Vers l'Alignement des Signaux Écrit et Sonore

Application pour la reconnaissance des expressions mathématiques

Sofiane MEDJKOUNE[1,2] — Harold MOUCHÈRE[1] — Simon PETITRENAUD[2] — Christian VIARD-GAUDIN[1]

[1] *LUNAM Université, Université de Nantes, laboratoire IRCCyN(IVC), France*
{sofiane.medjkoune, harold.mouchere, christian.viard-gaudin}@univ-nantes.fr
[2] *LUNAM Université, Université du Maine, laboratoire LIUM, France*
simon.petit-renaud@lium.univ-lemans.fr

RÉSUMÉ. *Dans cet article, nous rapportons de nouveaux résultats sur la reconnaissance des expressions mathématiques (EMs). Nous abordons cette problématique en considérant l'aspect bimodal de l'information : c'est à dire exploiter à la fois le signal de parole et celui de l'écriture manuscrite représentant la même EM. Ceci permet de disposer de plus de fiabilité lors d'un traitement automatique, d'autant plus que ces deux modalités s'avèrent être très complémentaires. Nous proposons d'aligner les deux modalités grâce à un classifieur de type réseau de neurones et en adoptant un apprentissage original des associations écrit-audio. Nous avons évalué le système proposé sur la base bimodale* HAMEX *d'EMs. Nous avons également confronté les résultats obtenus à ceux obtenus par notre précédent système et un système de référence basé uniquement sur l'écriture seule.*

ABSTRACT. *In this paper, we report some new results related to mathematical expression recognition. We tackle this problem, known to be very difficult, using multimodal information – handwriting and speech–. This bimodality aspect of the information provides greater reliability since the modalities in concern are very complementary. To combine the signals coming from both modalities in an efficient way, we propose an original learning process based on neural networks. This approach allows not only fusing both streams, but also the alignment of the signals coming from both modalities. The bimodal system is evaluated on real bimodal data from the HAMEX dataset and the obtained results are compared to a single modality (handwriting) based system.*

MOTS-CLÉS : *Expression mathématique, écriture manuscrite, parole, fusion de données, apprentissage bimodal*

KEYWORDS: *Mathematical expression, Handwriting, Speech, Data Fusion, Bimodal Learning*

CIFED 2014, pp. 341–356, Nancy, 18-21 mars 2014

1. Introduction

Avec l'avènement du progrès technologique, énormément d'outils d'interface homme-machine sont rendus accessibles. Faciliter la prise en main de tels dispositifs est de nos jours un objectif principal pour permettre de les rendre les plus accessibles et les plus intuitifs possibles. Depuis longtemps, la parole et l'écriture sont considérés comme les supports naturels et principaux de la transmission de l'information entre les humains. Utiliser ces modalités pour l'interaction humain-machine pourrait être d'un grand intérêt. Cette nouvelle génération d'outils d'interaction est encore plus utile pour le cas des langages graphiques tel que les expressions mathématiques (EMs). Les EMs sont caractérisées par leur structuration spatiale particulière rendant ainsi leur édition par l'écriture manuscrite ou à partir de leur description par la parole plus difficile et plus confuse que pour le cas d'un texte standard. Dans l'état de l'art, on rencontre des applications (basées sur le signal manuscrit en-ligne ou la parole) dédiées au langage mathématique, d'une part pour répondre au besoin évoqué plus haut et d'autre part pour les problématiques scientifiques très intéressantes que ce langage soulève. Les applications basées sur le tracé manuscrit en-ligne (Chan et Yeung, 2000 ; Tapia et Rojas, 2007) sont plus nombreuses et plus abouties que celles basées sur la parole (Elliott et Bilmes, 2007 ; Wigmore *et al.*, 2009). Toutefois, en considérant l'une où l'autre des modalités ces systèmes sont loin d'être à 100% fiables.

Les solutions proposées jusque là, considèrent les modalités sus-citées de façon disjointe et on ne recense pas de solution tirant profit de la complémentarité existante entre ces deux signaux. En effet, l'écriture manuscrite et la parole apportent des informations qui permettrait de résoudre les problèmes d'interprétation rencontrés suite à une analyse monomodale basée sur l'une ou l'autre des modalités tel que l'illustre la figure 1.

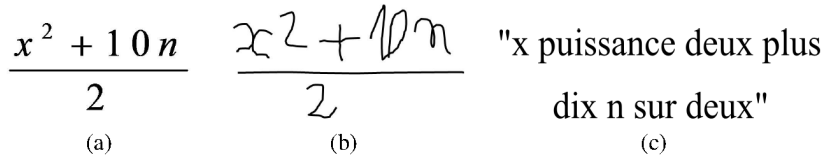

Figure 1. Exemple d'expression présentant des ambiguïtés liées à chacun des flux : (a) : vérité terrain ; (b) : son tracé manuscrit ; (c) : transcription possible

Dans ce travail nous proposons une solution explorant justement une analyse multimodale des expressions mathématiques. En effet, en considérant les limitations rencontrées par les systèmes mono-modaux existants (écriture manuscrite ou parole) et la forte complémentarité qui existe entre ces deux modalités (voir figure 1), nous présentons notre système bi-modal qui utilise conjointement les signaux de parole et d'écriture manuscrite en-ligne pour la reconnaissance automatique des expressions mathématiques. Ce travail découle des conclusions des précédents travaux que nous

avons menés sur la reconnaissance bimodale des EMs (Medjkoune *et al.*, 2013). En effet, nous avons proposé une architecture de reconnaissance des EMs bimodales dans laquelle la modalité audio vient en soutien à la modalité manuscrite en-ligne pour l'aide à la meilleure interprétation. Un obstacle majeur que rencontre notre système concerne la synchronisation des deux modalités en cause. L'objet de ce papier est donc de proposer une solution préliminaire où on cherche à procéder au meilleur alignement possible des deux modalités avant de les fusionner.

