La parole est à l'ordinateur

Vishwa Gupta, Matthew Lennig, Paul Mermelstein, Douglas O'Shaughnessy

Depuis les débuts de l'ordinateur dans les années 50, on a consacré beaucoup d'efforts à faciliter le dialogue homme-machine. A l'entrée, les claviers ont généralement remplacé les cartes perforées tandis que les écrans vidéo constituent une nouvelle forme de sortie, s'ajoutant à l'imprimé conventionnel. Aujourd'hui, on commence à utiliser le mode de communication le plus pratique: le dialogue direct avec l'ordinateur, à la fois en entrée et en sortie.

Cependant, comme la console de visualisation n'a pas éliminé les sorties imprimées, la communication parlée, plutôt que de remplacer les techniques existantes d'entrée-sortie, en augmente la diversité.

La technologie de la communication parlée recèle la promesse qu'un jour le téléphone va devenir le dispositif d'entrée-sortie à l'ordinateur le plus courant. On va pouvoir entrer l'information en répondant aux questions orales posées par l'ordinateur et la recueillir à l'aide d'un dialogue structuré qui va permettre de l'identifier. En communiquant directement avec la base de données, l'utilisateur d'un service de ce genre va pouvoir, par exemple, déterminer l'heure de départ d'un avion, savoir s'il reste des places, et aussi faire sa réservation et la faire porter à son compte. Bien que l'accès verbal direct à ce genre de service d'information n'existe pas encore, la technologie actuelle va offrir des formes simplifiées de ces systèmes prochainement.

La technologie de la reconnaissance et de la production de la parole va aussi améliorer les services de télécommunication actuels. De même que la transmission de textes a beaucoup de succès là où il y a des terminaux accessibles et du personnel qualifié, la transmission de messages verbaux promet des avantages similaires, sans la nécessité des terminaux. Le fait qu'un grand pourcentage des appels d'affaires ne parviennent pas à destination parce que le demandé est absent montre la nécessité d'un service de transmission de messages verbaux. Un service connectant automatiquement le demandeur à un système de messages automatique, lorsque la ligne du demandé est occupée ou sans réponse, pourrait grandement augmenter la satisfaction des clients.

En 1977, les Recherches Bell-Northern ont formé leur propre groupe de recherche sur les systèmes de communication de la parole. Avec le groupe de recherche sur la parole de l'INRS-Télécommunications (Université du Québec), le groupe des Recherches Bell-Northern étudie les relations entre la technologie existante et les applications possibles et identifie les problèmes techniques requérant une recherche plus poussée en vue de réaliser ces applications.

L'équipe de 10 membres des Recherches Bell-Northern et de l'INRS-Télécommunications se penche sur des problèmes de deux ordres principalement. Premièrement, le codage efficace de la parole pour la transmission numérique et deuxièmement, la conception de systèmes de communication homme-machine réalisables. Il s'agit dans le premier cas,de coder la parole de façon à minimiser la capacité de transmission nécessaire sans altérer la qualité de la parole. Les problèmes, dans le second cas, comprennent le développement de techniques de reconnaissance par l'ordinateur d'information verbale, la production de parole pour l'auditeur humain et la conception de protocoles de communication intégrant ces deux techniques pour des fins particulières. Afin d'obtenir l'expertise appropriée à une si large variété de sujets, l'équipe comprend des théoriciens en communications, des informaticiens,des linguistes et des psychologues.

Dans cet article, on traite des recherches actuelles des Recherches Bell-Northern et de l'INRS-Télécommunications dans le domaine de la synthèse et de la reconnaissance de la parole. Seuls les points importants sont présentés et le lecteur est renvoyé aux textes techniques pour plus de détails.

