Les défis en matière de traitement de la parole

Le but de cet exposé est de donner un état de l'art du traitement automatique de la parole et, plus particulièrement, de la reconnaissance de la parole, et de présenter les grands axes de recherche innovants dans le domaine. Actuellement, les axes de recherche porteurs sont l'accès à l'information par le contenu audio et les systèmes d'interaction homme-machine.

Présentation du conférencier

Mon nom est Laurence Devillers, je suis maître de conférences à l'Université de Paris-Sud, en France, et membre du LIMSI-CNRS. Le LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur) possède un département de recherche sur la communication homme-machine. Une des unités du département s'intéresse plus particulièrement au traitement de la langue parlée. Ses activités portent notamment sur la reconnaissance vocale, l'indexation audio et la création de systèmes de dialogue homme-machine. Au sein de ce groupe, je participe à diverses activités de recherche, tant au niveau acoustique que linguistique: décodage acoustico-phonétique, modèles de compréhension de la parole, systèmes de dialogue et, enfin, détection des émotions à partir d'indices prosodiques et lexicaux. J'ai aussi beaucoup d'intérêt pour les projets de création de matériaux langagiers et d'évaluation.

Je collabore notamment à une action «Corpus» qui vise à mettre en place une plate-forme de stockage des corpus provenant de différentes universités, qu'il s'agisse de corpus produits par des linguistes, des spécialistes du geste ou de l'oral, etc. Cet instrument commun fonctionnera à l'aide de standards d'annotation reconnus. Cela permettra aux acteurs de synchroniser leurs efforts et de réutiliser les matériaux linguistiques disponibles.

En matière d'évaluation, j'ai participé aux activités de l'action de recherches concertées ARC B2 de l'Aupelf-Uref relatives à la définition de métriques d'évaluation des systèmes de dialogue. J'ai aussi participé au montage du projet français *Technolangue* MEDIA/EVALDA qui a débuté en janvier 2003. MEDIA porte sur l'évaluation de la compréhension *hors* et *en* contexte du dialogue. Le projet a pour but de mettre en place une méthodologie d'évaluation pérenne et de mener une campagne d'évaluation auprès des principaux acteurs académiques et industriels français du domaine.

Des connaissances en progression rapide

Les premiers systèmes élaborés dans le secteur des technologies de reconnaissance de la parole ont été des systèmes de commande et de contrôle qui ne nécessitaient la reconnaissance que d'un petit nombre de mots isolés. Ils ne pouvaient fonctionner que dans un environnement très calme.

La dictée vocale, adaptée à la voix d'une personne, a été popularisée par des compagnies comme IBM (avec *ViaVoice*) ou Dragon Systems (avec *NaturallySpeaking*). Actuellement, les applications les plus importantes dans le domaine sont l'indexation des documents par le contenu audio et les systèmes de dialogue.

La reconnaissance de la parole est un domaine intrinsèquement pluridisciplinaire où interviennent des spécialistes de disciplines comme le traitement du signal, l'acoustique, la phonétique, la linguistique, l'intelligence artificielle et les sciences cognitives. Les modélisations les plus performantes en ce moment reposent sur des modèles statistiques produits à partir de corpus de données réelles.

Le signal de la parole est complexe. Cette complexité est due à plusieurs facteurs: à la coarticulation des sons avec les sons voisins (pas de coupure entre les mots), à des problèmes d'homomorphie (pour une même prononciation d'un mot plusieurs écritures sont possibles, par exemple en français de nombreuses lettres sont muettes), quatreaxes principaux de variabilité du signal sont dus aux locuteurs, aux styles de parole, au langage de l'application et à l'environnement de transmission.

De nombreux progrès ont été faits depuis 20 ans. Les recherches sur la reconnaissance de la parole portent maintenant sur des tâches de plus en plus difficiles comme la parole conversationnelle (conversations téléphoniques par exemple) ou sur les programmes radiotélévisés: fictions, documentaires ou journaux télévisés. Les enregistrements traités peuvent être obtenus dans des environnements de transmission sonore difficiles

(enregistrements sur des lieux publics, au téléphone). Les systèmes de reconnaissance de la parole sont capables de prendre en compte de plus en plus de variabilité inter-locuteur comme les accents régionaux ou étrangers.

