التعرف إلى المتكلم بالاعتماد على التعلم العميق

إعداد: جودي السراج – ماريسول العساف

إشراف: أ.د. عمار زقزوق

قسم هندسة الحاسوب – كلية الهندسة

الملخص:

يعتبر الكلام وسيلة اتصال لنقل معلومات مثل الجنس واللهجة والمشاعر وغيرها من الخصائص المميزة للمتكلم، والتي استخدمها الباحثون للتعرف إلى المتكلم بالاعتماد على التعلم العميق.

سنعرض في هذه المقالة أنواع تطبيقات التعرف إلى المتكلم، وماهي الخطوات الرئيسة لبناء نظام للتعرف إلى المتكلم بالاعتماد على تقنيات الذكاء الصنعي، وسنذكر أهم مقاييس تقييم هذه النظم.

الكلمات المفتاحية: الذكاء الصنعي – الشبكات العصبية – التعلم العميق – تطبيقات التعرف إلى المتكلم.

1- مقدمة:

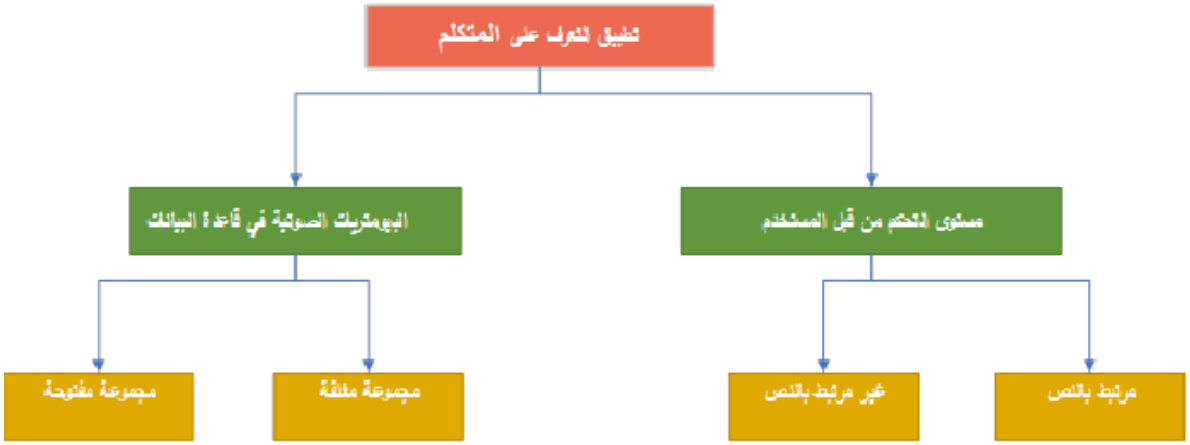
تحمل الإشارات الصوتية معلومات عن اللهجة والجنس والعواطف والسمات الفريدة الأخرى التي تدعى البيومترات الصوتية المتعلقة بالمتكلم، والتي تمكن من التمييز بين المتكلمين عند إجراء المكالمات الصوتية عبر الهواتف، على الرغم من أن المتكلم لا يكون حاضراً جسدياً.

تعتبر عملية التعرف إلى الهوية الصوتية (Speaker Identification: SI) مجالاً مهماً للبحث في العديد من التطبيقات، كالتعرف الجنائي لاكتشاف المشتبه فيهم (Campbell et al.,2009; Morrison et al., 2016)، والتحكم في الوصول إلى الحاسوب (Naik & Doddington, 1987)، وتحسين الأمان في الخدمات المصرفية والتسوق عبر الهاتف المحمول (Gomar, 2015).

2- تطبيقات التعرف إلى المتكلم:

تقسم تطبيقات التعرف إلى المتكلم (SI) إلى فئتين رئيستين كما هو موضح في الشكل (1).

تعتمد الفئة الأولى من تطبيقات SI على مطابقة البيومترات الصوتية للمتكلم مع قاعدة بيانات المتكلمين، وهي على مجموعتين: الأولى تكون مفتوحة يتم فيها المطابقة التامة ما بين المتكلم ونماذج المتكلمين في قاعدة البيانات ويتم الرفض في حالة عدم التطابق التام، والثانية هي مجموعة مغلقة يتم فيها دراسة درجة التشابه ما بين المتكلم ونماذج المتكلمين في قاعدة البيانات، وبالتالي تعاد هوية المتكلم الأكثر تشابهاً.

