تصنيف صوتيمات اللغة العربية واستخلاص واصفاتها باستعمال التعلّم العميق بهدف كشف أخطاء النطق

م. إلهام المفشي[1]

د. أميمة الدكاك[2] د. ندى غنيم[3]

الملخص

يعد كشف النطق الخاطئ من الأمور الهامة في نظم تعلّم اللغات بمساعدة الحاسوب Computer-Aided Language Learning (CALL)، حيث يساعد تحديدُ أماكن الخطأ في النطق متعلّم اللغة في الحصول على تقييم دقيق لصحة اللفظ. وقد لاقت هذه النظم اهتماماً كبيراً لأنها تمكن متعلّمي اللغة من تحسين قدراتهم اللغوية دون الحاجة للتواصل مع المختصين اللغويين بشكل مباشر، وذلك بالاستفادة من وسائل التعلّم الحديثة والتقنيات المتطورة. يهدف هذا البحث لإيجاد المنهجيات المناسبة في بناء نظام يساعد متعلّم اللغة على كشف أخطاء النطق وتصحيحها، وذلك بدراسة إمكانية تصنيف الصوتيمات وتمييزها آلياً وخاصة المتشابه والمشترك منها في مخرج الصوت وبعض صفات الحروف، إضافة لدراسة إمكانية تمييز واصفات النطق الكلامية بهدف استخدامها في كشف خطأ النطق وتحديد نوعه. وقد تم التركيز في البحث على اللغة العربية بهدف تقليص الهوة البحثية الموجودة بين التقانات اللغوية الداعمة للّغة العربية ومثيلاتها في اللغات العالمية والتي حققت تقدماً كبيراً في العديد من المجالات.

الكلمات المفتاحية: تعلّم اللغة بمساعدة الحاسوب، واصفات النطق الكلامية، التعرّف على الصوتيمات، شبكات التعلّم العميق.

[1] طالبة ماجستير، نظم معطيات كبيرة، المعهد العالي للعلوم التطبيقية والتكنولوجيا.

[2] مديرة بحوث. رئيسة قسم الاتصالات، المعهد العالي للعلوم التطبيقية والتكنولوجيا.

[3] استاذة مساعدة، كلية الهندسة المعلوماتية، الجامعة العربية الدولية.

Classifying Arabic phonemes and extracting their attributes using deep learning in view of mispronunciation detection

Eng. Elham Almfashi[(1)]
Dr. Oumayma Dakkak[(2)] **Dr. nada Ghneim[(3)]**

Abstract

Mispronunciation Detecting is an important issue in computer-aided language learning (CALL) systems, where locating errors in pronunciation helps the language learner to obtain an accurate assessment of pronunciation correctness, these systems have received great attention because they give language learners the possibility to improve their language proficiency without the need to direct communication with language specialists, by making use of modern learning methods and advanced technologies. This research aims to find appropriate methodologies in building a system that helps the language learner to detect and correct pronunciation errors, By studying the possibility of categorizing phonemes and distinguishing them automatically, especially the similar and common ones in the sound output and some character traits, in addition to studying the possibility of distinguishing speech descriptors in order to use them in detecting pronunciation error and determining its type. The focus of the research was on the Arabic language, with the aim of reducing the research gap that exists between the linguistic technologies that support the Arabic language and its counterparts in international languages, which have achieved great progress in many fields.

Keywords: Computer-Assisted Language Learning CALL, Speech articulatory features, Phoneme recognition, Deep neural networks.

[(1)]Master Student, Big Data, Higher Institute for Applied Science and Technology (HIAST).
[(2)]Research Manager .Head of Communications Department for Educational Affairs, Higher Institute for Applied Science and Technology (HIAST).
[(3)]Assistant Professor, Faculty of Information Technology Engineering, Arab International University.

المقدمة

حازت نظم تعلّم اللغات بمساعدة الحاسوب على قدر كبير من الاهتمام في السنوات الأخيرة بسبب التقدم الملحوظ في تقنيات الذكاء الصنعي والتعلّم الآلي، حيث تهدف هذه الأنظمة إلى استخدام الموارد الحاسوبية والتقنيات الحديثة في تسهيل عملية تعلّم اللغة وحل مشكلة صعوبة التواصل بشكل دائم بين مدرس اللغة ومتعلّمها. وقد تم إجراء العديد من الأبحاث في هذا المجال من أجل العديد من اللغات (الإنكليزية (K. Li et al., 2017)، والصينية (W. Li et al., 2016)، والألمانية، واليابانية، والعربية (Nazir et al., 2019)...). تناولت بعض هذه الأبحاث موضوع كشف النطق الخاطئ وتصحيحه بهدف تحسين النطق عند المتحدثين غير الأصليين للّغة، وتم العمل على التحقق من النطق وفق عدة مستويات بدءاً من مستوى المتحدث بهدف تقييم طلاقة الكلام عند النطق واستخدامه في اختبارات الكفاءة اللغوية المنطوقة، وانتهاءً بمستوى الصوتيم (phoneme) على مستوى كل وحدة صوتية في الكلام، يقدم هذا المستوى معلومات أكثر دقة عن مكان ونوع الخطأ الذي ارتكبه المستخدم مما يرفع من سوية عملية التعلّم. تم اقتراح العديد من المنهجيات لمعالجة مسألة التحقق من النطق على مستوى الوحدة الصوتية في العديد من اللغات، منها ما يعتمد على إعطاء درجات للثقة ومنها ما يعتمد على القواعد ومنها ما يعتمد على أساليب التصنيف والتعلّم العميق.

يهدف هذا البحث لدراسة إمكانية تصنيف الصوتيمات وتمييزها آلياً وخاصة المتشابه والمشترك منها في مخرج الصوت وبعض صفات الحروف إضافة لدراسة أثر استخدام مميزات النطق الكلامية للمساعدة في بناء نظام يساعد متعلّم اللغة على كشف أخطاء النطق وتصحيحها.

نتناول في تتمة هذه الورقة في قسم **الدراسة المرجعية** المنهجيات المتبعة في مجال البحث، ثم نشرح **منهجية البحث المتبعة** وما فيها من مراحل، ونعرض في قسم **الجزء العملي والنتائج** التجارب التي قمنا بها والنتائج التي حصلنا عليها، وأخيراً نعرض في قسم **الخاتمة والتوصيات** خلاصة البحث والآفاق المستقبلية له.

1. الدراسة المرجعية

تم العمل وفق عدة منهجيات لكشف الأخطاء الصوتية في النطق ويمكن تصنيف هذه المنهجيات حسب عدة عوامل (Chen & Li, 2017) كاعتمادها على وجود نظام تعرّف آلي على الكلام والحاجة لوجود معرفة باللغة الأصلية و لغة المتعلّم إضافة لقدرة المنهجية على كشف وجود خطأ في النطق وتحديد طبيعته ومكانه، وهي حسب هذه العوامل تقسم لمنهجيات تعتمد على درجة الثقة Confidence score-based، ومنهجيات تعتمد على القواعد Rule-based، ومنهجيات تعتمد على المصنفات Classifier-based، ومنهجيات تعتمد على التعلّم العميق Deep neural network-based.

