Using Online Handwriting and Audio Streams for Mathematical Expressions Recognition: a Bimodal Approach

Sofiane Medjkoune, Harold Mouchère, Simon Petitrenaud, Christian Viard-Gaudin

Using Online Handwriting and Audio Streams for Mathematical Expressions Recognition: a Bimodal Approach

Sofiane MEDJKOUNE[a,b], Harold MOUCHERE[a], Simon PETITRENAUD[b] and Christian VIARD-GAUDIN[a]

[a]LUNAM University, University of Nantes, IRCCyN, France
{sofiane.medjkoune, harold.mouchere, christian.viard-gaudin}@univ-nantes.fr
[b]LUNAM University, University of Le Mans, LIUM, France
simon.petit-renaud@lium.univ-lemans.fr

ABSTRACT

The work reported in this paper concerns the problem of mathematical expressions recognition. This task is known to be a very hard one. We propose to alleviate the difficulties by taking into account two complementary modalities. The modalities referred to are handwriting and audio ones. To combine the signals coming from both modalities, various fusion methods are explored. Performances evaluated on the HAMEX dataset show a significant improvement compared to a single modality (handwriting) based system.

Keywords: Mathematical expression, Handwriting, Speech, Data Fusion, Belief functions

1. INTRODUCTION

Mathematical language is an universal language. It allows sharing knowledge around the world regardless of the mother tongue of the people involved. The widespread use of mathematical equations exceeds largely the context of specialist users, particularly because of the expressive power of such a language. Actually, they are used in various fields: in science such as computer science, physics ..., or in human sciences such as economics, sociology, law and so on. The mathematical formalism is often used to modelize the behaviour of a particular phenomenon for prediction or just understanding. During the edition of documents containing such a type of language, specialized tools are needed. This is because mathematics are a bi-dimensional language. This means that the allowed directions of writing are in a 2D plane unlike the text edition which is from left to right (or from right to left) as shown on Fig.1. Consequently, it is more difficult to insert a mathematical expression (ME) than a standard text using common tools dedicated to the document formatting task.

Usually, to insert a ME in a document, editors based on a *WIMP* (Windows, Icons, Menu, Pointing) interaction are provided. The two most famous ones are LaTeX and *MathType*. They are designed to make the ME edition as userfriendly as possible. Their major drawback is the difficulty of handling such tools since new edition rules are needed to manage the 2D aspect of the ME. Beside of that, the ME edition remains time consuming.

Recent technological advances have allowed the development of new tracks relying on natural human-machine interaction modes [1]. These tracks concern handwriting and speech, which are very intuitive communication modes for human beings. Both of them offer an easier alternative to enter 2D structures on documents once systems in charge of handwriting and speech signals interpretation are set up.

Figure 1. Direction of composition of a (a) standard text, (b) mathematical expression

Thus, the handwriting recognition and the automatic speech transcription communities have been highly interested by the problem of graphical language (2D) interpretation, especially for the case of ME. Even though the 2D aspect is more perceptible in the case of handwriting modality than in the case of speech, it is very relevant to use both signals since they are very complementary. More precisely, from the handwriting point of view, the ME layout is implicitly contained in the signal (points coordinates), even if the exact positions of the basic units (strokes) with respect to each others can be fuzzy. The speech signal does not embed implicitly the 2D layout information of the ME, but requires a linguistic post processing step to be able to recover the 2D organisation.

That is how the problem of handwritten ME recognition has been widely investigated [2, 3]. The efforts made by the scientific community led to the development of several competitive systems. Nevertheless, these systems are not hundred percent reliable. In fact, there are some intrinsic limitations that cannot be overcome because of the nature of the handwriting signal (symbol and relation ambiguities). Most of the time, these confusions are not obvious to discern, even for an experienced observer who would look at the handwritten ME layout (Fig.2).

More recently, works on spoken ME recognition have emerged [4, 5]. Most of them rely on a classical automatic speech recognition (ASR) system that provides the basic automatic transcription of the speech signal. Then, this latter is sent to a parsing module to convert the simple text describing the ME (1D) into its mathematical language writing (2D) [5, 6]. Here again, the systems set up are far from being hundred percent reliable. In addition to the resulting errors during the recognition step (common to all ASR systems), the transition from the textual description of the ME to its 2D writing is not obvious at all (Fig.2). Figure 2 shows

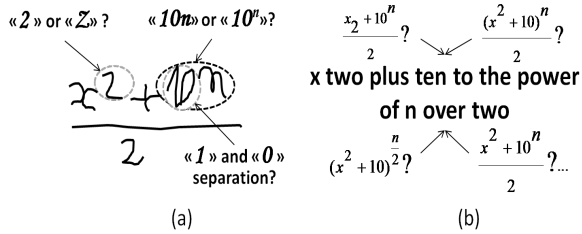

Figure 2. Examples of encountered problems in automatic MER with respect to (a) handwriting modality, (b) speech modality

some examples of cases where the two systems are in failure. In this example, we can observe the existing complementarity between the two modalities. The problems encountered by both modalities are of different kinds. So, the missing information in one modality is generally available in the other one.