La suite de cet article est organisée comme suit : dans la section 2 nous revenons sur la notion de langage mathématique. La section 3 est dédiée à la présentation de notre système et des différents modules le composant. En section 4 sont rapportés les résultats et leurs discussions. La dernière section rapporte les conclusions et les perspectives de ce travail.

2. Le langage mathématique

Du fait de leur appartenance à la catégorie des langages graphiques, les expressions mathématiques sont une structuration dans un espace 2D d'unités élémentaires, les symboles. Ces symboles sont organisés suivant différentes relations spatiales : *gauche/droit, haut/bas, indice/exposant, dans*. Ceci peut donner à l'EM une structure plus ou moins complexe. À partir de là, si l'on considère l'opération d'édition des EMs à partir de la modalité manuscrite ou de celle de la parole, en plus des problèmes classiques que rencontrent les systèmes d'interprétation automatique de tels signaux, certaines spécificités des EMs rendent cette tâche d'avantage compliquée. Cette complexité peut venir de deux origines : les difficultés liées à l'identification des symboles et celles dues à l'identification des relations spatiales liant les symboles.

2.1. *Difficultés au niveau des symboles*

• **Le caractère minuscule ou majuscule de certaines lettres de l'alphabet** rend parfois difficile l'association de la bonne étiquette au symbole dans le cas d'un signal manuscrit. C'est notamment le cas des lettres "c" et "o". Dans le cas de la modalité audio, cette ambiguïté peut exister ou non. En effet, dans le cas où la dictée est rigoureuse en précisant la casse de façon explicite, comme par exemple "petit c", "c majuscule", ..., la confusion n'existe pas. En revanche, lorsque cette précision est omise, l'ambiguïté est totale.

• **Les formes géométriques de certains symboles sont très proches.** En effet, à l'image des symboles "q" et "9" ou encore "(" et "c" , même s'ils sont sémantiquement différents, les tracés de certains symboles présentent des similitudes prononcées.

• **Les formes acoustiques de certains symboles sont très proches.** Du point de vue de la parole, il existe des symboles dont la prononciation est très similaire et fait intervenir quasiment la même séquence phonétique. Ceci est d'autant plus avéré si l'articulation du locuteur n'est pas des plus fiables. C'est le cas par exemple des symboles "a" et "1". Dans d'autres exemples, un symbole peut être confondu avec

une séquence de symboles, comme c'est le cas pour la lettre grecque "κ", prononcée *kappa*, et la concaténation des lettres "k", "p" et "a".

• **Le rôle de certains symboles change avec le contexte.** Les traits horizontaux peuvent être associés à des barres de fraction ou à des signes moins ou encore combinés avec un autre trait dans une disposition haut-bas pour former le signe d'égalité.

• **Certains symboles peuvent être des parties d'autres symboles.** Dans ce contexte, un tracé pouvant représenter un symbole donné se trouve être un morceau d'un symbole plus grand. C'est le cas du caractère "c" faisant partie du symbole "$\cos$".

2.2. *Ambiguïtés liées au caractère bidimensionnel des EMs*

Pour interpréter une EM, il ne suffit pas seulement d'identifier tous les symboles qui la composent. Il reste une autre tâche qui peut se montrer encore plus ardue. Il s'agit de retrouver le modèle spatial selon lequel ces symboles de base s'arrangent.

• **L'espace de recherche est très vaste.** Avant toute chose, le caractère 2D des EMs, autorisant des possibilités de disposition suivant les différentes directions, fait que l'espace des hypothèses de relation est nettement plus élargi, comparé à celui d'un texte standard. Combiné avec la variété et le grand nombre de symboles, cette propriété est encore plus contraignante dans le cas des EMs.

• **Les frontières séparant les différentes relations spatiales sont de nature floue.** Les relations liant les symboles d'une EM semblent bien nettes, du moins du point de vue théorique. Cela est loin d'être le cas dès qu'il s'agit de se positionner du point de vue pratique. Si différentes personnes écrivaient sur une feuille de papier une EM contenant une simple relation d'exposant (x^2 à titre d'exemple), on obtiendrait une grande variabilité dans le positionnement du "2" par rapport au "x". Cette variabilité traduit la liberté dans le geste ressentie par les scripteurs. Le contexte peut aussi avoir une influence : une équation contenant uniquement cette relation, ou au contraire qui serait une partie d'une EM globale ne mèneraient pas au même positionnement.

• **L'interprétation des relations dépend du contexte.** Ainsi que l'a noté Martin dans (Martin, 1971), résoudre l'ambiguïté présentée précédemment ne suffit pas pour déduire de façon certaine la relation. En effet, il faut composer avec le positionnement relatif des symboles. Ceci est d'autant plus vrai, quand on sait que la situation est plus complexe dès que plus de deux symboles sont impliqués.

• **La description orale (naturelle) des relations est parfois très vague et peu discriminante.** Du point de vue de la parole spontanée, il est assez souvent très difficile de prévaloir certaines relations au profit d'autres quand une description textuelle est disponible. En effet, si des règles de diction ne sont pas imposées, un texte décrivant une EM peut avoir différentes écritures en langage mathématique (Fateman, 1998).

À partir de là, l'identification automatique des symboles et l'extraction automatique de la structure d'une EM à partir de son tracé manuscrit en-ligne ou de la parole est une problématique très intéressante levant des challenges scientifiquement pertinents. Dans la section 3 est présenté notre système bi-modal de reconnaissance des EMs.

3. Architecture globale de reconnaissance des EMs proposée

L'architecture que nous proposons pour la reconnaissance bimodale des EMs est de type modulaire. La combinaison des deux flux est assurée par une fusion tardive. Il est de ce fait nécessaire de disposer de deux systèmes amonts experts (un pour l'écrit et un autre pour la parole). Ces derniers doivent être en mesure de fournir les décisions intermédiaires sur lesquelles va porter la fusion. La liaison entre ces deux systèmes est assurée par un module de fusion composé de plusieurs unités. La figure 2 donne le schéma de l'architecture proposée.