1.2 المنهجيات المعتمدة على درجة الثقة

يتم في هذه المنهجيات حساب درجة للثقة تعبر عن مدى قرب اللفظ المنطوق من اللفظ الصحيح ثم تتم مقارنة النتيجة مع عتبة محددة لأخذ القرار فيما إذا كان النطق صحيح أم لا. تم اعتماد عدة مقاييس للمقارنة ويعد مقياس صحة النطق Goodness Of Pronunciation (GOP) الأكثر استخداماً (Witt & Young, 2000) حيث يعتمد هذا المقياس على حساب نسبة احتمال likelihood ratio لتوافق الصوت المنطوق مع الصوت الصحيح. وقد تم العمل على تحسين جودة النموذج الصوتي المستخدم في هذه المنهجية باعتماد شبكات التعلّم العميق (DBN) Deep Belief Nets بدلاً من النموذج الغوصي (GMM)

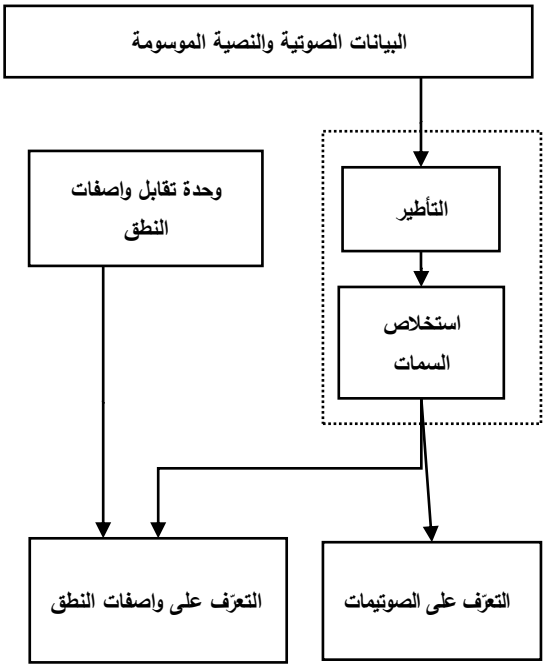
(Hu et (Hu et al., 2015) Gaussian mixture model al., 2013) . إن مقياس GOP حساس جداً لجودة النموذج الصوتي(Witt & Young, 2000)، كما أن عتبة القرار يتم تحديدها بالاعتماد على بيانات التدريب فقط مما جعل من الصعب اعتماد هذه المنهجية للتعميم من أجل طيف واسع من أخطاء النطق، وعلى الرغم من أن هذه المنهجية يمكنها الكشف عن الأخطاء لكنها غير قادرة على تشخيص نوع الخطأ الذي ارتكبه المتعلّم.

2.2 المنهجيات المعتمدة على القواعد

تتطلب هذه المنهجيات معرفة مسبقة بقواعد النطق الخاطئ وتتميز بقدرتها على تحديد مكان ونوع الخطأ الذي قام به المتكلم، إلا أنها محصورة بالأخطاء المعرفة مسبقاً ضمن القواعد حيث لا يمكن للنظام التعرّف على أخطاء جديدة غير معرفة مسبقاً. يتم بناء قواعد النطق الخاطئ يدوياً من قبل خبراء لغويين (Abdou et al., 2006; Harrison et al., 2008; Mao et al., 2018; Meng et al., 2007)، ويتم بناء نموذج صوتي للنطق الصحيح ثم يتم تطبيق القواعد لتوليد نماذج للنطق الخاطئ، بعد ذلك تتم المقارنة بين خرج النموذجين عند اختبار النطق للمتعلّم لتحديد صحة النطق. إن بناء قواعد النطق يتطلب وجود خبرة لغوية إضافة إلى أن هذه القواعد قد لا تغطي كل الحالات لذا تم تحسين هذه الطرق باستنتاج قواعد النطق الخاطئ بشكل أتوماتيكي باستخدام قواعد بيانات لمتعلّم اللغة L2 (Lo et al., 2010).

3.2 المنهجيات المعتمدة على المصنفات

تعتمد هذه المنهجيات على تصنيف الوحدات الصوتية باستخدام مصنفات مختلفة كمصنف آلة شعاع الدعم Support Vector Machine أو أشجار التصنيف Decision Trees أو الشبكات العصبونية Neural Networks. يتم استخلاص سمات مختلفة من الإشارات الصوتية في المرحلة الأولى لتكون دخلاً للمصنفات في المرحلة التالية. أثبتت مصنفات SVM فعاليتها في العديد من الأبحاث(Georgoulas et al., 2006) (Maqsood et al., 2016) (Wei et al., 2009). وفي (Hu et al., 2014) تم اقتراح استخدام مصنف a Neural Network based, Logistic Regression (NN) بالاعتماد على الشبكات العصبونية ومقارنته مع مصنف SVM ومنهجية GOP حيث أعطى نتائج أفضل في الحالتين. كما تم استخدام شبكات التعلّم العميق لتحسين جودة النموذج الصوتي المستخدم في استخلاص سمات الوحدات الصوتية قبل تدريب المصنف. وبصورة عامة فإن هذه المنهجية تعتبر أفضل من كل من المنهجيتين السابقتين إلا أنها بالمقابل تحتاج كمية كافية من بيانات التدريب.



الشكل (1) نموذج العمل المقترح

4.2 المنهجيات المعتمدة على التعلّم العميق في هذه المنهجيات يتم استخدام نوع خاص من الشبكات العصبونية وهي شبكات التعلّم العميق، لنمذجة التجريدات عالية المستوى في البيانات واستخلاص سمات تُحسّن دقة كشف أخطاء النطق كما في الأبحاث (Algabri et al., 2021; Hu et al., 2015; Hussain et al., 2020; K. Li et al., 2017). في بعض الأبحاث تم استخدام الشبكات العصبونية التلفيفية Deep Convolutional Neural Network (CNN) لاستخلاص السمات الممثلة للإشارات الصوتية من الصور الطيفية (spectrograms) (Nazir et al., 2019). وبهدف تحسين كشف خطأ النطق بالاستفادة من شبكات تعلّم مدربة مسبقاً تم العمل وفق طريقة نقل التعلّم transfer learning، فقد أثبتت التجارب أن هذه النماذج المدربة يمكن تعميمها على مهام عديدة كاستخدام نموذج لغة مدرب كمرحلة أولية في التصنيف في لغة أخرى (Duan, Kawahara, Dantsuji, & Nanjo, 2017) وهذا يوفر الوقت ويساعد بشكل كبير في حال وجود موارد لغوية محدودة (Duan, Kawahara, Dantsuji, & Zhang, 2017). كما تم استخدام شبكات التعلّم العميق في بعض الأبحاث لتحسين دقة النموذج الصوتي (Al-Marri et al., 2018; Shahin et al., 2014).

بالنسبة للسمات المستخدمة في الأبحاث فتنقسم إلى سمات صوتية يتم استخلاصها من البيانات مثل (MFCC, Mel spectrogram, statistical features…)، وتعتبر معاملات ميل الطيفيّة Mel Frequency Cepstral Coefficients الأكثر استخداماً وجدوى كونها تحاكي استجابة النظام السمعي للإنسان حيث أثبتت فعاليتها في العديد من الأبحاث، أما السمات النطقية فتم توظيفها حديثاً في مجال المعالجة الآلية للصوت، وتسمى واصفات النطق الكلامية وهي خصائص صوتية تصف آلية نطق الحروف ومخارجها وتستطيع أن تميز صوت عن صوت آخر (Noory, 2007)، وهذا يوفر لمتعلّم اللغة معلومات أدق عن صحّة نطقه ويساعده في تحديد موضع الخطأ وتصحيحه.

