذلك بالاعتماد على خواص تحويل فورييه السريع FFT على طول الإشارة الكامل لتسريع الحسابات وتقليل التعقيد

الزمني. ونظراً لأن عدد ملفات الإشارات الكلامية النظيفة يفوق عدد ملفات الاستجابة النبضية، فقد تمّ توليد الأزواج الصوتية عبر خلط عشوائي متكافئ (Random Uniform Shuffling) ، بحيث تُربط كل إشارة كلامية نظيفة باستجابة نبضية مختلفة تُختار من المجموعة المتاحة باحتمال متقارب. يتيح هذا الإجراء استخدام جميع ملفات RIR بشكل متوازن تقريباً، مما يولّد معطيات تدريبية متنوعة تمثّل بيئات صوتية متعددة. نتج عن هذه العملية زوج من الملفات لكل عينة صوتية: أحدهما يمثل الإشارة الكلامية النظيفة، والآخر يمثل نظيرها المشوب بالصدى، مما أفضى إلى تكوين 39,600 ملف صوتي بصيغة WAV في كل من مجموعات التدريب والتحقق والاختبار.

يُظهر الشكل (6.4) إحدى الملفات الصوتية الناتجة بعد إضافة الصدى، حيث تتوضع الإشارة الكلامية النظيفة (Clean Signal)في الجزء العلوي والإشارة المتأثرة بالصدى (Reverb Signal) في الجزء السفلي، يُلاحظ اتساع الذيل الزمني للإشارة الثانية نتيجة الانعكاسات الصوتية المتعددة التي تؤدي إلى تراكب العينات وتشويه وضوح الإشارة الأصلية.



الشكل 6.4- تمثيل لإشارة كلامية نظيفة وبعد إضافة الصدى

- بعد توليد الأزواج الصوتية المكوَّنة من الإشارات الكلامية النظيفة ونظيراتها المشوبة بالصدى، تمّ حساب طيف كل ملف من ملفات الأزواج السابقة باستخدام تحويل فورييه قصير المدى، استُخدمت نافذة من نوع Hanning بطول 1024 عينة مع خطوة hop length مقدارها 256 عينة، أي يوجد تراكب بمقدار 75% بين الإطارات المتتالية لمنع ضياع المعلومات، يُخزّن لوغاريتم مطال الطيف ضمن مصفوفة ثنائية البعد.

- بعد الحصول على مطال الطيف لملفات الأزواج الصوتية، تم تقييس القيم الطيفية Normalization لتكون ذات متوسط صفري وانحراف معياري يساوي واحد وذلك لضمان استقرار عملية التدريب وتقليل التباين بين العينات.

ثم خُزّنت النتائج بشكل مناسب لتدريب الشبكة العصبونية في ملفات Checkpointsبصيغة .pt باستخدام مكتبة PyTorch.

3.4- البرمجيات والأدوات المستخدمة

تم تنفيذ جميع مراحل هذا العمل باستخدام لغة البرمجة بايثون (Python) لما تتمتع به من مرونة عالية وتكامل واسع مع مكتبات الذكاء الصنعي ومعالجة الإشارات. اعتمدت الدراسة على مكتبة PyTorch لتصميم النموذج وتدريبه، كما استُخدمت مكتبة SoundFile لقراءة وكتابة الملفات الصوتية بصيغة WAV ومعالجتها في المجال الزمني، إلى جانب مكتبة Librosa لاستخراج التمثيلات الطيفيةSpectrograms. نُفِّذت جميع التجارب ضمن بيئة Python 3.10 على معالج مزوّد بوحدة معالجة رسومية (GPU)من شركة NVIDIA مدعومة بمكتبة CUDA التي تسمح بتنفيذ جميع العمليات الحسابية على GPU بدلاً من CPU، مما أتاح تسريع العمليات الحسابية الخاصة بالشبكات العصبونية وتقصير زمن التدريب بشكل ملحوظ. أتاح هذا التكامل بين البرمجيات والمكتبات إنجاز مراحل إعداد المعطيات، وتوليد الإشارات المشوبة بالصدى، وبناء النموذج وتدريبه ضمن بيئة موحدة وفعّالة.

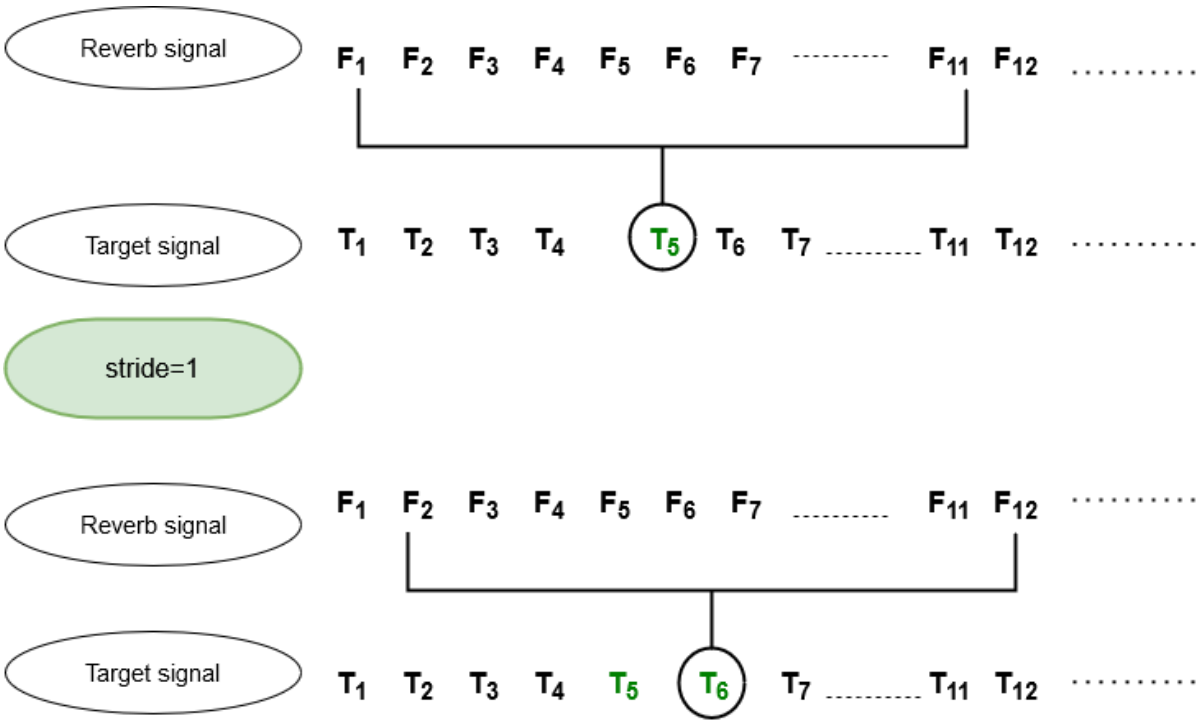
4.4- بنية النموذج

من أجل معالجة مشكلة الصدى الصوتي في الإشارات الكلامية، تم الانطلاق من فرضية مفادها أنّ التعامل الفعّال مع هذه الظاهرة يتطلّب نموذجاً قادراً على تمثيل السلاسل الزمنية الطويلة والتقاط الترابطات الديناميكية بين الإطارات المتتابعة. وفي هذا السياق، تُعدّ الشبكات العصبونية العودية RNNs من أكثر النماذج ملاءمةً لمثل هذه المشكلات نظراً لقدرتها على معالجة المعطيات المتتابعة. لذلك وقع الاختيار على نموذج [50] LSTM كنموذج أساسي (Base Model) بوصفه نموذجاً واعياً للسياق إذ يعتمد على آلية البوابات الداخلية Gating Mechanism التي تنظم تدفّق المعلومات بين الحالات الزمنية المختلفة وتتيح للنموذج الاحتفاظ بالمعلومات لفترات أطول.

1.4.4- مبدأ عمل النموذج

يقوم مبدأ عمل النموذج المقترح على معالجة الإطارات الطيفية بهدف تقدير الإطار المركزي النظيف بالاعتماد على الإطارات المجاورة له كما يوضح الشكل (7.4). يُطبَّق الحشو الصفري قبل عملية التنبؤ أثناء بناء النوافذ السياقية لضمان ثبات حجم الدخل عند أطراف الإشارة، بحيث تُعوَّض الإطارات المفقودة في بداية أو نهاية التسلسل بإطارات صفرية مكافئة . إذ يُقسَّم الطيف إلى نوافذ سياقية منزلقة Sliding Context Windows، تحتوي كل نافذة على 11 إطار طيفي متجاور يتم تمريرها إلى

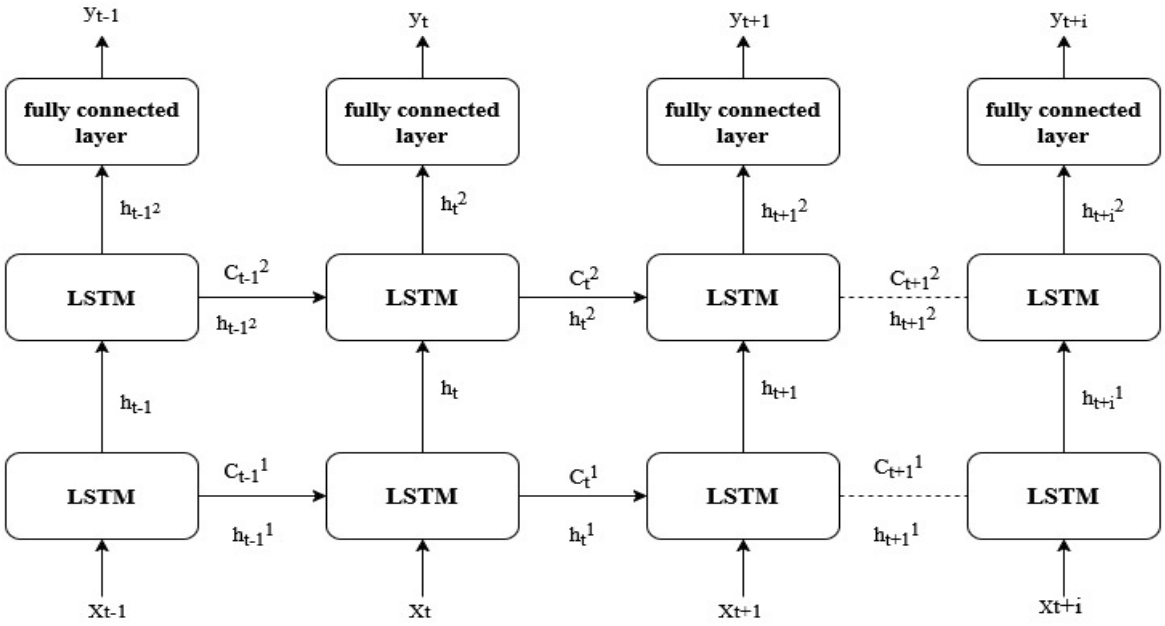
النموذج لتقدير الإطار المركزي منها. حيث تشير الرمز T إلى الإطار الهدف Target Frame بينما تمثّل F الإطارات المشوبة بالصدى Reverberant Frames.



الشكل 7.4-آلية استخراج الإطارات الهدف من الإشارات الصدى باستخدام نافذة انزلاقية بخطوة واحدة (Stride = 1)

2.4.4– بنية النموذج

يتمثل النموذج بشبكة عصبونية عودية تتألف من طبقة دخل وطبقتي LSTM و طبقة خرج Fully Connected-FC كما يظهر الشكل (8.4). إنّ دخل النموذج هو مطال طيف الإشارة بالمقياس اللوغاريتمي وتمرر بشكل سلاسل أشعة طول كل منها 512x11 وتمثل كل سلسلة نافذة سياقية تحوي 11 إطاراً طيفياً مجاوراً للإطار المستهدف، تمرّ هذه السلاسل عبر طبقتي LSTM تعملان على تحليل الترابطات الزمنية والترددية ضمن الإشارة، ثم يُسقط خرج الطبقة الثانية في فضاء خطي بعده 512 باستخدام طبقة FC التي تعيد بناء الإطار الطيفي المركزي النظيف بنفس عدد المركبات الترددية الأصلية.



الشكل 8.4-بنية النموذج

- **طبقة الدخل:**

اعتُمد لوغاريتم مطال طيف الإشارة الكلامية كدخل للنموذج، يحسب باستخدام تحول فورييه قصير الأمد. تردد التقطيع هو 16KHz والنافذة المستخدمة Hanning وطولها 64ms أي ما يعادل 1024 عينة أمّا نسبة التراكب 75% وبالتالي مقدار الإزاحة 16ms أي 256 عينة، يتم تطبيق تحويل فورييه قصير الأمد على المقطع الصوتي فيمثل الخرج ببيان ثلاثي الأبعاد، البعد الأفقي هو الزمن و البعد العمودي يمثل التردد أما البعد الثالث هو الطاقة ويعبر عنها بتدرج الألوان. طول المقطع الصوتي المستخدم هو 5s وتقابل 80000 عينة، وبالتالي يكون في حالتنا عدد الإطارات الزمنية $\frac{L-window_{length}}{hop_{size}} + 1 = \frac{80000-1024}{256} + 1 = 310$ أما عدد الترددات الموجودة في كل نافذة هو 512.

- **الطبقة الأولى LSTM**

تستقبل دخل مكوَّن من نافذة سياقية منزلقة Sliding context windows تضمّ 11 إطاراً طيفياً متجاوراً، بحيث يحتوي كل إطار على 512 مركّبة ترددية ناتجة عن تحويل STFT، تُبنى كل نافذة سياقية وفق انزياح إطار واحد (stride = 1) بحيث يكون الدخل على شكل شعاع flatten vector حول الإطار المستهدف و يمثّل سياق زمني—ترددي، ويُستخدم كدخل للشبكة. خرج هذه الطبقة 512 وحدة مخفية تمثّل الخصائص الزمنية—الطيفية المرتبطة بمكوّنات الصدى.

