HAMEX – a Handwritten and Audio Dataset of Mathematical Expressions

Solen Quiniou[†], Harold Mouchère[*], Sebastián Peña Saldarriaga[§], Christian Viard-Gaudin[*],
Emmanuel Morin[†], Simon Petitrenaud[‡] and Sofiane Medjkoune[*]
[*] *IRCCyN, Nantes, France*
Email: FirstName.LastName@univ-nantes.fr
[†] *LINA, Nantes, France*
Email: FirstName.LastName@univ-nantes.fr
[‡] *LIUM - EA 4023, Le Mans, France*
Email: simon.petit-renaud@lium.univ-lemans.fr
[§] *Synchromedia - Ecole de Technologie Supérieure, Montréal (Québec), Canada*
Email: spena@synchromedia.ca

***Abstract*—In this paper, we present HAMEX, a new public dataset, that contains mathematical expressions available in their on-line handwritten form and in their audio spoken form. We have designed this dataset so that, given a mathematical expression, its handwritten signal and its audio signal can be used jointly to design multimodal recognition systems. Here, we describe the different steps that allowed us to acquire this dataset, from the creation of the mathematical expression corpora (including expressions extracted from Wikipedia pages) to the segmentation and transcription of the collected data, via the data collection process itself. At present, the dataset contains 4 350 on-line handwritten mathematical expressions written by 58 writers, and the corresponding audio expressions (in French) spoken by 58 speakers. The ground truth is also provided for both the handwritten and audio expressions.**

***Keywords*-dataset; mathematical expressions; handwriting recognition; speech recognition; multimodality**

I. Introduction

We consider the problem of recognizing mathematical expressions using two different modalities: on-line handwriting and speech. While the problem of recognizing handwritten textual mathematical expressions has been studied before, the recognition of handwritten mathematical content and the use of audio content to assist with the recognition, however, presents a new set of challenges.

Handwriting and speech are the two most common interaction modalities for human beings. Each of them has specific features related to usability or expressiveness, and requires dedicated tools and techniques for acquisition and processing. In this respect, we are interested in the study of fusion strategies for a multimodal input system, combining on-line handwriting and speech, so that extended facilities or increased performances are achieved with respect to a single modality. The joint analysis of handwritten and spoken documents is a relatively new area of research, and only a few works have emerged concerning applications such as identity verification [1], whiteboard interaction [2], lecture note taking [3], and mathematical expression recognition [4], [5], [6]. To figure out the interest of working on both modalities, let us consider the simple handwritten mathematical expression displayed in Figure 1. The recognition

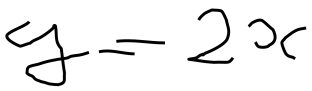


Figure 1. Example of an ambiguous handwritten mathematical expression.

of this expression is subject to various problems: the correct segmentation is not obvious, the label of the first symbol is ambiguous, and the spatial position of the last symbol is subject to different interpretations. Consequently, all the following mathematical expressions may correspond to the actual intentionof the writer: $y = 2x$, $y = 2^x$, $y - 2x$, $y = 20c$, $g = 2x$, etc. But, regarding the pronunciation of these mathematical expressions, they should be significantly different. For instance, the first expression could be pronounced “y equals two x", while the second one could be “y equals two (raised) to the power of x". To allow the exploration of such problems, the availability of databases that include both handwritten and audio signals is a requirement. This is the goal of the work presented in this paper. Indeed, we have collected a large dataset of mathematical expressions, which are all available with their on-line handwritten signal as well as their spoken audio signal. Furthermore, the handwritten and audio expressions are manually annotated with their ground truth.

The rest of this paper is organized as follows. The corpora of mathematical expressions is described in section II, while the handwritten and audio acquisition process is presented in section III. Then, section IV gives details on the handwritten and audio format of the expressions collected. Finally, some conclusions are drawn in section V.