Starting from this observation, in this paper we propose a bimodal system, exploiting both handwriting and speech modalities, in order to overcome the weaknesses of each modality taken separately. Thus, the remainder of this paper is organized as follows: in section 2 the specificities of the mathematical expression language are described. We briefly review the necessary background for the work proposed in this paper in sections 3 and 4. In section 5 we present our system. We will devote section 6 to the presentation of results and their analysis. Section 7 concludes this paper and gives perspectives of the current work.

2. MATHEMATICAL LANGUAGE SPECIFICITIES

Because they belong to the 2D graphical languages, the ME are a structure composed of elementary units called *symbols* in the bidimensionnal space. These symbols are spatially arranged according to different possible *relationships*. Thus, the spatial relationship between two symbols can be various (*left/right, up/down, subscript/superscript, inside*) giving rise to a possible complex layout.

On this basis, considering the handwriting or speech modalities for ME typing, beside of the classical problems which this kind of systems has to address, some ME specificities make this task more difficult. These difficulties can be of two types: those due to the symbols and those related to the spatial relationships.

- **At symbol level:**

  1- The total number of symbols is very large (≅ 220 various symbols) compared to a standard text (≅ 60). This makes the classification task harder.

  2- There are many confusions between the symbols. Since there are many symbols, there are also many similarities in the display of some symbols (see Fig.3-a).

  3- The role of a symbol depends on the context, for example if the sign '-' is encountered by the handwriting system, this sign can represent a *'minus sign'*, a *'fraction bar'* or a part of another symbol such as the *'equality sign'* (Fig.3-b).

  4- Concerning the speech modality, symbol pronounciation can be ambiguous (for example: *'m'* and *'n'*, *'x'* and *'s'*). This makes the automatic transcription of the speech signal more difficult.

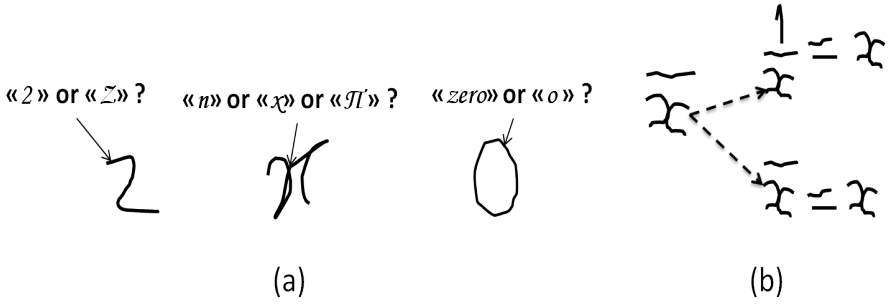

Figure 3. Examples of symbols confusion with respect to the handwriting modality due to (a) inter-symbols confusion (b) symbol role

- **At relation level:**

  1- The nature of relationships between symbols is fuzzy. This means that between the extreme cases (subscript, horizontal pair or superscript) which are easily identified, there is an ambiguity concerning the intermediate cases. This is illustrated on Fig.4-a.

  2- Another property is the relative symbol position as reported on Fig.4-b.

  3- Considering the speech modality, the language ambiguity during the dictation makes the choice of a relation instead of another one non trivial (Fig.4-c).

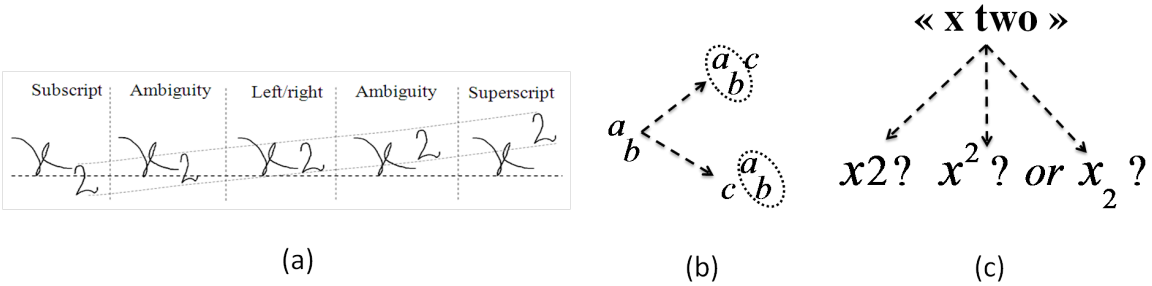

Figure 4. Confusions at relationship level due to (a) their fuzzyness nature (b) the dependence on the relative positions, (c) the audio description ambiguity

Thus, we can see that the problem of mathematical expression recognition rises very interesting challenges from a scientific point of view, in addition to the importance of the language itself as presented in the introduction section.