Des modèles spécifiques sont développés pour chaque type d'application. Par exemple, reconnaître des textes de journaux télévisés n'a rien de commun avec reconnaître le contenu d'un texte de dictée vocale. Le traitement de la langue spontanée nécessite de disposer de modèles différents, principalement pour le modèle de langage et le lexique. Par exemple, le lexique de reconnaissance doit tenir compte des interjections et hésitations qui n'existent pas dans les modèles de dictée. Il doit aussi contenir des formes agglutinées et fragmentées, puisqu'en mode de parole spontanée (ou relâchée), tous les phonèmes qui constituent une phrase ne sont pas toujours prononcés. Par exemple : « J'drais un billet d'train demain. » Il faut donc pouvoir prendre en compte les variantes de prononciation des mots « je voudrais » en « j'drais », « de train » en « d'train ». Pour chaque application, il faut également des modèles acoustiques adaptés au canal de transmission, par exemple pour le canal téléphonique. Des modèles spécifiques sont aussi en général développés pour les locuteurs féminins et masculins.

Évaluation, corpus et outils

Les progrès réalisés, ces dernières années, en reconnaissance de la parole sont principalement dus à l'existence de campagnes d'évaluation et à la disponibilité de corpus.

Les programmes d'évaluation de l'agence militaire américaine DARPA, programmes auxquels le LIMSI participe depuis le début des années 1990, permettent d'obtenir énormément de données et de confronter les technologies françaises à celles de laboratoires américains prestigieux. En 1996, DARPA classait le système du LIMSI au premier rang des logiciels de reconnaissance de la parole. En 1997 et 1998, il était classé parmi les trois meilleurs systèmes. De 1999 à 2002, DARPA classait de nouveau le système du LIMSI au premier rang. Les campagnes d'évaluation ont beaucoup évolué au fil des ans. Ainsi, l'enjeu, de 1992 à 1995, consistait principalement à dicter des textes de journaux contenant des vocabulaires de plus en plus grands. Depuis 1995, il s'agit de transcrire des émissions d'informations radiotélévisées de type HUB4 et des conversations téléphoniques de type HUB5. Actuellement, les évaluations portent sur la capacité des systèmes à faire des transcriptions enrichies. L'enjeu n'est plus seulement de reconnaître ce qui a été dit, mais aussi de reconnaître qui l'a dit, dans quelle langue et dans quel style.

Dans un système de transcription de la parole, deux traitements sont utilisés. Le premier fait une analyse acoustique du signal audio et le transforme en une suite de vecteurs de paramètres acoustiques. Le second est le module de reconnaissance, il transforme la suite de vecteurs acoustiques en une suite de mots. Il utilise trois types de modèles : des modèles acoustiques, un dictionnaire de prononciation et des modèles linguistiques. Les modèles acoustiques sont des modèles de Markov cachés qui permettent de construire des modèles de sons à partir de la succession des vecteurs acoustiques. Ces modèles de sons tiennent compte des distorsions temporelles du signal de parole et utilisent des densités de probabilité multigaussiennes par état. Le dictionnaire est une ressource qui contient toutes les variantes de prononciation, à chaque mot du système de reconnaissance sont associées plusieurs formes phonétiques. Les modèles linguistiques sont des modèles de probabilité de succession des mots, des modèles statistiques n-grammes (bigrammes, trigrammes, quadrigrammes, etc.).

L'évaluation est aussi un sujet important en France. Citons l'action Aupelf-B1 en 1997, qui portait sur l'évaluation de la capacité des systèmes à faire la transcription de journaux lus et, actuellement, la campagne ESTER, qui va se prolonger jusqu'en 2005. ESTER a pour but l'évaluation de la capacité de différents systèmes à déterminer le contenu et les langues d'émissions d'informations radiophoniques et télévisuelles, de même que les types de locuteurs prenant part aux conversations. Ce projet fait intervenir plusieurs laboratoires français.