2. منهجية البحث المتبعة

في هذا البحث تم اقتراح العمل وفق منهجية التعلّم العميق مع استخدام واصفات النطق الكلامية، إذ تتميز هذه الواصفات بفعاليتها في التطبيقات التي تتغير فيها الإشارات الصوتية بتغيّر المتكلم أو بوجود ضجيج في البيئة المحيطة وذلك مقارنة بالواصفات التقليدية، إضافة إلى أنها تساعد بشكل فعّال على تحديد خطأ النطق، حيث يتم اعتبار اللفظ خاطئاً عند وجود تغير في أحد مميزاته الأساسية، وهكذا تصبح عملية التعلّم أسرع وأفضل. وقد جرى التركيز على اللغة العربية نظراً لمكانتها وأهميتها بين اللغات بهدف تطوير أنظمة مساعدة لتعليمها. ويبين الشكل **(1)** منهجية العمل المتبعة.

تتضمن منهجية البحث الخطوات التالية:

- تحضير المعطيات ومعالجتها.
- بناء شبكات التعليم العميق وتدريبها.
- اختبار مدى كفاءة الشبكات المدربة.

1.3 تحضير المعطيات ومعالجتها

جرى العمل على معطيات من عدة مدوّنات سنتكلم عنها في فقرة **قواعد المعطيات المستخدمة للتدريب**. تتم في هذه المرحلة قراءة الملفات الصوتية للجمل المنطوقة، إضافة لقراءة معلومات تقسيم هذه الجمل إلى صوتيماتها (البيانات النصية الموسومة[1])، بعد ذلك يجري استخلاص السمات الصوتية لكل إطار زمني وفق نافذة زمنية بطول 25ms وانزياح زمني 10ms، حيث تعد الإشارة الكلامية مستقرة ضمن النافذة الزمنية السابقة.

جرى استخلاص معاملات MFCC (Mel Frequency Cepstral Coefficient) إضافة إلى معامل لوغاريتم الطاقة، ومشتقات هذه المعاملات من الدرجة الأولى والثانية (Δ, $\Delta\Delta$) لكل إطار صوتي. يتم حساب معاملات MFCC بالاعتماد على تحويل فوريه السريع (FFT) باستخدام مرشح تمرير حزمة كما في المعادلة (1) (Lu et al., 2003)

$$MFCC(t,k) = \sqrt{\frac{2}{N}}\sum_{n=1}^{N} log\ p_n \cos\left(k(n-0.5)\frac{\pi}{N}\right) \quad (1)$$

تمثل N عدد مرشحات تمرير الحزمة، p_n تمثل طاقة الخرج من المرشح رقم n في الزمن t، $k = 1,2,3 \dots L$ حيث L عدد معاملات MFCC، تضاف بعد ذلك معلومات السياق للإطار الزمني الحالي بمعدل 5 إطارات زمنية سابقة و5 إطارات زمنية لاحقة ليصبح طول شعاع المميزات 11 إطار زمني، وتم اعتماد هذا العدد بعد إجراء عدة تجارب (دون سياق، 3 إطارات سياق، 5 إطارات سياق) للتعرّف على الصوتيمات، والحصول على أفضل نتيجة باستخدام 11 إطار زمني. تتمثل إحدى المزايا المهمة لمعاملات MFCC أنها مميّزة للكلام المنطوق وهذا ما دفع لاستخدامها في أنظمة تعرّف الصوت.

بالنسبة لوحدة تقابل واصفات النطق، يجري فيها تحضير البيانات اللازمة عن طريق ربط كل واصفة مع الصوتيمات التابعة لها في قاعدة البيانات المدروسة وفق جدول تقابل محدد. نحصل في خرج هذه الوحدة على وسم لكل صوتيم يحدّد إذا كانت الواصفة موجودة أم لا في جملة الواصفات التي يتم تدريب النظام عليها.

2.3 تعرّف الصوتيمات

بعد استخلاص السمات الصوتية والتي سنقوم بذكر تفاصيلها في الجزء العملي لكل مجموعة معطيات، يتم إدخالها لوحدة التعرّف على الصوتيمات وتصنيفها باستخدام شبكة تعلّم عميق ذات الذاكرة الطويلة القصيرة الأمد Long-Short Term Memory (LSTM). جرى توظيف هذا النوع من الشبكات للتعامل مع البيانات المتسلسلة sequential data في العديد من التطبيقات وحققت نتائج جيدة فيها (Chao, 2015; Learning et al., 2011). تم إدخال مفهوم "بوابة" في هذه الشبكات لتحسين قدرة التذكر فيها، وتتكون الخلية العصبونية في ذواكر LSTM من بوابة إدخال input gate، وبوابة إخراج output gate، وبوابة نسيان forget gate. تتم في كل خلية العمليات التالية:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi})$$
$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$$
$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg})$$
$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho})$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh(c_t)$$

حيث:

x_t دخل الخلية في اللحظة t
h_t الحالة الخفية في اللحظة t
c_t حالة الخلية في اللحظة t
i_t, f_t, g_t, o_t قيمة بوابة الدخل والنسيان والخلية والخرج

[1] البيانات الموسومة تتضمن الحدود الزمنية لكل صوتيم ضمن الجملة المنطوقة (بدايته ونهايته) مع الوسم label المقابل له.

على الترتيب في اللحظة t
W_{ii} الوزن الذي يربط بين الوحدتين i,j ، b الانحياز
σ تابع sigmoid
$\odot$ جداء هارمارد Hadamard

قبل البدء بعملية التدريب يتم توحيد طول البيانات لكل تسلسل وفق أطول جملة منطوقة في كتلة البيانات المدربة وذلك بإضافة حشو صفري للتسلسل الزمني (Zero padding) يتكون من أصفار تضاف في بداية التسلسل أو نهايته.

خلال عملية التدريب وبفرض x_t تمثل سمات الدخل للصوتيمات المتتالية في إشارة الصوت، تُطبّق العمليات السابقة بشكل متكرر لحساب خرج الطبقة الخفية h_t ليتم استخدامها في حساب التوزع الاحتمالي للصوتيمات (احتمال توقع الصوتيم) عبر المعادلة (2):

$$\hat{y} = Wh_t + b \quad (2)$$

حيث W, b تمثل مصفوفة الأوزان والانحياز لطبقة الخرج الخطية.

يجري تدريب الشبكة لتحقيق أقل خسارة ممكنة لتابع الإنتروبية التقاطعية وفق المعادلة (3):

$$\text{Loss} = -\sum_{c=1}^{M} y_{o,c} \log(\hat{y}_{o,c}) \quad (3)$$

حيث:

M عدد الأصناف

log اللوغاريتم الطبيعي

c محدد ثنائي يدل على صحة التصنيف للصنف y

oضمن العينة

$\hat{y}$ احتمال توقع العينة o للصنف c

3.3 تعرّف واصفات الكلام

تعد واصفات الكلام من السمات المميّزة والهامة كونها تمثّل عملية نطق الكلام بشكل غني ومفيد، ويعبر عنها بمخارج الحروف وصفاتها. ما يميز هذه الصفات أنها تساعد في كشف النطق الخاطئ، حيث يؤدي وجود تغير في أحد هذه الواصفات لحدوث تغير في صوت الحرف وبالتالي خطأ في نطقه. كما أنها مشتركة بين اللغات، وهذا يُمَكّن من الاستفادة من قواعد بيانات صوتية من لغات مختلفة لتدريب نموذج صوتي لهذه الواصفات. يتم تدريب هذه الواصفات باستخدام مصنفات ثنائية، كون المسألة تهدف لتصنيف الإطارات الصوتية الحاملة للواصفة عن باقي الإطارات، حيث يحتمل كل إطار صوتي حالتين فقط لكل واصفة (وجود أو عدم وجود الواصفة). ويبين الجدول (1 و

مخارج الحروف	الحروف
بلعومي Pharyngeal	ع ح
حلقي Pharynx	ء ه ع ح غ خ
حنجرية Glottal	ء ه
لهوية Uvular	ق
فوق اللثة Post-alveolar	ج ش
وسط اللسان Middle_tongue	ج ش ي
اقصى اللسان Deep_tongue	ك ق
حافة اللسان Tongue_border	ض ل
طرف اللسان Tongue_tip	ط د ت ص س ز ظ ذ ث ن ر
بين الاسنان Interdental	ظ ث ذ

هذه الواصفات مع الأصوات المقابلة لها في اللغة العربية (Morsy et al., 2018).