Production et synthèse de la parole

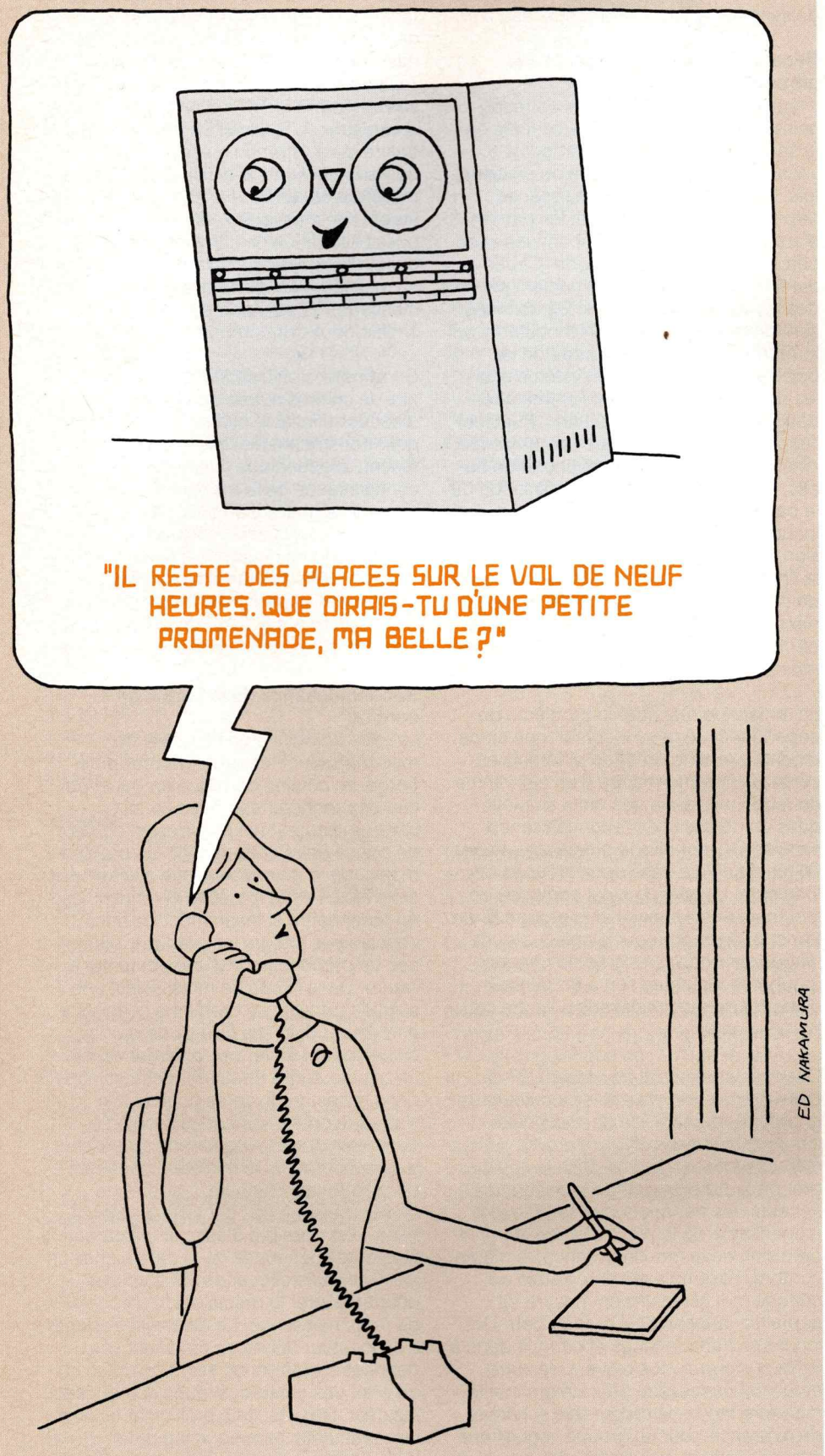
La variété des messages requis constitue le facteur principal pour déterminer la méthode de production de la parole utilisée. Lorsqu'on n'a besoin que de quelques messages, on peut enregistrer chaque message individuellement, le mettre en mémoire sous forme analogique ou numérique et le diffuser sur demande. Cependant, lorsque la capacité de mémoire de la machine est inférieure à ce qui est nécessaire pour enregistrer chaque message intégralement, il faut faire une synthèse de la sortie parlée à partir d'unités plus petites. Les sorties parlées s'appliquent aux systèmes de maintenance avec signaux d'avertissement verbaux, qui ne requièrent que quelques messages différents, aussi bien qu'à l'obtention de la version orale d'un texte complet, où l'enregistrement de chaque message séparément est impossible. Dans le cas d'exigences intermédiaires dans la variété des messages, il faut comparer le coût du codage de la parole sous une forme comprimée et de son décodage sur demande, avec les coûts réduits d'enregistrement. Si les messages doivent être non seulement intelligibles mais aussi paraître naturels, cela peut réduire encore le degré de compression des signaux de paroles qu'on peut atteindre.