Voici quelques performances indicatives des systèmes de reconnaissance sur des tests en anglais américain. Ces performances sont mesurées en taux d'erreur sur les mots

à partir de la comparaison entre le texte reconnu et le corpus de référence transcrit manuellement. Les systèmes font peu d'erreurs quand il s'agit de reconnaître des nombres (0,7 %). Le taux d'erreur est de 7 % (0,5 % pour les performances humaines) dans le cas de journaux dont le contenu est lu à haute voix. Dans le cas des journaux télévisés, le taux grimpe à 20 % d'erreur. Pour une conversation téléphonique, il est de l'ordre de 35 % contre 4 % pour un être humain. La machine est encore loin des performances humaines.

L'évaluation des systèmes de dialogue est beaucoup plus difficile que l'évaluation des systèmes de reconnaissance de la parole. Actuellement, il n'y a pas de consensus sur la manière d'évaluer un système : on fonctionne beaucoup avec des critères subjectifs. Il n'y a pas de standard d'annotation non plus. En France, l'objectif de la campagne MEDIA du projet Technolangue est d'amener les grands laboratoires et les entreprises du secteur à s'entendre sur des standards d'annotation et des façons d'évaluer la qualité des systèmes de dialogue.

Les modélisations les plus performantes en reconnaissance de la parole reposent sur des modèles statistiques qui nécessitent de grands corpus. On peut donner une idée de la taille des corpus utilisés pour la reconnaissance de grands vocabulaires. Le corpus audio pour entraîner les modèles acoustiques en américain contient 200 heures de documents sonores radiophoniques et télévisuels (ABC, CNN,…) soit un total de 1,9 million de mots issus de transcriptions fines, c'est-à-dire qu'il est possible de faire un bon alignement entre le signal audio et la transcription linguistique. Pour construire des modèles acoustiques en français et en allemand, des sources radio et télévisées ont été enregistrées (Arte, TF1, A2, FrInfo, FrInter). Le *Linguistic Data Consortium* (LDC) fournit aussi des données dans d'autres langues que l'anglais américain comme le mandarin. Pour entraîner les modèles de langage, des données textuelles issues de journaux ou des transcriptions commerciales sont utilisées. Pour donner un ordre d'idées, en américain, 790 millions de mots issus de journaux et 240 millions de mots de transcriptions commerciales ont été utilisés.

Verrous technologiques

Il n'existe pas de système universel de reconnaissance de la parole. Le développement d'une application et l'entraînement des modèles nécessitent toujours le recours à des données réelles. Cependant, l'annotation de ces données est très coûteuse. Le défi consiste donc à favoriser l'adaptation rapide des modèles acoustiques et des modèles de langage dont nous disposons, à permettre à la machine d'apprendre automatiquement des prononciations. Il faudra aller dans cette direction si on veut être capable de produire des modèles performants à moindre coût. Pour illustrer ce propos : lorsqu'on entraîne un modèle acoustique à l'aide d'un corpus de 500 heures transcrites automatiquement et d'un corpus d'une heure transcrite manuellement, on obtient les mêmes performances que quand on a recours à un corpus de 200 heures transcrites manuellement.

Le « challenge » est donc de trouver des technologies génériques et de minimiser la collecte de corpus transcrits.

Au-delà de la reconnaissance de la parole

Les travaux en reconnaissance automatique de la parole s'étendent aussi à l'identification de la langue et du locuteur. Les mêmes technologies sont utilisées. Les difficultés sont différentes suivant les langues et les locuteurs et varient grandement selon le type de parole. Comme nous l'avons déjà souligné, il est difficile de transcrire automatiquement des journaux télévisés, mais il est encore plus difficile de transcrire automatiquement des conversations téléphoniques de parole spontanée.

Pour donner un ordre des performances de reconnaissance du locuteur, la machine ne se trompe que dans 1 % des cas quand elle a eu deux minutes pour se familiariser avec la voix de chacun des locuteurs appartenant à un groupe de 630 personnes enregistrées dans d'excellentes conditions en studios. Son taux d'erreur est nettement plus élevé, à 40 %, dans le cas d'enregistrements téléphoniques spontanés. Cette hausse importante du taux d'erreur vient en partie du fait que la machine ne peut pas reconnaître deux personnes qui parlent en même temps, phénomène fréquent dans l'oral spontané. En fait, on estime que les gens parlent en même temps jusqu'à 30 % du

temps, souvent pour maintenir la conversation en donnant des signes d'accord.