الجدول (1) صفات الحروف في اللغة العربية

صفات الحروف	الحروف
الهمس Whisper	فحثه شخص سكت
الصفير Whistle	ص س ز
الاطباق Adhesion	ص ض ط ظ
مركبة Affricates	ج
الانحراف Deviate	ل ر
الاستعلاء Elevation	خص ضغط قظ
احتكاكي Fricatives	ص س ز ظ ذ ه ح ع غ خ ش ف ث
الرخاوة Softness	و ي ض ف غ خ ح ه ص س ش ذ ث ظ ز ا و ي المدية
التفشي Spreading	ش
الشدة Strength	أجد قط بكت
الخفاء Hiding	هاوي
التوسط Moderate	لن عمر
التكرار Repetition	ر
الاستطالة Prolongation	ض

الصوائت Vowels	ا و ي المدية
الصمت Silence	-

3. الجزء العملي والنتائج

تم إجراء التجارب باستخدام المخدم السحابي Google Colab (Google Colaboratory, n.d.) وذلك بهدف الاستفادة من قدرات التخزين والمعالجة الحسابية GPU التي تقدمها Google مجاناً، والتي تساعد في التنفيذ السريع عند تدريب شبكات التعلّم العميق. وقد استخدمت لغة بايثون لعمليات معالجة البيانات وتدريب الشبكات.

الجدول (2) مخارج الحروف في اللغة العربية

مخارج الحروف	الحروف
بلعومي Pharyngeal	ع ح
حلقي Pharynx	ء ه ع ح غ خ
حنجرية Glottal	ء ه
لهوية Uvular	ق
فوق اللثة Post-alveolar	ج ش
وسط اللسان Middle_tongue	ج ش ي
اقصى اللسان Deep_tongue	ك ق
حافة اللسان Tongue_border	ض ل
طرف اللسان Tongue_tip	ط د ت ص س ز ظ ذ ث ن ر
بين الاسنان Interdental	ظ ث ذ

تم العمل على Google Colab مع استخدام GPU حيث يقدم ذاكرة تخزين عشوائية RAM بحجم 12GB إضافة لحجم القرص الصلب 68GB، إضافة لذلك تم إجراء بعض عمليات المعالجة والتجارب على حاسب بمعالج intel core i7 وذاكرة 8GB.

1.4 قواعد المعطيات المستخدمة للتدريب Speech Corpora

تم العمل على ثلاث قواعد معطيات لاختبار منهجية العمل المقترحة اثنتين منها باللغة العربية وواحدة باللغة الإنجليزية وفيما يلي وصف مختصر عن كل منها:

i. Arabic Speech Corpus قاعدة بيانات النطق بالعربية MSA

هي قاعدة بيانات صوتية للغة العربية الفصحى تم بناؤها كجزء من أطروحة دكتوراة نوار حلبي (Halabi, 2016) في جامعة ساوثهامتون. وتحوي تسجيلات صوتية بأطوال زمنية مختلفة عددها 1913 مدتها حوالي 4 ساعات لمتحدث عربي الأصل مع النص المقابل لها. تم تقسيم هذه البيانات على مستوى الصوتيم (phoneme) واستخدمت لتركيب كلام منطوق آلياً ذي جودة عالية، كما يمكن استخدامها في تطبيقات صوتية أخرى، لم يسبق العمل على هذه البيانات لنفس هدف البحث، لكن تم توظيفها في أنظمة أخرى، منها نظام لنمذجة المدة الزمنية للكلام (Zangar et al., 2018)، ونظام تحويل نص لكلام بالاعتماد على التعلّم العميق (Fahmy et al., 2020). تضم هذه القاعدة 82 صوتيماً تم تقليصها في مجموعة تضم 38 صوتيماً بعد دمج الصوت والصوت المشدّد منه، ودمج الحركة وحرف المد التابع لها (مثل الفتحة والألف)، إضافة لاعتبار الأصوات الأجنبية مثل (P,v,G ,...) مع التشويش صوتيماً واحداً. ويوضح الجدول (3 هذه الصوتيمات.

ii. Texas Instruments Massachusetts Institute of Technology (TIMIT)

قاعدة معطيات صوتية معروفة وشائعة لمتحدثين باللغة الإنجليزية الأمريكية (Lamel et al., 1993). تم تطويرها من قبل شركة Texas Instruments (TI) ومعهد ماساتشوستس للتكنولوجيا (MIT) ومعهد ستانفورد للأبحاث (SRI). تضم TIMIT تسجيلات لـ 630 متحدث بثماني لهجات مختلفة لكل من الجنسين، قام كل منهم بقراءة عشر جمل كل جملة بطول 30 ثانية، لتكون المدة الإجمالية لجميع التسجيلات حوالي 5.4 ساعة. تم تقسيم هذه القاعدة إلى بيانات تدريب تضم 462 متحدث وبيانات

اختبار تضم بقية المتحدثين البالغ عددهم 168. تحوي هذه القاعدة 45 صوتيماً تم تقليصها في مجموعة تضم 39 صوتيم كما اقترح (Lee & Hon, 1989). استخدمت هذه القاعدة في العديد من الأبحاث كونها مقطعة وموسومة على مستوى الصوتيم بشكل يدوي.

iii. KACST Arabic Phonetic Database (KAPD)

هي قاعدة بيانات صوتية تم إنشاؤها من قبل مدينة الملك عبد العزيز للعلوم والتقنية King Abdul-Aziz City for Science and Technology (KACST) في سنة 2003. تحوي على تسجيلات صوتية لسبع متحدثين أصليين للّغة العربية لما يقارب 1.2 ساعة، تضم 35 صوتيماً. جرى العمل في هذا البحث على النسخة المطورة من القاعدة التي تم تطويرها وتقطيعها يدوياً على مستوى الصوتيم، بهدف استخدامها في تطبيقات التعلم الآلي والتنقيب عن المعطيات كما يوضح البحث (Seddiq et al., 2016). تنقسم القاعدة إلى مجموعة تدريب بنسبة 71.4% ومجموعة اختبار بنسبة 28.6%.