3. ONLINE HANDWRITTEN MATHEMATICAL EXPRESSION RECOGNITION

In this paper we are interested by online handwritten signals. This means that the handwriting recognition system receives a set of elementary strokes. These strokes are temporally ordered according to their time of acquisition. Each stroke is defined by a certain number of points bounded by a pen-down and a pen-up points. In this work, we will consider that a pen-up is present at the end of every symbol, which can be written with several strokes. These strokes are not necessarily consecutive, since some strokes can be delayed. The number of points depends on the temporal sampling rate of the digital pen, the speed of writing and of course on the length of the stroke. Mostly, before starting the recognition process itself, the input signal undergoes a preprocessing step [7]. It consists of spatially resampling each stroke using a constant rate. This preprocessing ensures consistency during the following processing steps, especially for the recognition step.

Chan and al. [2] have identified three independent steps that a handwritten ME recognition system has to include to achieve its task. The first step is the *segmentation* process in which the possible groups of strokes are formed. This stage is not trivial when as supposed here, interspersed symbols are authorized. Each *group* is called a segmentation hypothesis (*'hs'*). Ideally, each *'hs'* corresponds to a mathematical symbol. The recognition process is the second step. It aims to assign a symbol label (or a list of possible symbols) and a recognition score for each *'hs'*. The third step is the structural analysis. All the recognized symbols are used to deduce the final ME. This is done through a spatial-grammatical analysis. The main limit of the approaches consisting on optimizing each step alone is that the failure of one step will lead to the failure of the next one. Rhee and Kim reported in [8] a solution to reduce this error propagation which consists in the simultaneous optimization of the segmentation and recognition steps. In this case, the classifier is trained separately on isolated symbols. Furthermore, Awal and al. proposed a more global architecture [9]. The strengths of their system are the following. First of all, the recognition module is trained within the expressions and not longer uses an isolated symbol database. This allows a direct interaction between the different stages of the system (segmentation, recognition and 2D parsing). Secondly, during the segmentation step, a non-consecutive stroke grouping is allowed to form valid symbols. Finally, the structural analysis (2D parsing) is controlled by both symbol recognition scores and a contextual analysis (spatial costs). The handwritten MER sub-part used in our architecture will be largely based on Awal and al.'s system.

4. SPOKEN MATHEMATICAL EXPRESSION RECOGNITION

To achieve a MER task based on automatic speech recognition (ASR), two main modules are mainly needed [4, 5]. The first one achieves the automatic speech recognition task. The output of this module provides a textual description as reliable as possible (depending on the performance of the ASR) of the audio description. This text is composed of words written with alphabetic characters as they are recognized by the ASR system. This text is ideally a fair description of the ME (it also depends on the accuracy of the speaker who speaks out the ME). The second module is a parser, which processes the previous transcription in the 2D space to deduce the associated ME.

The automatic transcription in the global MER system is given by ASR system which is quiet similar to the one described in the case of handwriting modality. The main difference is the nature of the signal which is processed (acoustic one in this case). The recognition procedure involves three stages. During the first one, the acoustic signal is filtered and resampled, then a frame description is produced, where a feature vector is computed for each window of 25 ms with an overlap of 10 ms. The features are most of the time the cepstral coefficients and their first and second derivatives [10]. Segmentation into homogeneous parts is operated in a second step. Resulting segments are close to minimal linguistic units. The last step is to perform the decoding itself using models and tools learned within a training step (acoustical model, pronunciation dictionary and language model).

Parsing the resulting transcription from the previous module is a very hard task. In the rare existing systems [4, 5], the parsing is most of the time assisted by either introducing some dictation rules (for delimitating fraction's numerator and denominator for example) or using an additional source of information (such as using a mouse to point the position where to place the different elements). By adding such constraints, the editing process becomes less natural and far from what is expected from this kind of systems.

The work we report in this paper concerns the French spoken language. The task of speech recognition in our system is carried out by a system largely based on the one developed at the LIUM [10], which kernel is one of the most popular worldwide speech recognition systems (CMU-Sphinx)[11].

5. BIMODAL MATHEMATICAL EXPRESSION RECOGNITION

5.1 The concept of data fusion

The idea of multimodal human-machine interaction comes from the observation of the human beings' interaction. Usually, people simultaneously use many communication modes to converse. In so doing, the conversation becomes less ambiguous. The main goal of this work is to mimic this procedure to be able to set up a multimodal system dedicated to mathematical expressions recognition (MER).