- **الطبقة الثانية LSTM**

تستقبل خرج الطبقة الأولى (512 وحدة مخفية) وتعمل على تعزيز فهم الترابطيات الزمنية بعيدة المدى بين الإطارات. كما تسهم في تمثيل أعمق لبنية الصدى عبر التسلسل الزمني الكامل، وتُنتج 512 وحدة مخفية تُمرَّر إلى طبقة الخرج النهائية.

- **الطبقة الثالثة FC**

طبقة خطية تتألف من 512 وحدة خرج، بحيث تمثل كل وحدة مركبة ترددية ضمن الإطار المركزي المقدّر .تهدف إلى إعادة بناء الطيف النظيف بنفس البعد الترددي الأصلي ويُعتبر ناتج هذه الطبقة هو الطيف المنقّى الذي يعبر عن الإشارة الكلامية بعد إزالة تأثيرات الصدى.

تمّ تنفيذ البنية باستخدام مكتبة PyTorch، حيث يعتمد الانتشار الأمامي (Forward Pass) على تمرير متجهات تمثيل الإطارات عبر طبقتي LSTM متتاليتين، تليهما طبقة خطية تولّد الطيف النظيف المقدر.

5.4- مرحلة اختيار السمات الطيفية

جرى تحديد السمات المدخلة إلى النموذج بطريقة تجريبية. في البداية، تم استخدام سمات Mel-Spectrogram كما هو مطروح[23]، واختُبرت من خلال إعادة بناء الإشارة الأصلية بعد تحسين طيف ميل الناتج عن النموذج. غير أنّ التقييم السمعي أظهر وجود ضجيج واضح أثر سلباً بشكل كبير على جودة الإشارة الكلامية. ولحل هذه المشكلة قمنا بتغيير السمات إلى مطال الطيف Magnitude Spectrum.

ذكرنا أنّ الملفات الصوتية في مجموعة المعطيات LibriSpeech ذات مدة زمنية مختلفة، وبالتالي بعد تطبيق تحويل فورييه على كل ملف صوتي، سينتج لدينا مصفوفة ترتبط أبعادها بطول الملف وعدد نقاط تحويل فورييه. وبالتالي استخدام تابع Resize على مطال الطيف ليقوم بضغط أو تمديد الطيف على المحورين الزمني والترددي لتوحيد أبعاده يؤدي إلى تشويه السمات وإضافة ضجيج غير طبيعي، وقد تمّ اختبار هذا الجزء عملياً بعد تدريب النموذج، ومن ثم التحقق من هذا الأثر بتطبيق الدالة على طيف ناتج لملف صوتي، و إعادة بناء الإشارة الكلامية، حيث تبين وجود ضجيج ملحوظ في الإشارة المستعادة، وبالتالي كان الحل هو توحيد طول الملفات الصوتية المستخدمة إلى 5 ثوانٍ كما هو موضح مسبقاً.

[23] https://github.com/DiegoLeon96/Neural-Speech-Dereverberation

تجدر الإشارة أنّ منحني التدريب أظهر انخفاضاً ملحوظ في جميع مراحل اختبار النموذج على سمات ميل أو حتى مع استخدام تابع لتوحيد الأبعاد إلا أنّ النتائج الإدراكية السمعية كانت غير مرضية. وهذا يدل أن جميع خيارات السمات تمت على أساس اختبارات عملية تجريبية هدفت إلى الوصول إلى التمثيل الطيفي الأكثر ملاءمة لأداء النموذج في مهمة إزالة الصدى.

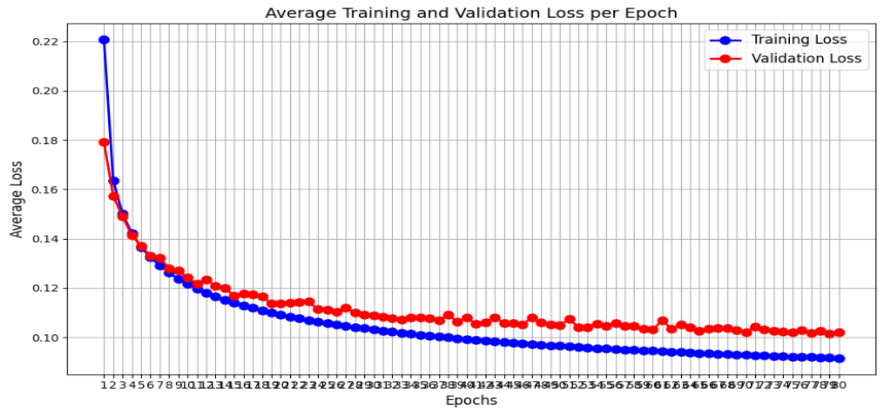
6.4- تدريب النموذج

بعد تصميم بنية النموذج المقترح واختيار السمات المناسبة، تمّ التدريب باستخدام مكتبة PyTorch وفق إعدادات تجريبية مدروسة لضمان استقرار عملية التدريب وتحقيق أداء أمثل.

استُخدمت دالة الخسارة من نوع Mean Squared Error-MSE لقياس الفرق بين الطيف الناتج عن النموذج والطيف النظيف المستهدف، تمّ اعتماد AdamW Optimizer لتحديث الأوزان، لما يتمتع به من ثبات عالٍ في التدرّج وقدرته على التعامل مع معطيات غير خطية ومعقدة، مع تعيين حجم الدفعة Batch Size إلى 32 عينة تدريبية في كل خطوة تحديث. حُدِّد معدل التعلّم الابتدائي Learning Rate بقيمة 1×10^{-4}، مع تطبيق آلية الضبط التلقائي لمعدل التعلّم Adaptive Learning Rate Scheduling عبر خوارزمية ReduceLROnPlateau، بحيث يتم تقليل معدل التعلّم إلى النصف عندما يتوقف التحقق (Validation Loss) عن التحسّن بعد مرور خمس دورات تدريبية متتالية 5 Epochs، وذلك حتى الوصول إلى الحد الأدنى لقيمة معدل التعلّم 1×10^{-6}. تجدر الإشارة أنّ اختيار موسطات النموذج Hyperparameter تمّ بشكل تجريبي.

نُفّذ التدريب حتى 80 دورة تدريبية (Epochs) كما يبين الشكل (9.4)،حيث يُلاحظ انخفاض خسارة التدريب انخفضت بشكلٍ مطّرد، مما يشير إلى قدرة النموذج على التعلّم الفعّال من المعطيات. كما يظهر انخفاض خسارة التحقق Validation Loss ما يدلّ على تعميم جيد للنموذج وعدم وجود علامات واضحة على التلبيق الزائد Overfitting حيث تم تتبّع أداء النموذج بشكل دوري لقياس جودة التعلّم، مع حفظ النماذج في نقاط مرحلية (Checkpoints) تتيح استئناف التدريب أو المقارنة بين نتائج الإصدارات المختلفة.

```
Training Summary:
Final Training Loss: 0.0915
Final Validation Loss: 0.1020
Best Training Loss: 0.0915
Best Validation Loss: 0.1013
```



الشكل 9.4- منحني تدريب النموذج

أُجريت جميع التجارب ضمن بيئة معالجة تعتمد على وحدة GPU لتسريع العمليات الحسابية وتقصير زمن التدريب، مع مراقبة منحنيات Training Loss و Validation Loss لضمان تجنّب ظاهرة التلبيق الزائد Overfitting وتحقيق أفضل أداء ممكن للنموذج.

7.4- مرحلة الاختبار

مجموعة معطيات الكلام التي استخدمت في مرحلة الاختبار باللغة العربية كون نموذج ASR-HuBERT[24] المعتمد يعمل على اللغة العربية وكما أشرنا نموذج حذف الصدى يجب أن يكون مستقلاً عن اللغة، لذلك تمّ الاختبار على Arabic Speech Corpus التي صممها الباحث السوري نوار الحلبي في جامعة ساوثمبتون مجموعة معطيات باللغة العربية ضمن إطار رسالة الدكتوراه. يبلغ حجم هذه المجموعة حوالي 1.5GB، وتتضمن تسجيلات صوتية بطول إجمالي يبلغ 3.7 ساعة، جميعها مسجلة بصوت متحدث واحد فقط. ونظراً لندرة مجموعات المعطيات الصوتية باللغة العربية، شكّلت هذه المجموعة خطوة مهمة جداً في مجال معالجة اللغة العربية، حيث تُعد مفيدة في مهام تركيب الكلام والتعرف عليه.

تمّ تقييم أداء النموذج المقترح في حالتين مختلفتين: الحالة الأولى باستخدام الإشارة المحسنة أي بعد إزالة الصدى والحالة الثانية باستخدام الإشارة الأصلية المشوبة بالصدى، وذلك بهدف تقييم فعالية إزالة الصدى على جودة الإشارة الكلامية وأداء أنظمة تعرف الكلام آلياً.

[24] https://huggingface.co/omarxadel/hubert-large-arabic-egyptian

- معيار جودة الكلام (Perceptual Evaluation of Speech Quality - PESQ)

أظهرت النتائج الواردة في الجدول (2.4) تحسّناً واضحاً في جودة الإشارات الكلامية بعد تطبيق النموذج، لا سيما في البيئات ذات الصدى العالي مثل Octagon و Great Hall مما يدل على قدرة النموذج في تقليل أثر الصدى وتعزيز وضوح الإشارة السمعية.

مجموعة المعطيات	مع صدىPESQ	بعد حذف الصدى PESQ
Great Hall omni	1.31	2.291
Class room omni	1.24	2.25
Octagon	1.49	2.56
Mardy	2.43	2.76

الجدول 2.4- قيم معيار جودة الإشارة الإدراكية (PESQ) مع صدى وبعد حذف الصدى.

- معيار معدل الخطأ في الكلمات (Word Error Rate-WER)

عند تقييم نفس النموذج باستخدام معيار معدل الخطأ في الكلمات بالتكامل مع نظام تعرّف الكلام آلياً ASR باللغة العربية، جاءت النتائج على خلاف المتوقع، إذ تبيّن أن أداء النموذج لم يكن بالمستوى المطلوب، حيث لوحظ ارتفاع في نسبة WER كما يبين الجدول (3.4):

مجموعة المعطيات	مع صدى WER	بعد حذف الصدى WER
Great Hall omni	11.27	23.24
Class room omni	16.20	30.28
Octagon	8.45	20.42
Mardy	7.04	11.97

الجدول 3.4- قيم معيار WER مع صدى وبعد حذف الصدى.

تُظهر هذه النتيجة تعارضاً واضحاً بين التحسّن المحقق في جودة الإشارة من الناحية الإدراكية PESQ وبين أداء نظام تعرف الكلام آلياً ASR، وقد يعود السبب في ذلك إلى أنّ نموذج LSTM على الرغم من نجاحه في تحسين الإدراك السمعي وجودة الكلام، إلا أنه يؤدي إلى تشويه بعض الخصائص الطيفية المهمة التي يعتمد عليها نموذج ASR في التمييز بين الكلمات والتعرف عليها بدقة، وتأتي هذه الملاحظة متوافقة مع ما أوردته بعض الدراسات السابقة في المجال، والتي أشارت إلى أنّه ليس بالضرورة أن تكون جميع نماذج تحسين جودة الكلام متوافقة مع نماذج ASR، بل إن بعضها قد يؤدي إلى نتائج عكسية من ناحية معدل الخطأ في الكلمات.

8.4- الخلاصة

يمكن القول إن النموذج المقترح حقق هدفه الأساسي في إزالة الصدى وتحسين الجودة الإدراكية، لكنه كشف أيضاً عن تحدٍّ بحثي مهم يتمثل في ضرورة مواءمة أهداف تحسين جودة الكلام مع متطلبات أنظمة التعرف الآلي على الكلام، وهو ما سيُؤخذ بالاعتبار في التصميم المستقبلي للنماذج الهجينة.

الفصل الخامس

تطوير نظام End-to-End لحذف الصدى الصوتي بالاعتماد على بنية Mamba بالتكامل مع ASR

يبدأ الفصل بعرض مجموعات المعطيات الصوتية المستخدمة في تدريب النموذج حذف الصدى الصوتي، ثم يوضح تفاصيل بنية النموذج وآلية عمله في المجال الزمني-الترددي، متبوعاً بنتائجه ثم يوّصف مرحلة التكامل مع نموذج HuBERT من أجل تطوير نظام End-to-End لتحسين أداء التعرف على الكلام .كما يتناول الفصل بيئة التنفيذ والبرمجيات المعتمدة في التدريب، ويستعرض بالتفصيل إعدادات التدريب والموسطات المستخدمة، وصولاً إلى تحليل نتائج الاختبارات.