2.4 مقاييس التقييم

تم إجراء التقييم على مستويين الأول هو مستوى الإطار الزمني frame والثاني هو مستوى الصوتيم phoneme

الجدول (3) صوتيمات قاعدة بيانات اللغة العربية

بيانات التدريب	Frame Accuracy		98.75
	Phoneme Accuracy		99.78
بيانات التحقق	Frame Accuracy		91.08
	Phoneme Accuracy		94.91
بيانات الاختبار	Frame	Accuracy	91.50
	Phoneme		95.62
	Frame	Precision	92.64
	Phoneme		96.34
	Frame	Recall	91.50
	Phoneme		95.62
	Frame	F1	91.50
	Phoneme		95.52

يتم التقييم على مستوى الصوتيم بالاعتماد على نتيجة تصنيف إطاراته المكونة له حيث يتم اعتماد صنف الإطارات الأغلبية majority. تم اعتماد مقاييس الضبط accuracy، والدقة precision، والاسترجاع recall، والتقييم الشامل f1، ومعدل خطأ الصوتيم Phoneme error rate (PER) عند عمل التجارب.

3.4 تعرّف الصوتيمات

تم في هذه المرحلة بناء نموذج تعرّف الصوتيمات وتدريب شبكة تعلّم عميق ذات الذاكرة الطويلة القصيرة الأمد LSTM في كل قاعدة معطيات.

a. تعرّف الصوتيمات في قاعدة بيانات النطق بالعربية MSA

تم استخلاص السمات الصوتية (24 معامل MFCC) إضافة إلى معامل لوغاريتم الطاقة، ومشتقات هذه المعاملات ال 25 من الدرجة الأولى والثانية (Δ, ΔΔ)، ليصبح لدينا شعاع سمات بطول 75 لكل إطار صوتي، وبعد إضافة 11 إطار زمني كمعلومات للسياق يصبح طول شعاع السمات 825 سمة لكل إطار زمني، بعد ذلك تم تدريب شبكة التعلّم العميق LSTM، وفق خوارزمية الأمثلة ADAM بمعامل تعلّم 0.001 (learning_rate) (تم اختياره بعد تجريب قيمتي 0.1 و 0.01 أيضاً)، وتم تقسيم بيانات التدريب بنسبة 60% training تدريب، 20% تحقق Validation، 20% اختبار testing. يبين الجدول 4 نتائج تدريب الشبكة بطبقة خفية واحدة حجمها 2048 عصبون بعد 20 دورة تدريب epoch.

الجدول (4) مقاييس التقييم للتعرّف على الصوتيمات في MSA باستخدام شبكة LSTM

الصوتيم	الترميز	الصوتيم	الترميز
ء	>	ف	f
ب	b	ق	q
ت	t	ك	k
ث	^	ل	l
ج	j	م	m

ح	H	ن	n
خ	x	هـ	h
د	d	و	w
ذ	*	ي	y
ر	r	ألف مدية	a
ز	z	ألف مدية جوارها مفخم	A
س	s	واو مدية	u0
ش	$	واو مدية جوارها مفخم	U0
ص	S	واو مدية ممالة	u1
ض	D	ياء مدية	i0
ط	T	ياء مدية جوارها مفخم	I0
ظ	Z	ياء مدية ممالة	i1

لاحظنا أن معظم الأخطاء في التعرّف كانت في الصوتيمات التي تتصف بصفة الشدة، حيث يكون زمنها قصيراً مقارنة بباقي الصوتيمات. إضافة لوجود لبس في التعرّف بين الصوتيمات المتشابهة مثل (ط–ت) (ض– د) (ظ–ذ)، كون هذه الأصوات تشترك بالمخرج وبعض صفات الحروف. فمثلاً الطاء والتاء تخرجان من طرف اللسان وتشتركان بصفة الشدة، إلا أن الطاء تتميز عنها بوجود الاستعلاء والإطباق فيها، بينما تتصف التاء بالهمس. بالنسبة لحروف المد لاحظنا فيها وجود خطأ تعرّف بين صوت حرف المد وصوت حرف المد الذي يجاوره حرف مفخم لتشابه النطق فيهما. أما بالنسبة للتشويش كانت نسبة التعرّف فيه منخفضة لعدم وجود عدد عينات كاف منه في بيانات التدريب والاختبار.

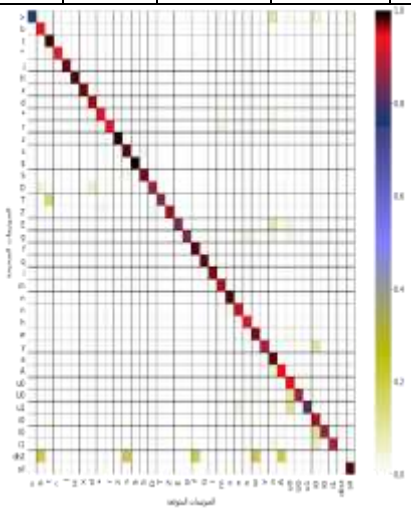
تتضح هذه الأخطاء في الشكل 2 لمصفوفة الدقة confusion matrix للصوتيمات.

b. تعرّف الصوتيمات في قاعدة بيانات TIMIT

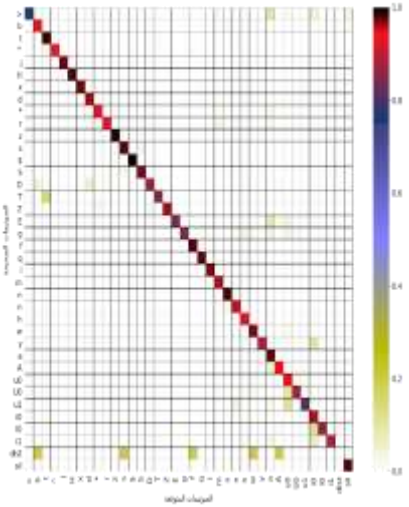
تم تدريب شبكة تعلّم عميق لتصنيف الصوتيمات 39 باعتماد مجموعة اختبار من 24 متحدث core test set من أجل عملية الاختبار إضافة لمجموعة الاختبار الكاملة المكونة من 168 متحدث. قمنا باستثناء جمل اللهجة المسمّاة SA من قاعدة البيانات قبل التدريب كما هو موصى به في (Lopes & Perdigao, 2011). تم استخلاص السمات الصوتية (13 معامل MFCC)، ومشتقاتها من الدرجة الأولى والثانية (Δ, ΔΔ)، ليصبح لدينا شعاع سمات بطول 39 لكل إطار صوتي، وبعد إضافة سمات 5 إطارات مجاورة سابقة ولاحقة يصبح طول شعاع السمات 429. تتكون شبكة التدريب من طبقتين خفيتين بحجم 2048، ويوضح **الجدول 5** نتائج تدريب هذه الشبكة.

الجدول (5) مقاييس التقييم لتعرّف الصوتيمات في TIMIT باستخدام شبكة LSTM

Core Testing		Testing		
Phon-eme	frame	Phon-eme	frame	
81.71	79.57	82.08	80.01	Accuracy
84.14	83.99	83.68	81.00	Precision
81.71	79.57	82.08	80.01	Recall



يوضح معدل خطأ الصوتيم (Phoneme error rate) PER لتعرّف صوتيمات TIMIT لمجموعة الاختبار core test set مقارنة مع بعض الأعمال.