Une des recherches actuelles consiste à tenter d'établir la qualité des messages produits, à l'aide de formes hautement comprimées de mots séparés. Des mots prononcés naturellement sont traités pour séparer l'information concernant le ton, ou fréquence de vibration des cordes vocales, de celle qui spécifie les sons individuels du discours. Lorsqu'il faut reproduire ces mots, on le fait avec des ajustements de ton et de durée qui facilitent la compréhension du message.

Quand une personne parle, elle modifie la forme acoustique de la plupart des sons pour les adapter à la structure d'une phrase particulière. Les syllabes d'un mot polysyllabique sont prononcées avec des degrés d'accentuation variés. Ce sont principalement la durée et le ton de la syllabe qui contrôlent l'accentuation. Le schéma de tonalité de la phrase est de plus modifié selon le mot ou le groupe de mots à accentuer pour marquer un sens particulier. Ainsi, pour répondre aux deux questions suivantes:

- "Reste-t-il des places sur le vol de neuf heures?"
- "Y a-t-il un vol après huit heures?", on peut utiliser la même phrase:
- "Il reste des places sur le vol de neuf heures."

Cependant, on pourrait rendre la réponse à la première question plus intelligible en mettant l'accent sur "Il reste des places", tandis que la réponse à la seconde



" RESTE-T-IL DES PLACES SUR LE VOL DE NEUF HEURES ? "

question pourrait être plus claire en accentuant "le vol de neuf heures" – et peut-être le mot "neuf" plus particulièrement. Lorsqu'un mot est accentué, il y a non seulement un changement sensible dans le ton, mais la durée augmente aussi, surtout la durée de la voyelle de la syllabe accentuée. En outre, les mots qui précèdent une pause, comme le dernier mot d'une phrase, sont généralement prononcés plus lentement. Sans ces modifications, la parole synthétique semble dénuée de sens et peut être difficile à comprendre.

Lorsqu'on prononce une suite de mots rapidement, le premier et le dernier son de chaque mot peuvent se modifier, selon les sons des mots qui précèdent et suivent. Cette influence est plus marquée dans le cas de mots-outils courts tels que les prépositions, comme le démontre par exemple, la prononciation différente de "to" dans "to Ottawa" et dans "to Toronto". Des règles particulières, appelées règles de coarticulation, permettent de prédire l'effet que des sons adjacents ont les uns sur les autres et de modifier la forme enregistrée des mots selon leur contexte.

Les techniques employées pour produire des messages par modification des formes enregistrées de mots prononcés naturellement donnent de meilleurs résultats que celles qui assemblent les sons constituant le discours.

Bien que seuls 40 sons de base ou phonèmes soient utilisés en anglais, ou en français, leur forme spécifique, une fois prononcés, varie grandement. Il faut peut-être environ 200 unités sonores (allophones) pour représenter toutes les variantes sonores importantes. De plus, il faut lier ces sons séparés par une méthode complexe de chevauchement et d'interpolation qui imite les modifications des organes phonatoires humains lorsque ceux-ci passent de la prononciation d'un son à celle d'un autre son. Bien que ces règles soient suffisamment connues pour permettre de produire une parole intelligible, la sonorité naturelle de la voix synthétique est limitée. Peut-être à cause de sa qualité artificielle, la voix synthétique requiert de l'auditeur une plus grande attention pour la compréhension du message.

Les circuits intégrés pour la synthèse de la parole présentement sur le marché, comme le dispositif Speak and Spell de Texas Instruments, ne modifient pas le mot prononcé, mais produisent le signal de parole à partir de sa forme comprimée d'une façon peu coûteuse à réaliser. Les mêmes circuits intégrés peuvent servir à produire des suites de mots ou des phrases, qu'elles soient enregistrées en unités entières ou assemblées sur demande à partir d'unités plus petites. Les dispositifs de ce genre produisent une

assez bonne qualité de parole pourvu que les paramètres sous-jacents soient correctement spécifiés.