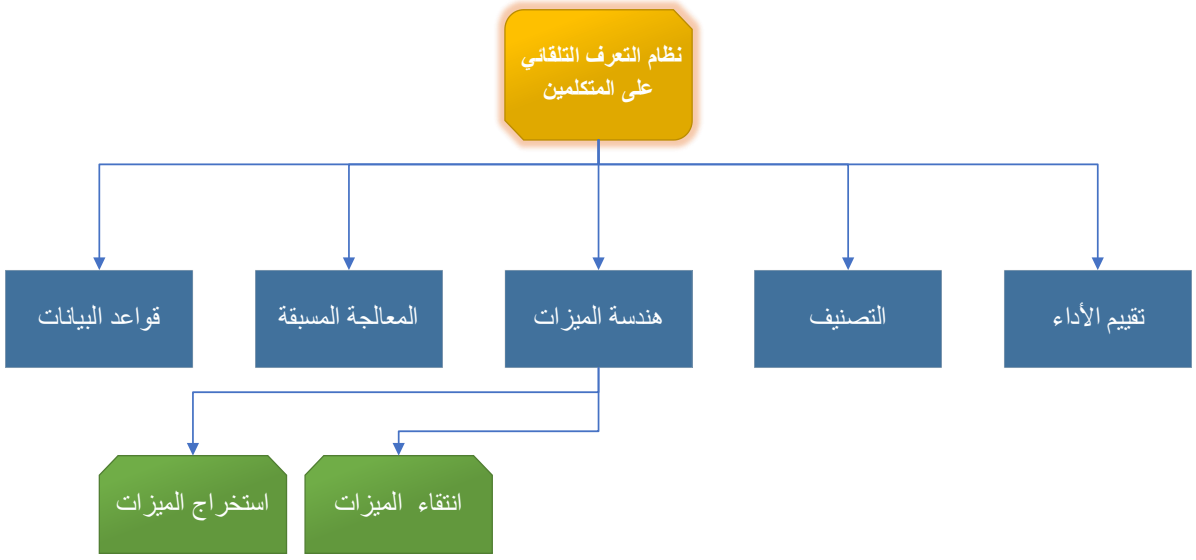


الشكل (1): تصنيف تطبيقات التعرف إلى المتكلم.

أما الفئة الثانية من تطبيقات SI تعتمد على مستوى التحكم من قبل المستخدم للتطبيق من خلال الأنظمة المعتمدة على النص المقروء أو المستقلة عنه، وبالتالي لا يوجد قيود على الكلمات أو الجمل التي يسمح للمتكلم التحدث بها، وتعتبر من أكثر الأنظمة صعوبة، وذلك لصعوبة تقليد عبارة غير معروفة أكثر من العبارة المعروفة.

3- أنظمة التعرف إلى المتكلم:

يتألف نظام التعرف إلى المتكلم باستخدام تقنيات الذكاء الصنعي من خطوات عدة يتم توضيحها في الشكل (2).



الشكل (2): خطوات بناء نظام للتعرف إلى المتكلم.

1.3- قاعدة البيانات:

تحوي عينات صوتية لمتكلمين مختلفين سيتم تدريب واختبار نظام التعرف إلى المتكلم عليهم، إذ يتم تحليل البيومترات الصوتية لكل متكلم في القاعدة وانشاء نموذج فريد له.