Les trois modules cités précédemment et identifiés par leurs couleurs respectives sur la figure 2 ont donc chacun une tâche particulière à accomplir. C'est ainsi que le module en charge de la modalité audio assure le traitement du signal correspondant à l'enregistrement de la dictée de l'EM pour fournir une (ou plusieurs) transcription(s) associée(s). Le module qui traite la modalité manuscrite se charge quant à lui de l'interprétation du signal manuscrit en-ligne pour fournir l'interprétation de l'EM. Pour finir, les modules de fusion dans cette architecture, assurent l'interaction entre les deux modalités audio et manuscrite. Ils permettent notamment d'extraire de l'information de la modalité audio pour l'introduire à différents niveaux dans la chaîne de traitement du signal manuscrit en-ligne. Dans la suite, nous allons procéder à la description des systèmes spécialisés (écrit et audio) utilisés. Ensuite nous décrivons les modules assurant la fusion et le mode de fonctionnement de notre système.

3.1. *Présentation du module de reconnaissance des expressions mathématiques manuscrites en-ligne*

Le problème de l'interprétation d'une EM manuscrite en-ligne peut être formulé comme étant un problème de minimisation d'une fonction de coût $Cout_{EM}$ à valeurs dans $\mathbb{R}$. Elle est définie par $Cout_{EM} : E_{traitsEM} \longmapsto \mathbb{R}$, où $E_{traitsEM}$ est l'ensemble de traits composant l'EM. Ce coût est choisi parmi ceux associés à chacune des interprétations possibles de l'EM données par l'ensemble $E_{solutionsEM}$. Cette fonction est déduite des coûts liés aux différentes étapes composant le système global (segmentation, reconnaissance, interprétation). Il s'agit du coût associé aux scores de reconnaissance des hypothèses de symboles $Cout_{reco}$ (qualifiant les étapes de segmentation et de reconnaissance) et du coût de l'interprétation, dit aussi coût structurel, $Cout_{struct}$. Ainsi $Cout_{EM}$ est une fonction des deux coûts précédents : $Cout_{EM} = f(Cout_{reco}, Cout_{struct})$. Pour répondre à cette problématique, nous utilisons le système proposé par Awal et al. (Awal *et al.*, 2012). L'architecture proposée s'inscrit dans la catégorie des systèmes opérant une optimisation globale des trois étapes segmentation-reconnaissance-interprétation. En effet, dans ce cas, l'obtention de la solution ($Cout_{EM}$ minimal) repose sur l'optimisation simultanée de la segmentation, de la reconnaissance de symboles, et de l'interprétation.

Dans la suite, une brève description de chaque module de ce système est donnée.

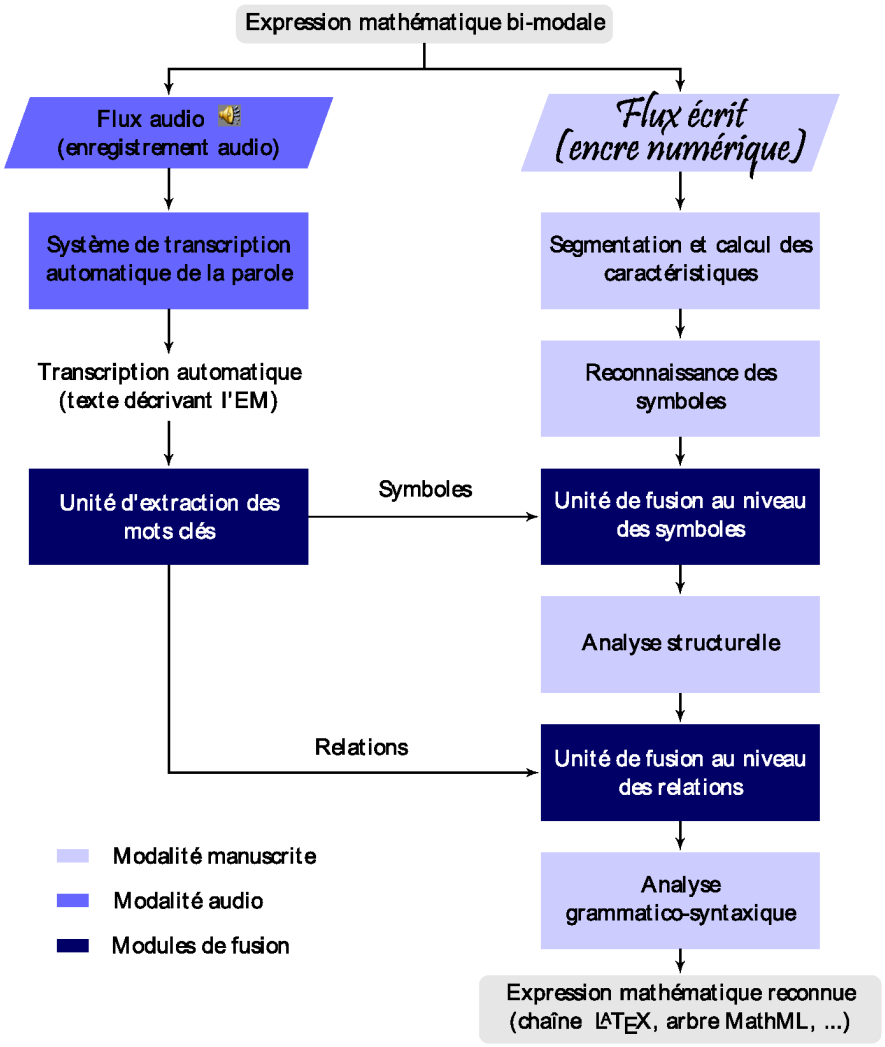

Figure 2. *Architecture globale proposée pour la reconnaissance des expressions mathématiques bi-modales*