1.5- مجموعات المعطيات المستخدمة

- مجموعة معطيات الكلام

شملت معطيات التدريب مقاطع صوتية نظيفة من متحدثين متنوعين وتغطي لغتين أساسيتين اللغة الصينية والانكليزية، بالنسبة للغة الصينية تمّ استخدام معطيات من مجموعات AISHELL-1[25] و AISHELL-2[26] و THCHS-30[27]، والتي تضم مئات المتحدثين وتغطي لهجات وسياقات مختلفة. أمّا بالنسبة للغة الإنجليزية، اعتُمدت معطيات من مجموعات EARS[28] و VCTK[29] و DNS-Challenge[30]، كما يبين الجدول (1.5):

Language	Speech dataset	Duration (hours)	#Speakers
Chinese	AISHELLI	25	240
	AISHELL II	178	987
	THCHS30	18	49
English	EARS	81	92
	VCTK	32	49
	DNS I challenge	94	1263

[25] AISHELL-1 - Linguistic Data Consortium
[26] https://www.aishelltech.com/aishell_2
[27] https://www.openslr.org/18/
[28] https://github.com/sp-uhh/ears_dataset
[29] https://service.tib.eu/ldmservice/dataset/vctk-corpus
[30] https://github.com/microsoft/DNS-Challenge

الجدول 1.5- مجموعة معطيات الكلام النظيف المستخدمة في CleanMel

تمّ اختيار 200 ساعة من مجموعات المعطيات السابقة وفق معيار الجودة DNSMOS. مما يضمن استخدام عينات صوتية عالية الجودة خالية من الضجيج والتشوه.

- مجموعة معطيات الاستجابة النبضية

تمّ استخدام مجموعة واسعة من الاستجابات النبضية من عدة مجموعات معطيات، بحيث تغطي نطاقاً واسعاً من أزمنة الصدى بين البيئات الصغيرة منخفضة الصدى والقاعات الكبيرة ذات الصدى المرتفع. يبين الجدول (2.5) تنوع هذه المجموعات من حيث عدد الاستجابات وزمن الصدى:

RIR dataset	#RIRs	RT60 range (second)
ACE	84	[0.33, 1.22]
AIR	204	[0.08, 4.50]
ARNI	1000	[0.51, 1.20]
dEchorate	594	[0.17, 0.75]
NaturalReverb	270	[0.02, 2.00]
REVERB	240	[0.25, 0.70]
RWCP	182	[0.10, 0.72]

الجدول 2.5 مجموعة معطيات RIRs المستخدمة في CleanMel

طُبقت هذه الاستجابات على 80% من معطيات الكلام لتوليد إشارات تحتوي صدى، في حين تمّ الإبقاء على 20% من المقاطع دون صدى لمحاكاة الحالات التي يكون فيها الميكروفون قريباً من مصدر الصوت. يهدف هذا التوزيع إلى تدريب النموذج على مزيج متوازن من الحالات الصوتية، مما يزيد من قدرته على التعميم وتحسين الأداء في ظروف التسجيل المختلفة.

- معطيات الضجيج

تمّ استخدام معطيات ضجيج متنوعة لتوليد إشارات واقعية تحاكي البيئات السمعية المختلفة. اعتمدت الدراسة على مصدرين رئيسيين للضجيج، إحداهما مجموعة DNS-Challenge الخاصة بـ Microsoft تحتوي على نحو 181ساعة من معطيات الضجيج المأخوذة من مجموعتي AudioSet وFreesound، وتشمل مجموعة واسعة من الأصوات البيئية مثل المحادثات الخلفية، ضجيج المكاتب والمطاعم، والأصوات الطبيعية وأصوات المواصلات. ومجموعة RealMAN و تتألف من 106ساعة من تسجيلات الضجيج البيئي الواقعي تمّ جمعها في 31 مشهد يومي، وتغطي بيئات داخلية (indoor) وخارجية (outdoor) ونصف خارجية (semi-outdoor) بالإضافة إلى مشاهد النقل والمواصلات [18] .

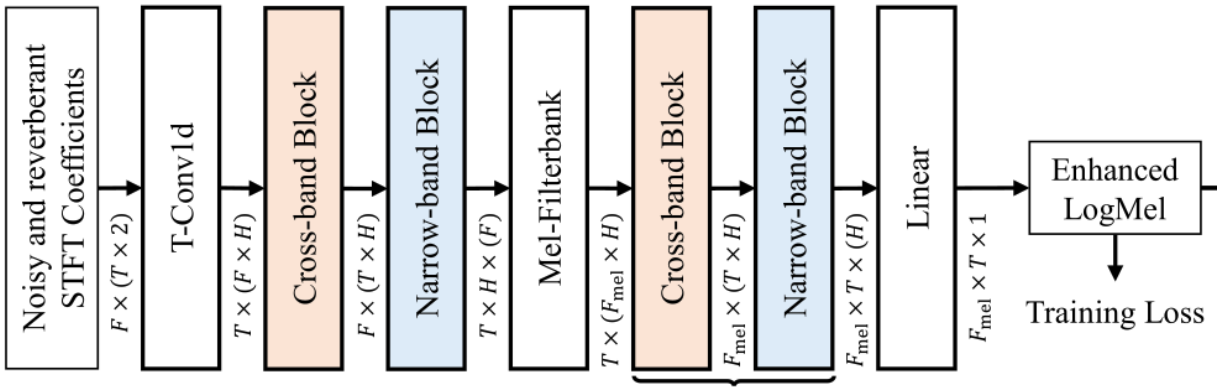
يضمن هذا التنوع الزمني والطيفي في مصادر الضجيج تغطية شاملة لأنماط الضجيج التي قد تواجه أنظمة معالجة الكلام في البيئات الواقعية، مما يعزز من قدرة النموذج على التعميم عند الاختبار في ظروف غير مرئية أثناء التدريب.

وجب التنويه أنّه تمّ تشكيل المعطيات المشوبة بالصدى بنفس الآلية التي جرت في الفصل السابق.

2.5- مبدأ عمل النموذج وبنيته

تمّ بناء النظام المطور بالاعتماد على نموذج CleanMel. يعتمد نموذج CleanMel على تحسين الإشارة الكلامية ضمن المجال الزمني–الترددي العقدي Complex Time-Frequency Domain، وذلك من خلال تطبيق تحويل فورييه قصير الأمد لتحويل الإشارة من المجال الزمني إلى المجال الترددي، ثمّ تمثيلها على شكل طيف ميل(Mel-Spectrogram) .

تتألف البنية العامة للنموذج [18] من طبقة دخل تليها كتل عريضة المجال Cross-Band Blocks، ثم كتل ضيقة المجال Narrow-Band Blocks تعمل في المجال الترددي الخطي. بعد ذلك، تُطبّق مرشحات ميل Mel Filterbanks لتحويل التمثيل الطيفي إلى مجال طيف الميل، وتُكرّر عملية المعالجة باستخدام البنية نفسها للكتل السابقة ولكن ضمن مجال ميل، لتنتهي بطبقة خرج خطية تنتج طيف الميل المحسَّن كما هو موضّح في الشكل (1.5):



الشكل 1.5- مخطط تمثيلي يوضح بنية النموذج CleanMel

1.2.5- المعالجة في المجال الترددي الخطي

- **مرحلة المعالجة المسبقة**

تبدأ عملية معالجة الإشارة الكلامية في نموذج CleanMel [18]بتحويلها من المجال الزمني إلى المجال الزمني–الترددي باستخدام تحويل STFT ، وذلك لاستخراج التمثيل الطيفي العقدي الذي يحتوي على المعلومات الزمنية والترددية معاً. تُأخذ الإشارة الكلامية بتردد تقطيع يبلغ 16KHz و يُطبّق عليها STFT باستخدام نافذة من نوع Hanning طولها 512 عينة ما يعادل

32 ms، مع خطوة إزاحة hop size قدرها 128عينة أي 8ms، مما يحقق نسبة تداخل overlap تساوي 75% بين النوافذ المتجاورة.

ينتج عن هذه العملية مخطط طيفي ثلاثي الأبعاد Spectrogram، يُمثَّل على شكل مصفوفة أبعادها $M \times N$ حيث M يمثل عدد النوافذ الزمنية المتراكبة وبما أن طول المقطع الصوتي المستخدم 4sec وهو ما يعادل 64000 عينة، فيكون عدد النوافذ (الأطر) $\frac{L-window_{length}}{hop_{size}} + 1 = \frac{64000-512}{128} + 1 = 497$، أما N فهو عدد الترددات المفيدة في كل نافذة هو 256 يمثل كل منها بقيمة عقدية تُقسم إلى جزأين حقيقي وتخيلي لتشكل معاً دخل الشبكة العصبونية في المراحل اللاحقة من النموذج.

- **طبقة الدخل**

تتكوّن طبقة الدخل من شبكة تلافيفية أحادية البعد 1D Temporal Convolution تُطبّق على المحور الزمني باستخدام نواة (Kernel) بحجم 5 أي يغطي خمس إطارات زمنية متجاورة ضمن كل تردد، تتمثل الوظيفة الأساسية لهذه الطبقة في استخلاص السمات المحلية من التمثيل الطيفي العقدي للإشارة الكلامية، بحيث تلتقط الترابطات الزمنية القصيرة المدى بين الإطارات المتجاورة.

ينتج عن هذه الطبقة تمثيل موسّع الأبعاد بالصيغة:

$$F \times T \times H \qquad (1.5)$$

حيث F تمثل عدد الترددات، T عدد الإطارات الزمنية.

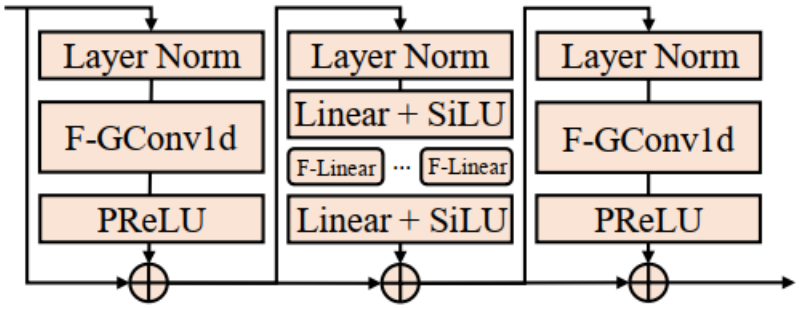
أما $H = 144$ فتمثل عدد السمات المضمّنة (Embedded Features) المستخلصة من كل تردد عبر خمس إطارات زمنية متجاورة.

يمرر خرج طبقة الدخل إلى مرحلة المعالجة في المجال الترددي الخطي والتي تتألف من نوعين من الكتل:

- **الكتل العريضة النطاق Cross-band block**

تختص بتعلم النمط الطيفي الكامل عبر جميع الترددات full-band spectral pattern ضمن إطار زمني واحد، حيث تتم معالجة كل إطار بشكل مستقل باستخدام ذات البنية لجميع الإطارات.

تعتمد بنية هذه الكتل على الشبكات التلافيفية الأحادية البعد، التي تُطبق على المحور الترددي بهدف التقاط الارتباطات الطيفية بين الترددات المختلفة داخل الإطار الواحد. كما يبيّن الشكل (2.5):



الشكل 2.5- بنية الكتلة عريضة النطاق.

تتكوّن كل كتلة عريضة المجال من سلسلة من الطبقات الفرعية تشمل:

- طبقة استنظام (Layer Normalization) تُستخدم لتثبيت التوزيع الإحصائي للسمات وتحسين استقرار عملية التدريب.
- طبقة تلافيفية Frequency-Grouped Conv1d(F-GConv1d) تعمل على تحليل الترابطات الطيفية على المحور الترددي باستخدام نواة صغيرة الحجم (kernel = 5) مع تطبيق تقنية Grouped Convolution(groups=8) التي تتيح تقسيم الترددات إلى مجموعات فرعية تُعالَج بشكل مستقل. تسهم هذه التقنية في تقليل عدد الموسطات والكلفة الحسابية مع الحفاظ على القدرة على التقاط الأنماط المحلية داخل كل مجموعة من الترددات المتجاورة، مما يمكّن النموذج من تمثيل البنية الطيفية الدقيقة للإشارة داخل كل إطار زمني.
- تابع تفعيل غير خطي من نوع PReLU (Parametric ReLU- PReLU لتعزيز التمثيل الطيفي.
- مسار اتصال مباشر Residual Connection يربط الدخل بالخرج للحفاظ على تدفق المعلومات وتقليل خسارة السمات عبر الطبقات.
- بالإضافة إلى ذلك، تتضمن الكتلة وحدة Full-band LinearGroup في المنتصف، تعمل على ربط جميع الترددات معاً لتعلّم الترابطات الكاملة للحزمة الترددية (Full-band dependencies) و تعتمد على تابع Sigmoid Linear Unit- SiLU.

تُنتج هذه الكتل تمثيلاً طيفياً بنفس أبعاد الدخل $F \times T \times H$ وبذلك تمكّن النموذج من تعلم الترابطات بين النطاقات الترددية بشكل فعّال.

- **الكتل ضيقة النطاق narrow-band Block**

بعد مرحلة المعالجة عريضة المجال التي تركز على تعلم الترابطات الطيفية بين الترددات ضمن الإطار الواحد، تنتقل المعالجة إلى الكتل ضيّقة النطاق التي تُعنى بنمذجة العلاقات الزمنية لكل تردد على حدة. تكمن الفكرة الأساسية لهذه الكتل في معالجة كل تردد بشكل مستقل، من خلال تحليل الإشارة الكلامية إلى مجموعة من التلافيف ضيّقة النطاق المستقلة ترددياً (Frequency-Independently Narrow-Band Convolutions) كما توضحه العلاقة (2.5) مما يساعد في استخلاص المعلومات الضرورية لإزالة الصدى وأيضاً يساهم في إزالة الضجيج.

$$Y(f,t) \approx S(f,t) * A(f,t) + E(f,t) \qquad (2.5)$$

حيث:

$Y(f,t)$ طيف الإشارة المشوبة بالصدى.

$S(f,t)$ طيف الإشارة النظيف.

$A(f,t)$ تمثّل طيف استجابة الغرفة النبضية.

$E(f,t)$ مركبة الضجيج.