II. Textual mathematical expression corpora

The main difficulty in building a corpus of textual mathematical expressions is to find realistic expressions from

the real world. Some approaches generate such a corpus from a grammar [7], but it supposes that the grammar used is representative of the language. Thus, the best way is to use genuine data. In our case, existing mathematical expressions are extracted from real documents from Wikipedia, which is an immense and free source of documents containing mathematical expressions. In addition to the expressions extracted from Wikipedia, we use simpler mathematical expressions generated with fewer symbols and simpler spatial relationships. Therefore, we create three corpora with different levels of complexity. Table I presents the general characteristics of these three textual corpora, whereas Table II gives details on the symbols composing each corpus vocabulary. It is worth noting that the WIKIEM-EXT vocabulary includes the WIKIEM vocabulary, which itself includes the CALCULATOR vocabulary. The creation of the corpora WIKIEM, WIKIEM-EXT, and CALCULATOR are described in greater details in the following sub-sections.

Table I
OVERVIEW OF THE TEXTUAL MATHEMATICAL CORPORA

Corpus	Number of expressions	Number of symbols	Size of the vocabulary
CALCULATOR	870	17 478	25
WIKIEM	1 740	17 020	56
WIKIEM-EXT	1 740	21 390	74

Table II
SYMBOLS IN THE VOCABULARY OF EACH CORPUS

Classes	CALCULATOR	WIKIEM	WIKIEM-EXT
Latin characters		$abcdefgi$ $knrsxyz$	$a \dots z$
Greek char.		$\alpha\beta\gamma\phi\pi\theta$	$\alpha\beta\gamma\phi\pi\theta$
Up. case char.		XY	XY
Digits	$0 \dots 9$	$0 \dots 9$	$0 \dots 9$
Operators	$+ - \pm \times / \div$	$+ - \pm \times / \div$	$+ - \pm \times / \div$
Equality op.	$= \neq < \leq > \geq$	$= \neq < \leq > \geq$	$= \neq < \leq > \geq$
Elastic op.		$\sum - \int \sqrt{}$	$\sum - \int \sqrt{}$
Set operators			$\in \forall \exists$
Functions		$\cos \sin \log$	$\cos \sin \log \lim$
Braces	()	()	()
Others	.	$. \rightarrow$	$. \rightarrow \dots \infty,$

A. *Corpus* CALCULATOR

The sub-corpus CALCULATOR consists of expresssions randomly generated, that model elementary arithmetic and comparison operations (the set of symbols used to generate this sub-corpus is given in Table II). The generation process is based on Prüfer codes [8]. A Prüfer code is a sequence $\mathbf{a} = \{a_1, a_2, \dots, a_{n-2}\}$ of $n-2$ integers representing a labeled tree of n nodes; each integer represents the label of a node, and the order in which they appear determines the connections between them. First, we generate a random Prüfer sequence corresponding to a binary tree. Then, we convert the Prüfer code into a tree using a standard algorithm. The leaf nodes in the tree are replaced by random integers, while the other nodes are replaced by binary operators. Figure 2 shows the generation process, leading to the expression $148 \div 85 < 2$.

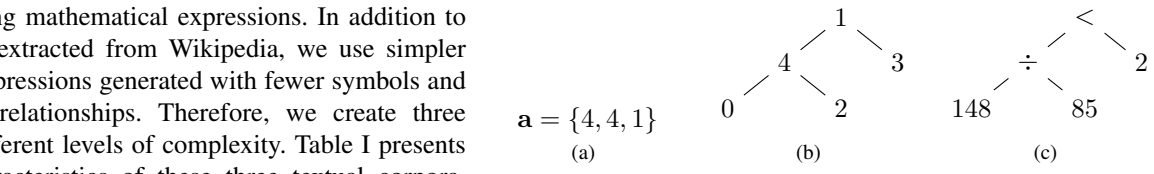


Figure 2. Generation of random expressions. (a) Prüfer code. (b) Labeled tree. (c) Expression tree.

If an equality operator is drawn, the expression is evaluated in order to make it true. Moreover, if an expression is ambiguous (e.g. $1+2\times3$), we force the operator precedence by adding parentheses: $1 + 2 \times 3$ becomes $1 + (2 \times 3)$.

B. *Corpora* WIKIEM *and* WIKIEM-EXT

To create the sub-corpora WIKIEM and WIKIEM-EXT, we use only the French version of the Wikipedia page collection (to avoid artificial repetitions of the same expressions in the duplicated pages of each language). First, we collect all the mathematical expressions in these pages by detecting the `math` tag. Terefore, we obtain 75 000 expressions composed of more than 1 100 000 symbols from more than 600 symbol classes. Then, we clean the expressions extracted in order to avoid very short ones (*e.g.* isolated symbols in plain text), rare symbol classes (only the first 220 classes have more than 100 samples), expressions just made of text (like function names), and very long expressions. Finally, we keep 59 000 expressions, the symbols of which belong to 210 different classes[1]. These expressions contain between 4 and 50 symbols (14.5 symbols on average). Moreover, each symbol class has more than 10 samples.