Générateur d'hypothèses

Il est en charge d'explorer les différentes hypothèses de symboles (h_s). Chaque h_s est une combinaison d'un ensemble de traits. Sur l'ensemble de toutes les hypothèses de segmentation, la plupart d'entre elles sont invalides. Dans le cas de ce système, le générateur est basé sur une variante d'algorithme de programmation dynamique $2D$. Cela lui procure la possibilité de proposer des combinaisons de traits temporel-

lement non successifs, contrairement à une programmation dynamique $1D$. Dans le cas du signal manuscrit en-ligne en général, et des EMs en particulier, ceci est très important pour être en mesure de capturer l'information liée à l'opération de correction par des traits retardés dans le temps (ajouter les points sur des lettres, étendre les symboles élastiques tels que la barre de fraction ou la racine carrée...). En revanche cette extension $2D$ étend l'espace de recherche des hypothèses de façon exponentielle. Des contraintes de proximité géométrique et du nombre maximal de traits constituant l'hypothèse sont utilisées pour réduire le nombre d'hypothèses sur lesquelles opérer la recherche.

Classifieur de symboles

La tâche du classifieur de symboles est de fournir, à chacune des hypothèses de symbole, une liste des meilleures classes avec leurs scores associés. La particularité de ce classifieur est le pouvoir de **rejet**. En effet, en plus des classes correspondant aux différents symboles considérés par le vocabulaire, une classe additionnelle est rajoutée. Elle correspond à un rejet d'une hypothèse de segmentation. En d'autres termes, le classifieur n'a pas tout le temps à se prononcer en faveur d'une classe correspondant aux symboles quelque soit l'exemple qui lui est présenté. En effet, dans le cas des EMs complètes, énormément de groupements invalides de traits sont proposées par le générateur des hypothèses de symbole. Ceci fait qu'un grand nombre d'hypothèses qui atteignent l'étage de classification ne correspondent à aucune classe de symboles. Ces hypothèses doivent être, dans le cas de ce système, étiquetées comme des segmentations invalides grâce à la classe de rejet.

Analyse structurelle

Elle consiste à extraire les informations spatiales associées à chacune des hypothèses de symbole mais aussi associées à des sous-expressions composées de deux ou plusieurs hypothèses de symbole (dans (Awal *et al.*, 2012), le nombre d'hypothèses est égal soit à 2, 3 ou 4). Le but est de construire des coûts relationnels intermédiaires qui vont contribuer au coût structurel final ($Cout_{struct}$). Ces informations spatiales sont définies à partir des lignes de bases (pour rendre compte de l'alignement) ainsi que des hauteurs (pour rendre compte des tailles) des hypothèses de symbole. ll s'agit donc d'informations associées aux boîtes englobantes des sous-expressions, une sous-expression pouvant être une simple hypothèse de symbole. C'est cette procédure de construction des coût relationnels qui est à l'origine de l'appellation de coût géométrique du coût final alloué à une interprétation en EM donnée du tracé.

Modèle de langage (Analyse syntaxique)

L'analyse structurelle est accomplie grâce à l'utilisation d'une grammaire $2D$. Cette dernière est une combinaison de deux grammaires $1D$ dont les règles sont appliquées chacune dans une direction (l'axe horizontal et l'axe vertical). L'application successive de ces règles se fait jusqu'à atteindre les symboles élémentaires composant l'EM. Ensuite, une analyse ascendante est opérée afin de construire l'arbre relationnel

de l'EM. Les règles de grammaire dans ce cas sont associées, chacune, à une relation spatiale et sont de ce fait appliquées aux entités composant ladite relation. De ce fait, au cours de la validation syntaxique, toutes les relations possibles entre ces entités sont explorées et validées par la grammaire en tenant compte des coûts relationnels. Finalement, la relation choisie est celle exhibant le coût minimal parmi toutes celles qui sont valides et en considérant l'interprétation globale de l'EM.

Dans les travaux qui sont rapportés ici, nous avons utilisé la grammaire définie au cours de l'éditions 2012 de la compétition *CROHME* sur la reconnaissance des EMs manuscrites en-ligne.

Décision

C'est à ce niveau que la recherche de l'EM solution est faite. Elle correspond à celle ayant le coût global ($Cout_{EM} = f(Cout_{reco}, Cout_{struct})$) minimal. Cette solution doit être celle qui utilise tous les traits, chacun étant utilisé une et uniquement une seule fois.

3.2. *Présentation du système de transcription automatique de la parole*

Le comportement attendu d'un tel système est de produire en sa sortie une description textuelle à partir du signal de parole enregistrant la dictée d'une EM (description de l'ME par un locuteur). Le système utilisé, dans notre cas, est en mesure de fournir un graphe de mots correspondant aux différentes possibilités de découpage (segmentation) du signal audio complet. Chacun des chemins reliant les nœuds de début et de fin du signal (modélisés par des silences) donne une transcription hypothèse possible de la dictée. La solution la moins coûteuse (au regard des scores de reconnaissance donnés par le modèle acoustique et du coût de l'interprétation du point de vue du langage donné par le modèle de langage) est celle qui est retenue comme meilleure solution. C'est cette solution qui est fournie par le système comme transcription automatique du signal d'entrée. Dans ce qui suit nous présentons brièvement le détail de ces modules.

Décodeur

Dans notre architecture globale, ce module est basé sur le décodeur *CMU Sphinx* (Chan *et al.*, 2007), qui est basé sur les Modèles de Markov Cachés. Ce dernier est l'un des *STAP* libres[1] les plus utilisés à travers le monde. Plus exactement, c'est $Sphinx$ 3.3 qui constitue le cœur de notre *STAP*. Cette version du décodeur est celle qui fournit la meilleure précision de reconnaissance possible (Ravishankar *et al.*, 2000).

Toutefois, les ressources exploitées par ce décodeur, et fournies avec $Sphinx$ 3.3, sont destinées à la transcription de la parole continue en langue anglaise. Il a fallu les adapter au cas du langage mathématique parlé en langue française.

