كيف يعمل نظام التعرف على الكلام؟



www.nasainarabic.net
@NasaInArabic NasaInArabic NasaInArabic NasaInArabic NasaInArabic

حين تتصل بإحدى الشركات الكبرى في وقتنا الحاضر، لا يرد عليك شخصٌ بل يجيب بدلًا عنه تسجيلٌ صوتيٌّ آليٌّ ويرشدك إلى ضغط أزرارٍ محددةٍ تنقلك إلى قائمة الخيارات، إلا أن بعض الشركات ذهبت إلى أبعد من ذلك حيث باستطاعتك نطق بعض الكلمات كأوامر للحصول على غايتك. إن النظام الذي يجعل هذه الأشياء ممكنةً هو أحد برامج التعرف على الكلام speech recognition program، ويُسمّى نظام الهاتف الآلي automated phone system.

يمكن أيضًا استخدام خاصية التعرف على الكلام في المنازل ومقرات العمل، حيث تُمكّن هذه البرمجيات المستخدمين على نطاقٍ واسعٍ من إعطاء الأوامر للحواسيب وتحويل الكلام إلى نصٍّ يُكتب كملفٍ في محرّر النصوص أو مستند نصٍّ في البريد الإلكتروني إضافةً إلى إمكانية الوصول إلى الأوامر الوظيفية مثل فتح ملفات أو الوصول إلى قوائم بواسطة أوامر صوتيةٍ. إضافةً إلى المجالات التي ذُكِرت آنفًا

فإن بعض البرامج تُخصص لمجالاتٍ محددةٍ مثل المجال الطبي أو النسخ القضائي.

من جهةٍ أخرى فإن ذوي الاحتياجات الخاصة ممن لا يستطيعون الكتابة يعتمدون على أنظمة التعرف على الكلام، فعند فقدان المستخدم القدرة على استخدام يديه، أو كان يعاني من ضعفٍ في النظر وليس من الممكن استخدام لوحة مفاتيح بريل (**Braille**)، فهذا النظام يسمح لنا باستخدام تعبيرٍ شخصيٍّ عن طريق أوامر صوتيةٍ إضافةً إلى التحكم في العديد من المهام الحاسوبية. تقوم بعض البرامج بحفظ بيانات كلام المستخدم بعد كلّ جلسةٍ، وبذلك تسمح للأشخاص الذين يعانون من تلكؤ في الكلام بشكلٍ متواصلٍ من الاستمرار بإعطاء الأوامر لحواسيبهم.

تُصنَّف البرامج الحالية إلى فئتين

- **مفردات قليلة/عدد كبير من المستخدمين:**

يُعدّ هذا النظام مثاليًّا لمجيب الهاتف الآلي حيث يستطيع المستخدم التحدث بعددٍ كبيرٍ من اللهجات وعينات الكلام، ويبقى النظام رغم ذلك قادرًا على فهمها في الغالب. يبقى هذا الاستخدام مقتصرًا على عددٍ قليلٍ من الأوامر والمُدخلات المتاحة مثل خيارات القوائم الأساسية أو الأرقام.

- **مفردات كثيرة/عدد محدود من المستخدمين:**

يعمل هذا النظام في أفضل حالاته في مجالات الأعمال حيث تشغل أعدادٌ قليلةٌ من المستخدمين هذا البرنامج على الرغم من أن هذه الأنظمة تعمل بدقةٍ عاليةٍ تصل إلى نسبة 85 بالمائة أو أكثر مع مستخدمٍ محترفٍ، وتحتوي على عشرات الآلاف من المفردات إلا أنه ينبغي تطويرها لتعمل بشكلٍ جيدٍ مع أعدادٍ قليلةٍ من مستخدمين مبتدئين حيث سيقل معدل الدقة بشكلٍ كبيرٍ مع أيّ مستخدمٍ آخر.