Nonetheless, this set of expressions is still too complex with regards to the symbol set, the variety of structures, and the size of some expressions, for instance. Thus, we define two subsets of expressions corresponding to two levels of difficulty and to two subsets of symbol classes. These two subsets correspond to the sub-corpora WIKIEM and WIKIEM-EXT (their characteristics were given in Tables I and II). Figure 3 gives examples of mathematical expressions extracted from Wikipedia and composing these two corpora.

$$\int_{-a}^{a} f(x)dx = 0 \quad \text{(a)} \qquad \lim_{t \to t_0} y(t) = +\infty \quad \text{(b)}$$

Figure 3. Examples of expressions of (a) WIKIEM. (b) WIKIEM-EXT.

[1]the special LaTeX symbols _ and ^ are counted as symbols here.

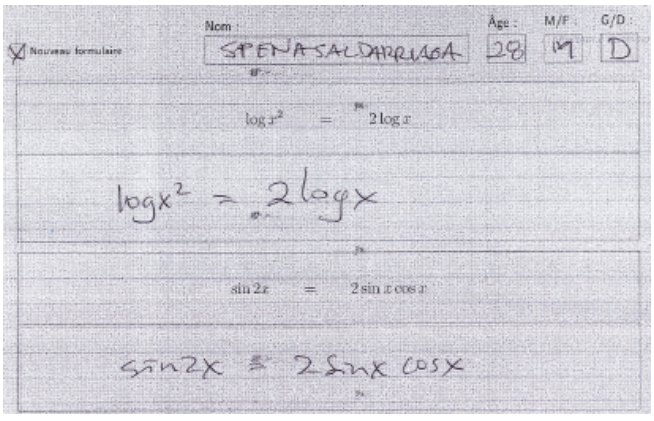
III. Data acquisition protocol

In order to collect the handwritten and audio mathematical expressions, the three corpora defined in section II are merged into one single corpus. Then, forms are automatically generated from this whole corpus to allow the expression collection from users. In the following sub-sections, we describe the handwriting and the audio acquisitions, respectively.

A. Handwriting acquisition

First, we designed a tool to automatically generate the collection forms, from the whole textual corpus (forms are produced in PDF format). Each form page contains 5 expressions (one from CALCULATOR and two from WIKIEM and two from WIKIEM-EXT), as well as input areas on which to write the expressions. A form contains 15 pages, thus corresponding to 75 expressions; one form is filled by one writer. Furthermore, on the first page of a form, the writer provides her/his name, age, sex, and handedness. Figure 4(a) shows part of the first page of a form, and Figure 4(b) shows the corresponding filled in part.

(a)

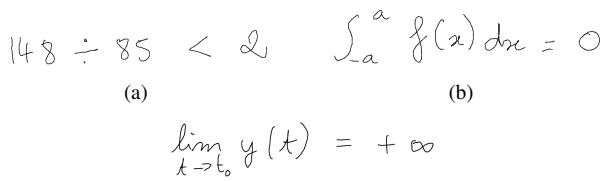


(b)

Figure 4. Example of a form page (a) Blank. (b) Filled in.

To collect the on-line handwritten mathematical expressions, a digital pen and paper solution is used. During the acquisition process, no constraint is given to the writers regarding the writing of the expressions.

At the end of the handwriting acquisition process, one INKML file (see sub-section IV-A for more details on this format) has been created for each writer (corresponding to a filled form). It contains the digital ink corresponding to the whole set of mathematical expressions she/he has written. Figure 5 gives examples of handwritten expressions of the dataset).



(a) (b)

(c)

Figure 5. Examples of handwritten mathematical expressions from (a) CALCULATOR. (b) WIKIEM. (c) WIKIEM-EXT.

