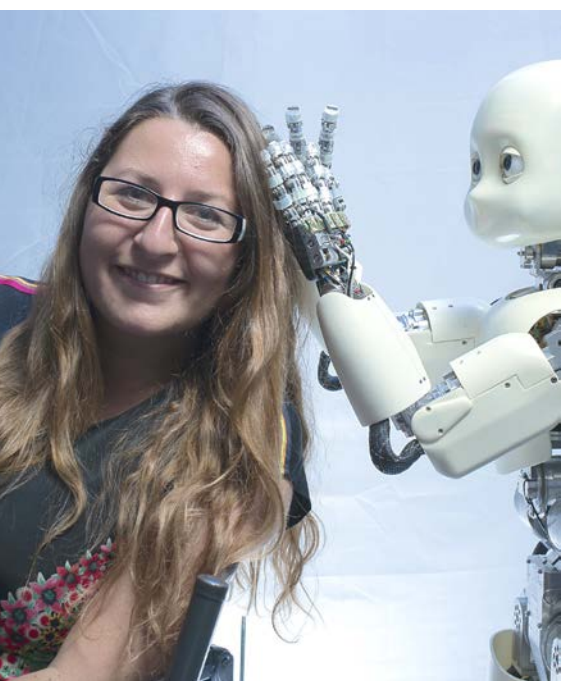


■ L'AUTEURE



Pascale FUNG est professeure d'électronique et de génie informatique à l'université des sciences et technologies de Hong Kong.

LES ROBOTS EMPATHIQUES
seront-ils un jour sympathiques ?
La réponse dépendra de la qualité atteinte dans le développement de logiciels d'empathie artificielle et du sens du mot « amitié ». Pour certains, l'amitié avec une machine est par définition impossible. Pour d'autres, elle sera évidente. Comment vivrons-nous avec nos amis les gentils robots ?



les exploite pour saisir l'état émotionnel de l'autre et pour lui faire comprendre le sien. Ces signaux sont notamment les mimiques faciales, tous les éléments du langage corporel (attitude, gestes...), le ton de la voix, mais aussi, évidemment, le contenu du discours. La mise au point d'un module empathique implique donc d'identifier les traits de la communication humaine exploitables par les machines, puis d'« entraîner » des algorithmes d'apprentissage automatique.

Du moins est-ce ainsi que s'y est prise mon équipe pour développer des machines capables non seulement d'identifier les mots et leur sens, mais aussi d'extraire plusieurs caractéristiques fondamentales du signal acoustique qu'est la voix. Ainsi, même si nous la considérons peu sous cet angle, la communication entre humains est aussi une forme de traitement du signal. Le cerveau est entraîné à détecter les émotions convoyées par la voix à partir des caractéristiques des signaux acoustiques qui leur sont associés, qu'il s'agisse de stress, de joie, de peur, de colère, de dégoût... Une personne joyeuse parlera plus vite et sur un ton aigu ; stressée, elle adoptera au contraire une voix plate et sèche. Doté d'un algorithme d'apprentissage automatique convenablement entraîné, un ordinateur peut, comme un cerveau, détecter ce genre de caractéristiques, un peu comme un détecteur de mensonges mesure la tension artérielle, les pulsations cardiaques et la conductivité de la peau.

Pour obtenir une machine ayant, par exemple, la capacité de détecter le stress, nous avons supervisé l'entraînement d'un algorithme en lui faisant analyser des enregistrements de la voix de personnes stressées.

Déceler la détresse dans la voix

C'est ainsi que nous avons commencé à apprendre aux machines à déceler la détresse dans des extraits vocaux d'étudiants de notre université, lesquels s'y connaissent puisque, entre eux, ils surnomment notre établissement l'« université du stress et de la tension de Hong Kong ». Pour obtenir ces extraits, nous leur avons posé une série de 12 questions de plus en plus stressantes et avons obtenu ainsi un premier corpus de signaux vocaux stressés en anglais, mandarin et cantonais. Dix heures d'enregistrement ont suffi pour que nos algorithmes reconnaissent le stress dans 70 % des cas où il est présent

dans la voix. Ce niveau de performance est comparable à celui des auditeurs humains.

Pendant que nous effectuions ces travaux, une autre équipe du groupe apprenait aux machines à reconnaître l'humeur exprimée par la musique d'un morceau de musique de variété, indépendamment des paroles. Contrairement à l'émotion, l'humeur est une ambiance qui reste généralement pendant toute la durée d'un morceau. Nos chercheurs ont commencé par rassembler quelque 5000 morceaux en tous genres dans les principales langues européennes et asiatiques. Les musicologues avaient d'ailleurs déjà classé plusieurs centaines de ces morceaux en 14 catégories d'humeur.

À l'aide d'ordinateurs, nous avons extrait de ces chansons 1000 caractéristiques fondamentales – par exemple l'énergie ou des propriétés acoustiques telles la fréquence fondamentale, les harmoniques... –, puis nous avons utilisé ce corpus pour entraîner 14 programmes « classificateurs », chacun spécialisé dans une humeur particulière. L'un d'eux décèle par exemple l'humeur mélancolique d'une chanson, un autre l'humeur joyeuse, etc. Comme ces 14 classificateurs peuvent être employés en même temps, leurs résultats peuvent être combinés, mais il arrive qu'ils se contredisent, par exemple que le détecteur d'humeur joyeuse trouve joyeuse une chanson mélancolique. Dans un tel cas, nous le réentraînons au cycle d'apprentissage suivant. D'anomalie en anomalie, nous avons ainsi fini par rendre performant le système formé des 14 détecteurs d'humeur. Nous avons ainsi obtenu une machine capable d'extraire l'humeur de n'importe quel morceau de musique juste en « l'écouter ».