لقد ابتُكِر نظام التعرف على الكلام منذ أكثر من عشر سنوات، وكان هذا النظام في ذلك الحين أمام خيارين إما التعامل مع الكلام المتقطع أو مع الكلام المستمر، فالمعروف عن هذا النظام أنه يستوعب الكلمات المنفصلة مع توقفاتٍ بسيطةٍ بين كلمةٍ وأخرى بسهولةٍ أكبر، إلا أن العديد من المستخدمين يفضلون التحدث بشكلٍ طبيعيٍّ على نحو الحديث العادي، لذا فقد تم العمل على تطوير الأنظمة الحديثة لتصبح جميعها تقريبًا قادرةً على فهم الكلام المستمر والمتسلسل.

تحويل الكلام إلى بيانات

يجب على الحاسوب القيام بعدّة خطواتٍ معقدةٍ من أجل تحويل الكلام إلى نصٍ مقروءٍ أو أوامر حاسوبيةٍ.

: يقوم نظام التحويل من الرقمي إلى التناظري ADC بترجمة أمواج صوت الشخص إلى بيانات رقمية عن طريق أخذ عينات الصوت، وكلما ارتفعت معدلات أخذ العينات والدقة ارتفعت معها الجودة.

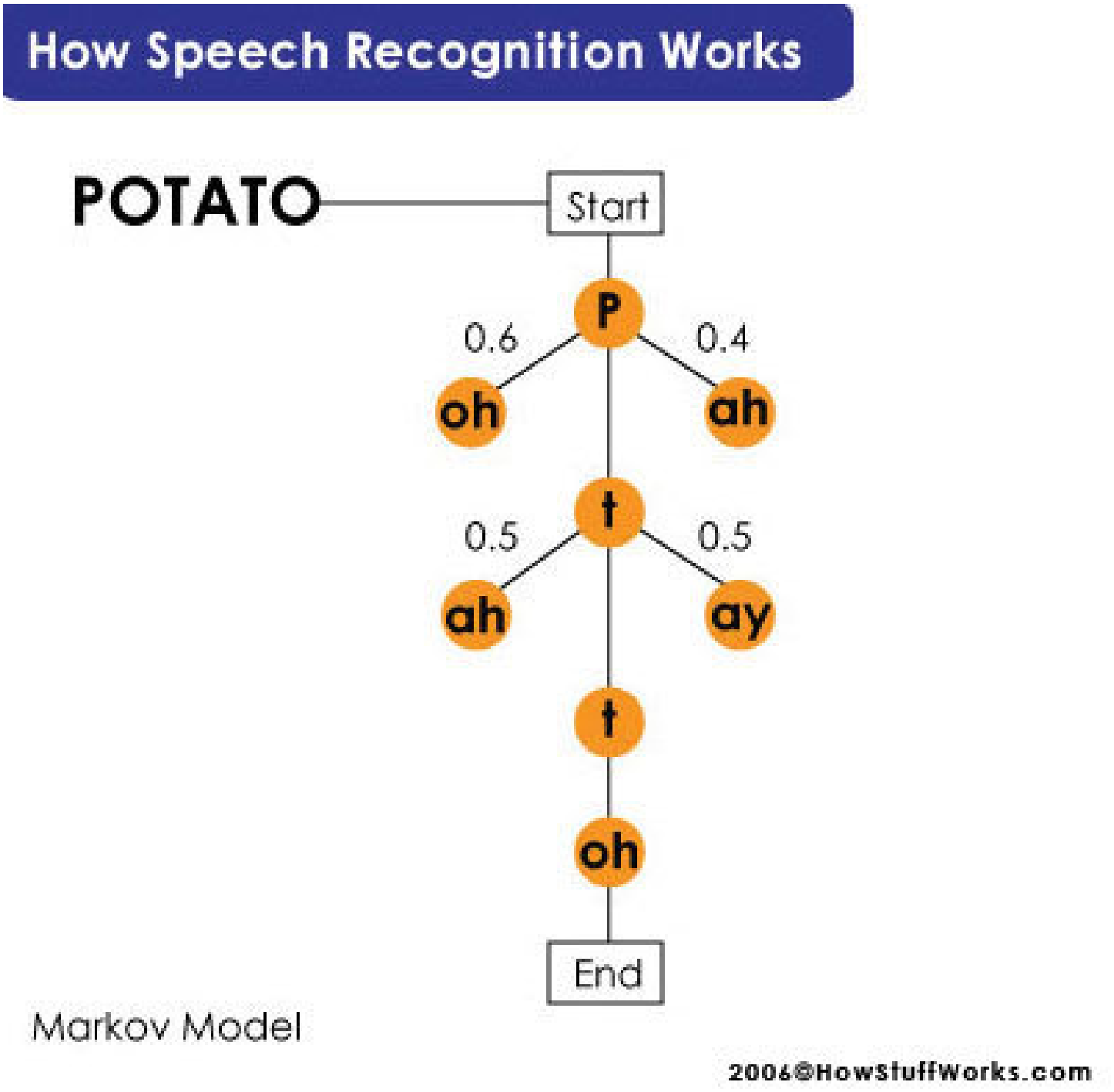
عندما تبدأ بالكلام فإنك تقوم بخلق اهتزازاتٍ في الهواء، ويقوم المحول التناظري إلى الرقمي (**analog-to-digital converter ADC**) بتحويل هذه الأمواج التناظرية إلى بياناتٍ رقميةٍ يستطيع الحاسوب التعامل معها، وللقيام بعملية أخذ العينات الرقمية للصوت تؤخَذ مستوياتٌ محددةٌ للأمواج عبر فواصل متساويةٍ ومتواترةٍ ثم يقوم هذا النظام بترشيح الصوت الرقمي لإزالة الضجيج غير المرغوب به، وفي بعض الأحيان يقوم بتقسيمه إلى حزمٍ مختلفةٍ من الترددات **frequency** (التردد هو الطول الموجي لأمواج الصوت، يُسمع من قبل البشر بطبقاتٍ صوتيةٍ مختلفةٍ). إضافةً إلى معالجة الصوت وضبطه بمستوىً ثابتٍ. وبما إن الأشخاص عادةً لا يتكلمون بنفس