Generally, data fusion methods are divided in three main categories [12, 13]: *early fusion* which happens at features levels; *late fusion* which concerns the intermediate decisions fusion and the last one is the *hybrid fusion* which is a mix of the two. Within each approach, three kinds of methods can be used to carry out the fusion process. Rules based approaches represent the first category, it includes methods using simple operators such as max, (weighted) mean or product. The second category is based on classifiers and the last one is based on parameter estimation.

Let us remember that we are interested in combining the information coming from handwritten and audio streams. These two signals are of heterogeneous natures. This prevent from considering an early strategy of fusion and suggest using a late one. In such a way, we will also make sure to use suitable recognition systems with respect to each modality. In addition, the matter of this paper is to explore some of the different possibilities of combination and the synchronization problem is not investigated here. In the following sections we describe the architecture of the proposed collaborative system.

5.2 Data fusion for mathematical expression recognition

The proposed architecture for bimodal mathematical expressions recognition (BMER) is presented in Fig.5. Its overall operation is as follows.

As input, the system receives a ME available at both speech and handwritten forms. The audio signal describing the ME is sent to the automatic speech recognition system (ref. section 4). This latter processes the audio signal and gives as an output an automatic transcription which is a text describing the ME. For each word of this text a recognition score is assigned. At this level, the recognition result has still only one dimension, as a standard text. Not all the words composing the textual description are useful in a mathematical language point of view. So, the automatic transcription should undergo a preprocessing step. It is the keyword extraction step. This procedure will be detailed in the next section. Meanwhile, the online handwritten signal is also processed by the module in charge of the handwriting recognition part. This signal consists on a sequence of strokes. As explained in section 3, the first step aims to form the basic symbols composing the ME, by grouping together the strokes belonging to the same symbol. Second, each group is assigned a list of labels and their corresponding scores. The set of formed symbols is then parsed in the 2D plan to define the ME layout. Practically speaking, the fusion process for MER can be carried out at two levels: during the recognition step (symbol level) which ensures a larger reliability at the symbol recognition level (Fig.3). The second level is the relational one. This is done during the structural analysis process and it aims to reduce the occuring confusion at this level (*cf.* Fig.4). This merging process is done thanks to the fusion units (grey boxes) in the architecture of Fig.5 which will be described below.

5.2.1 Keyword extraction from the audio transcription

The purpose of this step is to analyze the text describing the ME given by the audio system. As a result, two word categories are identified. The first one is composed of words which are useful for the MER process. They spot either symbols (such as: *'x', 'deux', 'parenthèses'*); relations (*'indice', 'exposant'*); or both (*'intégrale', 'racine'*). The second category of words includes all the other words (stopwords). These words are only used to make sense from a language point of view. Here, we consider the words from the first category, as *keywords*. A dictionary is built in such a way that each symbol and each relation is associated to one or more keywords.

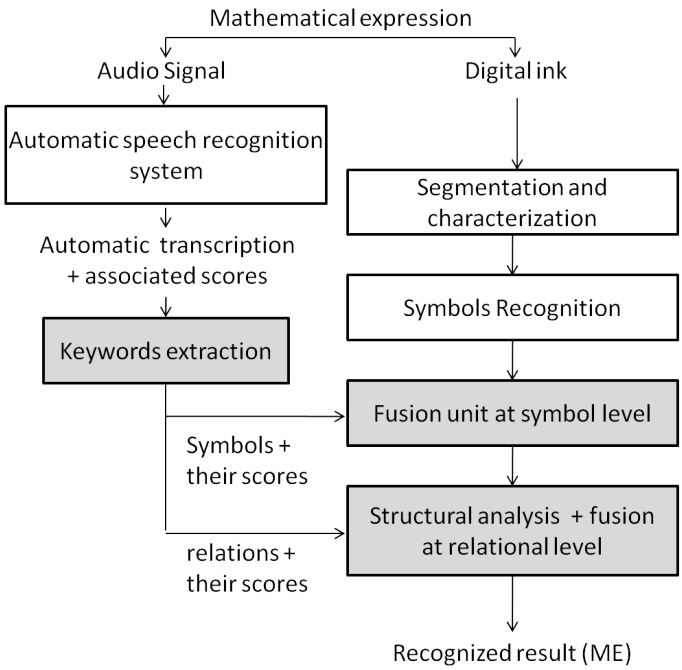

Figure 5. The collaborative architecture for complete mathematical expression recognition

For example if the word *'carré'* (which means *'squared'* in French) is existing in the transcription, the ME we are processing could contain the symbol *'2'* and the relation *'superscript'*. If any confusion concerning these symbols appears during the handwriting recognition, the fact that they are less ambiguous in the speech modality increases the confidence about them.