Depuis, nous nous sommes appuyés sur ces recherches pour fonder avec d'anciens étudiants la société Ivo Technologies de développement de machines empathiques utilisables à domicile. *Moodbox*, son premier produit, sera un centre de contrôle d'ambiance, qui réglera la musique et la lumière de chaque pièce en fonction des émotions de ses utilisateurs.

Toutefois, pour saisir l'humour, le sarcasme, l'ironie et d'autres caractéristiques complexes de la communication verbale interhumaine, une machine aura besoin de plus que juste un logiciel de reconnaissance des émotions véhiculées dans le son. Il lui faudra aussi comprendre la signification sous-jacente d'un discours et en comparer le contenu avec l'émotion qui le baigne.

Depuis les années 1980, des chercheurs ont développé des systèmes perfectionnés de reconnaissance de la parole, de sorte que cette technologie atteint sa maturité. Pour autant, il y a loin entre transcrire un discours et le comprendre ! Songez à la complexité de la chaîne d'événements cognitifs, neurologiques et musculaires qui se combinent quand deux personnes se parlent. L'un des interlocuteurs formule ses pensées, choisit des mots puis articule, tandis que l'autre reçoit les sons et les autres signaux envoyés, avant d'y déchoder le message qui y est enfoui.

La chaîne qui s'établit lors d'un dialogue entre un humain et une machine est plutôt la suivante : les ondes sonores associées à la parole sont d'abord numérisées, puis de nombreux paramètres caractérisant le discours sont extraits de la série numérique obtenue ; ensuite, un logiciel de reconnaissance de la parole convertit ces paramètres en mots et un déchodeur sémantique transforme la série de mots ainsi extraite en sens.

Au début de nos recherches sur les robots empathiques, nous nous sommes aperçus que des algorithmes similaires à ceux conçus pour analyser les émotions exprimées par des utilisateurs dans leurs commentaires en ligne pourraient nous aider. Ces algorithmes d'apprentissage automatique traquent la présence, dans leur contenu, de signaux révélateurs de la présence d'émotions. Des mots-clés tels que « chagrin » et « crainte » suggèrent la solitude. L'utilisation répétée d'expressions familières (« *Let's go !* », par exemple) révèle qu'une chanson est énergique. L'information donnée par le style du discours est aussi à prendre en compte : les réponses d'une personne sont-elles assurées et claires, ou hésitantes, parsemées de silences et de formulations non compromettantes ? Élaborées et détaillées, ou brèves et sèches ?

Au cours de nos recherches sur la reconnaissance de l'humeur dans la musique, nous avons entraîné des algorithmes à repérer des signaux émotionnels dans les paroles. Au lieu d'extraire des signatures sonores de chaque morceau de musique, nous avons tiré, à partir des paroles, des séries de mots – des *n*-grammes – dont nous nous sommes servis pour entraîner 14 classificateurs, chacun responsable de la détection d'une des 14 humeurs définies précédemment.

« Nous avons obtenu une machine capable d'extraire l'humeur de n'importe quel morceau de musique »

Nous avons aussi défini ce que nous nommons des « étiquettes de partie de discours », que nous avons prises en compte dans nos définitions des signatures de telle ou telle humeur. Sur la base d'une collection suffisante de jeux de ces paramètres extraits de discours humain, un ordinateur peut inférer des approximations statistiques des règles grammaticales de n'importe quelle langue. Ce sont ces règles élaborées au cours d'un entraînement qui permettent à des programmes tel *Siri* de comprendre à leur manière un discours ou encore à un programme tel *Google Translate* de traduire un texte.

Dès qu'une machine peut comprendre le contenu d'un discours, elle peut comparer ce contenu avec la façon dont il est émis. Un algorithme pourra détecter la contradiction entre le sens d'un discours et les signaux émotionnels qui l'accompagnent. C'est le cas, par exemple, quand la phrase « Je suis si heureux de passer le week-end à travailler » est accompagnée d'un soupir. Dans un tel cas, l'ordinateur calculera la probabilité que la déclaration humaine soit en réalité un sarcasme.

Intentions complexes

Afin de détecter les intentions plus complexes, une machine capable d'appréhender le contenu d'un discours et les émotions qui le baignent pourra aussi associer les informations recueillies à des données extérieures susceptibles d'aider à son interprétation. Si quelqu'un s'exclame « J'ai faim ! », elle tiendra compte, pour déterminer la meilleure réponse à apporter, de l'heure ou du lieu ainsi que de sa mémoire des usages et préférences de la personne. Au milieu du Sahara, elle s'abstiendra de proposer la recherche d'une « pizzeria dans le quartier » ; à la maison à l'heure du déjeuner, en revanche, elle pourrait répondre : « Souhaitez-vous que je vous prépare un en-cas ? »

Début 2015, des étudiants et des chercheurs postdoctoraux de mon laboratoire ont commencé à réunir nos différents modules de reconnaissance de la parole et de reconnaissance de l'émotion pour créer une machine empathique prototype que nous appelons *Zara la Supergirl*. Il a fallu des centaines d'heures de données pour entraîner *Zara*, mais aujourd'hui, ce programme tourne