Pour donner un ordre de performance en identification de la langue sur le corpus OGI de parole téléphonique en 11 langues (anglais, farsi, français, allemand, hindi, japonais, coréen, mandarin, espagnol, tamoul, vietnamien), à 10 secondes, le système a un taux d'erreur d'identification de la langue de 20%, à 45 secondes de parole, son taux d'erreur chute à 10% seulement, le corpus d'apprentissage étant de 24 heures d'enregistrement de bonne qualité, de 1 heure 30 minutes à 4 heures 30 minutes par langue.

Quelques enjeux et applications innnovantes

L'indexation audio représente un enjeu considérable dans le domaine de la reconnaissance de la parole. En effet, chaque pays produit et stocke chaque année des centaines de documents audiovisuels qu'il faudrait arriver à structurer et à organiser à moindre coût. Une des manières d'aborder la question consiste à indexer les documents audiovisuels de manière automatique, à l'aide de la reconnaissance de la parole. Les applications de cette technologie seront multiples: elle facilitera la recherche documentaire, permettra de produire de l'information à la demande, facilitera la surveillance des médias, etc.

Il faut d'abord, pour faire de l'indexation automatique, structurer le signal audio en tours de parole, segments musicaux, etc. Ensuite, le contenu audio est transcrit automatiquement avec un système de reconnaissance de la parole. Finalement, l'indexation peut se faire. Ceci signifie que les textes doivent avoir été annotés et balisés par des thèmes de référence. Le procédé consiste à utiliser des lemmatiseurs et des filtres sur les mots les plus connus, de manière à produire des modèles thématiques. La recherche de document se fait ensuite par une requête en langage naturel. Voici un exemple de la manière dont fonctionne la recherche de document. La requête de l'utilisateur subit un pré-traitement. Par exemple, la question: « Quelles sont les mesures prises contre la maladie de la vache folle ? » devient après lemmatisation « Quelle mesure prendre maladie vache folle ? » Cette requête est ensuite enrichie par l'ajout des termes les plus fréquemment retrouvés dans les documents du corpus les plus pertinents à la requête. Une deuxième requête ainsi constituée est lancée. Le moteur de recherche se sert de l'index pour extraire les documents audiovisuels les plus pertinents, c'est-à-dire ceux qui ont une similarité entre les termes de la requête et ceux des segments indexés.

Le LIMSI a mis au point un système d'indexation audio pour une application d'alerte: Audiosurf. Ce système est un prototype de laboratoire. Une personne s'inscrit à ce service en indiquant les thématiques qui l'intéressent. Tous les jours, Audiosurf indexe les journaux télévisés diffusés sur plusieurs chaînes de télévisions nationales (Fr2, Arte...). La base contient actuellement une centaine d'heures de journaux télévisés. Quand le système détecte la présence d'un thème, il envoie un courriel au client avec l'extrait de la vidéo du journal détecté. Par exemple, le client intéressé au thème *météorite* se verrait envoyer l'extrait du document vidéo correspondant au bulletin sonore suivant:
« Des scientifiques sont en train d'arriver sur place. Romuald Bonnant: «Un objet lumineux fonçant dans le ciel à plus de 200000 kilomètres/heure et puis, quelques secondes plus tard, le sol qui tremble. Voilà ce que les habitants de deux villages sibériens ont vu et ressenti lorsque tout à l'heure, en pleine nuit, une météorite a heurté la terre à pourtant plus de 100 kilomètres de chez eux. Une importante pierre céleste..." » Le document détecté correspond bien au thème recherché, malgré 20% d'erreurs de reconnaissance dans ce type d'enregistrement dues en majorité à des fautes d'accord grammatical, des homophones et des noms propres. Il a été montré que 20% d'erreurs de transcription est la limite pour ne pas dégrader les performances de détection des documents pertinents.

Un autre champ d'application intéressant de la reconnaissance de la parole est celui de la traduction parole-parole. Le défi consiste à faire de la reconnaissance de conversations dans des domaines très différents.
Ce type de système est évidemment difficile à réaliser. Une difficulté qui s'ajoute aux difficultés de traduction vient du fait que l'on part de documents oraux où il y a des erreurs de reconnaissance. On peut, par exemple, imaginer qu'un tel système de traduction fonctionnerait sur le mode du dialogue. Il serait possible à l'usager de discuter avec le système pour raffiner la traduction qu'il veut obtenir.
Le groupe TLP au LIMSI participe à un projet de ce type dans le cadre d'un projet européen du FP6.

