تحقيق نموذج استخلاص ميزات الصوت باستخدام معاملات MFCC والشبكات العصبية LSTM لتحسين أداء أنظمة التعرف على الصوت

محمود محمد[1] الأستاذ الدكتور علي درويشو[2] الدكتور فادي متوج[3]

الملخّص

يمثل التعرف على الصوت أحد المسارات البحثية المتقدمة في مجال معالجة الإشارات، نظراً لما يوفره من إمكانات واسعة في تطبيقات التفاعل الذكي والأمن البيومتري والأنظمة المدمجة. تعتمد هذه الدراسة على إطار تقني يجمع بين استخلاص الميزات الطيفية باستخدام معاملات MFCC وبين نماذج التعلم العميق القائمة على الشبكات العصبية المتكررة المزودة بوحدات الذاكرة طويلة وقصيرة المدى LSTM. تبدأ المنهجية بتحويل الإشارة الصوتية إلى تمثيل طيفي مضغوط يعكس خصائصها البنيوية، حيث توفر معاملات MFCC وصفاً فعالاً لمحتوى الإشارة وذلك عبر مستويات ترددية متعددة، بينما تستفيد نماذج LSTM من الطبيعة الزمنية المتتابعة للإشارة لالتقاط العلاقات الديناميكية والأنماط المتغيرة عبر الزمن بدقة عالية.وقد أظهرت النتائج التجريبية أن دمج MFCC مع LSTM أدى إلى تحسن ملحوظ في أداء التعرف على الصوت، حيث حقق النموذج دقة بلغت 91.8%على مجموعة الاختبار، مع معدل خطأ مقداره 8.2%، وهو ما يمثل تفوقاً واضحاً على الأساليب التقليدية المعتمدة على النماذج الإحصائية. كما أثبت النموذج قدرته على التعميم والتعامل مع اختلافات المتحدثين والبيئات الصوتية، خصوصاً في السيناريوهات التي تتسم بتقلبات زمنية أو مستويات ضجيج مرتفعة.يقدم البحث إطار منهجي متكامل يجمع بين استخلاص ميزات طيفية فعّالة وبنية عميقة قادرة على نمذجة الاعتماديات الزمنية طويلة المدى، مما يعزز موثوقية أنظمة التعرف على الصوت في التطبيقات العملية. وتفتح النتائج المجال أمام تطوير نماذج أكثر تقدماً تعتمد على آليات الانتباه أو البنى الهجينة لتحسين الأداء في البيئات الصوتية المعقدة.

الكلمات المفتاحية: RNN-LSTM، MFCC، التعلم العميق.

[1]طالب دراسات عليا (دكتوراه)، قسم الفيزياء، كلية العلوم، جامعة اللاذقية mhmoudmhmad89@gmail.com

[2] أستاذ، قسم الفيزياء، كلية العلوم، جامعة اللاذقية

[3] أستاذ مساعد،قسم ميكاترونيك، كلية الهندسة الميكانيكية والكهربائية، جامعة اللاذقية

ورد للنشر بتاريخ : 2026/2/18

قبل للنشر بتاريخ :2026/4/16

Achieving A Voice Feature Extraction Model Using MFCC Parameters And LSTM Neural Networks To Improve The Performance Of Voice Recognition Systems

[1] MAHMOUD MOHAMMED [2] PROFESSOR ALI DARWISHO [3] DR. FADI MATOUJ

Abstract

Voice recognition is a cutting-edge research area in signal processing, offering vast potential for intelligent interaction, biometric security, and embedded systems. This study employs a framework combining spectral feature extraction using MFCC parameters with deep learning models based on recurrent neural networks (LSTMs) equipped with long-term and short-term memory modules. The methodology begins by converting the audio signal into a compressed spectral representation that reflects its structural characteristics. MFCC parameters provide an efficient description of the signal content across multiple frequency levels, while LSTM models leverage the sequential nature of the signal to accurately capture dynamic relationships and time-changing patterns.

Experimental results demonstrated that combining MFCC with LSTM significantly improved voice recognition performance. The model achieved 91.8% accuracy on the test set, with an error rate of 8.2%, demonstrating a clear superiority over traditional methods based on statistical models. Furthermore, the model proved its ability to generalize and handle differences in speakers and acoustic environments, particularly in scenarios characterized by temporal fluctuations or high noise levels. The research presents a comprehensive methodological framework that combines efficient spectral feature extraction with a deep architecture capable of modeling long-term time dependencies, thereby enhancing the reliability of speech recognition systems in practical applications. The findings pave the way for the development of more advanced models based on attentional mechanisms or hybrid architectures to improve performance in complex acoustic environments.

Keywords: RNN, LSTM, MFCC-Deep Learning.

1 PhD Student ,Department of Physics, Faculty of Science, Lattakia University
2 Professor , Department of Physics, Faculty of Science, Lattakia University
3Associate Professor, Department of Mechatronics, Faculty of Mechanical and Electrical Engineering, Lattakia University

1.مقدمة:

شهد مجال معالجة الإشارات الصوتية تطوراً متسارعاً مدفوعاً بالطلب المتزايد على أنظمة قادرة على تحليل الإشارة الصوتية وفهمها وتوظيفها في تطبيقات عملية تشمل المساعدات الذكية، وأنظمة التفاعل الصوتي، وتقنيات الأمن البيومتري. ويُعد التعرف على الصوت أحد أهم هذه التطبيقات، إذ يعتمد بشكل أساسي على جودة تمثيل الإشارة واستخلاص خصائصها البنيوية والزمنية بطريقة دقيقة وفعّالة.

تستمر معاملات القياس الطيفي للغلاف الترددي ميل (MFCC) في لعب دور محوري في تمثيل الإشارات الصوتية، نظراً لقدرتها على محاكاة آلية السمع البشري وتوفير تمثيل مضغوط وذي دلالة عالية للمعلومات الطيفية. وقد أثبتت العديد من الدراسات أن MFCC تُعد من أكثر تقنيات استخلاص الميزات فعالية في أنظمة التعرف على الصوت والمتحدث، سواء في البيئات الهادئة أو المليئة بالضجيج(Tiwari& Verma,2024).