Pour mettre en relief les aspects du signal assemblé qui peuvent donner des résultats peu naturels, on compare la forme d'un message parlé naturel avec la même forme assemblée à partir d'unités distinctes. Comme il n'existe aucune théorie sur la sonorité naturelle de la parole,la plupart de la recherche se fait par tâtonnements et essais successifs. On part du principe que la transformation de sons individuels ou de groupes de sons en unités cohérentes s'effectue selon certaines règles et on les évalue dans un grand nombre de contextes. Les problèmes identifiés servent à modifier les règles jusqu'à l'obtention de résultats acceptables.

Bien qu'il ne soit pas encore possible de produire un texte complet de qualité suffisante pour le réseau téléphonique public, la production de ce genre de texte, sur un appareil de lecture pour handicapés visuels, a connu un grand succès. Un appareil existe déjà pour l'anglais et d'autres seront produits bientôt; cependant, aucun appareil du genre n'existe pour le français.

Pour répondre à ce besoin, nous établissons actuellement les règles de prononciation du français qui seront plus tard appliquées à un appareil de lecture de textes français.[1] Un lecteur optique sert à entrer une page dactylographiée ou imprimée dans l'appareil et un programme de reconnaissance des caractères transforme l'entrée en une suite de caractères codés accompagnés de signes de ponctuation. Une liste de règles de prononciation est consultée pour convertir les caractères en sons et en marques d'accentuation lexicale. Un analyseur de phrases élémentaire contrôle le schéma d'intonation en fixant les limites des phrases. Ces procédures forment un code de sons qui, une fois converti sous forme de paramètres, peut faire fonctionner un synthétiseur de parole.

La conversion du code de sons à la forme paramétrique fait appel à des règles imitant les contraintes du système phonatoire humain. Dans la production de la parole par l'homme, les sons sont limités par les caractéristiques physiques des organes phonatoires – la vitesse à laquelle ils peuvent changer de forme, par exemple. Le signal de parole final est obtenu en tant que sortie du synthétiseur. Actuellement, on peut simuler le système de synthèse entier, en démonstration.

Reconnaissance des mots et des phrases

Il faut entraîner les systèmes de reconnaissance des mots les plus simples en faisant réciter par l'utilisateur tout le vocabulaire du système une ou plusieurs fois. La capacité de reconnaissance dépend donc du locuteur et les paroles d'autres personnes sont reconnues avec beaucoup moins d'exactitude. Cependant, on peut éviter l'entraînement par des locuteurs individuels si les données sont produites par différents locuteurs qui forment un échantillon typique de la population qui utilisera le système et si les données représentent l'entrée attendue de la part de la population. Puisque l'accès universel constitue une exigence importante du système téléphonique public, les dispositifs de reconnaissance de la parole ne doivent dépendre ni du locuteur, ni des conditions de transmission, mais ils doivent localiser le début et la fin de chaque mot, malgré la présence de parasites ou de diaphonie. La bande passante réduite de la parole transmise par téléphone présente un problème bien moins sérieux.

En tentant de résoudre le problème de dépendance face au locuteur, une étude récente examine l'effet de différents accents sur la performance d'un algorithme de reconnaissance des mots employé dans un dispositif de reconnaissance simulé.[2] On recueille la prononciation de 20 mots anglais dans deux groupes de locuteurs. Le groupe A est composé de locuteurs anglophones et le groupe B, de francophones pour qui l'anglais est une langue seconde. La moitié de chaque groupe de locuteurs sert à entraîner le dispositif de reconnaissance, l'autre moitié, à le vérifier.

Quoique le dispositif ait reconnu 97 % des mots prononcés par les locuteurs du groupe A, sa précision de reconnaissance est tombée à 93 % avec les mêmes mots prononcés dans le cas du groupe B. La plus grande variation des accents des francophones semble être responsable de la diminution de la précision de la reconnaissance: le même nombre d'exemplaires enregistrés de chaque mot est moins en mesure de capter la variation plus grande parmi les locuteurs francophones et conduit donc à un taux d'erreur plus élevé. Une autre étude est nécessaire pour déterminer la meilleure façon de choisir des schémas de référence pour un groupe linguistique non homogène, comme les utilisateurs du téléphone par exemple.

