Résumé

L'utilisation d'informations supplémentaires conjointement à celles extraites du signal acoustique est une nouvelle méthode utilisée afin d'améliorer les performances et la robustesse des systèmes de reconnaissance automatique de la parole. De nombreux travaux sur la perception de la parole ont montré l'importance des informations visuelles dans le processus de reconnaissance chez l'homme. L'utilisation de données sur la forme et le mouvement des lèvres du locuteur semble donc être une voie prometteuse pour la reconnaissance de la parole.

Notre travail, dans le cadre de ce magister, concerne la mise en œuvre d'un système de reconnaissance automatique de la parole (RAP) audiovisuelle pour les chiffres arabes en milieu réel. Il s'agit d'une intégration des informations visuelles aux informations acoustiques.

Tout d'abord, se pose le problème du niveau dans lequel se fera la fusion : est-ce au niveau des données ou bien au niveau des résultats. Ensuite intervient le problème d'adaptation des contributions des deux modalités acoustique et visuelle.

Le système audiovisuel que nous avons mis en œuvre utilise les modèles de Markov cachés continus (CHMM) comme moteur de reconnaissance aussi bien pour la modalité acoustique que pour la modalité visuelle. Le résultat final de reconnaissance est obtenu après fusion des scores issus de chaque reconnaisseur. Les CHMM constituent l'approche la plus performante actuellement pour la RAP. L'estimation des paramètres des modèles par maximum de vraisemblance nécessite des procédures itératives, chaque itération mettant elle-même en jeu des parcours récursifs, visant à calculer la loi conditionnelle des variables cachées. La difficulté consiste à obtenir des algorithmes efficaces, interprétables de manière probabiliste.

La fusion des scores est basée sur l'utilisation de réseaux de neurones de type Perceptron MultiCouches (PMC)

Nous avons testé les performances de notre système sur un corpus audiovisuel des chiffres arabes en mode mono locuteur. Ce corpus a été enregistré par nos soins au niveau de notre laboratoire. Les paramètres auditifs utilisés sont les coefficients cepstraux dans l'échelle Mel (Mel Frequency Cepstral Coefficients) et les paramètres visuels sont eux basés sur la DCT (Discrete Cosine Transform). Les tests réalisés, dans un milieu réel, ont montré que de bonnes performances sont obtenues pour le système acoustique (TMBR = 69.33%) par rapport au système visuel (TMBR = 59.33%), elle meilleures pour la reconnaissance audiovisuelle (TMBR = 87%).

Les expériences réalisées nous ont permis de constater que l'information visuelle intégrée à l'information acoustique peut constituer une alternative pour augmenter la performance des systèmes de reconnaissance en milieu réel (nécessairement bruité).

Les performances de notre système restent à être évaluées dans un contexte plus étendu, par exemple un corpus de plus grande taille prononcé par plusieurs locuteurs. Nous comptons tester notre système dans le cas des signaux bruités à différents bruits. Nous pensons également utiliser d'autres types de fusion telles que la fusion au niveau des paramètres ou la fusion hybride.