5.2.2 Fusion units description

As presented in the beginning of this section, the goal of the fusion unit is to ensure the fact that the information (symbols and relations) taking part in the construction of the final solution of the ME layout is coming from the two modalities. This can be achieved thanks to various fusion methods. We summarize in the following, the different approaches we explored.

A **Fusion methods at symbol level (IFSL):** In this paper, we mainly explored rule based methods to achieve the task of fusion at symbol level. The details of these methods are given below.

First, let us define some notations which will be used within the remainder of this paper. Let $C = \{c_1, c_2 \dots c_N\}$ be the set of the N possible classes of symbols we consider in our system. This means that, both handwriting and audio systems choose symbol labels from this list. In other words, for each *group of strokes*, also called *segmentation hypothesis*, the handwriting system assigns a list of the possible labels with their respective scores (ranked according to their score). This list is a part of C. The speech recognition system does the same for each *segment* of the audio signal, which corresponds to the *segmentation hypothesis* in the ASR system. We define $d_{i,j}(x)$ as the score (before fusion) that the symbol to classify x is the class c_j with respect to the modality i ($i \in \{s, h\}$, s for the speech modality and h for the handwriting modality). After the fusion process, we denote the resulting score as $d_j(x)$.

Based on these notations, we focus on the methods used to obtain the scores after fusion.

A-1 **Weighted summation[12, 13]:** the score $d_j(x)$ is obtained by summing the weighted scores of that class (j) from the two modalities. Formally, this is given by the equation (1)

$$d_j(x) = \sum_{i=\{s,h\}} w_{i,j} d_{i,j}(x) \tag{1}$$

where $w_{i,j}$ is a weight applied to the recognition score of the class c_j from the modality i. These weights can be:

- whether the same in the two modalities; in which case we are just calculating the mean score (for example $w_{s,j} = w_{h,j} = 0.5$). No parameters are considered in this case.
- or depending on the global performances of the two systems taken alone (global symbols recognition rates)

$$\begin{cases} w_{h,j} = \dfrac{R_h}{R_h + R_s} \\ w_{s,j} = \dfrac{R_s}{R_h + R_s} \end{cases} \tag{2}$$

where R_h and R_s are the symbol recognition rates according to handwriting and speech modalities respectively. Only two parameters are considered here.

- or even depending on the performances of the two systems taken alone (symbols recognition rates) at a class level.

$$\begin{cases} w_{h,j} = \dfrac{R_{h,j}}{R_{h,j} + R_{s,j}} \\ w_{s,j} = \dfrac{R_{s,j}}{R_{h,j} + R_{s,j}} \end{cases} \tag{3}$$

where $R_{h,j}$ and $R_{s,j}$ are the symbol recognition rates according to handwriting and speech modalities respectively for the symbol class c_j. In this case, for each class there exists two parameters.

A-2 **Borda count method[12, 13]:** this method uses the rank of a class symbol instead of its score during the fusion process. This prevent from the existing problem of score normalization within the differents systems (modalities). For example, in our case, each modality (speech and handwriting) proposes a ranked list (according to the recognition scores) of the classes for a given hypothesis. During the fusion process, each class is assigned a number which is the sum of the ranks with respect to each modality. After that, a new ranking is done according to this number. The resulting list corresponds to the fused list.

A-3 **Belief functions theory:** before presenting the usage of the belief functions theory in the context of MER, let us briefly recall some notions of this theory [14]. Its aim is to determine the belief concerning different propositions from some available information. It is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities, and the combination of such degrees of belief when they are based on independent items of evidence.
Let Ω be a finite set, called frame of discernment of the experience. The concept of belief function is the representation of the uncertainty. It is defined as a function m from 2^Ω to $[0, 1]$ with $\sum_{A \in \Omega} m(A) = 1$. This quantity $m(A)$ gives the belief that is exactly allowed to the proposition A.
We call a focal element of m every element A that satisfies $m(A) > 0$. Various combination operators are defined in literature. In this work, we focus on the most used and optimal one [15]. It is the Dempster's combination rule. For two belief functions m_1 and m_2, we obtain $\tilde{m}$ thanks to the conjunctive binary operator:

$$\forall A \in \Omega,\ \tilde{m}(A) = \sum_{B \cap C = A} m_1(B) * m_2(C). \tag{4}$$

