La reconnaissance de la parole

Mahieddine DJOUDI

1 Introduction

La parole est certainement le moyen le plus direct et le plus naturel utilisé par l'homme pour échanger l'information. Depuis longtemps, la parole passionne les chercheurs qui pensent pouvoir arriver un jour à utiliser ce même moyen pour communiquer avec une machine.

2 Les avantages de la communication parlée

Les avantages de la communication orale homme-machine sont multiples. Ce mode de relation libère complètement l'usage de la vue et des mains et laisse l'usager libre de ses mouvements. La vitesse de transmission des informations est supérieure dans le sens homme-machine à celle qui permet l'usage du clavier. De plus la voix informe sur l'identité du locuteur et peut être par ailleurs transportée par des moyens simples existants, comme le réseau téléphonique. Ses avantages se traduisent par l'existence d'une grande variété d'applications liées à la reconnaissance automatique de la parole, à titre d'exemple on peut citer :

CAO : Commandes des machines.

Robotique : Guidage d'un robot.

Transport : Réservation automatique.

Postes : Demande de renseignements.

Médecine : Aide aux handicapés.

Bureautique : Machine à dicter à commande vocale.

Le développement de telles applications vise à améliorer le confort de l'utilisateur, à augmenter l'efficacité de la communication et à offrir de nouvelles possibilités.

3 Les difficultés de la reconnaissance vocale

La reconnaissance automatique de la parole est encore un problème difficile à résoudre. Cette situation résulte de la complexité du langage parlé. La parole présente un certain nombre de caractéristiques spécifiques qui demeurent très difficiles à analyser. Les sources de la variabilité et la complexité du message vocal sont essentiellement :

- l'environnement sonore et les conditions d'enregistrement : La présence du bruit et des perturbations apportées par le microphone ou le téléphone rendent le traitement du signal de parole plus complexes et compliquent davantage le problème.
- la variabilité interlocuteur : chacun d'entre nous parle différemment des autres, la prosodie, la prononciation et les timbres sont caractéristiques de l'age du sexe ou du milieu socio-culturel de l'individu.
- la variabilité intralocuteur : une phrase peut être prononcée par une même personne de différentes manières en fonction de son état physiologique et psychologique. Les paramètres qui rentrent en jeu sont par exemple l'émotion, la fatigue ou le rhume
- les variations contextuelles : la prononciation d'un mot dépend du contexte de sa production, en fonction qu'il est au début, au milieu ou en fin de phrase. Il peut s'agir d'anticipations ou de retard dans le positionnement des organes articulatoires. On parle de phénomènes de co-articulation.

— la complexité linguistique du langage parlé aux niveaux lexical, grammatical et sémantique. Quel vocabulaire utiliser et quelle structure syntaxique adopter pour traiter les particularités de la langue parlée? une langue pleine de distorsions et de sous-entendus.
De plus, lorsqu'on traite de la parole continue se pose le problème très délicat de séparation des mots [1].

4 Les techniques de reconnaissance

Pour réaliser des systèmes de reconnaissance, deux approches principales sont utilisées, l'une consiste à reconnaître globalement des mots séparés par des intervalles de silence, l'autre permet en revanche d'aborder le problème de la reconnaissance de la parole continue éventuellement dans un contexte multilocuteur. Nous allons décrire ces deux approches.

4.1 La méthode globale

Dans cette approche, l'unité de base est le plus souvent le mot considéré comme une entité globale. Cette méthode est caractérisée par le fait qu'une phase d'apprentissage est nécessaire pendant laquelle l'utilisateur prononce la liste des mots du lexique de son application.
Lors de la phase de reconnaissance, le mot à reconnaître est comparé à tous les mots de références du lexique. Le mot ressemblant le plus au mot prononcé est alors reconnu. Les avantages d'une telle approche sont d'une part l'indépendance vis-à-vis des particularités de la langue du fait de la phase d'apprentissage, et d'autre part, l'excellente capacité de reconnaissance pouvant atteindre les 99%. Néanmoins, le vocabulaire est assez limité et les systèmes sont le plus souvent monolocuteur, de plus la prononciation en mots isolés est peu naturelle. Les principales techniques utilisées par cette méthode sont la programmation dynamique, les modèles stochastiques de Markov et les modèles connexionnistes.