سُمِّيت هذه الكتل "ضيّقة النطاق" لأنها تعمل على معالجة كل تردد على حدة عبر الزمن، مع مشاركة نفس بنية الشبكة العصبونية لجميع الترددات، مما يقلل التعقيد الحسابي ويضمن ثبات المعالجة عبر النطاقات الترددية. وتعتمد البنية الداخلية لهذه الكتل على نموذج Mamba صُمِّم لالتقاط الترابطات الزمنية القصيرة والطويلة المدى.

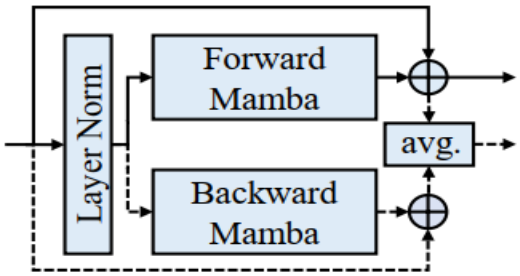
تستند هذه المقاربة إلى فرضية أن خصائص الصدى في الغرفة قد تبقى ثابتة لفترات زمنية طويلة، ما يتطلب آلية قادرة على تتبّع العلاقات بعيدة المدى، أو قد تتغير بسرعة نتيجة تغيرات في البيئة أو في موضع المتحدث، وهو ما يستلزم أيضًا القدرة على التقاط الترابطات قصيرة المدى.

تتضمن كل كتلة ضيّقة النطاق بالإضافة إلى طبقة التقييس مسارين متوازيين للمعالجة، كما يُظهر الشكل (3.5):

- مسار أمامي (Forward Mamba) يلتقط تطور الإشارة عبر الزمن في الاتجاه الأمامي.

- مسار خلفي (Backward Mamba) يعالج الإشارة في الاتجاه المعاكس.

ثم تُدمج مخرجات المسارين بواسطة متوسطهما الحسابي للحصول على تمثيل طيفي أكثر توازناً وثباتاً، مما يعزز جودة التمثيل النهائي للإشارة ويزيد من قدرة النموذج على إزالة الصدى والضجيج معاً بكفاءة عالية.



الشكل 3.5- بنية الكتلة ضيّقة النطاق

يساعد الجمع بين الكتل عريضة المجال وضيّقة النطاق في بناء نموذج ثنائي المرحلة (Two-stage Model) قادر على معالجة العلاقات الطيفية–الزمنية بشكل متكامل، حيث تتولى الكتل الأولى اكتشاف الأنماط عبر الترددات، في حين تلتقط الثانية التغيرات عبر الزمن داخل كل تردد، مما يؤدي إلى تحسين فعالية إزالة الصدى وتعزيز جودة الإشارة الناتجة.

2.2.5– المعالجة في مجال Mel

بعد الانتهاء من المعالجة عبر طبقة واحدة للنموذج في المجال الترددي الخطي ضمن الكتل عريضة وضيّقة النطاق، يتم تحويل الإشارة المحسّنة إلى مجال طيف ميل من خلال تطبيق مرشّحات ميل Mel Filterbank على الطيف الناتج . يُطبق مجموعة من المرشحات الترددية مكونة من 80 مرشح مثلثي تغطي المجال الترددي من 0 إلى 8 كيلوهرتز، يؤدي هذا التحويل إلى تقليل عدد الموسطات الترددية من 257 تردد إلى 80 تردد في مجال ميل، مما يقلل التعقيد الحسابي ويُحافظ في الوقت ذاته على المعلومات الطيفية الأكثر أهمية إدراكياً، نظراً لأن مقياس Mel مصمم لمحاكاة طريقة الأذن البشرية في إدراك الترددات.

وتكرر عملية المعالجة في تمثيل ميل باستخدام نفس البنية الموضحة مسبقاً من كتل Cross-band وNarrow-band، حيث يتكون هذا الجزء من 15 طبقة، ثم طبقة خرج خطية للحصول على طيف ميل المحسّن.

3.5– البرمجيات المستخدمة

يتطلب تشغيل مكتبة Mamba بيئة عمل قائمة على نظام التشغيل Linux لمرونته ودعمه الواسع لأدوات التعلم العميق، إلى جانب وحدة معالجة رسومية من نوع NVIDIA GPU لضمان الأداء العالي وتسريع العمليات الحسابية من خلال دعم مكتبة CUDA بإصدار 11.6 أو أحدث ، استخدمنا الإصدار 12.1. كما يعتمد النموذج بشكل أساسي على مكتبة التعلم

العميق PyTorch بإصدار 1.12 أو أحدث، استخدمنا الإصدار 2.2.0 لتنفيذ خوارزميات التدريب والاستدلال بفعالية. تمّ استخدام مكتبة mamba_ssm بإصدار 1.2.0.post1 الخاصة بنموذج Mamba . كما استُخدمت مكتبات إضافية لمعالجة الإشارة الكلامية مثل Librosa و SoundFile لاستخراج السمات الطيفية وقراءة الملفات الصوتية ومعالجتها ضمن المجال الزمني والترددي. نُفذت جميع التجارب ضمن بيئة Python 3.10 على نظام Linux مزوّد بوحدة معالجة رسومية NVIDIA GPU.

4.5- مرحلة التدريب

هدف النموذج هو تعلم قناع Mel Ratio Mask يطبق على الطيف المشوب بالصدى لإنتاج الطيف النظيف وفق العلاقة التالية:

$$\widehat{M}(f_{mel}, t) = \min\left(\sqrt{\frac{X_{mel}(f_{mel}, t)}{Y_{mel}(f_{mel}, t)}}, 1\right) \qquad (3.5)$$

حيث X_{mel} يمثل طيف ميل للإشارة النظيفة(الهدف)، Y_{mel} طيف ميل لإشارة الصدى الداخلة إلى النموذج

يُدرَّب النموذج على تقدير هذا القناع باستخدام متوسط الخطأ التربيعي (MSE) بين القناع الحقيقي والقناع الذي يولده النموذج، وذلك كهدف تدريبي مباشر يوجّه عملية التعلم. بعد التنبؤ بالقناع، يُعاد بناء Mel-spectrogram المحسن عبر ضرب مربع القناع بطيف ميل لإشارة الصدى ثم يُطبّق التحويل اللوغاريتمي وفق العلاقة:

$$\hat{X}_{logmel}(f_{mel}, t) = \log\left(\max\left\{\widehat{M}(f_{mel}, t)^2\, Y_{mel}(f_{mel}, t), \epsilon\right\}\right) \qquad (4.5)$$

حيث $\epsilon = 10^{-5}$ تُستخدم لتجنب القيم الصفرية أثناء التحويل اللوغاريتمي.

والخرج النهائي للنموذج هو طيف Mel اللوغاريتمي المحسَّن، الذي يتكامل مباشرةً مع نموذج Vocos[31]، وهو مولِّد صوتي عصبوني سريع (Fast Neural Vocoder) صُمّم لتوليد الموجات الصوتية من السمات الصوتية (Acoustic Features) المحسّنة.يعتمد Vocos في تدريبه على خوارزمية الشبكات التوليدية التنافسيةGenerative Adversarial Network-GAN، ويقوم بإعادة بناء الإشارة الكلامية بحيث تتم مطابقة الخصائص الطيفية بين ناتج نموذج CleanMel ومدخلات Vocos. يضمن هذا التكامل أن الطيف المحسَّن الناتج عن CleanMel-L-mask يمكن استخدامه مباشرةً كدخل لـ Vocos لإنتاج إشارة كلامية محسَّنة ذات جودة إدراكية عالية ووضوح طبيعي في المجال الزمني.

[31] https://github.com/gemelo-ai/vocos

1.4.5– موسطات النموذج

يعتمد نموذج CleanMel-L-Mask على بنية متعددة الطبقات تتألف من 16طبقة رئيسية من نوع CleanMelLayer، تدمج بين الكتل عريضة النطاق والكتل ضيقة النطاق وقد تمّ تحديد الموسطات البنيوية كما يلي :[18]

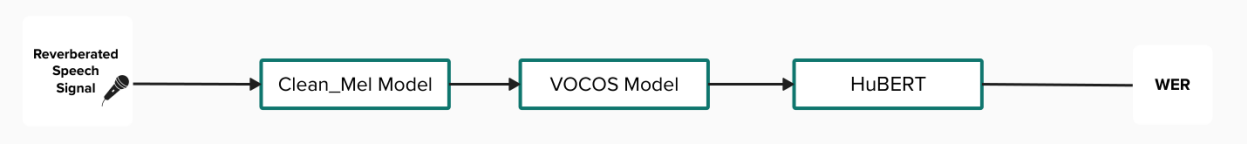
- عدد الطبقات الكلية 16 (طبقة واحدة في المجال الترددي الخطي و15 طبقة في مجال طيف ميل)
- عدد طبقات Mamba: 32 (طبقتين أمامية وخلفية ضمن كل كتلة ضيقة النطاق).
- Hidden dimension =144
- عدد الترددات في المجال الخطي :256
- عدد ترددات طيف ميل :80
- التلافيف الترددية (F-GConv1d) : حجم القناع kernel size = 5 و عدد المجموعات (groups) = 8، يوجد 32 طبقة F-GConv1d (طبقتين ضمن كل كتلة عريضة النطاق).
- 64 طبقة تقييس (3 طبقات في كل كتلة عريضة النطاق وطبقة واحدة في كل كتلة ضيقة النطاق).

2.4.5– موسطات التدريب

- تردد التقطيع 16KHz
- حجم الدفعة Batch size= 32
- عدد التكرارات Epochs =150
- خوارزمية التعلم ADAMW
- معدل التعلم يبدأ بـ 0.001 ويتناقص تدريجياً وفق العلاقة $lr = 0.001 \times 0.99^{epoch}$

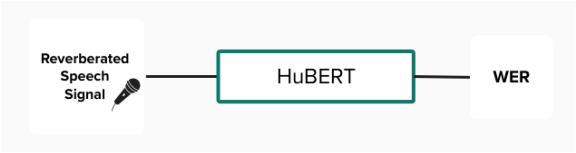
5.5– مرحلة الاختبار

تمّ تقييم أداء النموذج المقترح في حالتين مختلفتين: الحالة الأولى باستخدام الإشارة المحسنة أي بعد إزالة الصدى كما يُبين الشكل (4.5):



الشكل 4.5- مخطط تمثيلي يوضح آلية اختبار أداء النموذج بعد حذف الصدى.

والحالة الثانية باستخدام الإشارة المشوبة بالصدى، كما يُظهر الشكل (5.5).



الشكل 5.5- مخطط تمثيلي يوضح آلية اختبار أداء النموذج قبل حذف الصدى.

مجموعة الكلام: تم استخدام Arabic Speech Corpus المذكورة في الفصل السابق كمجموعة معطيات كلامية لاختبار النموذج في اللغة العربية.

مجموعة معطيات الصدى:

لتحليل قدرة النموذج على التعميم عبر بيئات صوتية مختلفة، تمّ اختبار أدائه على مجموعتين من معطيات الصدى الحقيقة RIRs، مع التنويه إلى أنّ مجموعتي معطيات الاختبار المستخدمة حالة CleanMel تمثل بيئات صوتية أعقد وأكثر تبايناً من مجموعات الاختبار الخاصة بنموذج LSTM.

- ✓ **المجموعة الأولى:** تمّ تجميعها في [51] وتُعدّ من المراجع المعتمدة في تقييم أداء أنظمة ASR في بيئات صوتية حقيقية ومتنوعة. تتكوّن هذه المجموعة من 325 استجابة نبضية حقيقية جُمعت من ثلاث مجموعات معطيات مختلفة هي:

  - RWCP Sound Scene Database
  - REVERB Challenge Database
  - Aachen Impulse Response Database (AIR)

- **نتائج الأداء على المجموعة الأولى:**

المعيار	مع صدى	بعد حذف الصدى
PESQ	2.125	2.680
WER	33	21

الجدول 3.5- نتائج اختبار نموذج CleanMel وفق معياري PESQ,WER على المجموعة الأولى.

✓ **المجموعة الثانية:** التي تمّ اختبار أداء النموذج عليها هي مجموع MIT IR Survey، تحتوي على 271 استجابة نبضية حقيقية تمّ تسجيلها في مواقع حقيقية ومتنوعة ضمن ولاية Massachusetts في (الولايات المتحدة الأمريكية)، مثل القاعات، الغرف، الممرات، المطاعم، والمساحات المفتوحة، كما ورد في [52] .

- **نتائج الأداء على المجموعة الثانية:**

المعيار	مع صدى	بعد حذف الصدى
PESQ	2.449	2.954
WER	72	31

الجدول 4.5- نتائج اختبار نموذج CleanMel وفق معياري PESQ,WER على المجموعة الثانية.

بالنظر إلى النتائج السابقة على مجموعتي معطيات الصدى، يُلاحظ أن النموذج حقق تحسناً واضحاً في جودة الإشارة الكلامية مقروناً بانخفاض كبير في معدل الخطأ في الكلمات، مما يعكس كفاءته في معالجة الصدى عبر بيئات صوتية متنوعة. كما تُظهر الفروقات بين المجموعات أن أداء النموذج يتأثر بطبيعة الصدى وخصائص البيئة الصوتية. وبذلك يمكن القول إن النموذج لم يقتصر على تحسين الجودة الإدراكية (PESQ) فحسب، بل ساهم أيضاً في تعزيز دقة أنظمة تعرف الكلام آلياً (ASR) في ظروف صوتية واقعية ومتنوعة.