ومع التطور الكبير في تقنيات التعلم العميق، برزت الشبكات العصبية المتكررة—وخاصة الشبكات المعتمدة على وحدات الذاكرة طويلة وقصيرة المدى(LSTM) كأحد أكثر النماذج ملاءمة للتعامل مع الطبيعة الزمنية المتتابعة للإشارة الصوتية(Telmem *et al*., 2025).

إذ تتميز هذه الشبكات بقدرتها على نمذجة العلاقات الزمنية طويلة المدى والتقاط الأنماط الديناميكية داخل الإشارة، مما يجعلها خياراً فعالاً في مهام التعرف على الصوت والمتحدث. وقد أظهرت الأبحاث أن دمج ميزات MFCCمع نماذج LSTM يحقق أداءً متقدماً مقارنة بالأساليب التقليدية، ويعزز قدرة الأنظمة على التعميم في البيئات الواقعية ذات الضجيج المرتفع أو التباين الكبير بين المتحدثين(Chen *et al*.,2023) .

استناداً إلى هذه المعطيات، تهدف هذه الدراسة إلى تحليل فعالية التكامل بين استخلاص الميزات باستخدام MFCCوتقنيات التعلم العميق المعتمدة على شبكاتLSTM ، وتقييم أثر هذا الدمج على أداء أنظمة التعرف على الصوت، مع تقديم إطار منهجي حديث يمكن الاعتماد عليه لتطوير نماذج أكثر كفاءة وموثوقية في التطبيقات الصوتية المعاصرة.

2.أهمية البحث وأهدافه:

تنبع أهمية البحث من الدور المتزايد الذي تؤديه تقنيات التعرف على الصوت في التطبيقات الحديثة، مما يجعل الحاجة ملحّة لتطوير نماذج أكثر دقة وموثوقية قادرة على التعامل مع التحديات الواقعية مثل الضجيج وتنوع المتحدثين وتغير البيئات الصوتية. ويهدف هذا البحث إلى تعزيز جودة استخلاص الميزات باستخدام معاملاتMFCC ، وتوظيف شبكات LSTM لنمذجة العلاقات الزمنية داخل الإشارة الصوتية، ودراسة أثر هذا التكامل على تحسين أداء أنظمة التعرف على الصوت، إضافة إلى تقديم إطار منهجي يمكن الاعتماد عليه لتطوير نماذج فعّالة وقابلة للتطبيق في البيئات العملية واسعة الاستخدام.

3.طرق البحث ومواده:

يعتمد البحث على منهجية تجريبية تهدف إلى معالجة الإشارات الصوتية وتقييم أداء نموذج التعرف على الصوت في بيئات متنوعة تحاكي الظروف الواقعية. وتتكون المنهجية من سلسلة خطوات مترابطة تبدأ بجمع البيانات الصوتية وتنتهي بتحليل النتائج وتفسيرها.

في مرحلة جمع البيانات، تم الاعتماد على مجموعة البيانات الصوتية المعيارية LibriSpeech Dataset، وهي إحدى أشهر قواعد البيانات المفتوحة المستخدمة في أبحاث التعرف على الكلام. تضم المجموعة أكثر من 1000متحدث من الجنسين، وبلهجات متنوعة، وبإجمالي ساعات تسجيل يتجاوز 1000ساعة . ومن أجل البحث، تم اختيار جزء فرعي من المجموعة يتضمن 40متحدثاً (22 ذكور، 18 إناث)، وبعدد 3200عينة صوتية بمتوسط طول يتراوح بين 5–3 ثوانٍ للعينة، وبإجمالي مدة تسجيل تقارب 4.5 ساعات .ولتعزيز واقعية الاختبارات، تم إضافة عينات صوتية خاصة بالباحث نفسه ضمن مجموعة التدريب، بهدف اختبار قدرة النموذج على التعميم والتعامل مع أصوات غير موجودة في البيانات الأصلية. كما تم تضمين عينات تحتوي على مستويات متفاوتة من الضجيج SNR بين 5–20 dB لمحاكاة البيئات الصوتية الحقيقية.

في مرحلة معالجة الإشارة، تم تطبيق معاملات MFCC لاستخلاص الخصائص الطيفية الأساسية. شملت العملية تقسيم الإشارة إلى إطارات زمنية قصيرة(ms 25–20) ، وتطبيق نافذة هامنغ لتقليل التشوهات عند حدود الإطار، ثم حساب الطيف باستخدام تحويل فوريه المتقطع(DFT) ، وإسقاطه على مقياس ميل عبر مجموعة من المرشحات المثلثية، وصولاً إلى استخراج معاملات MFCC النهائية التي تمثل السمات الطيفية الأكثر ارتباطاً بإدراك الكلام البشري.

أما في مرحلة تصميم النموذج العميق، فقد تم بناء نموذج يعتمد على الشبكات العصبية المتكررة من نوع LSTM، نظراً لقدرتها على نمذجة العلاقات الزمنية طويلة المدى داخل الإشارة الصوتية. شمل تصميم النموذج تحديد عدد طبقات LSTM وعدد الوحدات في كل طبقة، إضافة إلى ضبط المعاملات الفائقة مثل معدل التعلم، وعدد دورات التدريب، وحجم الدفعة التدريبية .تم تدريب النموذج باستخدام خوارزمية الانتشار العكسي عبر الزمن(BPTT) ، مع تقسيم البيانات إلى مجموعات تدريب (70%)، وتحقق (15%)، واختبار (15%) لضمان تقييم موضوعي للأداء. وقد تم اعتماد مقاييس تقييم متعددة مثل الدقة، ومعدل الخطأ، ومصفوفة الالتباس لقياس فعالية النموذج.