In our experiment, the beliefs functions are deduced from the recognition scores of symbols assigned by the specialized systems (handwriting recognition and speech recognition). These scores are normalized to be in the range $[0, 1]$. For example, let us consider H_{hyp} and S_{hyp} respectively a handwriting and speech hypotheses to be combined. The recognition processes in both modalities give recognition lists (symbol label with according score). The associated masses (beliefs) can be deduced as follows:

recognized labels list (with their associated scores)		**example of associated beliefs (masses)**
$for\ S_{hyp} : \begin{cases} s(x) = 0.62 \\ s(s) = 0.1 \end{cases}$		$for\ S_{hyp} : \begin{cases} m(x) = 0.62 \\ m(s) = 0.10 \\ m(\omega) = 0.28 \end{cases}$
	$\Rightarrow$	
$for\ H_{hyp} : \begin{cases} s(n) = 0.52 \\ s(x) = 0.46 \end{cases}$		$for\ H_{hyp} : \begin{cases} m(n) = 0.52 \\ m(x) = 0.46 \\ m(\omega) = 0.02 \end{cases}$

The score $d_j(x)$ for a class $'x'$ using this formalism is then equal to $\tilde{m}(x)$ calculated using Eq.4, where m_1 and m_2 are respectively the handwriting and speech masses.

B **Fusion method at relational level (IFRL):** the fusion at relational level is done during the spatial analysis phase. The parser in charge of this task, in the handwriting modality, explores all the possible relations for each group of elementary symbols proposed by the recognition module (after IFSL). For example if we consider the case of two symbols, the relations explored including only these two symbols can be: *left/right, superscript, subscript, above, under and inside* (if one of the symbols is the *'square root'* sign for example). For each explored relation a cost is associated (calculated from the geometrical properties of the symbols involved in this relation [9]). The relation which will be considered in the ME is the one having the smallest cost and satisfying the considered grammar. The fusion at this level is done by exploring the keyword list extracted as explained in section 5.2.1. If an explored relation exists in this keywords list, its cost is decreased, otherwise it is increased. This is expressed in Eq.5.

$$\tilde{RC}(R_i) = \begin{cases} \alpha_e \times RC(R_i) & \text{if the relation } R_i \text{ is present in both modalities.} \\ \alpha_p \times RC(R_i) & \text{otherwise.} \end{cases} \quad (5)$$

where $RC(R_i)$ and $\tilde{RC}(R_i)$ are respectively the relational costs before and after fusion for the relation R_i. α_e and α_p are two parameters to set. They are used to decrease the relational cost of the relations which are also available in the speech modality ($\alpha_e < 1$), and in the contrary, increase those that are missing in the audio stream ($\alpha_p > 1$).

6. EXPERIMENTAL RESULTS AND DISCUSSION

The bimodal nature of our experiments requires the use of a database which gives each ME in both modalities (handwritten and spoken). We used the *HAMEX* database [16] which is set up for that purpose (the audio part is currently available in the French language). The handwriting recognition task we used to run our experiments is the one we participated with for the second edition of the *Competition on Recognition of On-line Handwritten Mathematical Expressions* (CROHME2012*) [17]. Since the parameters of the fusion system (*cf.* equations 2, 3, 4 and 5) are experimentally tuned, we used a data set of *500* ME from the *HAMEX* train part to do that. The results reported here are from a data set of *519* ME of the *HAMEX* test part. These ME are chosen in such a

*http://www.isical.ac.in/~crohme/

way to satisfy to the the CROHME grammar (task 2). Finally, the models of the ASR system are trained on the whole speech data of the train part of *HAMEX*. Concerning the fusion process itself, in this first experiment exploiting the complete information (labels and scores), the selection of the speech segment to combine with the handwriting group of strokes (hypotheses) is done according to the labels intersection in the top N (N is set experimentally to 3). Thus a handwriting segmentation hypothesis is combined with a speech segmentation hypothesis only if there is a common label in the 3-best recognition lists from both modalities. In other words, the handwriting segmentation is processed by the handwriting recognition system. This latter provides a list of possible labels with their scores (*cf.* section 3). The speech segmentation hypothesis undergoes the same processing (*cf.* section 4). After that, at most three labels are considered in both lists and this couple segment (speech) group (handwriting) is considered for the fusion process only if there is a common label in these two sublists. The combination is done thanks to the methods presented in section 5.

We report in figure 6, the obtained results after the fusion process (at both symbol and relation levels) compared to baseline system which is the handwriting based system (first bar).

