AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

M^me^ : ADDARRAZI ILHAM

Soutiendra : le 26/11/2022 à 10H
Lieu : Département de Géologie – Salle de Réunion

Une thèse intitulée :
" Etude et conception d'un système de reconnaissance audio visuelle "

En vue d'obtenir le Doctorat

FD : Sciences et Technologies de l'information et de la Communication (STIC)
Spécialité : Informatique

Devant le jury composé comme suit :

	Nom et prénom	Grade	Etablissement
Président	ABERKAN MUSTAFA	PES	Faculté Polydisciplinaire - Taza
Directeur de thèse	SATORI HASSAN	PES	Faculté des sciences Dhar El Mahraz –Fès
Co-Directeur de thèse	SATORI KHALID	PES	Faculté des sciences Dhar El Mahraz –Fès
Rapporteurs	EL AZAMI IKRAM	PH	Faculté des sciences - Kenitra
	OUANAN MOHAMED	PES	Faculté des sciences -Meknès
	BOUMHIDI JAOUAD	PES	Faculté des sciences Dhar El Mahraz –Fès
Membres	EL ABDERRAHMANI ABDELLATIF	PH	Faculté des sciences Dhar El Mahraz –Fès
	LOQMAN CHAKIR	PH	Faculté des sciences Dhar El Mahraz –Fès
	MAHRAZ ADNANE	PH	Faculté des sciences Dhar El Mahraz –Fès

Résumé :

La reconnaissance automatique de la parole (RAP) est le processus de transformation des signaux acoustiques en une séquence de mots. Cependant, les performances des systèmes RAP dans un environnement bruyant sont insatisfaisantes. Par conséquent, un problème majeur qui persiste dans le système RAP est sa robustesse au bruit. Motivés par les mécanismes du cerveau humain, les mots prononcés peuvent être perçus à travers de multiples modalités sensorielles : auditif et visuelle. L'utilisation d'informations visuelles, en étudiant le mouvement des lèvres des locuteurs, fournit des données complémentaires pour décider la parole. L'ajout d'informations sous la forme des lèvres du locuteur semble être utile pour les problèmes de reconnaissance vocale dans un environnement bruyant. Ainsi, la reconnaissance audiovisuelle de la parole (RAVP) est un domaine qui vise à assumer la nature bimodale de la perception humaine de la parole en combinant des informations auditives et visuelles. C'est dans ce cadre que s'inscrit notre thèse, elle s'intéresse à la réalisation d'un système de reconnaissance audiovisuelle de la parole basant sur une modélisation statistique Markovien pour modéliser les dix chiffres Amazigh à base des unités phonétiques élémentaires. Pour tester la performance du système RAVP, la création d'une base de données audiovisuelle est nécessaire. Toutefois, la réalisation de ce corpus nécessite des données audio et vidéo qui ont été enregistrées simultanément (4000 fichiers audio et vidéo). En se basant sur ce corpus, nous avons réalisé un sous-système de la reconnaissance visuelle pour reconnaitre les dix chiffres Amazigh. Les caractéristiques visuelles de ce sous-système sont extraits avec la méthode TCD (La transformée en cosinus discrète) à partir de chaque image buccale pour être, après, modéliser par les modèles de Markov cachées (MCC). Concernant le sous-système de reconnaissance vocale de parole, des MMCs sont utilisées pour modéliser les dix chiffres Amazigh basant sur ses coefficients acoustiques MFCC (Mel-FrequencyCepstral Coefficients).La stratégie de fusion de décision est appliquée pour combiner les décisions des deux sous-systèmes. Afin de tester la performance de notre RAVP dans un milieu réel, nous avons étudié l'effet du bruit sur les dix premiers chiffres Amazighs dans des conditions bruyantes basé sur le rapport signal sur bruit (en anglais signal-to-noise ratio (SNR)). Nos expériences de tests ont été procédées avec les différents niveaux de bruit SNR (de 5 db à 25 db). Nos résultats atteignent une grande précision. Cette précision démontre que l'intégration des informations visuelles et acoustiques offre de meilleures performances que celles fournies par les systèmes uniquement audio et uniquement visuels.

Mots-clés : Système de reconnaissance vocale, lecture labiale, intégration multimodale, intégration tardive, modèle de Markov caché, algorithme de Viterbi, GMM, MFCC, DCT.

THE AUDIOVISUAL SPEECH RECOGNITION SYSTEM CONCEPTION AND IMPLEMENTATION

Abstract:

Automatic Speech Recognition (ASR) is the process of transforming acoustic signals into a sequence of words. However, the performance of ASR systems in a noisy environment is unsatisfactory. Therefore, a major problem that persists in the ASR system is its robustness to noise. Motivated by the mechanisms of the human nature, spoken words can be perceived through multiple sensory modalities: auditory and visual. The use of visual information, by interpreting the movement of the lips of speakers, provides additional information to decide what has spoken. The addition of information in the form of speaker's lips appears to be helpful for speech recognition issues in a noisy environment. Thus, Audiovisual Speech Recognition (AVSR) is a field that aims to assume the bimodal nature of human speech perception by combining auditory and visual information. In the same context, the aim of this thesis is the realization of an audiovisual speech recognition system based on a Markovian statistical modeling to recognize the first ten Amazigh digits based on elementary phonetic units. In order to test the performance of the AVSR system, the creation of an audiovisual database is necessary. However, the realization of this corpus requires audio and video data which were recorded simultaneously (around 4000 audio and video files). Based on this corpus, we created a visual speech recognition subsystem to recognize the first ten Amazigh digits visually. In this subsystem, the visual features are extracted using the DCT approach (Discret Cosine Transform) from each mouth region image. These features are modeled by the hidden Markov models (HMM). Regarding the speech recognition subsystem, HMMs are used to modeling the first ten Amazigh digits based on its acoustic coefficients MFCC (Mel-Frequency Cepstral Coefficients). The decision integration strategy is applied to combine the decisions of the two subsystems. In order to test the performance of our proposed AVSR under noisy conditions based on the signal-to-noise ratio (SNR)). Our test experiments were carried out with the different SNR noise levels (from 5 db to 25 db). The achieved results demonstrate that the integration of visual and acoustic information provide a good performance than that provided by audio-only and visual-only systems.

Keywords: Speech recognition system, lip-reading, multimodal integration, late integration, HMM, GMM, DCT, Face Detection, phoneme, viseme.