4.1.1 La programmation dynamique

La programmation dynamique est une méthode de résolution de problèmes d'optimisation exprimés par un ensemble de contraintes. En utilisant le principe de mise en correspondance optimale, la programmation dynamique permet de tenir compte des distorsions temporelles entre deux formes à comparer. Chaque mot du lexique est représenté par une suite de vecteurs $r(1), ..., r(j)$, chaque forme à reconnaître est représentée par $t(1), ...t(i)$. Il faut trouver un chemin de recalage qui à chaque vecteur de T, fait correspondre un vecteur R. Ce chemin devra prendre en compte les contraintes naturelles de la parole. Ensuite, parmi tous les chemins de recalages possibles, il faut choisir celui dont la somme des distances le long du chemin est minimale. L'algorithme permet d'éliminer rapidement des références lorsque des différences notables apparaissent au cours d'une étape quelconque de l'examen de la référence à reconnaître. Les avantages de la méthode sont d'une part son excellente capacité de reconnaissance et son indépendance vis à vis des particularités de la langue [5].

4.1.2 Les modèles de Markov

Dans les modèles stochastiques, les formes acoustiques des références sont représentées par un graphe sous forme d'une chaîne de Markov ou plus précisément par des modèles de Markov cachés HMM (Hidden Markov Model). Le graphe est composé d'un nombre fini d'états représentant les segments stables du signal, tandis que les variations spectrales sont modélisées par les arcs de transition. Ce graphe peut être vu comme un modèle de production d'un mot, ou' chaque transition est accompagnée de l'émission d'un vecteur de paramètres spectraux. A chaque état S_j sera associé une distribution de probabilité $P(a_i/S_j)$, probabilité de produire l'événement a_i sur une transition d'origine S_j, et à chaque arc sera associé une probabilité de transition $P(S_j/S_i)$, probabilité que le modèle passe de l'état S_i à l'état S_j en une seule transition. Les paramètres du modèle sont obtenus au cours de la phase d'apprentissage à

partir d'un nombre important d'énoncés du même mot. Ce qui donne à la méthode un avantage majeur de prise en compte de la variabilité du signal vocal. En revanche, s'appuyant sur une modélisation purement mathématique, ils ne permettent pas d'introduire de façon explicite des connaissances phonétiques [3].

4.1.3 Les modèles connexionistes

Ils sont fondés sur une modélisation des réseaux de neurones. Ces derniers possèdent des avantages forts intéressants tels que le parallélisme, le raisonnement à partir de données incomplètes et la capacité de généralisation. Nous assistons actuellement à un regain d'intérêt pour l'utilisation des modèles connexionistes en reconnaissance automatique de la parole même s'ils n'ont pas encore prouvé leur supériorité par rapport aux autres méthodes.

4.2 La méthode analytique

Cette approche tente de détecter et d'identifer des unités élémentaires (phonèmes, diphonèmes, syllabes) puis de reconnaître la phrase effectivement prononcée. La méthode fait apparaître plusieurs modules qui communiquent entre eux.

- Le module acoustique a pour rôle d'extraire les caractéristiques du signal de parole destinées aux module phonétique et prosodique.
- Le module prosodique sert à trouver les informations sur le rythme et l'intonation de la phrase.
- le module phonétique traduit la liste des indices en une suite d'unités phonétiques.
- Le module phonologique porte sur les phénomènes de la langue dont le contenu phonétique est modifié par les articulations rapides, les liaisons et les variétés dialectales.
- Le module lexical, à ce niveau interviennent les informations sur les mots qui composent la langue.
- Le module syntaxique renferme les règles de la grammaire qui permettent de décrire et d'analyser la langage en termes grammatical et fonctionnel et permet donc de définir toutes les séquences de mots acceptables.
- Le module sémantique permet de donner la signification de l'énoncé et le rejet des phrases syntaxiquement correctes n'ayant aucune interprétation.
- Le module pragmatique, utilisé en dialogue, permet de déterminer le sens de la phrase dans le contexte de l'application et de gérer l'historique du dialogue.