من أكثر القواعد المستخدمة هي قاعدة VoxCeleb، وهي قاعدة مجانية تحتوي على مقاطع صوتية ومقاطع فيديو لأشهر الشخصيات العامة والمشهورة على الانترنت، والتي تم تسجيلها بقنوات ثنائية أو أحادية وبنمط

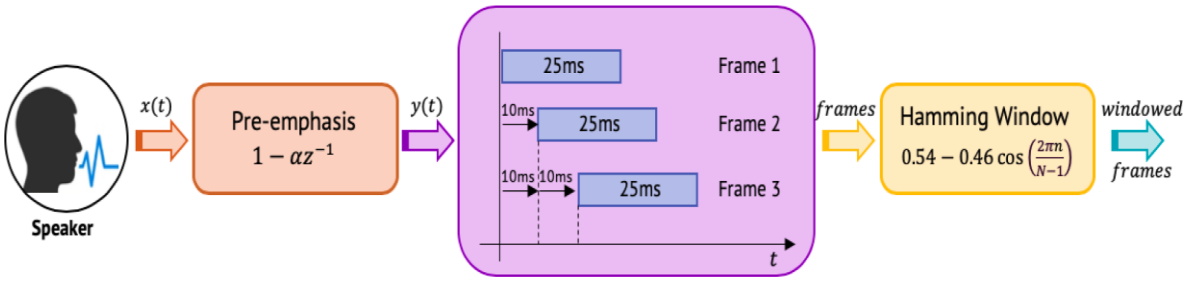
WAV، وقد تم مراعاة اللهجات المحتملة للغة الإنكليزية، عدد عينات القاعدة 153516 عينة صـــوتية ما بين ذكور وإناث.

2.3- المعالجة المسبقة للإشارة الصوتية:

هناك العديد من عمليات المعالجة التي تتم على الإشارة الصوتية من أهمها الآتي:

1.2.3- قسيم الإشارة الصوتية:

تسمح عملية تقسيم الإشارة الصوتية الى مقاطع زمنية (إطارات) بتحليل الإشارة واستخراج الميزات المختلفة مثل الترددات وطول الموجة وغيرها، وذلك لتلافي التغيرات التي تحدث للإشـــارة مع مرور الزمن (Gill, Kaur, & Kaur, 2010; Togneri & Pullella, 2011)، عادة يتم تقسـيم الإشـارة إلى إطارات بطول 25 ميلي ثانية وتداخل 10 ميلي ثانية بين الإطارات، الشكل (3).



الشكل (3): عملية تقسيم الإشارة الصوتية.

2.2.3- المخطط الطيفي للإشارة الصوتية (Spectrogram):

((y Badshah et al., 2019; Stolar, Lech, Bolia, & Skinner, 2017; Sun, Chen, Xie, & Gu, 2018)

هي عملية التمثيل ثنائي البعد للإشارة الصوتية، ويتم الحصول عليه بتطبيق تحويل فوريه على إشارة كل إطار من الإطارات الناتجة من مرحلة تقســيم الإشـــارة الصـــوتية، بحيث يمثل المحور (x) محور الزمن والمحور (y) محور التردد، وكل نقطة من نقاط المخطط الطيفي تمثل شـدة الإشـارة الصـوتية في الزمن والتردد المعين. يعتبر المخطط الطيفي للإشارة أحد أهم الأدوات لتحليل الإشارة الصوتية وفهما، إذ يوضح كيفية تغير شدة الإشارة عبر التردد والزمن.

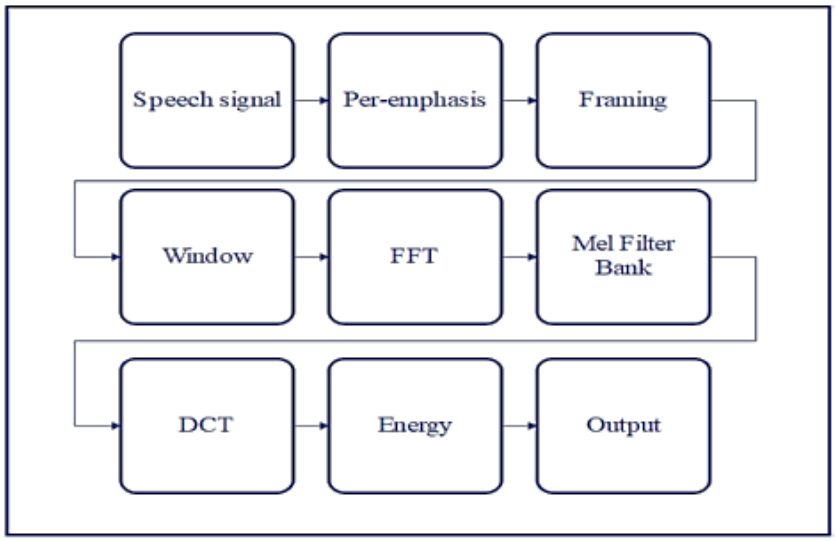
3.3- استخراج الميزات (Mujtaba et al., 2019):

تعتبر عملية اســـتخلاص الميزات من أهم خطوات بناء نظم التعرف على المتكلم لما لها من تأثير على نســـبة التعرف الكلية للنظام. سنورد فيما يلي إحدى أهم خوارزميات استخراج الميزات في مجال التعرف على المتكلم.

