Dossier sens artificiels : des machines tout ouïe

Des machines tout ouïe

Du bruit de fond ou des mots mal prononcés rendent souvent les systèmes de reconnaissance vocale inefficaces. Des chercheurs de l'IDIAP de Martigny se basent sur les propriétés de l'oreille humaine pour les améliorer.

PAR OLIVIER DESSIBOURG

Lorsqu'elles prêtent l'oreille, les machines ont l'ouïe plutôt fine. Elles parviennent à identifier leur interlocuteur et même retranscrire un dialogue. Pour autant toutefois que les personnes « entendues » parlent distinctement, ne soient pas stressées ou émues, et que le bruit de fond reste minime. Car dans ces cas, les systèmes de reconnaissance vocale, montrant leurs limites, deviennent durs d'oreille pour distinguer des mots et la font sourde lorsqu'il s'agit de reconnaître des voix. C'est pour combler ces lacunes que de nombreux groupes, dont un au Pôle de recherche national IM2 situé à l'Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) de Martigny, mènent des recherches visant à améliorer ces systèmes.

Analyse du spectre des fréquences

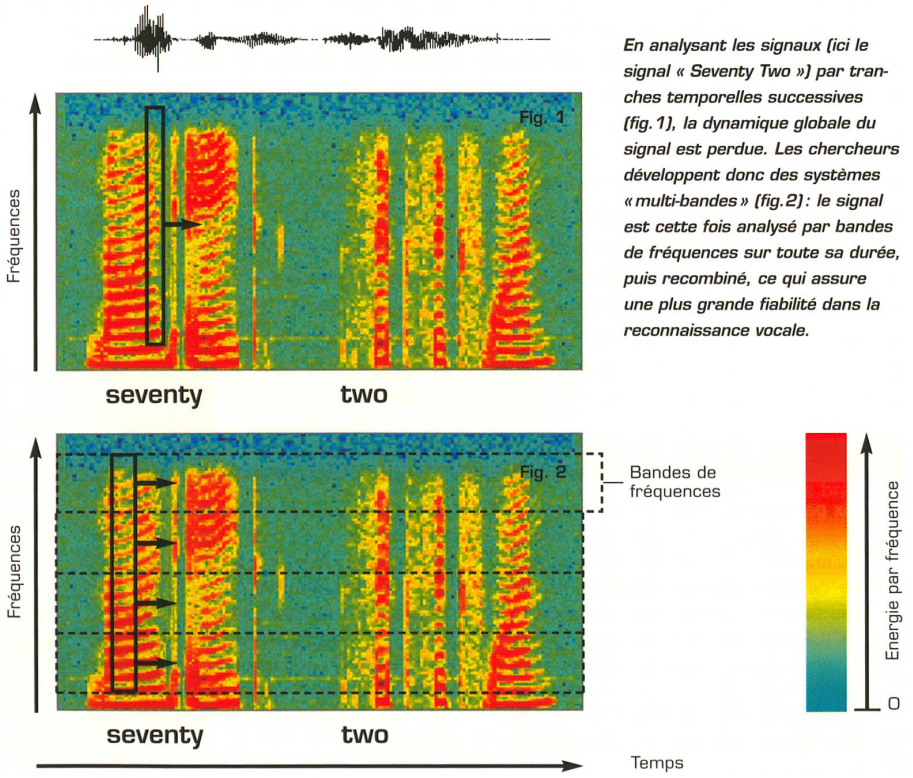
« Ceux-ci procèdent par analyse du spectre des fréquences », explique Hervé Bourlard, directeur. Autrement dit, tout signal sonore peut être caractérisé par les fréquences présentes (qui déterminent la hauteur des sons) ainsi que l'énergie transmise (qui correspond en quelque sorte au volume). Le système de reconnaissance en analyse l'une après l'autre des portions longues de 25 à 30 millisecondes (fig.1). Leur contenu est alors comparé avec des modèles statistiques d'un répertoire, qui représentent des phonèmes, soit des éléments sonores du langage articulé. En intégrant encore des règles lexicales et grammaticales, le tout est envoyé dans un « décodeur » qui reproduit la séquence de mots la plus probable.

« Or, commente M. Bourlard, c'est là une méthode statistique, basée non sur l' « intelligence » mais sur l'ignorance, soit l'extraction de connaissances à partir d'un ensemble d'exemples. » Et de citer deux cas à problèmes : « Lorsque qu'un mot prononcé est amputé ou lors d'une discussion bruyante, l'identification devient mauvaise ».

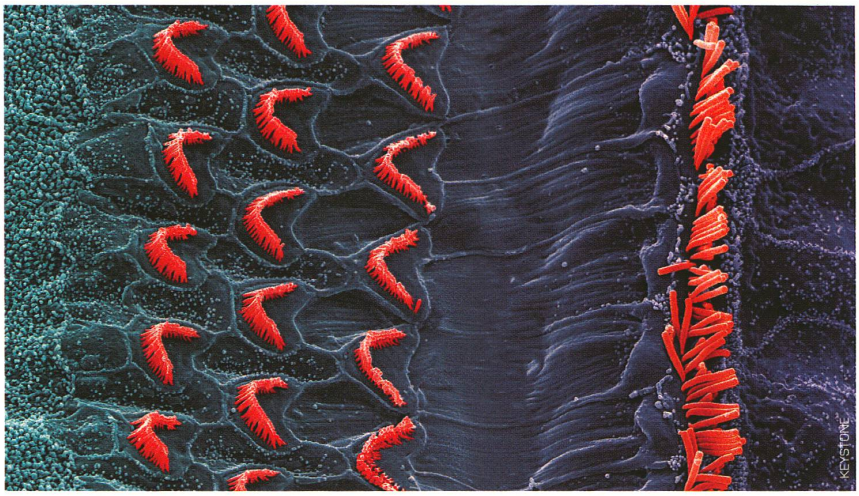
Une solution pour améliorer ces systèmes est de mieux tenir compte des propriétés de l'oreille humaine. « On est certain que, grâce aux millions de cils vibratiles de la cochlée qui, en vibrant, transmettent l'information sonore au nerf auditif, celle-ci effectue aussi une telle analyse spectrale », justifie-t-il. Ainsi, l'oreille n'ayant pas la même sensibilité pour toutes les fréquences sonores, les chercheurs ont par exemple réussi à pondérer aussi judicieusement les différents signaux de leurs spectres.