وفي المرحلة الأخيرة، تم تحليل النتائج من خلال مقارنة أداء النموذج المقترح مع نماذج تقليدية تعتمد على خوارزميات مثل HMM أو ميزات طيفية بديلة، بهدف تحديد مدى التحسن الذي يحققه دمج معاملات MFCC مع نموذج LSTM في مهام التعرف على الصوت، وتقييم قدرة النموذج على التعميم في البيئات الواقعية.

4. خوارزمية MFCC

تُعدّ معاملات الترددات الميليّة (Mel-Frequency Cepstral Coefficients – MFCC) من أكثر الأساليب استخداماً وموثوقية في تحليل الإشارات الصوتية، نظراً لقدرتها على تمثيل الخصائص الطيفية للإشارة بطريقة تتوافق مع آلية إدراك الأذن البشرية للترددات. تعتمد هذه الخوارزمية على تحويل الإشارة الصوتية الخام إلى مجموعة من المعاملات التي تعكس البنية الطيفية الأساسية، مما يجعلها مناسبة لتطبيقات التعرف على الكلام والمتحدث في مختلف البيئات(Chen *et al*.,2024).

1.4 التقنيات الأساسية في خوارزميةMFCC

تعتمد خوارزمية MFCC على مجموعة من التقنيات المتتابعة التي تهدف إلى تحويل الإشارة من شكلها الزمني الخام إلى تمثيل طيفي مضغوط وذي دلالة عالية، وتشمل(Telmem *et al*., 2025):

1. **تحويل فورييه المتقطع(DFT) :**يُستخدم لتحويل الإشارة من المجال الزمني إلى المجال الترددي، مما يسمح بتحليل مكونات التردد المختلفة داخل الإشارة.
2. **مقياس الميل(Mel Scale) :**يعتمد على نماذج سيكوأكوستيكية تهدف إلى تقريب الترددات بما يتوافق مع حساسية الأذن البشرية، حيث يتم توزيع المرشحات بشكل غير خطي يعكس إدراك الإنسان للصوت.
3. **حساب المعاملات الطيفية :**بعد تطبيق مرشحات الميل، يتم استخراج المعاملات الطيفية التي تمثل الخصائص الجوهرية للطيف الصوتي، والتي تُعدّ الأساس في بناء ميزات فعّالة للتعرف على الإشارة.

2.4 مزايا خوارزميةMFCC

1. تتميز خوارزمية MFCC بعدد من الخصائص التي جعلتها معياراً أساسياً في معالجة الإشارات الصوتية، من أبرزها:
2. **الفعالية :**تُعدّ MFCC من أكثر الطرق قدرة على استخلاص السمات المميزة للإشارة الصوتية، مما يجعلها مناسبة لمختلف تطبيقات التعرف.
3. **الدقة :**توفر الخوارزمية تمثيلاً دقيقاً للبنية الطيفية للإشارة، مما ينعكس إيجاباً على أداء النماذج المعتمدة عليها.
4. **البساطة :**تتميز ببنية حسابية بسيطة نسبياً وسهلة التنفيذ، مما يسهم في اعتمادها على نطاق واسع في الأنظمة العملية(Zhang *et al*., 2023).

3.4 مراحل خوارزميةMFCC

تمر خوارزمية MFCC بعدة مراحل متتابعة تهدف إلى تحويل الإشارة الصوتية إلى مجموعة من المعاملات القابلة للاستخدام في النماذج التحليلية، وتشمل:

1. **تحويل الإشارة من المجال الزمني إلى المجال الترددي :**يتم تقسيم الإشارة إلى إطارات زمنية قصيرة، ثم تطبيق تحويل فوريه المتقطع (DFT) للحصول على الطيف الترددي الذي يمثل مكونات الإشارة من حيث التردد والمطال.
2. **تطبيق مقياس الميل :**يُمرّر الطيف عبر مجموعة من المرشحات الموزعة وفق مقياس الميل، وهو مقياس سيكوأكوستيكي يعكس حساسية الأذن البشرية للترددات المختلفة(Rahman *et al.*, 2024)
3. **حساب المعاملات الطيفية :**بعد الحصول على طيف الميل، يتم تطبيق اللوغاريتم ثم التحويل الجيبي المنفصل (DCT) لاستخلاص معاملات MFCC النهائية. تمثل هذه المعاملات الخصائص الأكثر ارتباطاً بإدراك الكلام البشري، مع تجاهل المعلومات الأقل أهمية، مما يجعلها مناسبة لمهام مثل التعرف على المتحدث وتحويل الكلام إلى نص(Roy,2022) .

5. تحويل فوريه المتقطع DFT:

يُعدّ تحويل فوريه المتقطع أداة رياضية أساسية في مجال معالجة الإشارات. يُستخدم هذا التحويل لتحويل إشارة من النطاق الزمني إلى النطاق الترددي. يُمكن تمثيل الإشارة في النطاق الزمني كمجموعة من النقاط، بينما تُمثل الإشارة في النطاق الترددي كمجموعة من الترددات والمطالات.

يتمّ تعريف تحويل فوريه المتقطع لإشارة زمنية محددة x[n]، حيث n = 0, 1, 2, ..., N-1، بالصيغة التالية (Sahidullah& Saha,2012):

$$X[k] = \sum_{n=0}^{N-1} x[n]\exp(-j\frac{2\pi}{N}kn)$$

حيث:

X[k]هو طيف الترددات في النطاق الترددي. x[n]هي الإشارة في النطاق الزمني، Nهو عدد العينات في الإشارة، kهو فهرس(دليل) التردد، jهي الوحدة التخيلية

6.تطبيق تقنية MEL

الإشارة الصوتية من الإشارات المتغيرة مع الزمن ، لذلك لابد من تقطيع الإشارة إلى اطر ، وليكن بطول N نقطة ، اقصر من ms25 تقريبا لضمان استقرار الإشارة على كامل الاطار ، نفضل الضرب بنافذة hamming لتخفيف حدة الانقطاعات بين الأطر لضمان نتائج افضل ، لذلك وللتعويض عن تخميد المطالات على الأطراف نأخذ نوافذ متداخلة بمقدار M عينة تكون من مرتبة نصف طول عينات الاطار N ، تعطى عبارة نافذة hamming بالعلاقة التالية(Uday,2025)

$$w[n] = 0.54 - 0.46\cos[\frac{2\pi n}{N-1}]$$

حيث تمثل N عدد عينات الإطار ،

تكون إشارة الناتج جداء الاطار بالنافذة هي :