كما قمنا بتقييم أداء النموذج على مجموعة المعطيات Arabic Speech Corpus خالية من الصدى والضجيج وذلك للتحقق مما إذا كان النموذج يؤثر سلباً على خصائص الإشارة الكلامية في حال عدم وجود صدى وكانت النتائج كالتالي:

المعيار	معطيات نظيفة	بعد معالجتها عبر النموذج
PESQ	4.644	3.435
WER	6	11

الجدول 5.5- نتائج اختبار نموذج CleanMel وفق معياري PESQ,WER على معطيات نظيفة.

تُظهر هذه النتائج أنّه على الرغم من فعالية النموذج الواضحة في حالات وجود الصدى، إلا أنّ تطبيقه على معطيات نظيفة أدّى إلى تراجع في جودة الإشارة وفق معيار PESQ وارتفاع طفيف في معدل الخطأ في الكلمات WER.

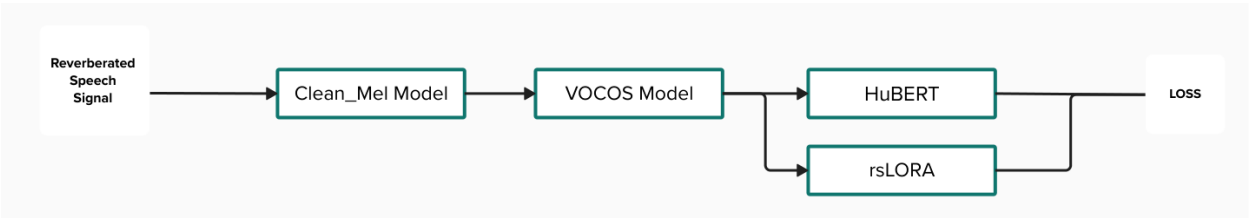
يُعزى هذا التراجع في الأداء عند تطبيق النموذج على معطيات نظيفة إلى أنّه تعلّم إزالة المكوّنات الطيفية المرتبطة بالصدى، وعند إدخال إشارة خالية من الصدى، يقوم النموذج بحكم آلية عمله بمحاولة تصحيح طيف لا يحتوي في الأصل على صدى، وبالتالي أثناء عملية تقدير القناع Mel Ratio Mask التي يعتمد عليها النموذج قد تؤدي إلى تخميد غير مقصود لبعض المكونات الترددية في الإشارات النظيفة، إذ يعامل النموذج بعض التباينات الطبيعية في الطاقة الطيفية على أنها آثار صدى. ونتيجة لذلك، تنخفض جودة الصوت الإدراكية (PESQ) وتزداد نسبة الأخطاء في التعرف WER.

استناداً إلى النتائج التجريبية السابقة، أظهر نموذج CleanMel-L-mask أداء مقبولاً في إزالة الصدى من الإشارات الكلامية مع الحفاظ على وضوحها السمعي وجودتها الإدراكية مقارنةً بالنموذج السابق LSTM، فعلى الرغم من أنّ نموذج LSTM قد حقق تحسّناً مقبولاً في جودة الإشارة، إلا أنّ تقييمه ضمن أنظمة تعرّف الكلام آلياً أظهر ارتفاع ملحوظ في معدل الخطأ في الكلمات بعد إزالة الصدى، لذلك، تمّ اختيار CleanMel كأساس للنهج المقترح اللاحق، نظراً لفعاليته في إزالة الصدى دون الخلل بالبنية الصوتية الجوهرية للإشارة الكلامية، مما يجعله الأنسب للدمج مع نموذج HuBERT في الإطار المتكامل المقترح من طرف إلى طرف (End-to-End).

6.5- بنية النموذج المقترح end to end

إنّ إطار العمل المطروح يعتمد على دراسة فعالية النُهج التكاملية من طرف إلى طرف (End-to-End) في تجاوز مفارقة إزالة الصدى وتعرف الكلام آلياً، من خلال مواءمة أهداف تحسين الإشارة مع الحصول على دقة تعرّف جيدة.

يبين الشكل (6.5) هيكلية معالجة شاملة من طرف إلى طرف (End-to-End) للنظام المقترح والذي يتألف من ثلاث مراحل رئيسية مترابطة:



الشكل 6.5- مخطط تمثيلي يوضح هيكلية النظام المقترح.

تبدأ العملية بإدخال الإشارة الكلامية المشوبة بالصدى إلى نموذج CleanMel، الذي يقوم بتحليل طيف الإشارة الكلامية وإزالة مكونات الصدى في المجال الزمني—الترددي، مما ينتج طيف ميل (Mel-Spectrogram) محسّن يتميز بنقاء طيفي أعلى. بعد ذلك، تُمرّر المخرجات إلى نموذج Vocos، لنحصل على إشارة كلامية مُعاد تكوينها خالية من الصدى في المجال الزمني.

ثمّ تُغذّى الإشارة المحسّنة إلى نموذج HuBERT المعزز بوحدات rsLoRA، الذي يتولّى تحويل الإشارة الكلامية إلى تمثيل نصي.

وأخيراً، يتم تقييم أداء النظام ككل عبر مقارنة النص الناتج بالنص المرجعي وحساب دالة الخسارة المناسبة، مما يسمح بتقدير فعالية النظام في إزالة الصدى وتحسين دقة تعرّف الكلام آلياً في بيئات صوتية حقيقية.

1.6.5– تهيئة النظام المقترح للتدريب

تمّ اعتماد نموذج تعرف الكلام آلياً HuBERT المتاح عبر منصة Hugging Face[32] ، وهو نموذج مُدرّب مسبقاً على معطيات كلامية ضخمة باللغة العربية، ونظراً للقيود الحسابية المرتبطة بإعادة تدريب نموذج ضخم من الصفر، تمّ اعتماد تقنية Low-Rank Adaptation (LoRA)، وبشكل خاص إصدارها المحسّن rsLoRA (Rank-stabilized LoRA)، التي تتيح إجراء Fine-tuning جزئي للنموذج من خلال إدخال مصفوفات منخفضة الرتبة إلى الأوزان الأصلية، مما يقلل عدد الموسطات القابلة للتحديث مع الحفاظ على كفاءة التعميم ودقة النتائج.

أُجريت تهيئة النظام وفق ما يلي:

- تمّ تجميد أوزان نموذجي CleanMel و Vocosبالكامل (Freeze Parameters) لمنع تحديثها أثناء التدريب، إذ يُستخدَمان فقط لإنتاج الإشارة الكلامية المحسّنة.
- تمّ تجميد أوزان HuBERT باستثناء الطبقات المخصصة لتطبيق rsLoRA ، وذلك بهدف توجيه عملية التعلم نحو تمثيلات صوتية تتلاءم مع المجال الجديد دون المساس بالمعرفة السابقة المكتسبة أثناء التدريب المسبق.
- أُدرجت وحدات rsLoRA على مستوى موسطات الانتباه الذاتي (Self-Attention)، وتحديداً ضمن مصفوفات الاستعلام (Query – Q)، والمفتاح (Key – K)، والقيمة (Value–V) ، وذلك بما يتوافق مع التوصيات الواردة في الورقة البحثية المرجعية الخاصة بـrsLoRA [44].

2.6.5– مجموعات معطيات التدريب

مجموعة معطيات اللغة العربية

نظراً للحاجة إلى مجموعة معطيات ضخمة باللغة العربية تحتوي على ملفات صوتية مصحوبة بالنصوص المقابلة لها، تمّ استخدام مجموعة Common Voice Corpus 14.0[33] الصادرة عن Mozilla، وهي من أكبر مجموعات المعطيات مفتوحة المصدر

[32] https://huggingface.co/omarxadel/hubert-large-arabic-egyptian
[33] https://commonvoice.mozilla.org/en/datasets

لتعلّم النماذج في اللغات الطبيعية. تتضمّن المجموعة ثلاث تقسيمات رئيسية :معطيات التدريب، التحقق، والاختبار، وفقاً للتوزيع التالي:

نوع المعطيات	عدد الملفات	المدة الزمنية (ساعة)
التدريب	28444	32.49
التحقق	10242	12.53
الاختبار	10463	12.63

الجدول 6.5- توزيع مجموعة معطيات اللغة العربية Common Voice Corpus 14.0

جميع الملفات محفوظة بصيغة MP3. بمعدل أخذ عينات 16KHz، تتميّز هذه المجموعة بتنوّع المتحدثين من حيث الجنس (ذكور وإناث) والفئة العمرية مما يعزز قدرة النموذج على التعميم والتعامل مع الأصوات المتنوعة.

قبل استخدام المعطيات النصية في مرحلة التدريب، تمّ تنفيذ عملية معالجة مسبقة للنصوص المرافقة للملفات الصوتية آلياً واختبارها على عدة جمل متنوعة، وذلك لضمان توافقها مع نموذج التعرّف الكلامي. تضمنت خطوات المعالجة ما يلي:

- إزالة الرموز والعناوين الإلكترونية مثل الروابط والبريد الإلكتروني.
- حذف علامات الترقيم والأرقام سواء المكتوبة بالعربية أو بالإنجليزية.
- توحيد أشكال الحروف العربية مثل استبدال (أ، إ، آ) بـ (ا)، وتحويل (ة) إلى (هـ)، و(ى، ئ) إلى (ي).
- إزالة جميع رموز التشكيل بما في ذلك الحركات (الفتحة، الضمة، الكسرة، التنوين، السكون، الشدة، إلخ)
- إزالة الأحرف اللاتينية والرموز غير العربية.
- حذف التكرارات الطويلة للأحرف غير الطبيعية (مثل "جمييييل" ← "جميل").
- توحيد المسافات البيضاء وحذف الفراغات الزائدة في بداية ونهاية الجمل.

تمّ تطبيق هذه الخطوات على جميع الجمل النصية المرافقة للمعطيات الصوتية في مجموعات التدريب والتحقق والاختبار، لضمان أن النموذج يتعلّم من معطيات لغوية نقية ومتناسقة. وقد أظهرت هذه العملية تحسّن واضح في استقرار التدريب ودقة التعرّف بفضل إزالة التباينات بين النص والملف الصوتي.

مجموعة معطيات الاستجابة النبضية

تمّ استخدام عدة مجموعات معطيات لاستجابات الغرف النبضية (RIRs) بهدف تمثيل بيئات صوتية متنوعة من حيث الحجم وشدة الصدى، بما يشمل الغرف الصغيرة والمتوسطة والقاعات الواسعة. شملت المجموعات المعتمدة في هذا النهج كلاً من:

CreatHall, ClassRoom, Octagon, MARDY, MIT IR,AIR RIR, Real RIR بالإضافة إلى Simulated RIR [51] لنمذجة استجابات افتراضية تحاكي ظروف صوتية يصعب قياسها ميدانياً لغرف صغيرة وكبيرة ومتوسطة الحجم.

- تمّ تقسيم معطيات الاستجابات النبضية وفق النسب التالية: 80% للتدريب و 10% للتحقق و 10% للاختبار.

وقد تمّ الحرص على أن تشمل كل مجموعة من هذه الأقسام أنواعاً مختلفة من البيئات الصوتية لضمان قدرة النموذج على التعميم عند التعامل مع إشارات جديدة غير مألوفة.

3.6.5– موسطات التدريب

- اعتمد التدريب على خوارزمية أمثلة AdamW Optimizer، مع ضبط معدل التعلّم الابتدائي 2×10^{-4} وتطبيق آلية جدولة تكيّفية لمعدل التعلّم تعتمد على خوارزمية Cosine Learning Rate Scheduler (تمّ اختيارها بشكل تجريبي) التي تقوم بتعديل معدل التعلّم وفق منحنى جيبي Cosine Function بحيث يُمنح النموذج في المراحل الأولى قيم مرتفعة لمعدل التعلّم لتسريع عملية التقارب نحو الحلول المثلى، ثم ينخفض تدريجياً مع تقدم الدورات التدريبية (Epochs)لضمان استقرار عملية التعلم مع مرور الزمن.

- استُخدمت خوارزمية Connectionist Temporal Classification (CTC) Loss[34] كدالة خسارة رئيسية في تدريب النظام، إذ تسمح بمطابقة النص الناتج مع الإشارة الكلامية المتغيرة الطول دون الحاجة إلى محاذاة زمنية مباشرة بين الإشارة الكلامية والنص المقابل لها، تُشير المحاذاة الزمنية (Temporal Alignment) إلى تحديد الإطارات الزمنية الدقيقة في الإشارة الكلامية التي يقابل كل منها رمز معين في النص (كحرف أو كلمة). إلا أنّ هذا التحديد عادة غير متاح في مهام تعرّف الكلام آلياً، لأن التسجيلات لا تحتوي على معلومات زمنية دقيقة تحدد مواضع الحروف داخل الإشارة. لذلك، تعتمد خوارزمية CTC على حساب جميع الاحتمالات الممكنة لمحاذاة الإشارة بالنص الصحيح، ثم تجمع هذه الاحتمالات لتقدير احتمال التوليد الصحيح للتسلسل النصي.

- بلغ عدد دورات التدريب عشرين دورة (20 Epochs)، مع تعيين حجم الدفعة التدريبية Batch Size=32 كما تمّ اعتماد استراتيجية Fine-Tuning تدريجية، حيث تُحدث فقط أوزان وحدات rsLoRA المضافة، بينما تبقى

[34] https://www.geeksforgeeks.org/nlp/connectionist-temporal-classification/

بقية طبقات نموذج HuBERT مجمّدة، مما يضمن استغلال المعرفة السابقة للنموذج وتقليل مخاطر التلبيق الزائد (Overfitting). جرى تهيئة وحدات rsLoRA وفق الإعدادات التالية:

- رتبة المصفوفة (Rank r = 32)
- معامل القياس ($\alpha = 128$)
- نسبة الإسقاط (Dropout = 0.1)

حيث بلغ عدد الموسطات القابلة للتدريب 4 ,718 ,592 موسط من أصل 320,203,437 موسط كلي في نموذج HuBERT، أي ما يعادل نسبة 1.47 % فقط من إجمالي موسطات النموذج.