Tout modéliser ?

Pourtant, rien ne sert de vouloir tout modéliser. « Les avions ne battent pas des ailes, mais volent tout de même, car l'homme n'a extrait de ses observations que les règles principales de l'aéronautique », compare M. Bourlard. De même, il faut déterminer parmi toutes les propriétés de l'oreille hu-



En analysant les signaux (ici le signal « Seventy Two ») par tranches temporelles successives (fig. 1), la dynamique globale du signal est perdue. Les chercheurs développent donc des systèmes « multi-bandes » (fig. 2) : le signal est cette fois analysé par bandes de fréquences sur toute sa durée, puis recombiné, ce qui assure une plus grande fiabilité dans la reconnaissance vocale.



Les cils vibratiles de la cochlée (en rouge) transmettent l'information sonore au nerf auditif.

maine celles qui sont pertinentes pour de tels systèmes. « Or, nous sommes encore au stade des avions qui battraient des ailes … » Et le chercheur d'indiquer que les travaux des psycho-acousticiens sont parfois utiles pour trouver de nouvelles pistes (lire encadré à droite).

Par ailleurs, son équipe planche sur un autre problème : « En analysant le signal par tranches temporelles indépendantes, on perd sa dynamique dans le temps, très importante dans l'oreille humaine », explique-t-il. C'est notamment elle qui permet l'écoute d'une voix dans un environnement bruité. Mais, « il a été observé que chaque cil vibratile vibre à une fréquence propre, et que ces cils semblent se « concerter » pour cibler et recombiner sur la durée l'information sonore voulue ».

Cette découverte a mis la puce à l'oreille des scientifiques, qui ont réussi à simuler cette parade : « Au lieu d'analyser le signal par portion de temps, nous le faisons sur des bandes de fréquences (fig. 2) ». Cela permet, dans le bruit, de détecter les canaux qui contiennent une information fiable et, par corrélation, de recombiner seulement les données fréquentielles les plus utiles. Ce système original, appelé « multi-bandes », fait la fierté de l'IDIAP et a déjà donné lieu à des applications, comme les « salles de réunions intelligentes » (lire ci-dessous). ■

DES SALLES DE RÉUNIONS INTELLIGENTES

Dix personnes, devant leur micro, discutent dans une salle de réunion. Qu'importe le brouhaha, les systèmes de reconnaissance vocale multi-bandes comme ceux qui sont développés à l'IDIAP permettent de cibler et écouter la voix d'une seule personne. Tout comme l'oreille humaine lors d'une discussion dans un cocktail. De même, lors d'une conférence téléphonique, les participants, dont les portraits apparaîtraient sur un site Internet relié au système de reconnaissance, pourraient à tout moment connaître l'identité de la personne qui parle. Mais les chercheurs de l'IDIAP visent déjà plus loin, puisqu'ils ont ajouté la reconnaissance visuelle afin de créer de véritables « secrétaires multi-modaux » de conférences.

L'influence du stress

Les systèmes de reconnaissance vocale sont déjà utilisés dans certains dispositifs d'identification biométrique de personnes (dans les banques p.ex.). Ils fonctionnent à environ 95% . Mais il suffit que le locuteur soit stressé ou ému pour que sa voix change subtilement et que le système se trompe parfois. Des défauts que cherchent à améliorer une équipe du Département de psychologie de l'Université de Genève.

« Les ingénieurs misent sur les algorithmes pour perfectionner les systèmes, ils ne prennent pas en considération la source des changements, estime le professeur Klaus Scherer. Dans le projet EMOVOX, nous essayons de comprendre vraiment ce qui se passe afin de donner plus de poids aux paramètres acoustiques qui changent le moins sous le stress ou l'émotion ». Comme il est difficile d'imposer de tels états psychiques à des personnes pour analyser les modifications dans leur voix, les chercheurs ont enregistré et analysé une centaine d'hommes dans des tâches induisant du stress, et extrait 99 paramètres acoustiques (fréquence fondamentale, déviations par rapport à celle-ci, amplitude du pic d'attaque du signal, etc.).

« Les paramètres stables chez une personne deviennent alors des critères d'identification », explique le professeur. Il faut toutefois encore, pour que ce processus soit totalement valable, que ces paramètres soient assez différents entre tous les individus…

Les premiers résultats, qui viennent d'être présentés à Genève au congrès Eurospeech des technologies de la parole, confirment les attentes des chercheurs, qui ont pu déterminer de tels paramètres. Ces données doivent maintenant être répliquées, car cette approche est différente et novatrice. « La qualité de la voix est un aspect souvent négligé dans ces systèmes de reconnaissance. Nos résultats pourraient être utiles aux ingénieurs », conclut Klaus Scherer. O. D.