La mise au point d'un filtre destiné à réduire le bruit revient, du point de vue mathématique, à projeter chaque point de la série temporelle s_k sur l'hypersurface caractéristique de la série x_k. Première difficulté : le signal vocal exempt de bruit x_k est évidemment inconnu. Pourtant, on parvient à identifier le bruit parasite, car il a une origine stochastique et ne favorise donc aucune direction dans l'espace des phases reconstruit, alors que le signal vocal x_k, que l'on recherche, contient des composantes déterministes (les phonèmes). Seconde difficulté : on ne connaît pas la dimension de l'espace des phases original associé au système, la reconstruction de l'espace des phases soulève donc des difficultés.

Finalement, on assimile la réalisation du filtre à un problème de minimisation, qui se résout par l'intermédiaire de l'algorithme suivant : à partir de la série temporelle s_k, on détermine les états du système dans l'espace des phases reconstruit ; chaque état est représenté par un point dans l'espace des phases reconstruit et on identifie tous les points dont la distance est inférieure à un certain seuil (ce seuil doit toutefois être supérieur aux fluctuations dues au bruit, c'est-à-dire à la largeur du nuage de points) ; on définit les directions principales sur lesquelles sont distribués ces points (la distinction entre la composante déterministe (x_k) et la composante stochastique (n_k) se produit au cours de cette opération) ; on projette chaque point dans l'hypersurface identifiée à partir des directions principales définies lors de l'étape précédente ; le cycle entier est répété jusqu'à l'obtention de la convergence. À chaque itération, les points se rapprochent de l'hypersurface associée à la composante déterministe, c'est-à-dire à un phonème. Chaque phonème occupe une hypersurface dans une région distincte de l'espace des phases reconstruit.

Appliquons maintenant le filtre à une série temporelle caractéristique de la voix humaine enregistrée par un micro. Le signal vocal est composé de phonèmes, dont la durée varie entre 50 et 150 millisecondes. Ils sont constitués de sous-phonèmes de quelques millisecondes qui s'y répètent. Le signal semble plus ou moins périodique et, grâce à ce

L'évolution des populations de filtres

Les systèmes de reconnaissance automatique de la parole sont en plein essor, mais ils se heurtent à une difficulté : leurs performances chutent dès que les conditions acoustiques changent (bruits de fond, réverbération, changement de locuteur, débit de paroles accéléré, etc.)

Pour résoudre ce problème, la stratégie prédominante consiste à enregistrer ce que l'on nomme des corpus d'apprentissage dans des conditions variées. Par exemple, *bonjour* est prononcé par plusieurs individus et dans différents environnements sonores et on «apprend» au système de reconnaissance que les signaux enregistrés dans chacun des cas correspondent à *bonjour*. Le corpus contient ainsi des signaux, plus ou moins bruités, auxquels on attribue une «signification». En pratique, quand le système de reconnaissance reçoit un signal, il le compare à ceux contenus dans le corpus, retient le plus similaire et donne la signification correspondante. Avec l'augmentation de la puissance de calcul et de la taille de la mémoire des ordinateurs, on fournit désormais des quantités considérables de données dans ces corpus. Malheureusement, les sources de variabilité qui réduisent les performances d'un système de reconnaissance sont nombreuses et imprévisibles, et ces corpus d'apprentissage ne permettent pas toujours aux systèmes de s'adapter aux variations acoustiques.

Une alternative consiste à développer des techniques de filtrage et d'adaptation au bruit ambiant. L'effet du bruit sur le signal de parole est variable : des pics apparaissent, ou au contraire, disparaissent, ce qui aplatit le spectre, ou la bande de fréquences se modifie, etc. En effet, les bruits ont des caractéristiques propres (ils sont périodiques ou impulsifs, étendus sur une large bande de fréquences ou non, etc.) et ne peuvent être traités de la même manière. Les techniques de filtrage sont très nombreuses et souvent plus efficaces sur tel ou tel type de bruit. Certaines sont fondées, par exemple, sur des méthodes statistiques qui séparent le bruit du signal de la parole, ou sur l'utilisation de plusieurs microphones. D'autres tentent de reproduire les capacités inégalées de l'oreille humaine à faire abstraction du bruit. Elles s'inspirent donc du fonctionnement de l'oreille humaine, qui serait fondé sur le traitement indépendant de sous-bandes de fréquences. Ainsi le décodage d'un message linguistique se ferait en parallèle, dans différentes bandes de fréquences, puis en mettant en commun les informations obtenues par chaque analyse. Ces différents types de filtres, encore pour la plupart au stade expérimental, n'ont pas obtenu de résultats suffisamment convaincants pour être intégrés dans les systèmes de reconnaissance.

Nous avons développé un système de reconnaissance de la parole capable de s'adapter aux changements de l'environnement acoustique : il s'automodifie dans le temps. Nous nous sommes inspirés des organismes vivants, capables de sélectionner les informations utiles et de modifier leur traitement pour s'adapter aux aléas de leur environnement. Nous avons exploré les possibilités offertes par un type d'algorithme, nommé algorithme évolutionnaire (dont les plus connus sont les algorithmes génétiques), inspiré de l'évolution naturelle formulée par Darwin. L'algorithme évolutionnaire fait évoluer des populations d'individus dans un environnement, en favorisant la survie et la reproduction des individus les mieux adaptés à cet environnement.

Dans notre cas, les individus sont des filtres, une population de filtres, soumise aux signaux vocaux bruités. Ces filtres traitent le signal de la parole pour transformer des données bruitées en données le moins bruitées possible. Ils sont codés dans un génotype composé de «gènes» : on considère qu'un filtre est consitué de plusieurs gènes dont les combinaisons varient d'un filtre à l'autre. Lorsqu'un signal de parole bruité arrive au système de reconnaissance, il est filtré par chaque individu de la population. Les individus les plus performants, c'est-à-dire ceux qui ont permis au système de reconnaître au mieux le signal de parole, sont sélectionnés. Les filtres les moins performants sont éliminés. Puis, on fait «se reproduire» les meilleurs filtres par des mécanismes de «croisement» et de «mutation» des génotypes, pour engendrer une nouvelle génération de filtres. Les croisements échangent des fragments de deux génotypes ; les mutations remplacent un 1 d'un génotype en un 0, ou *vice versa*. En itérant cette opération un certain nombre de fois, sur plusieurs générations, nous obtenons une population de filtres de plus en plus performants. Ainsi, à chaque fois que de nouvelles conditions acoustiques apparaissent, le système est capable de s'auto-adapter pour ne pas faire chuter ses performances.

Anne SPALANZANI, Laboratoire de communication langagière et interaction personne-système, Institut d'informatique et de mathématiques appliquées de Grenoble.