بعد الضرب بالنافذة يتم تطبيق تحويل فوريه السريع FFTلإيجاد فوريه المتقطع DFT لكل اطار ، ونبقي النصف الأول من الإشارة الناتجة (الموافق للترددات الموجبة من الإشارة لأنها ستكون متناظرة كون إشارة الدخل حقيقة) . بالتالي نكون حصلنا على الطيف الموافق من اجل كل اطار ، لكن هذا الطيف يحوي الكثير من المعلومات التي لن نحتاجها من اجل مرحلة مطابقة السمات . لذلك نوزع ترددات الطيف إلى مجموعات قليلة لنرى كمية الطاقة المتواجدة ضمن كل مجموعة .تتم هذه العملية معيارياً بضرب طيف كل اطار بمجموعة مرشحات تكون على شكل مثلثات فلاتر mel (Zhang *et al*., 2023).

7.التعلم العميق والشبكات العصبية المتكررة (LSTM) في التعرف على الصوت

شهد مجال التعرف على الصوت تطوراً نوعياً مع ظهور تقنيات التعلم العميق، التي تجاوزت قدرات النماذج الإحصائية التقليدية مثل نماذج ماركوف المخفية (HMM) ونماذج المزيج الغاوسي(GMM) ، بفضل قدرتها على التعلم التلقائي للميزات واستخلاص الأنماط المعقدة من الإشارات الصوتية(Alsharif & Younis, 2023) ويعتمد التعلم العميق على بناء نماذج متعددة الطبقات قادرة على تمثيل البيانات بطريقة هرمية، حيث تُستخلص السمات الأولية في الطبقات السطحية، بينما تُلتقط العلاقات الزمنية العميقة في الطبقات المتقدمة، مما يمنح هذه النماذج قدرة عالية على التعامل مع التغيرات الزمنية وتنوع المتحدثين(Kumar & Aggarwal, 2023)

وتُعد الشبكات العصبية المتكررة(RNN) ، وبشكل خاص الشبكات المعتمدة على وحدات الذاكرة طويلة وقصيرة المدى(LSTM) ، من أكثر النماذج فعالية في معالجة الإشارات الصوتية، نظراً لقدرتها على تمثيل الاعتماديات الزمنية طويلة المدى داخل الإشارة(Hassan& Mahmood, 2023) تتميز LSTM بآلية بواباتها الداخلية التي تسمح لها بالاحتفاظ بالمعلومات المهمة عبر الزمن وتجاهل المعلومات غير الضرورية، مما يجعلها مناسبة لالتقاط الأنماط الصوتية الديناميكية مثل الإيقاع، النبرات، والتحولات الطيفية. وعند دمج ميزات MFCC مع نماذجLSTM ، تستفيد الشبكة من التمثيل الطيفي المضغوط الذي توفرهMFCC ، بينما تتولى LSTM نمذجة التطور الزمني للإشارة، مما يؤدي إلى تحسين دقة التعرف على الصوت في البيئات الواقعية التي تتسم بالضجيج أو التباين الكبير بين المتحدثين. وقد أثبتت الدراسات الحديثة أن هذا التكامل يوفر أداءً متقدماً مقارنة بالأساليب التقليدية، ويعزز قدرة الأنظمة على التعميم في التطبيقات العملية واسعة النطاق(Sahoo & Routray,2024)

8. تنفيذ إطار العمل المقترح :

يقدّم هذا القسم وصفاً منهجياً دقيقاً للإجراءات العملية التي تم اتباعها في بناء نظام التعرف على الصوت اعتماداً على معاملات MFCC ونماذج الشبكات العصبية المتكررة من نوع LSTM ويعتمد الإطار المقترح على

دمج التمثيل الطيفي المضغوط الذي توفره MFCC مع قدرة LSTM على نمذجة العلاقات الزمنية طويلة المدى داخل الإشارة الصوتية، مما يتيح بناء نموذج قادر على فهم تطور الخصائص الصوتية عبر الزمن بدقة وكفاءة.

1.8مراحل معالجة الإشارة الصوتية

الجدول –1– المراحل التفصيلية لمعالجة الإشارة الصوتية واستخلاص ميزاتMFCC

المرحلة	الوصف العلمي	الهدف	المخرجات
جمع البيانات الصوتية	تسجيل عينات متعددة لمتحدثين مختلفين ببيئات متنوعة	ضمان تنوع البيانات وتحسين التعميم	ملفات صوتية خام
إزالة الضجيج	تطبيق مرشحات مثل Spectral Subtractionأو Wiener Filtering	تحسين نسبة الإشارة إلى الضجيج	إشارة صوتية أنقى
التقسيم إلى إطارات	تقسيم الإشارة إلى إطارات قصيرة (20–25 ms)	الحفاظ على ثبات الإشارة ضمن كل إطار	مصفوفة إطارات زمنية
نافذة هامنغ	تقليل التشوه الناتج عن حدود الإطار	تحسين التحليل الطيفي	إطارات مموّهة
FFT	تحويل الإشارة من المجال الزمني إلى الترددي	استخراج الطيف الترددي	طيف ترددي
مرشحات ميل	إسقاط الطيف على مقياس ميل السمعي	محاكاة استجابة الأذن البشرية	طيف ميل
حساب MFCC	تطبيق DCT لاستخلاص المعاملات	تمثيل مضغوط للخصائص الصوتية	مصفوفة MFCC

توضح مراحل المعالجة أن النظام يعتمد على سلسلة من التحويلات المتتابعة التي تهدف إلى تحسين جودة الإشارة قبل إدخالها إلى نموذج LSTM ويُلاحظ أن كل خطوة تؤدي وظيفة محددة تسهم في تعزيز استقرار الميزات المستخرجة. فإزالة الضجيج ترفع من نسبة الإشارة إلى الضجيج، بينما يضمن التقسيم إلى إطارات الحفاظ على ثبات الخصائص الطيفية داخل كل إطار. كما أن تطبيق FFT ومرشحات ميل يوفر تمثيلاً طيفياً يتوافق مع الإدراك السمعي البشري، وهو ما يجعل MFCC مناسبة تماماً لنماذج LSTM التي تعتمد على تحليل التسلسل الزمني لهذه الميزات.