السرعة لذا يجب ضبط الصوت ليطابق قالب عينات الصوت المخزَّن سابقًا في ذاكرة النظام.

في المرحلة التالية تُقسم الإشارة إلى مقاطع صغيرةٍ يكون حجمها بحجم أجزاءٍ من المئة أو حتى من الألف من الثانية في حالة الأصوات الساكنة الانفجارية **plosive consonant sounds**. تُنتَج أصوات التوقف الساكنة **Stop consonant** لدى الإنسان عن طريق إعاقة تدفق الهواء في المجرى الصوتي كما هو الحال مع حرفي "p" وَ "t". بعدها يطابق البرنامج هذه الأجزاء مع المقاطع الصوتية **phonemes** في لغةٍ ملائمةٍ حيث أن الفونيم هو أصغر عنصرٍ في اللغة وهو تمثيل للأصوات التي ننتجها ونجمعها معًا لإنتاج تعابير واضحةٍ. يوجد في اللغة الإنكليزية 40 مقطعًا صوتيًّا بينما يزيد هذا العدد أو يقل في باقي اللغات.

تبدو الخطوة التالية بسيطةً للوهلة الأولى إلا إنها صعبة الإنجاز حيث تتضمن هذه المرحلة التركيز عند عملية البحث لتتعرف على الكلام، حيث يقوم البرنامج باستقصاء مقاطع الصوت في السياق والمقاطع المجاورة لها، ويقوم بتفعيل خريطة المقاطع الصوتية المرتبطة بالسياق من خلال نماذج إحصائيةٍ معقدةٍ ويقارنها مع مكتبةٍ ضخمةٍ من الكلمات والعبارات والجمل المألوفة، ثم يحدد هذا البرنامج ما يقوله المستخدم ويقوم بإخراجه كنصٍّ أو أمرٍ حاسوبيٍّ.