Pour illustrer l'utilisation de la reconnaissance des mots dans le réseau public, on a conçu et simulé l'enregistrement et la diffusion de messages parlés.[3] Tout usager du téléphone peut s'adresser au système de messages et enregistrer ou recevoir un message à partir d'une série de questions verbales posées par l'ordinateur et des réponses verbales de l'usager. On demande d'abord à l'usager son numéro d'identification,puis s'il désire enregistrer ou recevoir un message. S'il désire enregistrer un message, on lui demande le numéro de destination du message et, après vérification, il peut établir son message. L'expéditeur a la possibilité de revoir son message et au besoin de le changer. S'il désire recevoir un message, on lui indique combien de messages lui sont adressés et il peut choisir ceux qui l'intéressent.

On appelle signalisation secondaire l'emploi de chiffres autres que le code d'accès initial et le numéro de téléphone pour transmettre de l'information sur le réseau téléphonique. L'emploi de la reconnaissance de la parole comme technique de signalisation secondaire a un avantage: le système téléphonique demeure toujours compatible avec la transmission de la voix tandis qu'un autre signal spécial peut ne pas être correctement transmis par un ou plusieurs composants du réseau téléphonique.

Reconnaissance de la parole en continu

La mise au point d'un dispositif de reconnaissance capable d'accepter la parole en continu constitue un but important des recherches.[4] Avec un tel système, le locuteur n'a plus besoin de faire de pause entre chaque mot. Un problème majeur de la reconnaissance à la fois des mots isolés et de la parole en continu est de reconnaître suffisamment les sons constituants malgré les grandes différences de prononciation d'une personne à l'autre. Dans le cas de mots isolés, une solution partielle au problème consiste à enregistrer diverses formes de prononciation pour refléter ces différences. On ne peut toutefois appliquer cette solution dans le cas du discours continu. On s'attaque présentement au problème dans le cadre d'un contrat de recherche subventionné par le ministère canadien de la Défense nationale. L'approche choisie dans ce cas consiste à déterminer s'il est possible d'estimer les caractéristiques générales de la parole chez un locuteur particulier et de les employer pour améliorer la précision du dispositif de reconnaissance. Le dispositif s'adapte à un locuteur donné en modifiant ses données de référence selon l'estimation qu'il fait des caractéristiques du nouveau locuteur. Des résultats préliminaires obtenus avec cette technique indiquent que le taux d'erreur pour les chiffres peut être réduit d'un facteur 3 à l'aide d'estimations

constituées de seulement trois mots.[5] On s'attend à ce que cette recherche apporte des améliorations substantielles à la performance de systèmes tels que les dispositifs de reconnaissance de chaînes de chiffres représentant un discours continu et qui peuvent fonctionner à partir d'un vocabulaire limité.

Les applications à la communication parlée entre l'homme et l'ordinateur abondent, surtout grâce à l'évolution rapide des puces de grande capacité. Il y a quelques années, les techniques de synthèse et de reconnaissance de la parole semblaient n'avoir qu'un intérêt limité. Elles représentent maintenant des solutions économiques à des problèmes de communication réels. Les nouvelles réalisations ont à leur tour donné un nouvel élan à la recherche de meilleure performance technique dans des tâches plus complexes. Dans quelques années vont être réalisés des systèmes à microprocesseurs de haute qualité pour la synthèse et la reconnaissance de la parole continue. Le défi actuel est de s'assurer que ces nouvelles possibilités servent efficacement à améliorer la communication homme-ordinateur. Actuellement, on ne possède qu'une expérience limitée sur la façon d'exploiter cette nouvelle voie de communication parlée. La conception de systèmes répondant à cette exigence constitue un objectif de recherche important pour les prochaines années.