وجب التنويه أنّ اختيار موسطات rsLORA ومعدل التعلم لم يكن عشوائياً أو قائماً على تجارب تدريب متكررة مكلفة زمنياً، بل استُخدم إطار عمل Optuna-A HyperParameter Optimization framework (HPO) يعتمد على منهجية بحث ذكية بدلاً من التدريب الكامل لكل مجموعة من القيم المحتملة .

يقوم Optuna بتدريب النظام المقترح على عيّنة محدودة من معطيات التدريب وتقييم أدائه على عيّنة من معطيات التحقق في كل تجربة (Trial) باستخدام معيار WER، وتتمثّل مهمته في اختيار التوليفة المثلى من القيم التي تحقق أقل معدل خطأ ممكن.

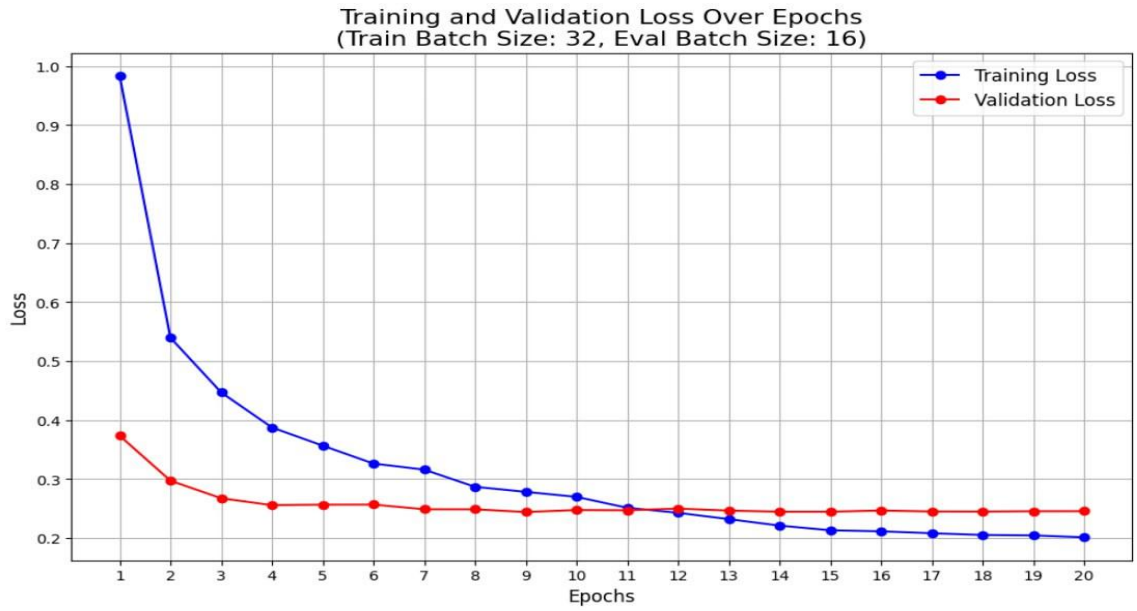
في هذا الإطار، تمّ تحديد فضاء بحث للموسطات الأساسية شمل القيم التالية:

- رتبة المصفوفة $r \in \{8,16,32\}$
- معامل القياس $\alpha \in \{16، 32، 64، 128\}$
- معدل التعلّم (Learning Rate) ضمن المجال $[1 \times 10^{-4}, 3 \times 10^{-4}]$

استُخدمت عيّنة من 1000 مثال تدريب و 400 مثال للتحقق و 20 دورة تدريبية في كل تجربة بهدف تسريع عملية البحث، بينما ظلّ التدريب النهائي على مجموعات المعطيات الكاملة.

أسفرت النتائج عن التوليفة المثلى التالية: r = 32 مع $\alpha = 128$ ومعدل تعلم 2×10^{-4} وتحقق عندها WER=0.3153 تمّ اعتماد هذه الإعدادات لتدريب النموذج الكامل.

بعد ضبط موسطات التدريب تمّ تقييم الأداء بشكل دوري بعد كل دورة تدريبية (Epoch) باستخدام معيار Word Error Rate (WER)، والذي يمثّل النسبة المئوية للأخطاء في التعرف على الكلمات مقارنة بالنص المرجعي. وقد تمّ اعتماد مبدأ حفظ أفضل نموذج عند الوصول إلى أدنى قيمة لـ WER على مجموعة التحقق (Validation Set) . لضمان اختيار النسخة المثلى من النموذج. وبعد تحليل منحنيات الأداء الموضحة في الشكل (7.5)، تبيّن أن أفضل أداء تحقق عند الدورة التدريبية رقم 17، حيث استقر كلٌّ من Training Loss وValidation Loss عند مستويات منخفضة وثابتة ولذلك تمّ اعتماد النموذج المحفوظ (Checkpoint 17) لإجراء مرحلة الاختبار النهائي على مجموعات المعطيات المختلفة.



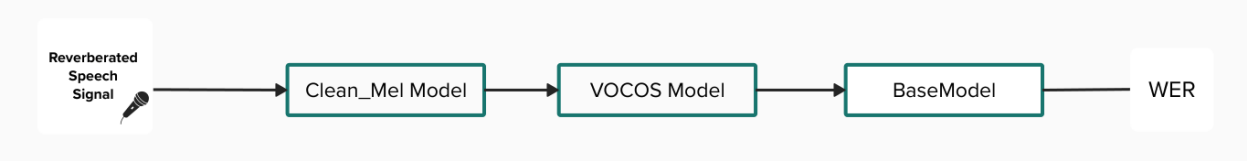
الشكل 7.5- منحنيي الخسارة للتدريب والتحقق عبر الدورات التدريبية.

4.6.5– مرحلة الاختبار

من أجل تبيان أثر التحسين تمّ الاختبار وفق مرحلتين:

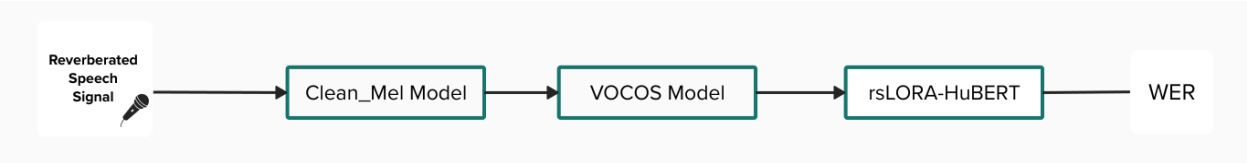
المرحلة الأولى

- نموذج CleanMel مع HuBERT قبل التدريب (سنسميه BaseModel لتسهيل التمييز) كما يبين الشكل (8.5):



الشكل 8.5- مخطط تمثيلي يوضح بنية النظام المقترح-حالة BaseModel

- نموذج CleanMel مع HuBERT بعد التدريب (سنسميه rsLORA-HuBERT) كما يبين الشكل (9.5):



الشكل 0 9.5- مخطط تمثيلي يوضح بنية النظام المقترح-حالة rsLORA-HuBERT

المرحلة الثانية بدون CleanMel

- نموذج HuBERT قبل التدريب (BaseModel) كما يبين الشكل (10.5):

الشكل 10.5- مخطط يوضح آلية اختبار BaseModel بدون حذف الصدى

- نموذج HuBERT بعد التدريب (rsLORA-HuBERT) كما يبين الشكل (11.5):

الشكل 11.5- مخطط يوضح آلية اختبار rsLORA-Hubert بدون حذف الصدى

تمّ تقييم أداء النظام المقترح على مجموعات الاختبار لكل من الكلام والاستجابة النبضية، ونظراً لما أُشير إليه سابقاً من أنّ أداء النموذج يتأثر بطبيعة الصدى وخصائص البيئة الصوتية، فقد جرى اختبار أدائه بشكل منفصل على عدد من مجموعات الاستجابة النبضية، ثم تمّ تنفيذ تجربة إضافية على مجموعة معطيات استجابة نبضية جديدة بالكامل لم تُستخدم في التدريب أو التحقق، وهي BUT Speech@FIT Reverb Database[35] وذلك لاختبار قدرة النظام على التعميم (Generalization) في بيئات صوتية غير مألوفة بالنسبة له.

وقد أظهرت النتائج ما يلي:

[35] https://speech.fit.vut.cz/software/but-speech-fit-reverb-database

مجموعة المعطيات	WER% مع حذف الصدى		WER% بدون حذف الصدى	
	BaseModel	rsLORA-HuBERT	BaseModel	rsLORA-HuBERT
TestSet	64.58	23.55	66.54	24.84
GreatHall	63.97	23.98	62.33	30.91
RealRIR	65.41	25.24	67.66	30.37
MIT IR	64.22	24.43	63.56	32.16
BUT Speech	85.25	47.35	95.44	73.21

الجدول 7.5- نتائج الاختبار للنظام المقترح End-to-End جميع الحالات.

كما تمّ الاختبار على مجموعة معطيات الاختبار الكلامية بدون إضافة أي نوع من الضجيج والصدى لدراسة تأثير النظام المقترح على المعطيات وكانت النتائج كالتالي:

مجموعة المعطيات	WER% مع CleanMel		WER% بدون CleanMel	
	BaseModel	rsLORA-HuBERT	BaseModel	rsLORA-HuBERT
Clean TestSet	61.60	23.31	55.19	21.01

الجدول 8.5- نتائج اختبار النظام المقترح على معطيات نظيفة.

5.6.5– مناقشة النتائج

تُظهر النتائج الواردة في الجدول () تفوّق النظام المقترح CleanMel + rsLoRA-HuBERT بوضوح على نموذج الأساس (CleanMel +BaseModel)في جميع مجموعات المعطيات التي تحتوي على صدى .فقد انخفض متوسط معدل الخطأ في الكلمات (WER) بشكل ملحوظ بعد إزالة الصدى، حيث انخفض مثلًا في مجموعة RealRIR من 67.66% إلى 30.37% قبل استخدامCleanMel ، ثم إلى 25.24% بعد دمج النموذج المقترح، مما يدل على تحسين مزدوج ناتج عن التكامل بين مرحلتي المعالجة المسبقة (CleanMel) والتعرّف المحسّن rsLoRA-HuBERT.

يتضح أيضاً أن النظام حافظ على أداء مستقر عبر مجموعات مختلفة من الاستجابات النبضية GreatHall، MIT IR، وهو ما يعكس قدرة جيدة على التعميم Generalization ضمن بيئات صوتية متنوعة.

أما على مجموعة BUT Speech@FIT Reverb Database، وهي مجموعة غير مرئية تماماً للنموذج أثناء التدريب، فقد بقي النموذج قادراً على تقليل WER من 95.44% إلى47.35%، وهو تحسن كبير رغم تعقيد هذه المعطيات.

أما بالنسبة لتجربة المعطيات النظيفة Clean TestSet فقد لوحظ ارتفاع طفيف في معدل الخطأ بعد تطبيق CleanMel من 21.01%إلى 23.31% وهو سلوك متوقع ناتج عن تعديل طفيف في الخصائص الطيفية التي يراها النموذج ضرورية للتعرّف، إلا أن هذا الانخفاض محدود ولا يشير إلى تدهور فعلي في الأداء.

بشكل عام، تشير هذه النتائج إلى أن:

- الجمع بين CleanMel لمعالجة الصدى و rsLoRA-HuBERT للتعرّف المحسّن حقق توازنًا فعّالا بين جودة الإشارة وأداء نظام ASR .
- النموذج المقترح يتفوّق بوضوح في بيئات الصدى المعقّدة، مع تأثير محدود وغير جوهري على الإشارات النظيفة.

يُعدّ هذا النهج خطوة مهمة نحو بناء نظام End-to-End Robust ASR قادر على العمل بكفاءة في ظروف صوتية واقعية.

7.5- الخلاصة

تناول كل من الفصلين الرابع والخامس الجوانب التطبيقية والتجريبية للبحث، بدءاً من بناء مجموعة المعطيات الصوتية ومعالجتها، مروراً بتصميم النماذج الأولية لاختبار فعالية الأساليب المختلفة في معالجة الصدى، وصولاً إلى تطوير النظام المقترح المتكامل من طرف إلى طرف (End-to-End) .

في المرحلة الأولى، تمّ تطوير نموذج أولي قائم على الشبكات العصبونية العودية من نوع LSTM لمعالجة الإشارة الكلامية في المجال الطيفي، حيث أظهر النموذج قدرة على تحسين جودة الإشارة الإدراكية من خلال إزالة مكونات الصدى بنجاح، إلا أن نتائجه من حيث معدل الخطأ في الكلمات (WER) كانت محدودة، ما دلّ على أن إزالة الصدى لم تكن متوافقة بالكامل مع متطلبات أنظمة التعرّف الآلي على الكلام.

أما في المرحلة الثانية، فقد تمّ اعتماد نموذج CleanMel-L-Mask المبني على معمارية Mamba، والمصمم لمعالجة الإشارة الكلامية في المجال الزمني–الترددي. وقد أظهر هذا النموذج كفاءة أعلى في الحفاظ على البنية الطيفية الدقيقة للإشارة وتحقيق توازن بين إزالة الصدى والحفاظ على خصائص الكلام الطبيعية، مما جعل نتائجه تتفوّق إدراكياً بشكل واضح على نموذج LSTM مع تحسين ملحوظ في جودة الإشارة عبر مختلف البيئات الصوتيّة.