التعرف على الكلام والنمذجة الإحصائية



"كيف يعمل التعرف على الكلام، نموذج ماركوف Markov model ".

حاول القائمون على أنظمة التعرف على الكلام جعل هذه الأنظمة تطبق مجموعةً من القواعد النحوية واللغوية للكلام، فإذا كانت الكلمات المحكية السليمة لفظيًا تحقق مجموعةً من القواعد فعندئذٍ يستطيع البرنامج تحديد ماهية هذه الكلمات. تتضمن اللغة البشرية العديد من الاستثناءات في قواعدها حتى عند التحدث بشكلٍ مستمرٍ. يمكن أن تقوم اللكنات واللهجات المحلية وأساليب الكلام أو السلوكيات بتغيير طريقة نطق بعض الكلمات والعبارات، تخيّل أن شخصًا ما من بوسطن **Boston** وهو يقول كلمة "**barn**" التي تعني إسطبل، سوف لن يظهر حرف "r" على الإطلاق وستكون الكلمة على وزن "John"، فعلى سبيل المثال ينطق بعض الناس عبارة "**I'm going to see the ocean**" دون إظهار جميع الحروف فتكون النتيجة "**I'm goin' da see tha ocean**" حيث يدمجون عدة كلماتٍ معًا أو يحذفون بعض الحروف، وبذلك لن تتمكن الأنظمة المعتمدة على القواعد من النجاح في عملها بسبب عدم إمكانيتها من التعامل مع هذه المتغيرات، وهذا يفسر عدم إمكانية أنظمة التعرف على الكلام البدائية من معالجة الكلام المستمر حيث كان على المتحدث نطق كلّ كلمةٍ على حدة بشكلٍ

منفصلٍ واضعًا فواصل زمنيةً فيما بينها.

أما اليوم فإن أنظمة التعرف على الكلام تستخدم أنظمة نمذجةٍ إحصائيةٍ **statistical modeling systems** معقدةً وفعالةً تتضمن هذه الأنظمة الاحتمالات والتوابع الرياضية وتستخدمها لتحديد النتيجة النهائية المحتملة. ووفقًا لجون غاروفولو **John Garofolo**، مدير فريق الكلام في مختبر تكنولوجيا المعلومات في المعهد الوطني للمعايير والتكنولوجيا **National Institute of Standards and Technology** فإن النموذجين الأكثر سيطرةً في هذا المجال هذه الأيام هما نموذج ماركوف المخفي **The Hidden Markov Model** والشبكات العصبية **neural networks**، حيث تتضمن هذه النماذج وظائف رياضيةٍ معقدةٍ ويعتمد عملها الأساسي على أخذ المعلومات المعروفة إلى النظام واستنتاج المعلومات المخفية.

يُعتبر نموذج ماركوف المخفي الأكثر شيوعًا لذا سوف نلقي نظرةً قريبةً عليه. في هذا النموذج يكون كلّ مقطعٍ صوتيٍّ بمثابة رابطٍ في سلسةٍ والسلسلة هنا هي الكلمة، تتشعب هذه السلسلة في اتجاهاتٍ مختلفةٍ فيحاول البرنامج مطابقة الصوت الرقمي مع المقطع الصوتي الذي غالبًا سيأتي بعده وخلال هذه العملية يقوم البرنامج بتحديد النتيجة المحتملة لكلّ مقطعٍ صوتيٍّ معتمدًا على تركيبها المعجمي وتدريب المستخدم لها، تكون هذه العملية أكثر تعقيدًا مع العبارات والجمل حيث يتحتم على البرنامج معرفة بداية ونهاية الكلمة، وأفضل مثالٍ على ذلك هي عبارة ""**recognize speech** وتعني التعرف على الصوت حيث ستبدو كأنها **"wreck a nice beach"** عندما تُلفظ بشكلٍ سريعٍ وبالتالي تغير معناها تمامًا. على البرنامج تحليل المقاطع الصوتية باستخدام العبارة التي تأتي قبلها من أجل قراءتها بشكلٍ صحيحٍ، وكمثالٍ على ذلك تحليل العبارتين:

r eh k ao g n ay z s p iy ch
"recognize speech"
r eh k ay n ay s b iy ch
"wreck a nice beach"