1.3.3- خوارزمية MFCC (Mel Frequency Cepstral Coefficients):

تعتبر خوارزمية MFCC من أهم خوارزميات اســـتخراج الميزات، والتي تمر بمراحل عدة بعد عملية المعالجة المســـبقة للإشـــارة الصـــوتية وتقســـيمها إلى إطارات (Framing)، أولها تطبيق نافذة هامينغ (Window

Hamming)، وذلك للحصـول على طيف الاتسـاع لكل إطار، ثم تحويل فوريه السـريع (FFT) لاسـتخراج عناصـر التردد في المجال الزمني، بعدها يتم التحويل إلى طيف ميل (Mel Filter Bank)، وأخيراً نطبق تحويل جيب التمام المتقطع (Discrete Cosine Transform: DCT) للحصول على المعاملات، الشكل (4).



الشكل (4): خطوات خوارزمية MFCC.

تسـتخدم الأداة البرمجية YAAFE (Mathieu, Essid, Fillon, Prado, & Richard, 2010) لاسـتخراج الميزات، وهي مجموعة أدوات مفتوحة المصـدر تدعم لغات C، MATLAB، Python، وتمثل الميزات المختارة كدخل لخوارزميات التصنيف في خطوات بناء نظام التعرف إلى المتكلم المطلوب.

4.3- التصنيف:

في هذه المرحلة، يتم بناء نموذج تصنيف بالاعتماد على بيانات التدريب باستخدام خوارزمية التعلم الآلي.

1.4.3- الشبكات العصبية الاصطناعية:

تعرف الشبكات العصبية الاصطناعية (Artificial Neural Networks: ANN) على انها نماذج حسابية ذات قدرة على تجميع البيانات وتنظيمها وتعلم المعلومات المعممة من خلال الاعتماد على المعلومات المتوازية. تتكون الشـبكة العصـبية من طبقات (Layers) مكونة من وحدات معالجة تسـمى بالخلايا (Neurons) متماثلة تتصـل فيما بينها عن طريق ارسـال الإشـارات الموزونة إلى بعضـها البعض، فيتم في مرحلة التعلم تحديث المعاملات لتحقيق شرط معي، وذلك عن طريق ارسال استخدام مجموعة التدريب (Learning Set) التي تستخدم لحساب نسبة الخطأ.

1.1.4.3- الشبكة العصبية متعددة التطبيقات:

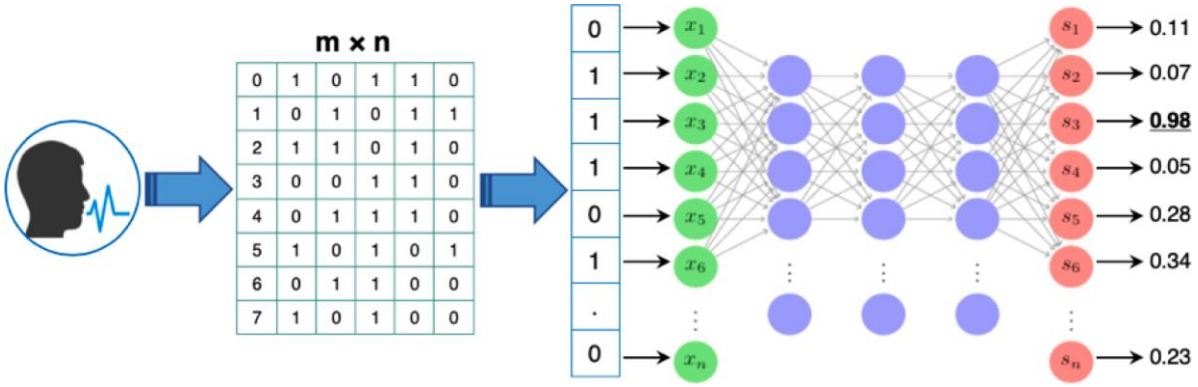
تتألف هذه الشـبكة العصـبية من طبقة واحدة أو طبقات عدة من الطبقات المخفية. إن فائدة الطبقات المخفية هو اكتشـاف المزايا (Features) الموجودة في الإشـارات الداخلة Yاليها. توجد العديد من الطرائق لتدريب الشـبكات، ومن أشـهرها هي الانتشـار الأمامي (forward-propagation) والانتشـار العكسـي (back-propagation).