الشكل (2) مصفوفة الدقة للصوتيمات

الجدول (5) مقارنة خطأ تعرّف الصوتيم فيMIT مع نماذج مختلفة

العمل	PER(%)
CenterNet-DLA (Algabri et al., 2020)	15.89
PYTORCH-KALDI (Ravanelli et al., 2018)	13.8
CNN (Abdel-hamid et al., 2014)	20.36
HMM (Bhowmik & Mandal, 2018)	22.39
LSTM(هذا العمل)	18.92

نجد بالمقارنة أن أداء شبكة LSTM عند تعرّف الصوتيمات في TIMIT أفضل من كل من HMM وCNN. بينما حصل العمل في PYTORCH-KALDI على أفضل أداء وذلك باستخدام مجموعة مختلفة من الشبكات مثل Li-GRU وMLP، واستخدام مجموعة مختلفة من السمات مثل MFCC وFBANK وfMLLR، كما حقق العمل باستخدام شبكة CenterNet نتيجة جيدة بالاعتماد على تقنية كشف الأغراض.

c. تعرّف الصوتيمات في قاعدة بيانات KAPD

تم اختيار 10% من بيانات التدريب لعملية التحقق كما في (Algabri et al., 2021) واستخلاص السمات الصوتية (13 معامل MFCC) إضافة إلى معامل لوغاريتم الطاقة، ومشتقات المعاملات من الدرجة الأولى والثانية (Δ, ΔΔ)، مع إضافة سمات الإطارات المجاورة (5 سابقة و5 لاحقة)، تم استخدام هذه السمات لتدريب شبكة تعلّم عميق LSTM بطبقة واحدة عدد العصبونات فيها 1024. فيما يلي الجدول (6 نتائج التدريب لبيانات الاختبار لهذه الشبكة.

الجدول (6) مقاييس التقييم للتعرّف على الصوتيمات في KAPD باستخدام شبكة LSTM

Phoneme	Frame	
90.67	85.87	accuracy
91.37	88.76	precision
90.67	85.87	recall
90.23	86.00	f1

يبين الجدول 8 معدل خطأ الصوتيم في KAPD مقارنة مع (Algabri et al., 2021)

الجدول (7) معدل خطأ الصوتيم في KAPD

المنهجية	PER (%)
AFD-Obj(YOLOv3-tiny-1S)	10.84
PD-Obj (YOLOv3-tiny-2S)	5.63
LSTM(هذا العمل)	9.33

إن النظام PD-Obj حقق معدل خطأ أقل وذلك باستخدام تقنية كشف الأغراض بالاعتماد على الصور الطيفية للإشارة الصوتية، بينما تفوقت شبكة LSTM على نظام AFD-Obj الذي يعتمد على واصفات النطق للكشف عن الأغراض بهدف تصنيف الصوتيمات.

مما سبق نجد أن أداء شبكة LSTM في عملية التصنيف جيد، لكن تحتاج الشبكة لسمات صوتية إضافية

قادرة على دعم عملية التصنيف من جهة، وكشف وجود خطأ النطق من جهة أخرى.

4.4 تعرّف واصفات الكلام

تم بناء نموذج للتعرّف على صفات ومخارج الحروف، كون هذه الواصفات تعطي معلومات أدق عن اللفظ وتساعد في معرفة خطأ النطق وتحديد مكانه.

تتم في البداية عملية تقابل بين كل صوتيم والواصفة المقابلة له بحيث تمثل جميع إطارات الصوتيم النماذج الحاملة للواصفة المدروسة، وتمثل إطارات باقي الصوتيمات النماذج غير الحاملة للواصفة، بعد ذلك يتم إدخال سمات الإطارات الصوتية التي تم استخلاصها في المرحلة السابقة مع الوسم الذي تم الحصول عليه من عملية التقابل لشبكة تعلّم عميق ذات مصنف ثنائي binary DNN classifier، تميز وجود الواصفة أو عدمها لكل إطار زمني. تم تدريب المصنفات بالاعتماد على التعلّم المتعدد المهام Multi-task learning. تعتمد هذه الطريقة في التعليم على تدريب شبكة على عدد من المهام المرتبطة، بمشاركة موسطات النموذج المدرب وطبقاته الخفية Hard parameter sharing، مع وجود طبقة خرج منفصلة لكل مهمة، ويتم تطبيق تابع أمثلة محدد للحصول على أقل خطأ ممكن. تم اختيار موسطات الشبكات المدربة بالاعتماد على إطار العمل البرمجي التحسيني Optuna(*Optuna*, n.d.) الذي يساعد في البحث بشكل تلقائي عن القيم المثلى للموسطات للحصول على أفضل أداء للشبكة.

i. تعرّف واصفات الكلام في TIMIT

قمنا باعتماد واصفات الكلام للغة الإنجليزية المستخدمة في العمل (Karaulov & Tkanov, 2019) وهي 28 واصفة، وتم استخدام سمات مرحلة تعرّف الصوتيمات (462 سمة) مع وسم الواصفة المقابل لكل إطار صوتيم.

بعد ذلك غذّيت شبكة تعلم عميق ذات مصنف ثنائي بهذه السمات. تتكون شبكة التعلم المدرّبة من ثلاث طبقات خفية بعدد مخارج 28، يمثل كل مخرج سمة مختلفة. ويبين الجدول 8 نتائج التصنيف على مستوى الإطار الصوتي مقارنة مع (Algabri et al., 2021; Karaulov & Tkanov, 2019)، حيث نجد تقارب النتائج الوسطية لهذه الواصفات مع تفوق نموذجنا بنسبة 0.47% و 0.1% على الترتيب، مع ملاحظة كون نتائج بعض الواصفات أفضل في نموذجنا المدرب (Continuant، Voiced،Silence)، إضافة لدقة تصنيف عالية بالنسبة للصمت Silence في هذا العمل.

ii. تعرّف واصفات الكلام في KAPD

استُخدمت نفس الواصفات المستخدمة في العمل (Seddiq et al., 2019) وهي 31 واصفة، وتم تدريب شبكة تصنيف ثنائية لهذه السمات بشكل مشابه لما تم عمله مع القاعدة السابقة. ويبين الجدول (9 نتائج التصنيف على مستوى الصوتيم باستخدام مقياس F1 بالمقارنة مع(Algabri et al., 2021) . نلاحظ تفوق النموذج المدرب بالاعتماد على المصنفات الثنائية بفارق وسطي 4%~، وتم الحصول على نتائج أفضل بالنسبة ل 66% من الواصفات. يعود السبب في كون نتيجة التعرف أفضل في هذا العمل اعتمادنا على المعطيات الموسومة مسبقاً في تحديد وسم الإطار الصوتي للواصفات، بينما يتم في نظام AFD-Obj تحديد الوسم بناء على الحدود الزمنية للصوتيمات المستنتجة من خرج خوارزمية **YOLO**، ما يقلل من دقة صحة الوسم للإطار الصوتي في عملية التدريب.

iii. تعرّف واصفات الكلام في قاعدة بيانات اللغة العربية

بعد اختبار تصنيف واصفات الكلام على كل من TIMIT و KAPD ومقارنتها مع منهجيات أخرى وإثبات فعاليتها، قمنا ببناء نموذج تعرّف على الواصفات في الجدول (1 و

مخارج الحروف	الحروف
بلعومي Pharyngeal	ع ح
حلقي Pharynx	ء هـ ع ح غ خ
حنجرية Glottal	ء هـ
لهوية Uvular	ق
فوق اللثة Post-alveolar	ج ش
وسط اللسان Middle_tongue	ج ش ي
اقصى اللسان Deep_tongue	ك ق
حافة اللسان Tongue_border	ض ل
طرف اللسان Tongue_tip	ط د ت ص س ز ظ ذ ث ن ر
بين الاسنان Interdental	ظ ث ذ

باستخدام قاعدة بيانات اللغة العربية. تم تقسيم البيانات إلى 40% تدريب، 30% اختبار، 30% تحقق . ويبين الجدول 10 موسطات الشبكة المدربة.