وانطلاقاً من هذه النتائج، تمّ تصميم النظام المقترح End-to-End الذي يدمج بين نموذج CleanMel-L-Mask لمعالجة الإشارة الكلامية، مع تدريب نموذج HuBERT لتعرّف الكلام، بهدف بناء سلسلة معالجة متكاملة تجمع بين تحسين جودة الإشارة وتحسين أداء التعرّف على الكلمات.

اعتمد النظام المقترح في تدريبه على rsLoRA التي سمحت بتحديث أقل من 1.5% من موسطات نموذج HuBERT الضخم، مع الحفاظ على كفاءته العامة وتجنب فرط التلبيق. وقد أظهرت النتائج التجريبية أنّ هذا النهج المتكامل نجح في تقليل معدل

الخطأ في الكلمات بشكل واضح على مجموعات معطيات صدى واقعية ومتنوعة، بما في ذلك مجموعات لم تُستخدم أثناء التدريب مما يؤكد قدرة النظام على التعميم في بيئات صوتية غير مألوفة.

كما أظهرت الاختبارات على معطيات نظيفة أن النموذج لا يسبب تدهوراً كبيراً في جودة الكلام، ما يدل على قدرته على التكيّف مع ظروف التسجيل المختلفة.

ختاماً، يمكن القول إنّ النظام المقترح قد حقق الأهداف الأساسية للبحث، حيث نجح في بناء بنية متكاملة لمعالجة الصدى وتعرف الكلام آلياً في بيئات حقيقية، وقدم دليلاً عملياً على أنّ الجمع بين نماذج تحسين الإشارة الحديثة CleanMel وتقنيات التخصيص منخفض الرتبة rsLoRA ضمن أطر التعلم العميق يمكن أن يشكل نهجاً واعداً لتطوير أنظمة تعرف كلام صلدة Robust End-to-End ASR في اللغات المختلفة، بما فيها اللغة العربية.

الفصل السادس

الخاتمة والآفاق المستقبلية

يُختتم هذا البحث بعرض أبرز المساهمات النظرية والتطبيقية التي تمّ تحقيقها، إلى جانب الإشارة إلى الآفاق المستقبلية الذي يمكن أن تُبنى على النتائج المتحصلة ولم تُغطى في عملنا. تناولنا في هذا البحث واحدة من أعقد المسائل في معالجة الإشارة الكلامية، وهي مسألة حذف الصدى في بيئة أحادية القناة بهدف تحسين أداء أنظمة التعرف على الكلام آلياً، حيث تمّ اختبار مجموعة من النماذج وتقييمها ضمن ظروف تركيبية وواقعية لقياس مدى فعاليتها في التوفيق بين تحسين جودة الإشارة وخفض معدل الخطأ في الكلمات WER.

أظهرت نتائج التجارب أنّ النماذج التقليدية، مثل نموذج LSTM، قادرة على تحسين جودة الإشارة من منظور إدراكي، لكنها لا تؤدي بالضرورة إلى رفع دقة أنظمة التعرف، بل قد تُحدث تشوهات طيفية تؤثر سلباً على خصائص الكلام المفيدة لأنظمة ASR، ومن هنا جاءت أهمية المقاربة التكاملية التي تم اقتراحها في هذا البحث بالاعتماد على بنية Mamba، حيث أتاح الدمج بين نموذج حذف الصدى و نموذج تعرف الكلام آلياً HuBERT بناء نظام End-to-End قادر على مواءمة هدفَي تحسين الإشارة والتعرّف في آنٍ واحد، وهذا ما انعكس على انخفاض لقيم WERمع الحفاظ على جودة إدراكية جيدة حسب مؤشر PESQ. كما أثبتت النتائج أنّ البُنى المعتمدة على نماذج فضاء الحالة الانتقائية Selective State Space Models توفر معالجة أكثر كفاءة للسلاسل الزمنية الطويلة مقارنةً بالشبكات العودية، بفضل قدرتها على الاحتفاظ بالمعلومات السياقية على مدى زمني واسع مع تعقيد حسابي منخفض، وهو ما يجعلها خياراً واعداً لتطبيقات المعالجة الكلامية المستقبلية.

إنّ مسألة حذف الصدى هي مسألة ذات جوانب معقدة ولاسيما في حالتنا المتخصصة برفع دقة أنظمة تعرّف الكلام آلياً، مما يمهّد لتوجهات بحثية متقدمة لم تُغطَّ بعد، يمكن أن تشمل:

- توسيع قاعدة المعطيات لتضم تسجيلات واقعية متعددة البيئات واللغات لتقييم قابلية تعميم النموذج.
- دمج نموذجَي حذف الصدى ونموذج HuBERT في إطار تدريبي موحّد خطوة مهمة بما يتيح استكشاف فعالية الضبط المشترك بين النماذج (وهو ما لم يُنفّذ بسبب محدودية الموارد).
- دمج نماذج حذف الصدى مع أنظمة تعرف الكلام الفورية Streaming ASR لمعالجة الإشارات في الزمن الحقيقي.

- تطوير آليات تقييم موحّدة تجمع بين مؤشرات جودة الإشارة ومؤشرات دقة التعرف لقياس فعالية النظام الكلية بصورة أدق.

جميع النقاط والنتائج التي تمّ التوصّل إليها في هذا البحث ستكون منطلقاً لاهتماماتنا المستقبلية الرامية إلى تطوير حلول أكثر شمولاً وواقعية لمسألة حذف الصدى بما يعزّز أداء أنظمة التعرّف على الكلام آلياً. وعلى الرغم من النتائج الإيجابية المحققة، ما تزال هذه المسألة قابلة للبحث والتطوير، إذ إنّ الوصول إلى نظام متكامل قادر على التكيّف بمرونة مع مختلف ظروف الصدى والضجيج يتطلّب المزيد من التجريب والتحسين.

المراجع

[1] T. v. Waterschoot, "Deep, Data-Driven Modeling of Room Acoustics: Literature Review and Research Perspectives," in *Forum Acusticum 2025*, Spain, 2025.

[2] Y. Haneda ،S. Makino و Y. Kaneda ،"Common acoustical pole and zero modeling of room transfer functions "،*IEEE transactions on speech and audio processing* ، pp. 320-328 ، .1994

[3] G. Vairetti ،E. D. Sena ،M. Catrysse ،S. H. Jensen ،M. Moonen و T. v. Waterschoot ،"A Scalable Algorithm for Physically Motivated and Sparse Approximation of Room Impulse Responses With Orthonormal Basis Functions "،*IEEE/ACM Transactions on Audio, Speech, and Language Processing* ،pp. 1547-1561 .2017

[4] J. Rämö ،V. Välimäki و B. Bank ،"High-Precision Parallel Graphic Equalizer "،*IEEE/ACM Transactions on Audio, Speech, and Language Processing* ،pp. 894-1904 .2014

[5] S. Tervo ،J. Pätynen ،A. Kuusinen و T. Lokki ،"Spatial Decomposition Method for Room Impulse "،*Journal of the Audio Engineering Society* ،pp. 17-28 .2013

[6] F. Pinto و M. Vetterli ،"Space-Time-Frequency Processing of Acoustic Wave Fields: Theory, Algorithms, and Applications "،*IEEE Transactions on Signal Processing* ، pp. 4608-4620 .2010

[7] D. P. Jarrett, E. A. Habets and P. A. Naylor, "Theory and Applications of Spherical Microphone Array Processing," *Springer*, 2017.

[8] A. Krokstad ،S. Strom و S. Sørsdal ،"Calculating the acoustical room response by the use of a ray tracing technique "،*Journal of Sound and Vibration* ،pp. 18-125 .1968 ،

[9] T. Funkhouser ،N. Tsingos ،I. Carlbom ،G. Elko ،M. Sondhi ،J. E. West ،G. Pingali ،P. Min و A. Ngan ،"A beam tracing method for interactive architectural acoustics "،*The Journal of the acoustical society of America* ، pp. 739-756 .2004

[10] M. R. Bai, "Application of BEM (boundary element method)-based acoustic holography to radiation analysis of sound sources with arbitrarily shaped geometries," *The Journal of the Acoustical Society of America*, p. 533–549, 1992.

[11] S. Bilbao, B. Hamilton, J. Botts and L. Savioja, "Finite volume time domain room acoustics simulation under general impedance boundary conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 161-173, 2015.

[12] J. Heymann ،"Neural network based spectral mask estimation for acoustic beamforming "، *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* ،pp. 196-200 .2016

[13] C. Quan و X. Li ،"SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation "،*IEEE/ACM Transactions on Audio, Speech, and Language Processing* ،pp. 1310-1323 .2024

[14] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon and K. Lee, "Real-time denoising and dereverberation with tiny recurrent U-net," *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5789-5793, 2021.

[15] L. Wang, W. Wei, Y. Chen and Y. Hu, "D²Net: A Denoising and Dereverberation Network Based on Two-branch Encoder and Dual-path," *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* pp. 1649-1654, 2022.

[16] Y. Fu ،Y. Liu ،J. Li ،D. Luo ،S. Lv ،Y. Jv و L. Xie ،"UFORMER: a UNet based dilated complex and real dual-path conformer network for simultaneous speech enhancement and dereverberation "،*ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* ،pp .2022 ،7421-7417 .

[17] L. Zhao ،W. Zhu ،S. Li ،H. Luo ،X.-L. Zhang و S. Rahardja ،"Multi-Resolution Convolutional Residual Neural Networks for Monaural Speech Dereverberation "، *IEEE/ACM Transactions on Audio, Speech, and Language Processing* .2024

[18] N. Shao ،R. Zhou ،P. Wang ،X. Li ،Y. Fang ،Y. Yang و X. Li ،"CleanMel: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR "،*arXiv preprint arXiv:2502.20040* .2025

[19] L. Xi, C. Szu-Jui and H. J. H.L, "Dual-Path Minimum-Phase And All-Pass Decomposition Network For Single Channel Speech Dereverberation," *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 1261-1265, 2024.

[20] N. Shao ،R. Zhou ،P. Wang ،X. Li ،Y. Fang ،Y. Yang و X. Li ،"How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR "،*arXiv preprint arXiv:2201.06685* .2022

[21] I. GoodfellowK, Y. Bengio and A. Courville, Deep learning, Cambridge: MIT press, 2016.

[22] N. Küh, M. Schemmer, M. Goutier and G. Satzger, "Artifcial intelligence and machine learning," *Electronic Markets 32.4,* 2022.

[23] E.¨. A. P. Habets ،Single-and Multi-Microphone Speech Dereverberation ،Electrical Engineering. Technische Universiteit Eindhoven. .2007

[24] J. D. Vallado ،T. P. AA de Lima و S. Netto ،"Feature analysis for the reverberation perception in speech signals "،*IEEE International Conference on Acoustics, Speech and Signal Processing* ،pp. 8169-8173 .2013

[25] Y. Zhao ،Z.-Q. Wang و D. Wang ،"Two-stage deep learning for noisy-reverberant speech enhancement "،*IEEE/ACM transactions on audio, speech, and language processing* ،pp. 53-62 .2018

[26] T. Ochiai ،a. Iwamoto ،M. Delcroix ،R. Ikeshita ،H. Sato ،S. Araki و h. Katagiri ، "Rethinking processing distortions: Disentangling the impact of speech enhancement errors on speech recognition performance "،*EEE/ACM Transactions on Audio, Speech, and Language Processing* .2024

[27] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *nature 521.7553,* pp. 436-444, 2015.

[28] C.olah, "Understanding LSTM Networks-colah's blog," 27 August 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[29] V. SINGH, "The Sigmoid Function: A Key Component in Data Science," 28 May 2025. [Online]. Available: https://www.datacamp.com/tutorial/sigmoid-function?dc_referrer=https%3A%2F%2Fwww.google.com%2F.

[30] M. Grootendorst, "A Visual Guide to Mamba and State Space Models," 19 feb 2024. [Online]. Available: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state.

[31] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv:2312.00752v2,* 2024.

[32] D. Bergmann, "state space model-IBM," 7 july 2025. [Online]. Available: https://www.ibm.com/think/topics/state-space-model#1190488337.

[33] A. Gu, K. Goel and C. R.´e, "Efficiently Modeling Long Sequences with Structured State Spaces," *arXiv preprint arXiv:2111.00396,* 2021.

[34] A. Gu ،s. Johnson ،K. Goel ،K. Saab ،T. Dao ،A. Rudra و C. Ré ،"Combining Recurrent, Convolutional, and Continuous-time "،*Advances in neural information processing systems* ،pp. 572-585 .2021

[35] P. Hegde, "Mamba architecture : A Leap Forward in Sequence Modeling-Medium," 11 February 2024. [Online]. Available: https://medium.com/@puneetthegde22/mamba-architecture-a-leap-forward-in-sequence-modeling-370dfcbfe44a.

[36] "HuBERT Model-GeeksforGeeks," 23 july 2025. [Online]. Available: https://www.geeksforgeeks.org/nlp/hubert-model/.

[37] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for," *n Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4171-4186, 2019.

[38] W.-N. Hsu ،B. Bolte ،Y.-H. H. Tsai ،K. Lakhotia ،R. Salakhutdinov و A. Mohamed ، "HuBERT: Self-Supervised Speech Representation "،*IEEE/ACM Transactions on Audio, Speech, and Language Processing* ،pp. 3451-3460 .2021

[39] W. Chen ،X. Chang ،Y. Peng ،Z. Ni ،S. Maiti و S. Watanabe ،"Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute "،*arXiv preprint arXiv:2306.06672* .2023

[40] "What is Parameter-Efficient Fine-Tuning (PEFT)?," GeeksforGeeks, August 2025. [Online]. Available: https://www.geeksforgeeks.org/artificial-intelligence/what-is-parameter-efficient-fine-tuning-peft/.

[41] A. Aghajanyan, L. Zettlemoyer and S. Gupta, "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning," *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, 2021.