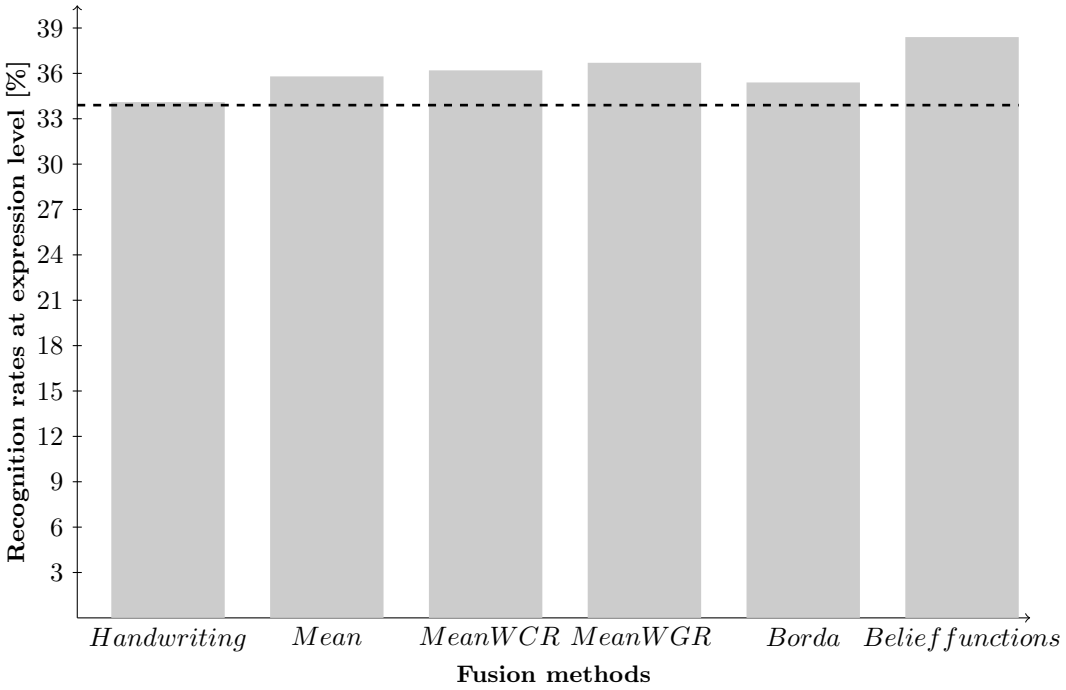

Figure 6. Recognition rates at expression level before (Handwriting) and after fusion
Handwriting: baseline system; Mean: weights are set to 0.5 in Eq.1; MeanWGR: weights from Eq.2; MeanWCR: weights from Eq.3; Borda: for using Borda count as a fusion method; Belief function: fusion thanks to belief functions Eq.4

As expected, the use of the bimodal aspect of the ME helps to improve the recognition rate at expression level. The best fusion configuration is the one based on the belief functions theory. This is consistent with the literature about the performances of such a theory dealing with the combination problem in the context of multimodal systems. The exploration of the mean weighted methods, with various weights, showed that a good weighting of the scores coming from both modalities is important, since it allows to deal with the problem of score normalization in the two modalities and take into account the performances of the systems at elementary level (symbols). The Borda count combination method, even unsensitive to the score normalization problem, does not show a good behaviour in our case since the information from the ASR system is not rich enough (often only one hypothesis per segment).
We present in figure 7 a real example of the obtained results. This example shows an expression which is only well recognized considering the belief functions aproach. In this example, the first two strokes (the red and

Figure 7. Real example of a contribution of the bimodal processing; (a) the ground-truth ME, (b) its handwritten version, (c) the recognized result without fusion, (d) the automatic transcription of its spoken description

black ones in figure 7-b) belong to the same symbol in the ground truth. However during the handwritten signal recognition, combining both of these strokes into the same symbol hypothesis leads to its missclassification. Indeed, the classifier suggests that this segmentation is not valid and assign a high score for rejection label. Beside of that, the *'x'* label is ranked second in the classifier list of labels propositions, even with a very low score. When we use the fusion process, this segmentation hypothesis is combined with the audio segment containing also the *'x'* label as a recognition hypothesis. In other words, knowing that scores are in the range $[0, 1]$, according to handwriting classifier, this segmentation hypothesis is rejected with a score of 0.84 and can be an *'x'* with a score of 0.15. Unfortunately, the belief functions fusion method aside, all the other methods do not allow to recover the right label. This is mainly due to the fact that in the audio segment also, there is a conflict between the labels *'s'* (with a score of 0.48) and *'x'* (with a score of 0.45). The belief functions based method, by modeling a part of ignorance in both modalities as presented in Eq. 4, helps to make the *'x'* label score enough high to rank it as a first hypothesis and include it during the structural analysis process.

The table 1 gives a comparison at lower levels (strokes and symbols recognition rates) between the best fusion configuration and the baseline system (handwriting based).

Table 1. Comparaison of the performances of the handwriting recognition system with the belief functions fusion based system

Evaluation level in [%]	*stroke classification rate*	*symbol recognition rate*	*expressions recognition rate with*		
			exact match	*1 error at most*	*2 errors at most*
handwriting based system	80.05	82.93	34.10	46.44	49.52
fusion based system	83.40	85.40	38.34	50.10	53.37

Table 1 shows that the benefit of the fusion process concerns both low level (strokes) and high level which is the complete ME. The results concerning the complete ME with one or two errors (either in symbol recognition or in relation interpretation), suggest that there is still scope for aditional contribution of the fusion process (38.34% of ME are completely recognized without any error and 50.10% of ME are well recognized if one error at most is allowed).