1. cmusphinx.sourceforge.net

Modèles acoustiques

Pour ce qui est des **modèles acoustiques**, nous avons eu recours à ceux mis en œuvre au sein du *LIUM* (Esteve *et al.*, 2010). Ces modèles acoustiques sont appris grâce aux outils fournis dans le projet *CMU Sphinx*. La boîte à outils *Sphinx Train* est utilisée à cette fin. L'apprentissage des modèles se fait sur des signaux de parole ayant une transcription vérité terrain. Cette transcription doit également être phonétisée. Cette description par des phonèmes de la transcription doit également être alignée de façon précise au signal de parole. Toutes ces tâches sont réalisées grâce aux outils de *Sphinx Train*.

Dictionnaire de prononciation (de phonétisation)

Dans le cas de notre application, la reconnaissance des EMs, le vocabulaire considéré est extrait du corpus audio de la base *HAMEX* (Quiniou *et al.*, 2011). En effet, nous avons extrait l'ensemble des mots disponibles dans les transcriptions des signaux audio de la partie apprentissage de la base *HAMEX*. Chacun des 423 mots extraits est décrit par la séquence phonétique qui le compose. Les différentes prononciations possibles de chaque mot sont considérées.

Modèle de langage

Pour définir le modèle de langage relatif à notre application, nous avons utilisé les outils disponibles dans *Sphinx Train* (le CMU Statistical Language Modeling (SLM) Toolkit[2]). Il s'agit de modèles de type $n - gram$, où dans notre cas $n = 3$ (modèles $tri - gram$). Dans ce genre de modélisation stochastique du langage, l'historique du mot est représenté par les $n - 1$ mots qui le précédent. Avec cette formalisation, il y a suffisamment d'information pour guider de façon efficace la tâche d'un $STAP$.

Dans notre cas, les données textuelles d'apprentissage composant le corpus ayant servi à l'estimation du modèle de langage sont issues de deux sources :

– La première est donnée par la partie apprentissage de la base *HAMEX*. Plus précisément, nous avons utilisé le texte correspondant aux transcriptions vérités terrains des expressions mathématiques dictées d'apprentissage ($2\ 925\ EMs$) de la base *HAMEX*.

– La deuxième source est un corpus synthétique. Ce dernier est construit en considérant environ $8\ 230$ chaînes LaTeX d'EMs réelles extraites du web. Cela a pour but d'enrichir la base d'apprentissage en terme de variabilité des EMs. Il s'agit en fait de construire à partir des chaînes LaTeX différentes dictées synthétiques possibles pour chacune d'elles. Pour cela nous avons développé un générateur de transcriptions synthétiques $Tex2Texte$. Celui-ci génère, à partir de la description sous forme de chaîne LaTeX un arbre de type *MathML* qui est par la suite analysé (à partir du terminal le plus à gauche) pour former, de façon aléatoire, une description textuelle possible de l'EM. Comme indiqué plus haut, plusieurs descriptions possibles de chacune des EMs sont générées (par tirage aléatoire).

2. http ://www.speech.cs.cmu.edu/SLM_info.html

3.3. *Présentation des modules de combinaison des deux modalités*

Comme nous l'avons présenté sur la figure 2, le module de fusion est principalement composé de trois unités. La première est directement connectée au système de transcription automatique de la parole et ne communique pas de façon directe avec le module de reconnaissance du tracé manuscrit, il s'agit de l'**unité d'extraction de mots clés**. Les deux autres unités (**unité de fusion au niveau symboles** et **unité de fusion au niveau relations**) sont quant à elles intégrées dans le système en charge de l'interprétation du tracé manuscrit. Elles sont donc en interaction directe avec ce système. L'unité d'extraction de mots clés se charge d'extraire l'information pertinente de la modalité audio pour la transmettre aux deux unités de fusion (symboles et relations).

Dans des travaux précédent (Medjkoune *et al.*, 2012), nous avons adopté une première approche, que nous avons qualifiée de fusion par approche "sac de mots" où nous avions considéré uniquement l'information de présence ou d'absence des mots clés issus de la transcription du signal audio pour venir désambigüiser si possible le signal manuscrit. Une extension de cette approche, présentée dans (Medjkoune *et al.*, 2013), vient enrichir le processus de fusion en apportant des informations additionnelles telles que les positions et les scores des hypothèses dans chacune des modalités et ne plus se connecter uniquement de l'information de présence ou d'absence des symboles. Dans la nouvelle approche que nous proposons ici, nous cherchons à affiner les associations des hypothèses en écrits et des hypothèses en audio afin de tendre vers une solution où la fusion concerne bien les même informations dans les deux modalités. Nous qualifions cette méthode de **fusion par alignement**. Nous donnons dans la suite une description plus détaillée de chacune des étapes assurant la fusion par le biais de de cette méthode.

3.4. *Extraction des mots clés*

Le vocabulaire utilisé au sein du langage mathématique regroupe deux catégories de mots. La première est la classe de mots qui sont utiles du point de vue du langage considéré (EMs). Ces mots concernent soit des *symboles*, soit des *relations*, ou encore les deux à la fois. Ils sont appelés **mots clés**. La seconde classe de mots comprend tous les autres termes. Ils ne sont nécessaires que pour donner un sens (une structure correcte), du point de vue linguistique, à la description textuelle de l'EM. Il sont identifiés par la terminologie de **mots vides**. En effet, ils n'apportent pas d'information pertinente au problème de reconnaissance des EMs. De ce fait, la première tâche de cette première unité du module de fusion a pour rôle d'identifier les mots clés parmi tous les mots de la description textuelle de l'EM. À partir de ces mots clés, la tâche suivante consiste à opérer une analyse sémantico-syntaxique (le sens des mots et leur association sont importants) pour extraire les symboles et relations qui y sont inclus. Ces derniers sont par la suite communiqués aux unités en charge du processus de fusion lui-même. En résumé, ce module a pour objectif de rechercher des sous-ensembles de symboles et de relations qui font partie de l'EM (au sens de la parole). Cela revient à

construire un **dictionnaire** "Français-EMs" qui, à chaque mot (ou ensemble de mots) du vocabulaire associe une liste de symboles et/ou une liste de relations.