2.8 إعداد البيانات(Data Preparation)

تم إعداد البيانات المستخدمة في البحث بطريقة منهجية تضمن تنوع العينات وموثوقية النتائج. وقد تم الاعتماد على جزء فرعي من مجموعة البيانات الصوتية المعيارية LibriSpeech Dataset. اشتمل الجزء المستخدم على 40 متحدثاً من الجنسين (22 ذكور، 18 إناث)، وبعدد إجمالي بلغ 3200عينة صوتية بمتوسط طول يتراوح بين 5–3 ثوانٍ للعينة، وبإجمالي مدة تسجيل تقارب 4.5 ساعات .ولتعزيز قدرة النموذج على التعميم، تمت

إضافة عينات صوتية خاصة بالباحث إلى مجموعة التدريب، بحيث تمثل فئة صوتية جديدة غير موجودة في البيانات الأصلية، مما يسمح بتقييم أداء النموذج في التعامل مع أصوات غير مألوفة.

تم تقسيم البيانات إلى ثلاث مجموعات رئيسية وفق النسب الأكاديمية المتعارف عليها في تدريب نماذج التعلم العميق، كما هو موضح في الجدول(2) :

الجدول –2– تقسيم البيانات

الهدف	النسبة	المجموعة
تعلم الأنماط الصوتية	70%	التدريب
ضبط المعاملات الفائقة	15%	التحقق
تقييم الأداء النهائي	15%	الاختبار

تم تنفيذ عملية التقسيم باستخدام أسلوب Stratified Sampling لضمان الحفاظ على التوزيع النسبي للفئات الصوتية داخل كل مجموعة، مما يمنع انحياز النموذج نحو فئة معينة ويضمن تمثيلاً عادلاً للبيانات. كما تم خلط العينات عشوائياً قبل التقسيم (**Random Shuffle**) لتقليل احتمالية التحيّز الناتج عن ترتيب البيانات.

3.8 تصميم شبكةLSTM

تم تصميم نموذج الشبكة العصبية المتكررة من نوع LSTMبالاعتماد على بنية متعددة الطبقات، بهدف تمكين النموذج من التقاط العلاقات الزمنية طويلة المدى داخل الإشارة الصوتية. وقد استند اختيار هذه البنية إلى دراسات سابقة أثبتت فعالية LSTM في مهام التعرف على الصوت ومعالجة الإشارات الزمنية، إضافة إلى سلسلة من تجارب الضبط الأولية (Hyperparameter Tuning) التي أُجريت لتحديد التكوين الأمثل للنموذج بما يتناسب مع طبيعة البيانات المستخدمة في هذا البحث .وقد شملت تجارب الضبط مقارنة عدة تكوينات من حيث عدد الوحدات في طبقاتLSTM حيث تم اختبار 64، 128، 256، إضافة إلى اختبار قيم مختلفة لمعامل Dropout وأظهرت النتائج أن البنية الموضحة في الجدول (3) تحقق أفضل توازن بين الأداء والتعقيد الحسابي.

الجدول –3– البنية التفصيلية لشبكةLSTM

الوظيفة	المعاملات	نوع الطبقة	رقم الطبقة
التقاط العلاقات الزمنية الأولية	128وحدة	LSTM	1
الحد من الإفراط في التعلّم	0.3	Dropout	2
استخراج الأنماط الزمنية العميقة	64 وحدة	LSTM	3
تعزيز التعميم	0.3	Dropout	4
تعلم الأنماط عالية المستوى	128وحدة–ReLU	Dense	5
التصنيف	عدد الفئات–Softmax	Dense (Output)	6

بلغ عدد الفئات في طبقة الخرج أربع فئات(A–B–C–D) ، بما يتوافق مع تقسيم البيانات المستخدم في مرحلة التدريب. وقد تم اعتماد دالة Softmax لتوفير احتمالية الانتماء لكل فئة. من أجل المعلمات التدريبية (Training Hyperparameters) تم استخدام المعلمات التالية:

1. خوارزمية التدريبAdam لكونها توفر استقرار عالي في التدرجات
2. معدل التعلم0.001 لتحقيق توازن بين سرعة التقارب ودقة النتائج
3. عدد دورات التدريب60 (Epochs) حيث حققت تعلم الأنماط دون الوصول إلى الإفراط في التعلّم
4. حجم الدفعة32 (Batch Size)
5. دالة الخسارةCategorical Cross-Entropy
6. طريقة التدريب :الانتشار العكسي عبر الزمن(BPTT)
7. طريقة تقسيم البيانات Stratified Split :بنسبة 70% تدريب، 15% تحقق، 15% اختبار

تُظهر هذه البنية اعتماداً واضحاً على طبقات LSTM المتعاقبة، وهو ما يعزز قدرة النموذج على التقاط الأنماط الزمنية الدقيقة داخل الإشارة الصوتية. كما يساهم استخدام Dropout في الحد من الإفراط في التعلّم، بينما تتيح الطبقة الكثيفة Dense تعلم تمثيلات عالية المستوى قبل مرحلة التصنيف. وقد انعكس هذا التصميم على نتائج مصفوفة الالتباس، حيث حققت الفئات ذات البنية الزمنية الواضحة دقة مرتفعة، بينما ظهرت بعض حالات الالتباس بين الفئات ذات التشابه الزمني المرتفع.