Un autre axe très porteur de la reconnaissance de la parole est celui des systèmes d'interaction homme-machine.

Où désirez-vous aller ?

Voici un exemple d'un système de dialogue développé au LIMSI dans le cadre d'un programme européen. Il s'agit d'un système utilisé pour donner des renseignements sur les horaires de train de la SNCF. Le lexique de ce genre d'application multilocuteurs est assez restreint. Il compte seulement 2 000 mots.

Système : Quel renseignement désirez-vous ? Locuteur : Je voudrais aller de Paris à Avignon demain à 14 h 30. *Vous avez un train direct de Paris à Avignon à 15 h 30 arrivant à 18 h 50 céans le 12 octobre. Qu'en pensez-vous ? Voulez-vous réserver ?* Quel est le prix en première classe non-fumeur avec une réduction Carissimo ? *Le prix du billet Paris-Avignon est de 301 francs en première classe avec une réduction Carissimo. Souhaitez-vous une autre information ?* Je souhaiterais aller de Grenoble à Clermont-Ferrand à Noël. Jeudi 25 décembre. *À quel moment de la journée voulez-vous partir ?* À 14 h 35. *Pardon je ne vous ai pas compris, pourriez-vous répéter svp ?* À 14 h 30. *Vous avez un train partant de Grenoble à 14 h 08, arrivant à Clermont-Ferrand à 20 h 53 avec un changement à Paris, gare de Lyon, le 25 décembre*.

Une version de ce système d'information sur les horaires de train de la SNCF est maintenant accessible au grand public par téléphone. Une autre fonctionne à partir d'un kiosque multimodal qui a été évalué à la gare Saint-Lazare auprès de vrais voyageurs, mais reste encore un prototype de laboratoire. Pour le protocole de test, chaque sujet devait effectuer quatre transactions. Quatre-vingt-treize p. cent de celles-ci ont été réalisées avec succès. La perception du public par rapport au logiciel s'est avérée excellente : 87 % des sujets ont dit trouver plus agréable de transiger avec une borne compréhensive et agréable à l'oreille plutôt qu'avec les bornes actuelles.

Un autre axe innovant du domaine est la détection des émotions. Par exemple, il serait intéressant que la machine conversant avec le client de l'extrait suivant puisse discerner son inquiétude : « Voilà, voilà, mon problème est le suivant : ça m'indique "compte verrouillé". » Cela pourrait l'amener à changer de stratégie de communication ou, tout simplement, à transmettre l'appel à un opérateur humain.

Parmi les applications innovantes dans le domaine de l'interaction orale, il faut citer aussi les systèmes de communication homme-homme médiatisée par la machine. Ce type de système pourrait être capable d'observer, de reconnaître et de comprendre des communications et des comportements interpersonnels et de rendre des services aux usagers. On pourrait par exemple imaginer l'existence d'un petit système complice adapté à chaque propriétaire. Le système aurait une interface multicapteur et multimodale robuste rendant possible la saisie de sons et d'images dans un environnement bruité et la possibilité de stocker et d'organiser l'information. Un utilisateur dialoguant avec ce système pourrait, par exemple en réunion, retrouver l'information, identifier les personnes dans la salle, etc.

En conclusion, des progrès considérables en reconnaissance de la parole ont été faits au fil des ans, mais nous sommes encore loin des performances de l'être humain. La difficulté vient des nombreux niveaux de variabilité du signal de parole : conditions acoustiques, divers accents, parole superposée, parole spontanée, etc. La production d'un système universel de reconnaissance de la parole n'est pas encore pour demain. Nous devons également faire face à de nombreux défis quand vient le temps de créer des applications allant au-delà de la reconnaissance de la parole, comme celles de la traduction parole-parole ou celles de la compréhension de conversation entre deux personnes. Ces domaines innovants sont des enjeux scientifiques pour l'avenir et nous sommes encore loin de réaliser ce que les films de science-fiction nous promettaient voilà plusieurs années.

Laurence Devillers,
maître de conférences à l'Université de Paris-Sud et membre du LIMSI, France.