الجدول 8 نتائج تصنيف شبكة DNN لواصفات الكلام في TIMIT

LAS-MTL	YOLOv3-Tiny-2S	Our Work	
95	91.05	90.47	Alveolar
90	89.69	89.09	Anterior
98	97.12	96.85	Approximant
98	97.70	98.02	Bilabial
99	93.73	97.52	Central
97	94.13	94.18	Close
88	88.97	89.86	Consonantal
89	91.37	94.07	Continuant
95	96.03	96.28	Fricative
95	93.33	91.66	Front
99	98.67	99.37	Glottal
99	98.88	98.54	labiodental
99	98.21	97.64	Lateral approximant
97	90.28	90.13	Mid
99	97.59	97.63	Nasal
97	97.60	97.31	Non sibilant fricative
98	96.09	94.91	Open
99	99.60	99.56	palatal
99	99.18	98.79	postalveolar
98	94.99	94.73	Round
99	99.5	99.37	Sibilant affricate
98	97.97	98.3	Sibilant fricative
80	96.79	100	Silence
97	95.03	97.91	Stop
97	89.63	90.09	Tense
99	98.37	98.77	Velar
84	90.86	93.47	Voiced
92	91.31	92.2	vowel
95.5	95.13	95.60	Average

الجدول (9) نتائج مقياس F1 لشبكة DNN لتصنيف واصفات الكلام في KAPD

Our work	YOLOv3-Tiny-3S	
99.4	92.9	Affricative
96.1	98.9	Alveodental
.499	94.3	Alveopalatal
96.9	99	Anterior
99.3	94.1	Aspirated
97.6	90.8	Bilabial
99.2	99.8	Consonant
97.8	99.4	Continuant
93.3	98.3	Coronal
97.7	90.4	Emphatic
97.2	99	Fricative
98.4	93.3	Glottal
96.9	92	High
97.9	83.3	Interdental
99	72.9	Labiodental
99.5	96.6	Labiovelar
99.5	87.3	Lateral
99.2	97.3	Nasal
99.6	94.5	Palatal
99.1	96.1	Pharyngeal
98.2	93.6	Plosive
96.5	96.6	Rounded
99.1	97.2	Semivowel
99.3	99.8	Short
99.8	99.9	Silence
99.1	87.4	Trill
98.2	97.1	Unvoiced
98.8	92.6	Uvular
99.4	94.8	Velar
98.2	99.7	Voiced
99.2	99.9	Vowel
98.4	94.5	Average

الجدول 10 موسطات شبكةDNN لتعرّف واصفات الكلام في MSA

مميزات الدخل			825
عدد الطبقات الخفية			3
خصائص الطبقات	رقم الطبقة	عدد الوحدات في الطبقة	تابع التفعيل
	1	1920	Relu

الخفية	2	1949	Sigmoid
	3	498	Relu
تابع الأمثلة	Adam		
Optimizer	Learning rate 6.87e-05		

يبين الجدول 12 نتائج التصنيف التي حصلنا عليها حيث بلغت دقة التصنيف الوسطية لبيانات الاختبار 97.54 وهي نسبة تعرّف عالية تسمح باستخدام خرج المصنف كدخل لمرحلة تالية يتم فيها كشف خطأ النطق بالاعتماد على هذه الواصفات. نجد أقل نسبة تعرف في واصفات الحروف الصوتية Vowels وحروف الرخاوة Softness لتنوع الأصوات التابعة لها مما يزيد من معدل الخطأ فيها، ويجعلنا بحاجة لطريقة فعالة أكثر في التمييز كاستخدام مرحلة سابقة تكشف عن واصفات أكثر تحديداً، وتضيق مجال التعرف بالنسبة لهذه الأصوات. أيضاً نجد نسبة تعرف عالية بالنسبة لكل من (Spreading، Affricates، Prolongation، Post alveolar) كونها صفات مميزة عند النطق بها، عدد أصواتها محدد وهذا يسهل تمييزها في الإشارة الصوتية أيضاً.

الجدول 11 نتائج تصنيف شبكةDNN لواصفات الكلام في قاعدة بيانات اللغة العربية

Test	Valid	Train	
96.63	96.23	97.44	Whisper
99.07	99.38	99.45	Whistle
95.28	96.66	98.41	Adhesion
99.85	99.8	99.76	Affricates
98.17	98.18	97.92	Deviate
94.72	95.81	97.85	Elevation
97.58	97.07	97.08	Fricatives
93.68	93.86	92.68	Softness
99.86	99.83	99.85	Spreading
96.79	95.06	96.26	Strength
99.78	99.65	99.67	Hiding
96.02	95.34	95.46	Moderate
98.54	98.57	99.01	Repetition
99.69	99.6	99.68	Prolongation
91.39	91.77	90.5	Vowels
94.08	95.34	97.01	Silence
99.44	98.89	98.74	Pharyngeal
98.26	96.84	97.2	Pharynx
98.95	98.14	98.62	Glottal
99.49	99.37	99.64	Uvular
99.78	99.74	99.69	Post alveolar
98.05	98.4	98.09	Middle tongue
99.28	98.92	99.37	Deep tongue
98.91	98.15	98.26	Tongue border
94.51	95.18	95.47	Tongue tip
99.03	98.68	99.53	Interdental
96.86	97.11	96.92	Labial

1. الخاتمة والتوصيات

مما سبق نجد أن استخدام مصنفات ثنائية لتدريب واصفات الكلام أعطى نتائج أفضل مع معظم الواصفات في جميع قواعد المعطيات التي تم الاختبار عليها، وذلك باستخدام شبكة متعددة المخارج تقوم بعملية التصنيف بشكل تشاركي بالاستفادة من وجود نفس السمات التي يتم تدريب الشبكة عليها لكل واصفة، وبالتالي يمكن استخدام هذه الواصفات في كشف النطق الخاطئ بعد التعرّف بشكل جيد على الصوتيمات، حيث يتم تحديد وجود خطأ النطق وتحديد نوعه عن طريق واصفات الكلام التي تم التعرّف عليها، وهذا يعطي متعلّم اللغة تغذية راجعة تفيده في تصحيح نطقه للأصوات. على سبيل المثال عند نطق صوت الطاء بدون الاستعلاء والإطباق اللازمين سيكون صوتها أقرب للتاء، وهذا يعد خطأ نطقي، لذا عند تعرّف الواصفات وبعد كشف وجود الخطأ، سنلاحظ عدم تحقق صفتي الاستعلاء والإطباق، فنكون قد حددنا سبب الخطأ وأرشدنا المتعلّم لما يجب فعله لتصحيح خطأ اللفظ.

native training data based on DNN transfer learning. IEICE Transactions on Information and Systems, E100D(9), 2174–2182. https://doi.org/10.1587/transinf.2017EDP7019

11. Fahmy, F. K., Khalil, M. I., & Abbas, H. M. (2020). A Transfer Learning End-to-End Arabic Text-To-Speech (TTS) Deep Architecture. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12294 LNAI (pp. 266–277). https://doi.org/10.1007/978-3-030-58309-5_22
12. Georgoulas, G., Georgopoulos, V. C., & Stylios, C. D. (2006). Speech sound classification and detection of articulation disorders with support vector machines and wavelets. 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, 2199–2202.
13. Google Colaboratory. (n.d.). https://colab.research.google.com/
14. Halabi, N. (2016). Modern Standard Arabic Phonetics for Speech Synthesis. School of Electronics and Computer Science.
15. Harrison, A. M., Lau, W. Y., Meng, H. M., & Wang, L. (2008). Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. Ninth Annual Conference of the International Speech Communication Association.
16. Hu, W., Qian, Y., & Soong, F. K. (2013). A new DNN-based high quality pronunciation evaluation for computer-aided language learning (call). Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, August, 1886–1890.
17. Hu, W., Qian, Y., & Soong, F. K. (2014). A new Neural Network based logistic regression classifier for improving mispronunciation detection of L2 language learners. Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, ISCSLP 2014, 245–249. https://doi.org/10.1109/ISCSLP.2014.6936712
18. Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Communication, 67, 154–166.