لكن لنرى هل هي فعلًا بهذا الشكل من التعقيد؟ إذا كان البرنامج يتضمن **60000** كلمةٍ وهي مقدرة البرامج الحالية، فإن عبارةً مكونةً من ثلاث كلمات سوف تأخذ **216** تريليون من الاحتمالات، وبهذا حتى الحواسيب الخارقة لا يمكنها البحث من دون مساعدة. وتأتي هذه المساعدة في شكل برنامج تدريبٍ، حيث يقول جون غاروفولو: "تحتاج الأنظمة الإحصائية إلى العديد من بيانات التدريب النموذجية من أجل الوصول إلى الأداء الأمثل إذ يتطلب في بعض الأحيان آلاف الساعات من الكلام البشري المنقول ومئات الميغابايتات من النصوص، وبعدها تستخدم بيانات التدريب هذه في تشكيل نماذج سمعيةٍ للكلمات وقوائم لها بالإضافة إلى شبكات الاحتمالات المتعددة الكلمات. إن عملية اختيار وتجميع وتحضير بيانات التدريب هذه تتطلب شيئًا من الحرفية لكي يستوعبها النظام وكيف تتناغم نماذج النظام مع تطبيقاتٍ معينةٍ. مثل هذه التفاصيل هي التي تصنع الفارق بين الأنظمة ذات الأداء الجيد والأخرى ذات الأداء الضعيف حتى لو استخدمت نفس الخوارزميات الأساسية".

بينما يقوم مطورو البرمجيات الذين يزودون أداء النظام بالمفردات الأساسية بإجراء العديد من هذه التدريبات، يجب أيضًا على المستخدم النهائي تكريس بعض الوقت لتدريب النظام. في مجال العمل يجب على مستخدمي البرنامج الأوليين تكريس **10** دقائق من وقتهم كحدٍّ أدنى في التكلم مع النظام لتدريبه بنماذج كلام خاصة بهم ويجب أيضًا تدريبه للتعرف على مصطلحاتٍ وكلماتٍ مركبةٍ خاصةٍ بالشركة. هناك بعض الإصدارات الخاصة من برامج التعرف على الكلام لأقسام الطب والقضاء تم التدرب عليها مسبقا تحتوي على مصطلحاتٍ عامةٍ تُستخدم في تلك المجالات.

نقاط الضعف وعيوب نظام التعرف على الكلام



تساعد مكبرات الصوت عالية الكفاءة التي تلغي تأثير الضجيج بزيادة دقة نظام التعرف على الكلام

لا يوجد بالفعل نظام تعرفٍ على الكلام مثاليٌّ **100%** فهناك العديد من العوامل التي من شأنها التقليل من كفاءته ودقته، حيث يُعتبر البعض منها من الأمور المهمة التي ينصب عليها الاهتمام والتطوير كالتحسينات التقنية، أما البعض الآخر فيُهمل إن لم يُصحَّح بشكلٍ كاملٍ من قِبل المستخدم.

انخفاض معدل الإشارة بالنسبة إلى الضجيج

يحتاج البرنامج إلى سماع الكلمات المنطوقة بوضوحٍ إلا أن وجود أيّ ضجيجٍ إضافيٍّ داخل الصوت سوف يتضارب ويتداخل مع هذه الكلمات، حيث يمكن أن يتولد هذا الضجيج من العديد من المصادر بما في ذلك الضجيج العالي في الخلفية داخل المكتب أو الدائرة، لذا يجب على المستخدمين العمل في غرفةٍ هادئةٍ مستخدمين مكبرات صوتٍ ذات كفاءةٍ عاليةٍ موضوعةٍ بالقرب من أفواههم قدر الإمكان. إن كروت الصوت رديئة النوعية، المزودة بمنفذٍ لمكبرات الصوت من أجل إرسال الإشارة إلى الحاسوب، لا تملك حمايةً قويةً ضد الإشارات الكهربائية المتولدة من قِبل عناصر الحاسوب الأخرى وهذا بدوره يؤدي إلى إنتاج طنينٍ أو صفيرٍ ضمن الإشارة.