7. CONCLUSIONS AND FURTHER WORK

In this work we presented a new approach for mathematical expressions recognition. It is based on bimodal processing. The modalities involved are speech and handwriting. In this first experiment on the use of the full information from the two specialized systems, we showed the added value of a such processing. This can be seen either at expression level or in lower levels (strokes and symbols).

In a future work, we plan to improve the choice of the couple (speech hypothesis segment, handwriting hypothesis group) to be fused, by exploiting the temporal information in both modalities. The final goal is to reach the best possible synchronization between the two streams. Another interesting point to explore is the use of word lattice from the ASR system, which can provide more information for a considered speech segment. Beside of that, the context of the symbol or relation is still not used, we believe that this can improve also the accuracy of the global system.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. They also thank the French Region Pays de la Loire for funding this work under the DEPART project http://www.projet-depart.org/.

REFERENCES

[1] Karray, F., Alemzadeh, M., Saleh, J. A., and Arab, M. N., "Human-computer interaction: Overview on state of the art," *International Journal on Smart Sensing and Intelligent Systems (IJSSIS)* **1(1)**, 137–159 (2008).

[2] Chan, K. F. and Yeung, D. Y., "Mathematical expression recognition: A survey," *International Journal of Document Analysis and Recognition* **3(1)**, 3–15 (2000).

[3] Zanibbi, R. and Blostein, D., "Recognition and retrieval of mathematical expressions," *International Journal on Document Analysis and Recognition (IJDAR)* **15(4)**, 1–27 (2011).

[4] Fateman, R., "How can we speak math?," tech. rep., University of California at Berkeley (2011).

[5] Wigmore, A., Hunter, G., Pflugel, E., Denholm-Price, J., and 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)* **28(2)**, 177–189 (2009).

[6] Elliott, C. and Bilmes, J., "Computer based mathematics using continuous speech recognition," in [*[CHI 2007 Workshop on Striking a C[h]ord: Vocal Interaction in Assistive Technologies, Games, and More]*], (2007).

[7] Tapia, E. and Rojas, R., "A survey on recognition of on-line handwritten mathematical notation," tech. rep., Free University of Berlin (2007).

[8] Rhee, T. H. and Kim, J. H., "Robust recognition of handwritten mathematical expressions using search-based structure analysis," in [*[Proc. of Int. Conf. on Frontier in Handwriting Recognition (ICFHR)]*], 19 – 24 (2008).

[9] Awal, A., Mouchère, H., and Viard-Gaudin, C., "Towards handwritten mathematical expression recognition," in [*[Proc. of Int. Conference on Document Analysis and Recognition (ICDAR)]*], 1046 –1050 (2009).

[10] Deléglise, P., Estève, Y., Meignier, S., and Merlin, T., "Improvements to the lium french asr system based on cmu sphinx: what helps to significantly reduce the word error rate?," in [*[Interspeech 2009]*], (2009).

[11] "Cmu sphinx system." http://cmusphinx.sourceforge.net/html/cmusphinx.php. (Accessed on july, 18th, 2012).

[12] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S., "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems* **16(6)**, 345–379 (2010).

[13] Thiran, J. P., Marqués, F., and Bourlard, H., [*[Multimodal Signal Processing - Theory and Applications for Human-Computer Interaction]*], Elsevier (2010).

[14] Shafer, G., [*[A Mathematical Theory of Evidence]*], Princeton University Press (1976).

[15] Denoeux, T., "Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence," *Artificial Intelligence* **172(2)**, 234–264. (2008).

[16] Quiniou, S., Mouchère, H., S., P. S., Viard-Gaudin, C., Morin, E., Petitrenaud, S., and Medjkoune, S., "Hamex - a handwritten and audio dataset of mathematical expressions," in [*[Proc. of Int. Conference on Document Analysis and Recognition (ICDAR)]*], *Document Analysis and Recognition, International Conference on* , 452–456 (2011).

[17] Mouchère, H., Viard-Gaudin, C., Kim, D. H., Kim, J. H., and Garain, U., "Icfhr2012: Competition on recognition of online handwritten mathematical expressions (crohme 2012)," in [*[Proc. of Int. Conf. on Frontier in Handwriting Recognition (ICFHR)]*], (2012).

[18] Medjkoune, S., Mouchère, H., Petitrenaud, S., and Viard-Gaudin, C., "Handwritten and audio information fusion for mathematical symbol recognition," in [*[Proc. of Int. Conference on Document Analysis and Recognition (ICDAR)]*], 379–383 (2011).