Un tel dictionnaire n'est pas disponible a-priori. Il convient donc de le construire à partir du corpus des EMs audio disponibles pour l'apprentissage. Pour chaque EM, chaque unité lexicale reconnue comme un mot clé est traduite en un équivalent de un ou plusieurs symboles et/ou relations.

A ce stade, les informations extraites de la transcription automatique du signal audio se présentent sous la forme d'une double liste de symboles et relations, celles-ci sont exploitées par les unités de fusion associées. Dans la suite nous exposons ces unités en question.

3.5. *Fusion d'information au niveau symboles*

Dans ce cadre, nous cherchons à favoriser une mise en correspondance des modalités qui tienne compte de leur alignement temporel. En effet, il n'est pas seulement question de considérer la liste de symboles extraite de la transcription automatique du signal audio, mais d'avoir un découpage du signal complet en segments. Chaque segment représente une hypothèse (ou une liste d'hypothèses) de symbole(s). L'objectif est par la suite de trouver les associations groupement écrit (hypothèse de segmentation en écrit : regroupement de traits) et segment audio associé (hypothèse de segmentation en audio : portion du signal audio) tel qu'illustré en figure 3. L'association n'est valide que si le groupement écrit et le segment audio impliqués représentent la même information (même symbole mathématique dans ce cas). La finalité de cette association est d'être en mesure, une fois cette étape accomplie, d'appliquer une technique de fusion (à base de fonctions de croyance ici) pour tirer profit de l'aspect bimodal de l'information. Dans la suite est décrite la façon de procéder à l'alignement des hypothèses de symboles écrit/audio à fusionner.

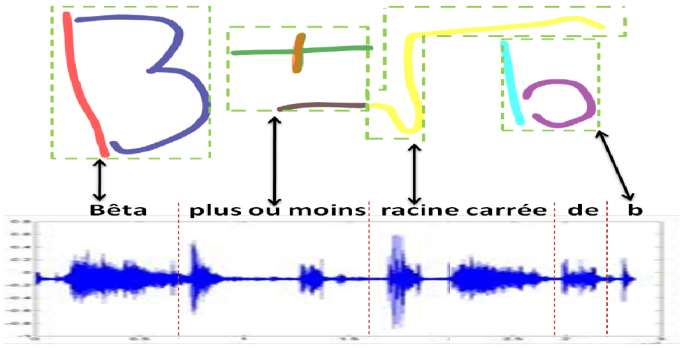

Figure 3. *Exemple d'association groupements/segments pour deux partitions écrit et audio données et représentant la même EM*

3.5.1. *Choix de l'association groupement/segment à retenir.*

Nous confions, dans notre approche, la responsabilité du choix de l'appariement entre un groupement de traits et l'ensemble des segments audio à un réseau de neurones. Celui-ci reçoit en entrée les scores des différentes classes provenant d'une part du classifieur du système en charge de la modalité manuscrite pour un groupement écrit donné et d'autre part les scores provenant du classifieur audio recevant en entrée un segment audio considéré. Les caractéristiques en entrée du classifieur assurant l'alignement sont la concaténation des listes précédentes. À partir de là, il élabore en sortie une décision pour considérer comme valide ou non l'appariement. Pour une hypothèse de symbole en écrit, une fois que l'ensemble des segments audio a été balayé, seul le meilleur appariement est conservé pour être envoyé au module de fusion. Ce classifieur, tout comme celui de l'écrit (duquel on s'est inspiré), est appris de façon globale. Cela revient à dire, qu'une fois une proposition de partition de l'EM complète est faite, chaque hypothèse de segmentation est analysée en considérant l'association qui lui a donnée naissance. Dans le cas d'une mauvaise proposition de segmentation (au niveau des traits de la solution finale) ou d'une mauvaise association des groupements écrits/segments audio, cet exemple engendre deux information à apprendre. Dans un premier temps, la proposition de segmentation en question est apprise comme un exemple à rejeter. À coté, les bonnes associations impliquant les bonnes hypothèses de symboles en écrit (en se basant sur les traits de la proposition de segmentation en question) et les hypothèses de segments audio associées sont également apprises comme les bons exemples.

3.6. *Fusion d'information au niveau relations*

Pour la fusion au niveau des relations, ce sont les coûts de reconnaissance des relations spatiales liant les différents symboles du système écrit qui vont être modifiés. Rappelons que les coûts fournis par le système écrit aux différentes relations sont du type géométriques, ceux-ci sont très dépendants de la taille et des dispositions des symboles. Cela rend leur normalisation ardue. En tenant compte de cette contrainte pratique, nous avons exploré une fusion par approche "sac de mots". Une transformation linéaire, donnée par l'équation 1, est appliquée en considérant les coûts des relations alloués par la modalité manuscrite. En effet, au moment de l'analyse structurelle par le système de reconnaissance du tracé manuscrit, pour chaque couple d'hypothèses de symboles, toutes les relations les liant sont explorées et des coûts géométriques sont calculés. Chaque coût est par la suite réévalué et l'ordre des relations est de ce fait susceptible d'être changé et cela en consultant la liste des relations possibles $LRel$ issues de la modalité audio (de la totalité du signal décrivant l'EM). Ceci a pour effet de pénaliser les relations absentes dans la transcription en augmentant leurs coûts. Les coûts des relations présentes en audio sont au contraire abaissés pour les favoriser lors de l'analyse grammaticale. On obtient alors :

$$Cout(R/j, LRel) = \begin{cases} Coef_r \times Cout_{ecrit}(R/j) & si\ R \in LRel \\ Coef_p \times Cout_{ecrit}(R/j) & sinon, \end{cases} \quad [1]$$

où : $Cout_{ecrit}(R/j)$ et $Cout(R/j, LRel)$ sont les coûts avant et après fusion pour que la sous-expression j utilise la relation de type R, $Coef_r$ et $Coef_p$ sont respectivement les coefficients de rehaussement ($Coef_r < 1$) et de pénalisation ($Coef_p > 1$).