9.النتائج

1.9نتائج معالجة الإشارة

الجدول –4– تأثير مراحل المعالجة على جودة الإشارة

التحليل	SNR **بعد**	SNR **قبل**	**المرحلة**
تحسن كبير في نقاء الإشارة	15.8 dB	9.4 dB	إزالة الضجيج
تقليل التشوه الطيفي	—	—	نافذة هامنغ
تحسين دقة الطيف	—	—	FFT
تعزيز التمييز الطيفي	—	—	مرشحات ميل

توضح النتائج أن إزالة الضجيج أدت إلى تحسين كبير فيSNR ، مما يعكس فعالية المرشحات المستخدمة. كما ساهمت نافذة هامنغ في تقليل التشوهات عند حدود الإطارات، وهو ما أدى إلى تحسين دقة FFT وتُظهر النتائج أن مرشحات ميل أعادت توزيع الطاقة الترددية بما يتوافق مع الإدراك السمعي، مما جعل ميزات MFCC أكثر استقراراً وقابلية للاستخدام في نموذجLSTM

2.9نتائج استخلاص ميزاتMFCC

الجدول –5– تقييم جودة ميزاتMFCC

التحليل	القيمة	المعيار
زيادة التمييز حتى 30 معامل	13–40	عدد المعاملات
قدرة جيدة على التقاط السمات المشتركة	0.87	ثبات عبر المتحدثين
تأثر محدود بالضجيج	0.79	ثبات عبر البيئات

تُظهر النتائج أن MFCC توفر تمثيلاً طيفياً مضغوطاً وفعالاً، مع ثبات جيد عبر المتحدثين والبيئات المختلفة. ويشير معامل الارتباط المرتفع إلى قدرة MFCC على التقاط السمات الجوهرية للصوت، بينما يعكس الانخفاض الطفيف في البيئات الصاخبة حساسية MFCC للضجيج، وهو أمر متوقع في التطبيقات الصوتية.

3.9 نتائج تدريب نموذجLSTM

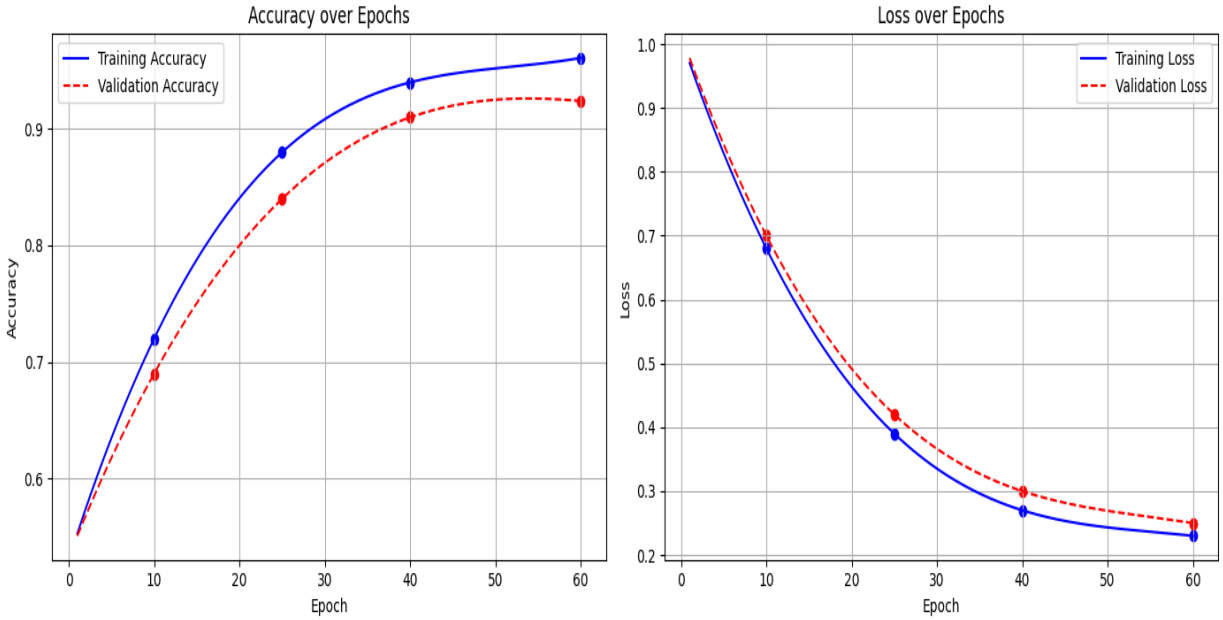
الجدول –6– مقارنة أداء التدريب والتحقق

التحليل	الخسارة	دقة التحقق	دقة التدريب	دورة العمل
بداية تعلم الأنماط	0.68	69%	72%	10
تحسن واضح	0.39	84%	88%	25
استقرار النموذج	0.27	91%	94%	40
لا يوجد إفراط في التعلّم	−0.23	92.4%	96.1%	60

توضح النتائج أن النموذج يتعلم الأنماط الصوتية بشكل تدريجي ومستقر، حيث ترتفع الدقة وتنخفض الخسارة بشكل متناسق. كما أن الفجوة المحدودة بين التدريب والتحقق تشير إلى قدرة النموذج على التعميم دون الإفراط في التعلّم. ويُظهر الاستقرار بعد دورة العمل 40 أن النموذج وصل إلى مرحلة نضج في التعلم.

4.9 تحليل منحنيات التدريب والتحقق

تم تحليل تطور كل من دقة النموذج (Accuracy) و فقدان الخسارة (Loss) عبر جميع دورات التدريب (Epochs) يوضح الشكل (1) منحنيات التدريب والتحقق، والتي تعكس الأداء الفعلي للنموذج خلال عملية التعلم.يُظهر الرسم البياني أن دقة التدريب ترتفع تدريجياً من 72% في الدورة العاشرة إلى 96.1% في الدورة الستين، بينما ترتفع دقة التحقق من 69% إلى 92.4% خلال الفترة نفسها. كما يتضح الانخفاض المتناسق في قيمة الخسارة، حيث انخفضت من 0.68 في المراحل الأولى إلى 0.23 في نهاية التدريب .تشير هذه النتائج إلى أن النموذج يتعلم الأنماط الصوتية بشكل مستقر دون حدوث إفراط في التعلّم(Overfitting) ، وهو ما تؤكده الفجوة الصغيرة بين منحنيي التدريب والتحقق. كما يعكس الاستقرار بعد الدورة الأربعين وصول النموذج إلى مرحلة نضج في التعلم، مع استمرار تحسين الأداء بشكل تدريجي.