Références

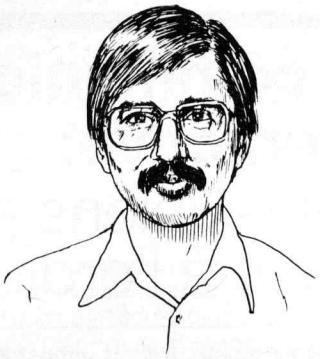
1. D. O'Shaughnessy, M. Lennig, P. Mermelstein et al: Simulation d'un lecteur automatique du français. *12ième journées d'Etude sur la Parole*, Groupement des Acousticiens de Langue Française, Montréal, Canada, 1981, p 315

2. V. Gupta, P. Mermelstein: Effects of accent on the performance of an isolated word recognizer. *J. Acoust. Soc. Am.* Vol. 68, 1980, p S86

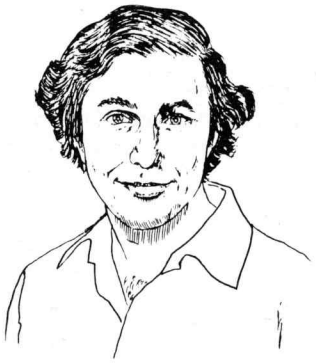
3. P. Mermelstein, V. Gupta: An experimental voice messaging system controlled by word recognition. *International Symposium on Computer Messaging Systems*, Avril 1981, Ottawa, Canada.

4. M.J. Hunt, M. Lennig, P. Mermelstein: Experiments in syllable-based recognition of continuous speech. *International Conference on Acoustics*, Speech and Signal Processing, Denver, CO, 1980, p 880

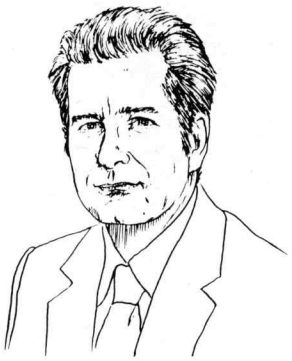
5. M.J. Hunt: Speaker adaptation for word-based speech recognition systems, J. Acoust. Soc. Am. Vol. 69, 1981, p S41



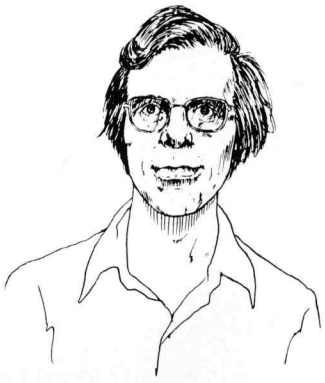
Vishwa Gupta, né en Inde, il obtient son doctorat en génie électrique en 1977 à l'Université Clemson, où il travaille ensuite comme instructeur invité pendant un an. Il se joint en 1978 au service de recherche sur la communication parlée des Recherches Bell-Northern et travaille au système de stockage et de retrait des messages, à la reconnaissance des mots isolés et des chaînes de chiffres.



Matthew Lennig, membre du personnel scientifique de Recherches Bell-Northern, à Montréal, est aussi professeur invité à l'INRS-Télécommunications, Université du Québec ainsi que conférencier à l'Université McGill, à Montréal. Il obtient son baccalauréat à l'Université de Princeton et poursuit ses études supérieures à l'Université de Pennsylvanie où il reçoit un PhD en linguistique en 1978. Il se joint ensuite aux Recherches Bell-Northern. Ses intérêts actuels en recherche comprennent l'étude de la variation linguistique dans le langage familier, la reconnaissance et la synthèse de la parole et la compréhension du langage naturel.



Paul Mermelstein est directeur de la recherche sur les communications orales aux Recherches Bell-Northern, à Montréal. Il est aussi professeur invité à l'INRS-Télécommunications, Université du Québec, et professeur auxiliaire à l'Université McGill. Né en Tchécoslovaquie, il obtient son baccalauréat à l'Université McGill et poursuit ses études supérieures au Massachusetts Institute of Technology où il obtient un doctorat en génie électrique. De 1964 à 1967, il occupe des postes de recherche en communication orale aux Bell Laboratories et aux Haskins Laboratories. Ses intérêts actuels en recherche se portent sur l'application de modèles de communication orale humaine pour améliorer la communication avec les machines.



Douglas O'Shaughnessy est professeur à l'INRS-Télécommunications, Université du Québec, et professeur auxiliaire au département de génie électrique de l'Université McGill, à Montréal. Natif de New York, il fait toutes ses études universitaires au Massachusetts Institute of Technology où il obtient un PhD en génie électrique et en informatique en 1976. Après avoir occupé un poste au MIT pendant un an, il se joint à l'INRS-Télécommunications. Ses intérêts actuels de recherche comprennent la communication parlée homme-machine et le codage numérique de signaux de conversation.