4. Résultats expérimentaux

Dans cette section nous allons présenter quelques résultats expérimentaux associés à notre système de reconnaissance d'EMs bi-modales. Nous commençons d'abord par présenter les données utilisées pour le test, ensuite les performances des systèmes mono-modaux sont données. Finalement, les performances du système complet de fusion sont rapportées et analysées.

4.1. *Données utilisées*

Pour être en mesure de faire l'apprentissage du classifieur en charge de l'alignement écrit/audio, il est nécessaire de disposer de la vérité terrain au niveau alignement des deux modalités. En effet, l'apprentissage global du classifieur a pour vocation d'apprendre à rejeter deux types d'exemples :

– Les exemples dont l'hypothèse de segmentation formulée à l'écrit est mauvaise. Nous remédions à ce cas en considérant le classifieur écrit déjà entraîné à reconnaitre les mauvaises segmentation grâce à la classe de rejet. Le classifieur de fusion aura dans ce cas à copier ce rejet formulé en écrit dans sa propre classe de rejet.

– Les exemples issus d'un mauvais alignement écrit/audio, abstraction faite de la qualité de l'hypothèse de segmentation de l'écrit et de la reconnaissance qui lui est associée. Ce type d'exemples requiert la disponibilité de la vérité terrain sur les associations groupement écrit/segment audio. Cette association peut être obtenue automatiquement tant qu'une classe de symboles n'apparait qu'une fois au maximum dans l'EM considérée. Dans le cas contraire, il faut opérer cet alignement en annotant manuellement les données.

Dans la mesure où notre base n'est pas annotée à ce niveau, nous présentons ici, une expérimentation préliminaire qui ne considère que les EMs qui ne contiennent pas la même classe de symbole à plusieurs reprises (une classe est présente une seule fois au maximum dans l'EM). Cette considération permet, comme indiqué, de générer automatiquement l'alignement vérité terrain.

Les données utilisées pour cette expérimentation sont extraites de la base $HAMEX$ en ne gardant que celles respectant la contrainte de présence une fois au maximum de chaque classe de symboles. Elles sont au nombre de 731 EMs pour l'apprentissage (549 pour l'entraînement et 182 pour la validation des poids optimaux) et 239 EMs pour le test. Cette nouvelle sous-base de test est du fait de cette contrainte moins complexe que la totalité de la base $HAMEX$. En effet, même si toutes les relations spatiales et tous les symboles pilotant notre système sont présents, la taille moyenne en symbole des EMs est plus faible relativement aux EMs de $HAMEX$ (5 symboles en moyenne, avec la plus courte écrite avec 3 symboles et la plus longue

avec 12 symboles contre 13 pour $HAMEX$ avec 3 symboles pour la plus courte et 28 pour la plus longue).

4.2. *Performances des systèmes mono-modaux*

Le système de reconnaissance des EMs manuscrites en-ligne appris sur la base d'apprentissage de *CROHME 2012* (Mouchère *et al.*, 2012), possède les performances résumées sur le tableau 1. Ces expérimentations sont conduites sur l'ensemble des 519 EMs de $HAMEX$ en considérant uniquement leur version manuscrite.

Niveau d'évaluation	Taux traits	Taux symboles	Exp. sans err.	Exp. à 1 err. près	Exp. à 2 err. près
Taux de reco. [%]	80.05	82.93	34.10	46.44	49.52

Tableau 1. *Performances du système de reconnaissance des EMs manuscrites*

D'après le tableau 1, seules 34.1% des EMs sont complètement reconnues par le système basé sur la modalité manuscrite. Toutefois, si une seule erreur en étiquette de symbole ou de relation est tolérée, on arrive déjà à 46.44% de bonne reconnaissance et à presque la moitié des EMs de bien reconnues si deux erreurs sont autorisées (49.52%). Cela suggère qu'une information extérieure pourrait faire basculer l'issue de la reconnaissance pour beaucoup d'EMs mal reconnues parce qu'une ou deux erreurs sont présentes.

Le système audio en charge d'apporter l'information extérieure pour aider la reconnaissance possède un taux de reconnaissance au niveau de tous les mots de 90.07% et un taux de reconnaissance de 97.21% au niveau des mots clés (test sur la partie audio de l'ensemble de la base de test de $HAMEX$).

Les performances du système de TAP sont très élevées, notamment en ce qui concerne la reconnaissance des mots clés. Cette propriété combinée avec les performances du système de reconnaissance du signal manuscrit renforce d'avantage l'hypothèse de complémentarité audio-écrit, c'est ce que nous allons vérifier dans la suite.

4.3. *Résultats du système proposé*

Le classifieur utilisé ici pour l'alignement est un perceptron multi-couches (PMC) avec une seule couche cachée, ayant $2 * Nbclasses + 1 = 113$ neurones sur la couche d'entrée ($Nbclasses$ est le nombre de classes de symboles, égal à 57 ici), 100 neurones sur la couche cachée et $Nbclasses + 1$ neurones sur la couche de sortie.

Sur le tableau 2 sont rapportés les résultats de cette étude comparés aux performances du système de référence (système manuscrit seul) et à celles de la meilleure configuration de fusion de notre système précédent à base de l'approche sac de mots (Medjkoune *et al.*, 2013).