[42] V. Lialin, V. Deshpande, X. Yao and A. Rumshisky, "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning," *arXiv preprint arXiv:2303.15647*, 2023.

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *In International Conference on Learning Representations*, 2022.

[44] D. Kalajdzievski ،"A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA "، *arXiv preprint arXiv:2312.03732* .2023

[45] R. Stewart and M. Sandler, "Database of omnidirectional and B-format room impulse responses," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 165-168, 2010.

[46] M. Jeub, M. Schafer and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th international conference on digital signal processing*, 2009.

[47] R. A. University, "Aachen Impulse Response Database," Institue of Communication System, 2012. [Online]. Available: https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/.

[48] J. Y. Wen, N. D. Gaubitch, E.¨ A. Habets, T. Myatt and P. A. Naylor, "EVALUATION OF SPEECH DEREVERBERATION ALGORITHMS USING THE MARDY," in *IWAENC*, Paris,France, 2006.

[49] M. Maciejewski, G. Wichern, E. McQuinn and J. L. Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[50] Y. Zhao, D. Wang, B. Xu and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[51] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer and S. Khudanpur, "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition," *IEEE international conference on acoustics, speech and signal processing*, pp. 5220-5224, 2017.

[52] J. Traer و J. H. McDermott ،"Statistics of natural reverberation enable perceptual "،*space. Proceedings of the National Academy of Sciences* .2016

[53] J. G. Beerends ،A. P. Hekstra ،A. W. Rix و M. P. Hollier ،"Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment "،.pp. 765-778 .2002

الملحقات

- **خوارزمية MFCC:** تعتبر خوارزمية MFCC من أهم خوارزميات استخلاص السمات وذلك لكونها تستلهم تصميمها من طبيعة النظام السمعي البشري، ولا تقتصر على معالجة رياضية صرفة للإشارة. فقد أثبتت دراسات علم السمع أنّ الأذن لا تدرك الترددات على مقياس خطي؛ إذ يكون الإنسان أكثر حساسية للفروق الصغيرة في طبقة الصوت عند الترددات المنخفضة مقارنة بالترددات المرتفعة، ولذلك يبدو الفارق بين 200 و300 هرتز إدراكياً أكبر من الفارق بين 9000 و9100 هرتز رغم تطابق الفارق العددي (100 هرتز) في الحالتين.

تبدأ الخوارزمية بمعالجة أولية للإشارة الكلامية ثم تقسيمها إلى أطر زمنية قصيرة من رتبة 20ms، يلي ذلك ضرب كل إطار بنافذة Hamming ، ثم التحويل إلى المجال الترددي باستخدام تحويل فورييه السريع ، بعدها يتمّ استخدام مقياس ميل الترددي Mel frequency وتجميع الطاقة ضمن حزم المرشحات. أخيراً تحويل التجيب المتقطع Discrete Cosine Transform (DCT) لنحصل على موسطات ميل التي تُعدّ تمثيلاً مضغوطاً وفعّالاً للخصائص الطيفية ذات الصلة بالإدراك البشري.

قبل البدء بخوارزمية استخلاص السمات تُحوَّل إشارة الصوت التماثلية إلى إشارة رقمية، ونظراً لأنّ طيف الإشارة الكلامية يظهر تخامداً ملحوظاً عند الترددات العالية نتيجة تأثير الإشعاع عند الشفاه، يُطبّق مرشح تمرير مرتفع من الدرجة الأولى لتعزيز هذه الترددات. تعطى علاقة هذا المرشح بالمعادلة التالية:

$$y[n] = x[n] - k.x[n-1], 0 < k < 1 \qquad (1)$$

حيث $y[n]$ خرج المرشح

$X[n]$ دخل المرشح

k من رتبة 0.95 تقريباً.

وبما أنّ الإشارة الكلامية تصنف من الإشارات المتغيرة مع الزمن، إلا أنّه يمكن اعتبارها شبه مستقرة خلال فترات زمنية قصيرة، ولذلك لا بدّ من تقسيم الإشارة إلى أطر بطول يتراوح بين 20ms-25ms وذلك لضمان استقرار الإشارة الكلامية على كامل الإطار، ثمّ بعد ذلك تطبق نافذة على كامل الإطارات بحيث تحقق أنها ذات قيم منخفضة عند الحواف ومرتفعة في المنتصف، نفضل نافذة Hamming وذلك لتخفيف حدة الانقطاعات بين الأطر، وللتعويض عن تخميد المطالات على الأطراف نأخذ نوافذ متداخلة بمقدار M عينة ولذلك تُضرب الإطارات بنافذةHamming.

تعطى عبارة نافذة Hamming بالعلاقة الرياضية:

$$w[n] = 0.54 - 0.46 * \cos\left[\frac{2\pi n}{N-1}\right] \; for\; n = 1,2, \dots, N-1 \qquad (2)$$

حيث تمثل N عدد عينات الإطار.

ثمّ يتم الانتقال إلى المجال الترددي لاستخلاص المركبات الترددية وذلك باستخدام تحويل فورييه السريع Fast Fourier Transform (FFT) وهو خوارزمية لتطبيق تحويل فورييه المتقطع على الإطارات بعد تطبيق نافذة هامنغ عليها، والهدف من هذه الخطوة هو حساب الطيف الترددي الموافق لكل إطار زمني، ومن الجدير بالذكر أنه نبقي النصف الأول من الطيف الناتج الموافق للترددات الموجبة لأنها ستكون متناظرة حول محور الترددات كون إشارة الدخل حقيقية. تبين العلاقة المعادلة المستخدمة في حساب تحويل فورييه على N نقطة.

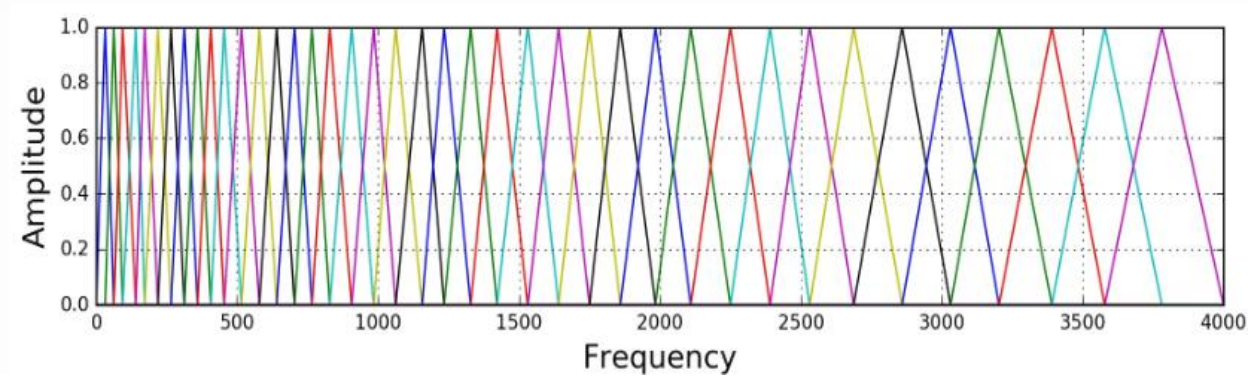
$$X_n = \sum_{k=0}^{N-1} x_k \, e^{\frac{-2\,\pi jkn}{N}} \quad , n = 0,1,2, \dots \dots, N-1 \qquad (3)$$

حيث X_n مركبات الطيف الترددي بينما x_k تمثل العينات الزمنية للإشارة، N عدد نقاط تحويل فورييه.

بعد الحصول على الطيف الترددي $X(f,t)$ ، يمكن تمثيل المعلومات الطيفية بطريقتين رئيسيتين: إما مطال الطيف أو طاقة الطيف Power Spectrum وهو التمثيل الأكثر استخداماً في التطبيقات العملية، لأنه يعكس كمية الطاقة الكامنة في كل تردد.

في هذه الدراسة، جرى اعتماد طاقة الطيف كأساس لبناء الـMel-Spectrogram، وذلك انسجاماً مع معظم أنظمة التعرف على الكلام (ASR) والنماذج العصبونية الحديثة، التي صُممت لتلقي مدخلات تعتمد على طاقة الطيف بدلاً من مطال الطيف.

وتتمثّل الخطوة التالية بتطبيق بنك مرشحات ميل(Mel Filter Banks) على طيف الطاقة، وتُعرَّف هذه المرشحات على هيئة دوال مثلثية موزَّعة وفق مقياس ميل (Mel Scale) كما يبين الشكل (1) بهدف محاكاة الإدراك غير الخطي للأذن البشرية؛ إذ تتميّز الأذن بقدرة تمييز أعلى عند الترددات المنخفضة مقارنةً بالمرتفعة. وبهذا تصبح الخوارزمية أكثر توافقاً مع خصائص السمع البشري، حيث تُعطي وزناً أكبر للفروق الطفيفة في النطاقات المنخفضة، بينما تقلّ حساسية التمييز تدريجياً عند الترددات العالية. ويُبنى كل مرشح مثلثي بحيث تبلغ استجابته القيمة العظمى (1) عند التردد المركزي للمرشح، ثم تنخفض الاستجابة خطياً حتى تصل إلى الصفر عند الترددين المركزيين للمرشحين المجاورين.



الشكل 01- بنك مرشحات ميل

بعد ذلك تُضرَب استجابة كل مرشح مثلثي بطيف طاقة الإطار، ثم تُجمَع القيم الناتجة للحصول على الطاقة ضمن كل مرشح. وبانتهاء هذه العملية نحصل على عدد من الموسطات يساوي عدد المرشحات المستخدمة، وهو ما يعكس توزيع الطاقة عبر الحزم الترددية

بعد ذلك نطبق تابع اللوغاريتم على هذه الموسطات ثم نطبق تحويل (DCT)، حسب العلاقة التالية:

$$C_n = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} (\log E_m) \cos\left(\frac{\left(m-\frac{1}{2}\right)\pi n}{M}\right) \qquad (4)$$

حيث: M عدد المرشحات المثلثية.

اعتمدت الورقة البحثية عدد مرشحات 80 وهذا الرقم ليس عشوائياً بل هو قرار تصميمي مدروس يتناسب مع أهداف الدراسة كون النماذج الحديثة المستخدمة القائمة على شبكات عصبونية عميقةMamba، لا تتعامل مع المدخلات عالية الأبعاد فحسب، بل تستفيد منها لاستخلاص أنماط أكثر تعقيداً وتفصيلاً من المعطيات على عكس النماذج التقليدية التي تعتمد على عدد مرشحات أقل بسبب محدودية قدرتها على التعامل مع المدخلات عالية الأبعاد. فضلاً عن كونه استخدام 80 مرشح يوفر مخطط طيفي (Mel-spectrogram) عالي الدقة .هذا التمثيل الغني بالتفاصيل يسمح لل Vocoder (وهو النموذج المستخدم لتحويل من Mel المحسن إلى إشارة كلامية لتكون دخلاً لنموذج ASR) بإعادة بناء إشارة كلامية أقرب إلى الأصل وأنقى وأكثر طبيعية ومن جهة أخرى بما أن تردد التقطيع هو 16KHz، فإن النطاق الترددي الفعال هو من 0 إلى 8KHz وإنّ توزيع 80 مرشح على هذا المجال يسمح بالتقاط تفاصيل دقيقة ليس فقط في الترددات المنخفضة المهمة للصوائت Vowels، ولكن أيضاً في الترددات العليا المهمة للصوامت Consonants وبالتالي عدد أقل من المرشحات سيجعل رؤية النموذج لهذه الترددات العالية أكثر ضبابية.

باختصار، قرار استخدام 80 مرشح هو مقايضة مدروسة: فهو يزيد من التفاصيل الطيفية بشكل كبير، وهو ما تستطيع النماذج العصبونية الحديثة الاستفادة منه لتحقيق دقة أعلى، مع الحفاظ على الفائدة الأساسية لمقياس ميل المتمثلة في ضغط المعلومات والتركيز على الجوانب الإدراكية المهمة للصوت.

- **معيار تقييم جودة الكلام الإدراكي PESQ** [53]

يُعدّ معيار تقييم جودة الكلام الإدراكي (Perceptual Evaluation of Speech Quality – PESQ) خوارزمية موضوعية (Objective Algorithm) تُستخدم لقياس جودة الإشارة الكلامية في أنظمة الاتصالات. وقد تمّ اعتماد هذا المعيار دولياً من قبل الاتحاد الدولي للاتصالات (International Telecommunication Union-ITU) ضمن التوصية P.862 عام 2001.

يتطلّب معيار PESQ إشارتين لإجراء عملية القياس: الإشارة المرجعية (Reference Signal) وهي إشارة الكلام الأصلية النظيفة، أمّا الإشارة الثانية فهي الإشارة الناتجة عن مرور الإشارة الأصلية عبر نظام الاتصال (مثل شبكة الهاتف أو الإنترنت)، أو بعد معالجتها بواسطة خوارزمية تحسين أو تحويل صوتي، أي بعد أن تتعرض لأي شكل من أشكال التغيير أو التشويه.

يأخذ معيار PESQ قيمته ضمن نطاق يتراوح تقريباً بين 0.5–و4.5 في صورته الخام، حيث تعبّر القيم المنخفضة عن جودة إشارة متدهورة جداً، بينما تعبّر القيم المرتفعة عن جودة إدراكية عالية للإشارة. وعند تحويل هذه القيم إلى مقياس متوسط درجات الرأي (Mean Opinion Score – MOS) المعتمد في توصية ITU، تُقيد النتائج ضمن المجال من 1.0 إلى 4.5، بحيث تمثل 1.0جودة رديئة جداً و 4.5 جودة ممتازة.