الكلام المتداخل

تواجه النظم الحالية العديد من الصعوبات في فصل الكلام الذي يصدر بنفس الوقت من عدة مستخدمين، وهنا يقول جون غاروفولو: "إذا حاولت توظيف تقنية التعرف على الكلام في الحوارات والمقابلات التي يقاطع فيها الأشخاص بعضهم البعض الآخر أو التكلم بينما يقوم شخص آخر بالكلام فسوف تحصل بلا شك على نتائج مزريةٍ للغاية".

الاستخدام المفرط لطاقة الحاسوب

يتطلب تفعيل النماذج الإحصائية المستخدمة في التعرف على الكلام معالجًا حاسوبيًّا للقيام بالعديد من مهام المستوى العالي، واحدٌ من أسباب استخدام هذه النوعية من المعالجات هو الحاجة لتذكر كلّ مرحلةٍ من مراحل البحث للتعرف على الكلمة في حال حاجة النظام للقيام بعملية الرجوع بالمسار للوصول إلى الكلمة الصحيحة. إن أسرع حاسوبٍ شخصيٍّ يُستخدم هذه الأيام لا يزال يواجه صعوباتٍ في معالجة الأوامر أو العبارات المعقدة ويقلل ذلك من زمن الاستجابة بشكل ملحوظٍ، إضافةً إلى أن المفردات التي يحتاجها البرنامج تأخذ مساحةً كبيرةً على القرص الصلب. لحسن الحظ فإن حجم القرص وسرعة المعالج هما العاملان الحاسمان لعملية التطور السريع لهذه التقنية لأن الحواسيب المستخدمة بعد عشر سنواتٍ من الآن ستستفيد من زيادةٍ هائلةٍ في هذين العاملين.

الكلمات المتجانسة أو التجانس

التجانس هو عبارةٌ عن كلمتين تختلفان بالإملاء والمعنى وتتشابهان بالصوت نفسه، مثالٌ على ذلك "**air" and "There" and "their"**, **"be" and "bee", ""heir"**. في الحقيقة لا توجد طريقةٌ لجعل برنامج التعرف على الكلام قادرًا على التميز بين هذه الكلمات بالاعتماد على الصوت وحده. إلا أن التدريب الواسع للأنظمة والنماذج الإحصائية التي تأخذ بالاعتبار في سياق الكلمة تحسّن الأداء بشكلٍ كبيرٍ.

مستقبل أنظمة تمييز الكلام

كان أول ظهورٍ لأنظمة التعرف على الكلام قبل اختراع الحاسوب بخمسين سنةً حيث عمد أليكسندر غراهام بيل **Alexander Graham Bell** إلى تجربة نقل الكلام إلى زوجته التي كانت صماء، وكان يأمل في الأساس اختراع جهازٍ يحول الكلام المسموع إلى صورٍ مرئيةٍ يستطيع الشخص الأصم تفسيرها، فقام بعمليةٍ تنتج من خلالها صور مخططاتٍ طيفيةٍ **spectrographic** للأصوات، ولكن زوجته لم تكن قادرةً على فك شيفرتها، وفي النهاية قاد مسار هذا البحث إلى اختراعه للهاتف.