On peut observer sur le tableau 2 que, même si le problème est simplifié en ne prenant que des EMs ayant une seule fois au maximum chaque étiquette de symbole, la méthode de fusion s'appuyant sur un alignement grâce à un classifieur est celle qui

Taux de reco.(%)	Taux traits	Taux symboles	exp. sans err.	exp. à 1 err. près	exp. à 2 err. près
Écrit seul	86.16	88.93	41.84	74.89	76.15
Ancien système	90.95	93.06	59.83	77.4	77.8
Nouveau système	92.18	92.49	62.76	79.9	79.9

Tableau 2. *Comparaison des performances du système de reconnaissance des EMs manuscrites en ligne sans fusion avec la meilleure configuration de fusion de notre précédent système et du système actuel.*

assure la meilleure performance au niveau EMs complètes (que l'on autorise ou pas d'erreurs). L'analyse plus fine des performances du système de REM manuscrites sans fusion et avec les meilleures configurations de fusion montre que la version actuelle de notre système est celle qui permet une plus grande amélioration des performances à tous les niveaux (étiquettes des traits, symboles, expressions si une ou deux erreurs sont autorisées).

5. Conclusions et perspectives

Dans ce papier, nous avons abordé le problème de la reconnaissance d'expressions mathématiques dans le cadre d'un traitement bi-modal. Les modalités considérées sont le flux audio et celui de l'écriture manuscrite en ligne. Nous avons utilisé une transcription automatique du signal de description de l'expression à reconnaître afin d'aider à la désambiguïsation de la reconnaissance au niveau du système REM. Les résultats obtenus sont meilleur que ceux obtenus dans les versions précédentes dans notre système et vont dans le sens des conclusions que nous avions énoncées alors. On arrive donc à améliorer les performances du système global de près de 50% par rapport au système écrit seul et de près de 5% par rapport à l'ancienne version de notre système. Un autre point important est lié à l'apport de la fusion aux deux niveaux expressions et relations confirmant ainsi les complémentarités qu'intuitivement on prévoyait.

Dans la suite nous prévoyons d'étendre notre cette étude à toute la base HAMEX en proposant une annotation complète de celle-ci au niveau alignement. Nous envisageons également d'aller plus loin dans l'exploitation des techniques de classification pour la fusion des flux écrit-audio. En plus de cela, nous n'avons jusqu'à présent pas exploité l'information de contexte des symboles et relations au sein des deux modalités.

Remerciements

Ce travail s'inscrit dans le cadre du projet DEPART porté par la région Pays de La Loire http ://www.projet-depart.org/.

6. Bibliographie

Awal A.-M., Mouchère H., Viard-Gaudin C., « A global learning approach for an online handwritten mathematical expression recognition system », *Pattern Recognition Letters*, 2012. In Press, Corrected Proof.

Chan A., Evandro G., Rita S., Mosur R., Ronald R., Yitao S., David H.-D., Mike S., *The Hieroglyphs : Building Speech Applications Using CMU Sphinx and Related Resources*, http ://speech.tifr.res.in/tutorials/sphinxDocChan070111.pdf. 2007.

Chan K. F., Yeung D. Y., « Mathematical Expression Recognition : A Survey », *International Journal of Document Analysis and Recognition*, vol. 3(1), p. 3-15, 2000.

Elliott C., Bilmes J., « Computer Based Mathematics Using Continuous Speech Recognition », *Proc. of Int. Conf. CHI 2007 Workshop on Striking a C[h]ord : Vocal Interaction in Assistive Technologies, Games, and More*, 2007.

Esteve Y., Deléglise P., Meignier S., Petitrenaud S., Schwenk H., Barrault L., Bougares F., Dufour R., Jousse V., Laurent A. *et al.*, « Some recent research work at lium based on the use of cmu sphinx », *les actes de CMU SPUD Workshop, Dallas (Texas)*, 2010.

Fateman R., « How can we speak math », *Journal of Symbolic Computation*, 1998.

Martin W. A., « Computer input/output of mathematical expressions », *Proceedings of the second ACM symposium on Symbolic and algebraic manipulation*, SYMSAC '71, ACM, New York, NY, USA, p. 78-89, 1971.

Medjkoune S., Mouchère H., Petitrenaud S., Viard-Gaudin C., « Using Speech for Handwritten Mathematical Expression Recognition Disambiguation », *Proceedings of 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR-2012)*, Bari, Italie, p. 1-6, 2012.

Medjkoune S., Mouchère H., Petitrenaud S., Viard-Gaudin C., « Multimodal Mathematical Expressions Recognition : Case of Speech and Handwriting », *in* M. Kurosu (ed.), *Human-Computer Interaction. Interaction Modalities and Techniques*, vol. 8007 of *Lecture Notes in Computer Science*, Las Vegas, États-Unis, p. 77-86, July, 2013. Région Pays de la Loire, Projet DEPART.

Mouchère H., Viard-Gaudin C., Kim D. H., Kim J. H., Garain U., « ICFHR2012 : Competition on Recognition of Online Handwritten Mathematical Expressions (CROHME 2012) », *Proc. of Int. Conf. on Frontier in Handwriting Recognition (ICFHR)*, p. 811-816, 2012.

Quiniou S., Mouchere H., Saldarriaga S., Viard-Gaudin C., Morin E., Petitrenaud S., Medjkoune S., « HAMEX - A Handwritten and Audio Dataset of Mathematical Expressions », *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, p. 452-456, 2011.

Ravishankar M., Singh R., Raj B., Stern R. M., « The 1999 CMU 10x real time broadcast news transcription system », *Proc. DARPA Workshop on Automatic Transcription of Broadcast News*, Citeseer, 2000.

Tapia E., Rojas R., « A Survey on Recognition of On-Line Handwritten Mathematical Notation », 2007.

Wigmore A., Hunter G., Pflugel E., Denholm-Price J., Binelli V., « Using Automatic Speech Recognition to Dictate Mathematical Expressions : The Development of the TalkMaths Application at Kingston University. », *Journal of Computers in Mathematics and Science Teaching (JCMST)*, vol. 28(2), p. 177-189, 2009.