يتم إدخال المدخلات إلى طبقة الإدخال لتمر إلى الطبقات المخفية، وتتم معالجة هذه الإدخالات تم تمرر وتعالج مرة أخرى في طبقة الإخراج، وتسمى خوارزمية التدريب هذه بالانتشار الأمامي (forward–propagation). أما في خوارزمية تدريب الشبكة عن طريق الانتشار العكسي (back–propagation)، فيتم تقديم أنماط الإدخال للتدريب إلى طبقة الإدخال للشبكة، تقوم الشبكة بعد ذلك بنشر نمط الإدخال من طبقة إلى طبقة حتى يتم إنشاء نمط الإخراج بوساطة طبقة الإخراج. يتم حساب الخطأ في حال ظهور نتائج غير مرغوبة، ثم يتم نشره للخلف من طبقة الإخراج إلى طبقة الإدخال، وتستمر العملية لحين الحصول على النتائج المطلوبة.

5.3- الشبكات العصبية العميقة:

أصبح التعلم العميق طريقة مثيرة للاهتمام وقوية لتعلم الآلة مع تطبيقات ناجحة في العديد من المجالات، مثل معالجة اللغة الطبيعية. والتعرف إلى الصور والأحرف المكتوبة بخط اليد والمتحدث، والرؤية الحاسوبية. إذ أظهر التعلم العميق نجاحاً في التعرف إلى الكلام وتحديد هوية المتحدث على الطرائق التقليدية، مثل تلك التي تستخدم معاملات تردد ميل للتعرف إلى المتحدث باستخدام نماذج خليط غاوسي.

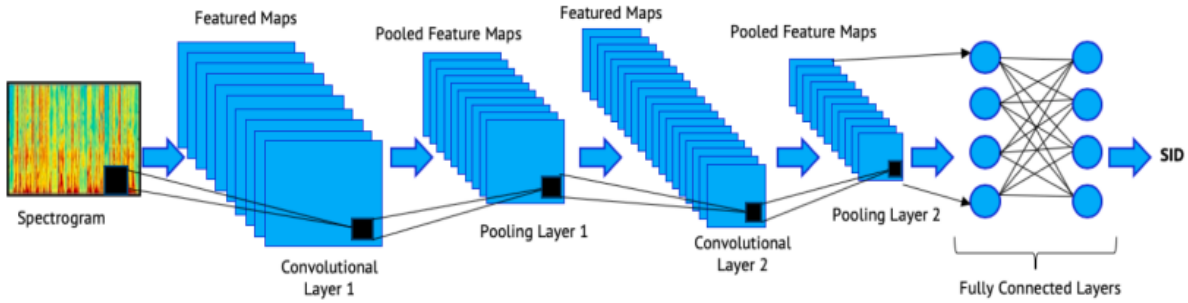
إن التعلم العميق يوفر طريقة أكثر تكيفاً من خلال استخدام الشبكات العصبية العميقة التي تتعلم الخصائص من بيانات الإدخال، وبالتالي فإنها تجعل الحاسوب قادراً على اتخاذ القرار. تعتمد تقنية التعلم العميق على الطرائق التي تستند إلى تعلم تمثيل البيانات، الشكل (5).



الشكل (5): معمارية الشبكة العصبية العميقة.