خلال عدة عقودٍ طور العلماء طرائق تجريبيةٍ للتعرف الآلي على الكلام، ولكن ما أعاقهم هو القدرة المحدودة للحوسبة المتوفرة في تلك الفترة من الزمن، في التسعينيات أصبحت قدرة الحواسيب فعالةً بشكلٍ كافٍ لجعل أنظمة التعرف على الكلام متوفرةً للمستهلك العادي، أما الأبحاث الحالية فيمكن أن تؤدي إلى ظهور تكنولوجيا شبيهةٍ إلى حدٍّ ما بإحدى حلقات مسلسل ستار تريك **"Star Trek"**.

تمتلك وكالة مشاريع أبحاث الدفاع المتقدمة دارب (**The Defense Advanced Research Projects Agency (DARPA**) ثلاثة فرقٍ من الباحثين تعمل على برنامج استخدام لغةٍ عالميةٍ مستقلةٍ (**Global Autonomous Language Exploitation (GALE**)، سوف يعمل هذا البرنامج على توريد المعلومات من وكالات أخبارٍ أجنبيةٍ وبرامج إذاعيةٍ وحتى صحفٍ ومن ثم يقوم بترجمتها، ومن المؤمل إنتاج برمجياتٍ يمكنها الترجمة فوريًا بين لغتين وبدقةٍ تصل إلى 90 بالمئة. يقول غاروفولو: "تقوم وكالة داربا بتمويل مبادرة **R&D** التي تُدعى ترانستك **TRANSTAC** ، وذلك لتمكين الجنود من التواصل بشكلٍ أفضل مع المدنيين الذين لا يجيدون اللغة الإنكليزية". إضافةً إلى ذلك فإن هذه التكنولوجيا سوف تُكرَّس بلا شكٍّ للأعمال المدنية ومن ضمنها المترجم العالمي الشامل الذي سيبقى تحقيقه بعيد المنال على المدى القريب لأنه من الصعب إنشاء نظامٍ قادرٍ على الجمع بين الترجمة الآلية والتكنولوجيا المعتمدة على الأصوات. ووفقًا لمقالة نشرتها شبكة **CNN** مؤخرًا فإن مشروع غايل **GALE** هو داربا هارد "**DARPA hard**" الذي يعني أنه من الصعب إنشاء مثل هذا النظام حتى مع وجود معايير داربا الدقيقة والسبب هو صعوبة خلق نظامٍ قادرٍ على التعامل بشكلٍ مثاليٍّ مع العديد من العقبات مثل اللغات العامية واللكنات واللهجات بالإضافة إلى ضجيج الخلفية. يمكن للتراكيب القواعدية المختلفة المستخدمة في اللغات الحدّ من هذه المشكلة، فعلى سبيل المثال تستخدم اللغة العربية في بعض الأحيان كلمةً مفردةً لإيصال أفكارٍ تامةٍ قد تعادل جملةً كاملةً في اللغة الإنكليزية.

قد يتحول مفهوم التعرف على الكلام في المستقبل إلى استيعاب الكلام، إن النماذج الإحصائية التي تسمح اليوم للحاسوب بتحديد ما يقوله الشخص ربما ستسمح في المستقبل بفهم المعنى المراد الكامن وراء الكلام. على الرغم من وجود قفزةٍ كبيرةٍ في القدرة الحاسوبية والتعقيد في البرمجيات فإن هناك جدلًا بين الباحثين على أن الجهود المبذولة في تطوير أنظمة التعرف على الكلام ربما تتجه من مجال الحواسيب إلى فروع الذكاء الاصطناعي الفعلي.

إذا كنا قادرين في هذه الأيام على التكلم مع الحواسيب فإننا بعد **25** سنةً ربما سنكون قادرين على إجراء حواراتٍ

- التاريخ: 2018-09-29
- التصنيف: تكنولوجيا

#نظام التعرف على الكلام

المصادر

- الصورة
- HowStuffWorks

المساهمون

- ترجمة
  - لايا البشلاوي
- مُراجعة
  - كرار زيني
- تحرير
  - رأفت فياض
- تصميم
  - رنيم ديب
- نشر

- أمل أحمد