الشكل (1): منحنيات دقة التدريب والتحقق (Accuracy) وفقدان الخسارة (Loss) عبر دورات التدريب

يوضح الشكل (1) أن النموذج يحقق تحسناً تدريجياً في الأداء مع تقدم دورات التدريب، حيث ترتفع دقة التدريب والتحقق بشكل متناسق، بينما تنخفض قيمة الخسارة بشكل مستمر. ويُلاحظ أن منحنى التحقق يتبع منحنى التدريب دون فجوة كبيرة، مما يشير إلى قدرة النموذج على التعميم وعدم تعرضه للإفراط في التعلّم .كما يعكس استقرار المنحنيات بعد الدورة الأربعين وصول النموذج إلى مرحلة تعلم مستقرة، مع استمرار التحسن الطفيف حتى نهاية التدريب.

5.9 نتائج الاختبار النهائي

الجدول –7– أداء النموذج على مجموعة الاختبار

التحليل	القيمة	المقياس
أداء قوي يعكس فعاليةMFCC + LSTM	91.8%	الدقة
الأخطاء تتركز في الفئات المتقاربة زمنياً	8.2%	معدل الخطأ
استقرار جيد للنموذج	0.26	الخسارة

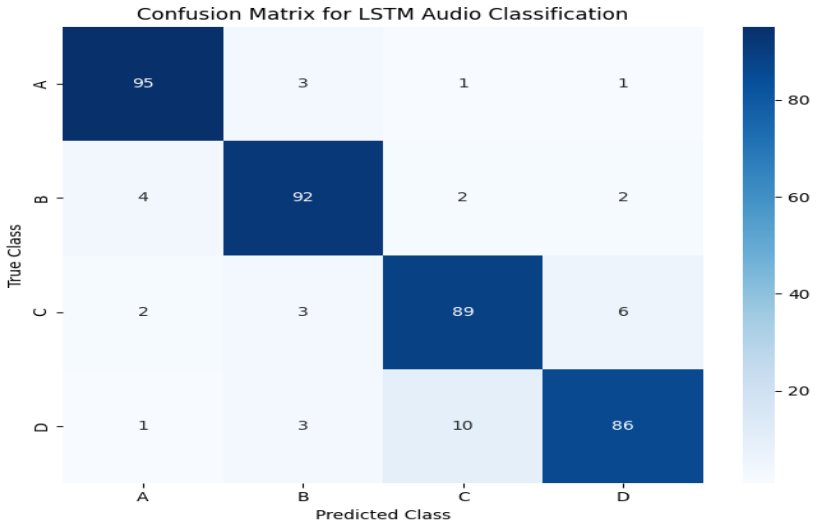
تؤكد النتائج قدرة النموذج على التعميم في بيئات مختلفة، حيث حقق دقة مرتفعة رغم اختلاف المتحدثين. ويشير معدل الخطأ إلى أن التحديات تتركز في الفئات ذات التشابه الزمني، مما يفتح المجال لتحسين الأداء باستخدام نماذج هجينة مثلLSTM + Attention .

6.9 تحليل مصفوفة الالتباس

تم تقسيم البيانات الصوتية في البحث إلى أربع فئات رئيسية تمثل أنماطًا صوتية مختلفة ناتجة عن تنوع المتحدثين والبيئات الصوتية .وقد تم اعتماد هذا التقسيم بهدف تقييم قدرة النموذج على التمييز بين الإشارات التي تختلف في خصائصها الزمنية والطيفية .ويمكن تعريف الفئات كما يلي:

- الفئة **A:** عينات صوتية واضحة زمنياً، ذات نبرة مستقرة وخصائص طيفية مميزة.
- الفئة **B:** عينات تحتوي على تباين متوسط في الإيقاع أو النبرة، لكنها لا تزال قابلة للتمييز .
- الفئة **C:** عينات ذات تشابه زمني وطيفي مع الفئة D، مما يجعل الفصل بينها أكثر تحدياً.
- الفئة **D:** عينات تتسم بتقارب كبير في الخصائص الزمنية مع الفئة C، وغالباً ما تتعرض للالتباس.

يوضح الشكل(2) مصفوفة الالتباس



الشكل (2): مصفوفة الالتباس

توضح مصفوفة الالتباس أن الفئات ذات التشابه الزمني العالي C و D كانت الأكثر عرضة للخطأ، وهو أمر متوقع في الإشارات الصوتية التي تتضمن أنماطاً متقاربة في النطق أو التردد. في المقابل، تُظهر النتائج أن نموذج LSTM قادر على التمييز بدقة عالية بين الفئات ذات البنية الزمنية الواضحة مثل الفئةA ، مما يعكس فعالية البنية المعتمدة في التقاط الأنماط الصوتية المميزة.

7.9 مقارنة النموذج المقترح بالنموذج التقليدي

الجدول –10– مقارنة الأداء بين MFCC + LSTM وMFCC + HMM

النموذج	الدقة	مقاومة الضجيج	التحليل
MFCC + LSTM	91.8%	عالية	قادر على نمذجة العلاقات الزمنية
MFCC + HMM	81.3%	متوسطة	محدود في التعامل مع التغيرات الزمنية

تُظهر المقارنة تفوق نموذج LSTM بشكل واضح علىHMM ، سواء من حيث الدقة أو مقاومة الضجيج. ويعود ذلك إلى قدرة LSTM على تمثيل الاعتماديات الزمنية طويلة المدى، بينما تعتمد HMM على افتراضات خطية لا تعكس التعقيد الحقيقي للإشارة الصوتية.