1.5.3- الشبكة العصبية الالتفافية:

تأتي تسمية هذا النوع من الطبقات من عملية الطي أو الالتفاف الرياضية، إذ تطبق في طبقات الثقافية مرشحات، ويُعرف أيضاً بـــ (kernel) بعدد N يحدد حسب النتائج أثناء عملية التدريب، ويمكن تمثيلها بشكل مصفوفة، من شأنه تحديد وجود سمات أو أنماط معينة في الصورة الأصلية. يكون حجم المُرشح صغيراً ليمسح مصفوفة الإدخال بشكل كامل ويطبق العمليات الحسابية بغية استخراج السمات (Features). يُعاد ضبط قيم المُرشح خلال عملية التدريب الدورية، وعند تدريب الشبكة، توظف الطبقات المخفية الأولى في استخراج السمات البسيطة والواضحة مثل الحواف في الاتجاهات المختلفة، ومع التعمق أكثر في الطبقات المخفية في الشبكة، تزداد درجة تعقيد السمات التي يجب تحديدها واستخراجها. يبين الشكل (6) معمارية الشبكة العصبية الالتفافية (CNN).



الشكل (6): معمارية الشبكة العصبية الالتفافية.

6.3- تقويم الأداء:

هناك تباين كبير في مقاييس الأداء المستخدمة لأنظمة التصنيف، ولكن من أكثرها استخداماً الدقة (Accuracy)

1.6.3- مقاييس النتائج والتحقق منها:

المقاييس التي نعتمد عليها في هذه الدراسة في تنبؤ الشبكة (Accuracy) هي التحقق من صحة النظام عبر مقاييس (FAR FRR). معدل القبول الخاطئ (FAR) هو النسبة المئوية لحالات تحديد الهوية التي يتم فيها قبول الأشخاص غير المصرح لهم بشكل غير صحيح، أما معدل الرفض الخاطئ (FRR) فهو النسبة المئوية لحالات تحديد الهوية التي يتم فيها رفض الأشخاص المصرح لهم بشكل غير صحيح.

4- الخاتمة:

مما تقدم، نجد أن هذه المقالة تقدم دراسة لتحديد الهوية الصوتية من خلال تقنيات الذكاء الاصطناعي. تحديد الهوية الصوتية هو عملية استخراج هوية المتحدث بناءً على الملامح الصوتية لصوتهم. تناقش المقالة تطبيقات التعرف إلى المتكلم والأنظمة التي تتيح لنا المقدرة على التعرف إليه. كما تسلط الضوء على كيفية التصنيف بالاعتماد على بيانات التدريب، ومناقشة مختلف الشبكات العصبية.

5- المراجع:

1- Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009). Forensic speaker recognition. IEEE Signal Processing Magazine, 26(2), 95-103: http://refhub.elsevier.com/S0957-4174(21)00032-4/h0135.

2- Naik, J., & Doddington, G. (1987). Evaluation of a high performance speaker verification system for access Control. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87. (Vol. 12, pp. 2392-2395): IEEE.

3- Gomar, M. G. (2015). System and method for speaker recognition on mobile devices. In: Google Patents.

4- Islam, M., & Rahman, M. (2009). Improvement of text dependent speaker identification system using neuro-genetic hybrid algorithm in office environmental conditions. arXiv preprint arXiv:0909.2363.

5- Revathi, A., & Venkataramani, Y. (2009). Text independent composite speaker identification/verification using multiple features. In 2009 WRI World congress on computer science and information engineering (Vol. 7, pp. 257-261): IEEE.

6- Larcher et al. (2014) and Tirumala et al. (2017). Text-dependent speaker verification: Classifiers, databases and RSR2015 - ScienceDirect.

7- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286. Scopus preview - Scopus - Welcome to Scopus.

8- (Georgescu, Ionescu, & Popescu, 2019; Wang, 2020. http://refhub.elsevier.com/S0957-4174(21)00032-4/h0280.

9- (Gill, Kaur, & Kaur, 2010; Togneri & Pullella, 2011). http://refhub.elsevier.com/S0957-4174(21)00032-4/h0295.

10- (Hwang, Park, & Chang, 2016). http://refhub.elsevier.com/S0957-4174(21)00032-4/h0390.

11- Hochreiter and Schmidhuber (1997). http://refhub.elsevier.com/S0957-4174(21)00032-4/h0375.