References

1. Abdel-hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional Neural Networks for Speech Recognition. 22(10), 1533–1545.
2. Abdou, S. M., Hamid, S. E., Rashwan, M., Samir, A., Abd-Elhamid, O., Shahin, M., & Nazih, W. (2006). Computer aided pronunciation learning system using speech recognition techniques. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2(June 2014), 849–852.
3. Al-Marri, M., Raafat, H., Abdallah, M., Abdou, S., & Rashwan, M. (2018). Computer Aided Qur'an Pronunciation using DNN. Journal of Intelligent and Fuzzy Systems, 34(5), 3257–3271. https://doi.org/10.3233/JIFS-169508
4. Algabri, M., Mathkour, H., Alsulaiman, M. M., & Bencherif, M. A. (2021). Deep Learning-Based Detection of Articulatory Features in Arabic and English Speech. https://doi.org/10.3390/s21041205
5. Algabri, M., Mathkour, H., Bencherif, M. A., Alsulaiman, M., & Mekhtiche, M. A. (2020). Towards Deep Object Detection Techniques for Phoneme Recognition. IEEE Access, 8, 54663–54680. https://doi.org/10.1109/ACCESS.2020.2980452
6. Bhowmik, T., & Mandal, S. K. Das. (2018). Manner of articulation based Bengali phoneme classification. International Journal of Speech Technology, 21(2), 233–250. https://doi.org/10.1007/s10772-018-9498-5
7. Chao, L. (2015). Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition. 65–72.
8. Chen, N. F., & Li, H. (2017). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016, 1–7. https://doi.org/10.1109/APSIPA.2016.7820782
9. Duan, R., Kawahara, T., Dantsuji, M., & Nanjo, H. (2017). Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection. August, 42–46.
10. Duan, R., Kawahara, T., Dantsuji, M., & Zhang, J. (2017). Articulatory modeling for pronunciation error detection without non-

Content-based audio classification and segmentation by using support vector machines. Multimedia Systems, 8(6), 482–492. https://doi.org/10.1007/s00530-002-0065-0

29. **Mao, S., Wu, Z., Li, X., Li, R., Wu, X., & Meng, H. (2018). Integrating Articulatory Features into Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech. Proceedings - IEEE International Conference on Multimedia and Expo, 2018-July.**
https://doi.org/10.1109/ICME.2018.8486462
30. **Maqsood, M., Adnan Habib, H., Nawaz, T., & Zeeshan Haider, K. (2016). A Complete Mispronunciation Detection System for Arabic Phonemes using SVM. IJCSNS International Journal of Computer Science and Network Security, 16(3), 30.**
31. **Meng, H., Lo, Y. Y., Wang, L., & Lau, W. Y. (2007). Deriving salient learners' mispronunciations from cross-language phonological comparisons. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 437–442.**
32. **Morsy, H., Shahin, M., Aljohani, N., Shoman, M., & Abdou, S. (2018). Automatic Speech Attribute Detection of Arabic Language. International Journal of Applied Engineering Research, 13(8), 5633–5639.**
33. **Nazir, F., Majeed, M. N., Ghazanfar, M. A., & Maqsood, M. (2019). Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes. IEEE Access, 7, 52589–52608. https://doi.org/10.1109/ACCESS.2019.2912648**
34. **Noory, mhd jawad. (2007). Arabic phonology.**
35. **optuna. (n.d.). https://optuna.readthedocs.io/en/stable/index.html**
36. **Ravanelli, M., Parcollet, T., & Bengio, Y. (2018). The PyTorch-Kaldi Speech Recognition Toolkit. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May, 6465–6469. https://doi.org/10.1109/ICASSP.2019.8683713**
37. **Seddiq, Y., Alotaibi, Y. A., Selouani, S. A., & Meftah, A. H. (2019). Distinctive phonetic features modeling and extraction using deep neural networks. IEEE Access, 7, 81382–81396. https://doi.org/10.1109/ACCESS.2019.2924014**

https://doi.org/10.1016/j.specom.2014.12.008

19. **Hussain, F., Ehatisham-ul-haq, M., Baloch, N. K., & Ishmanov, F. (2020). Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features. June. https://doi.org/10.3390/electronics9060963**
20. **Karaulov, I., & Tkanov, D. (2019). Attention model for articulatory features detection.**
21. **Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N. https://doi.org/1993STIN...9327403G**
22. **Learning, M., Liu, P., Qiu, X., & Huang, X. (2011). Recurrent Neural Network for Text Classification.**
23. **Lee, K. F., & Hon, H. W. (1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. In IEEE Transactions on Acoustics, Speech, and Signal Processing (Vol. 37, Issue 11, pp. 1641–1648). https://doi.org/10.1109/29.46546**
24. **Li, K., Qian, X., & Meng, H. (2017). Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks. IEEE/ACM Transactions on Audio Speech and Language Processing, 25(1), 193–207. https://doi.org/10.1109/TASLP.2016.2621675**
25. **Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. H. (2016). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2016-May, 6135–6139. https://doi.org/10.1109/ICASSP.2016.7472856**
26. **Lo, W.-K., Zhang, S., & Meng, H. (2010). Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. Eleventh Annual Conference of the International Speech Communication Association.**
27. **Lopes, C., & Perdigao, F. (2011). Phoneme Recognition on the TIMIT Database. In Speech Technologies (Vol. 1, pp. 285–302). InTech. https://doi.org/10.5772/17600**
28. **Lu, L., Zhang, H. J., & Li, S. Z. (2003).**

38. **Seddiq, Y., Meftah, A., Alghamdi, M., & Alotaibi, Y. (2016). Reintroducing KAPD as a Dataset for Machine Learning and Data Mining Applications. 2016 European Modelling Symposium (EMS), 70–74. https://doi.org/10.1109/EMS.2016.022**
39. **Shahin, M., Ahmed, B., McKechnie, J., Ballard, K., & Gutierrez-Osuna, R. (2014). A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech. Fifteenth Annual Conference of the International Speech Communication Association.**
40. **Wei, S., Hu, G., Hu, Y., & Wang, R. H. (2009). A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models. Speech Communication, 51(10), 896–905. https://doi.org/10.1016/j.specom.2009.03.004**
41. **Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30(2–3), 95–108.**
42. **Zangar, I., Mnasri, Z., Colotte, V., Jouvet, D., Houidhek, A., Zangar, I., Mnasri, Z., Colotte, V., Jouvet, D., Houidhek, A., Zangar, I., Mnasri, Z., Colotte, V., Jouvet, D., & Houidhek, A. (2018). Duration modeling using DNN for Arabic speech synthesis To cite this version : HAL Id : hal-01889917 Duration modeling using DNN for Arabic speech synthesis.**