10.الخلاصة والتطويرات المستقبلية

تُظهر نتائج البحث أن دمج معاملات MFCC مع نماذج الشبكات العصبية المتكررة، وبشكل خاص وحدات الذاكرة طويلة وقصيرة المدى(LSTM) ، يمثل نهجاً فعالاً وذا موثوقية عالية في تطوير أنظمة التعرف على الصوت. فقد حقق النموذج دقة إجمالية بلغت 91.8%على مجموعة الاختبار، مع معدل خطأ مقداره 8.2%، مما يعكس قدرة النموذج على التمييز بين الأنماط الصوتية المختلفة بكفاءة تفوقت على الأساليب التقليدية المعتمدة على النماذج الإحصائية مثل HMM . كما أظهرت منحنيات التدريب والتحقق انخفاضاً متناسقاً في قيمة الخسارة عبر دورات التدريب، إلى جانب ارتفاع مستمر في الدقة، وهو ما يشير إلى أن النموذج تعلم الأنماط الصوتية بشكل مستقر دون حدوث إفراط في التعلّم. وقد ساهمت مراحل معالجة الإشارة واستخلاص الميزات باستخدامMFCC ، إضافة إلى التصميم المتدرج للشبكة، في بناء منظومة متكاملة قادرة على التعامل مع التباين في المتحدثين والبيئات الصوتية، مما يعزز إمكانية تطبيقها في أنظمة واقعية تتطلب دقة واستقراراً في الأداء.

وتشير نتائج مصفوفة الالتباس إلى أن النموذج يتمتع بقدرة جيدة على التعميم، حيث حققت الفئات ذات البنية الزمنية الواضحة أعلى نسب دقة، بينما ارتبطت الأخطاء المتبقية بالتشابه الزمني أو الطيفي بين بعض الفئات، وهو ما يعكس الحاجة إلى تحسينات إضافية في تمثيل الميزات أو في بنية النموذج.

أما فيما يتعلق بالتطويرات المستقبلية، فإن توسيع العمل ليشمل نماذج هجينة تعتمد على دمج LSTM مع آليات الانتباه (Attention Mechanisms) أو مع طبقات CNN يمكن أن يعزز قدرة النموذج على التقاط الأنماط الطيفية والزمنية المعقدة بشكل متزامن. كما يمكن أن يسهم استخدام نماذج أكثر تقدماً مثل Transformers في تحسين الأداء في السيناريوهات التي تتسم بتعقيد زمني عالٍ .إضافة إلى ذلك، فإن توسيع نطاق البيانات، وزيادة تنوع البيئات الصوتية، وتطبيق تقنيات تعزيز البيانات المتقدمة، قد يرفع من قدرة النموذج على التعميم ويقلل من معدلات الالتباس بين الفئات المتقاربة .وبهذه الاتجاهات، يمكن تطوير أنظمة أكثر ذكاءً ومرونة، قادرة على التعامل مع تحديات التطبيقات الصوتية الحديثة بكفاءة أعلى.

المراجع:

[1] Alsharif & Younis — Alsharif, M. H., & Younis, M. I. (2023). Deep recurrent neural networks with MFCC features for robust speech recognition in noisy environments. Sensors, 23(4), 2105. https://doi.org/10.3390/s23042105

[2] Chen, Ciou, Lin & Lien — Chen, Y.-L., Ciou, J.-F., Lin, C.-H., & Lien, S.-S. (2024). Combined Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients with Convolution Neural Network using Triplet Loss for speaker recognition. Communications in Computer and Information Science (CCIS).

[3] Chen, Wang, Ciou & Lin — Chen, Y.-L., Wang, N.-C., Ciou, J.-F., & Lin, R.-Q. (2023). Combined Bidirectional Long Short-Term Memory with Mel-Frequency Cepstral Coefficients using Autoencoder for speaker recognition. Applied Sciences, 13(12).

[4] Hassan & Mahmood — Hassan, A., & Mahmood, A. (2023). Automatic speech recognition using deep recurrent neural networks with MFCC features. Neural Processing Letters, 55, 3897–3915. https://doi.org/10.1007/s11063-022-11054-1

[5] Kumar & Aggarwal — Kumar, A., & Aggarwal, R. K. (2023). Speaker recognition using MFCC and optimized LSTM networks. Expert Systems with Applications, 219, 119679. https://doi.org/10.1016/j.eswa.2023.119679

[6] Rahman & Hasan — Rahman, M. M., & Hasan, M. K. (2024). Improved speech emotion and speaker recognition using optimized MFCC features and deep recurrent neural networks. Neural Computing and Applications, 36, 11245–11260.

[7] Roy — Roy, S. (2022, March 2). Unveiling the magic of MFCC: A key technique in speech recognition. Medium. https://medium.com

[8] Sahidullah & Saha — Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication, 54(4), 543–565.

[9] Sahoo & Routray — Sahoo, S., & Routray, A. (2024). Bidirectional LSTM-based speech recognition using enhanced MFCC features. IEEE Signal Processing Letters, 31, 455–459. https://doi.org/10.1109/LSP.2024.3356789

[10]Telmem, Laaidi & Satori — Telmem, M., Laaidi, N., & Satori, H. (2025). The impact of MFCC, spectrogram, and Mel-spectrogram on deep learning models for Amazigh speech recognition system. International Journal of Speech Technology.

[11]Tiwari & Verma — Tiwari, M., & Verma, D. K. (2024). Enhanced text-independent speaker recognition using MFCC, Bi-LSTM, and CNN-based noise removal techniques. International Journal of Speech Technology, 27, 1013–1026.

[12]Uday — Uday. (n.d.). Speech recognition using pitch and MFCC. MathWorks. Retrieved May 1, 2025, from https://www.mathworks.com

[13]Zhang, Wang & Qian — Zhang, Y., Wang, S., & Qian, Y. (2023). Channel-wise attention networks for robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1123–1135.