5 Le décodage acoustico-phonétique

De toutes les opérations décrites par les différents modules de l'approche analytique, la transformation du signal vocal en une suite d'étiquettes phonétiques est la plus fondamentale. Toute erreur à ce niveau augmente considérablement l'indéterminisme des traitements ultérieurs. Le décodage acoustico-phonétique est lié à deux aspects importants en reconnaissance de la parole : la représentation paramétrique et l'identification phonétique.
Les techniques utilisées en décodage phonétique peuvent être classées selon deux approches :

1. Une approche de reconnaissance des formes qui consiste à affecter des étiquettes à des segments grâce à des critères de proximité. On retrouve dans cette approche la programmation dynamique, les modèles de Markov et les réseaux de neuronnes.
2. Une approche basée sur la reconnaissance de traits qui à l'inverse de l'approche précedente, ne nécessite pas d'apprentissage, elle permet, par contre, la prise en contexte. Elle s'effectue souvent de deux étapes, la segmentation du signal en unités et l'identification des segments en utilisant les traits. La modélisation des connaissances se fait le plus souvent à l'aide d'un système expert.

6 Les résultats actuels

Compte tenu de la complexité du message vocal, aucun système de reconnaissance actuel ne peut traiter la parole spontanée dans n'importe quelle situation de communication. Tous imposent des contraintes d'utilisation : prononciation en mots isolés ou en parole continue, utilisation mono- multi-locuteur ou indépendante du locuteur, limites sur l'étendu du vocabulaire ou sur la grammaire.

Compte tenu de ces contraintes voici quelques exemples de systèmes qui existent actuellement :

- le système développé par ATT peut reconnaître, de manière indépendante du locuteur, des chiffres prononcés continuement à travers le réseau téléphonique avec un taux de reconnaissance de 98,6%. Des résultats du même ordre sont observés pour les lettres de l'alphabet.
- Dans le cadre du programme américain DARPA, pour la reconnaissance, en mode continue et en contexte monolocuteur d'un vocabulaire de 1000 mots, le système BYBLOS de BBN et celui de Lincoln Laboratory du MIT ont obtenu une performance évaluée en 1990 à plus de 98%.
- Le système SPHINX développé à Carnegie Mellon University a obtenu un pourcentage de 95.5 % de bonne reconnaissance en fonctionnement indépendant du locuteur, pour des locuteurs masculins.
- IBM a présenté en 1985 aux USA un système TANGORA capable de reconnaître un vocabulaire de 5000 mots en mode "mots isolés", 20 000 mots en 1987, 50 000 mots en parole contine en 1989. Les résultats comparables sont enregistrés par le système Dragon Writer. Nous enregistrons aussi la sortie récente des systèmes "speech server" d'IBM et "Dragon Dictate". Les performances des ces systèmes sont médiocres au départ (ordre 50%), mais croissent progressivement pour atteindre un taux de l'odre de 95% dans un mode de fonctionnement indépendant du locuteur.
- IBM France travaille sur la reconnaissance de 200 000 mots dictés syllabe par syllabe.
- Le laboratoire LIMSI-CNRS est l'auteur du premier système franĉais qui dut développé en 1980 et qui reconnait 5000 mots isolés. La sociéte VECSYS a développé à partir des travaux du LIMSI une carte d'interface vocale qui permet de reconnaitre des phrases construites à partir de plusieurs centaines de mots prononcés de manière continue, en mode monolocuteur [4], [2].

L'évaluation des performances des systèmes de reconnaissance de la parole est à la fois cruciale et délicate. Le problème se pose dans le choix du corpus d'apprentissage et/ ou d'évaluation. Le plus souvent on utilise du texte lu, avec une syntaxe toujours correcte, des mots appartenants au lexique et enregistrés sans bruits ni begaiement. Ces limites rendent difficile la comparaison des systèmes entre eux. Les recherches dans le domaine continuent pour améliorer le taux de reconnaissance et réduire les contraintes imposées à l'utilisateur.

Références

[1] Calliope. *La parole et son traitement automatique.* Masson, 1989.

[2] J.J Mariani. Reconnaissance vocale : l'ordinateur manque d'à-propos. *Sciences et avenir, Hors série no 86*, Mars/avril 1992.

[3] A. Markov. *The Theory of Algorithm.* US Dpt of Commerce, 1954.

[4] G. Perennou. La reconnaissance vocale. *Le courrier du CNRS no 80*, Fevrier 1993.

[5] H. Sakoe and S. Chiba. Dynamic programming algorithm optimisation for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, February 1978.