B. Audio acquisition

The forms used to collect the audio mathematical expressions are created in a similar way to those for the handwriting acquisition, but without the input areas (see subsection III-A). Here, each page displays only one expression. Moreover, 74 additional pages are added to each form, with each page containing one of the vocabulary symbols (according to Table II). Thus, each speaker has to utter 75 expressions and 74 isolated symbols in French, as well as her/his name, age, gender, and mother tongue (at the beginning of the recording).

To collect the audio mathematical expressions, we chose to use a conventional recorder and an external microphone. We used a Lem DO21B[2] microphone, and a Marantz PMD 661 recorder. Recording is done in mono mode, at 48kHz. Before the recording, we ensure that the speaker knows how to pronounce each symbol. But, during recording, the speaker is free to pronounce the expressions as she/he wants, to ensure a natural diversity of the audio expressions. Examples of possible pronunciations are given in Table III, for the expression $(a + b) \times c$.

Table III
DIVERSITY OF PRONUNCIATIONS

Reference	Possible pronunciations
$(a + b) \times c$	a plus b times c
	a plus b multiplied by c
	a plus b in brackets, times c

At the end of the audio acquisition process, one audio file (in a classical WAV format) has been created for each speaker. It contains the whole set of mathematical expressions uttered by her/him, in French.

[2]http://www.lemindus.com/do21b.php

IV. HANDWRITTEN AND AUDIO MATHEMATICAL EXPRESSIONS COLLECTED

The set of mathematical expressions collected is further broken down into a training set and an evaluation set. Their characteristics are given in Table IV.

Table IV
HANDWRITTEN AND AUDIO MATHEMATICAL EXPRESSIONS COLLECTED

Base collected	Number of expressions	Total duration of audio expressions	Number of writers or speakers
Training	2 925	8 h	39
Evaluation	1 425	4 h	19

The format of the handwritten and of the audio mathematical expressions is described in the following subsections. The HAMEX dataset can be downloaded from http://www.projet-depart.org/, under the "Resources" category (after registration).

A. Handwritten data format: INKML

As stated in sub-section III-A, the mathematical expressions written by a given writer are grouped in one file according to the INKML standard from the W3C[3], which is based on an XML structure. The digital ink corresponding to each expression is then extracted and saved in a new INKML file. An INKML file mainly contains three kind of information:

- the ink: a set of traces made of points;
- the symbol level ground truth: the segmentation and label information of each symbol of the expression;
- the mathematical ground truth: the MATHML[4] structure of the expression.

The two ground truth information (at the symbol level, and the mathematical one) are entered manually. Furthermore, some general information is added in the file:

- the channels (here, X and Y);
- the writer information (identification, handedness, age, sex, etc.), if available;
- the LaTeX ground truth (without any reference to the ink, to easily render it).

The INKML format enables to make references between the digital ink of the expression, its segmentation into symbols and its MATHML representation. Listing 1 shows an example of an INKML file for the expression $a < \frac{b}{c}$, containing 5 symbols for a total number of 6 strokes (two for the 'a', and one for the other symbols).

It can be seen that the `traceGroup` with identifier xml:id="8" has references to the 2 corresponding strokes of symbol 'a', as well as to the MATHML part with identifier xml:id="A". Thus, the stroke segmentation of a symbol can be linked to its MATHML representation.

[3]http://www.w3.org/2003/InkML

[4]http://www.w3.org/1998/Math/MathML

```
<ink xmlns="http://www.w3.org/2003/InkML">
 <traceFormat>
  <channel name="X" type="decimal"/>
  <channel name="Y" type="decimal"/>
 </traceFormat>

 <annotation type="writer">w123</annotation>
 <annotation type="truth">$a&lt;\frac{b}{c}$</annotation>
 <annotationXML type="truth" encoding="Content-MathML">
  <math xmlns="http://www.w3.org/1998/Math/MathML">
   <mrow>
    <mi xml:id="A">a</mi>
    <mrow>
     <mo xml:id="B">&lt;</mo>
     <mfrac xml:id="C">
       <mi xml:id="D">b</mi>
       <mi xml:id="E">c</mi>
     </mfrac>
    </mrow>
   </mrow>
  </math>
 </annotationXML>

 <trace id="1">985 3317, ..., 1019 3340</trace>
       ...
 <trace id="6">1123 3308, ..., 1127 3365</trace>
 <traceGroup xml:id="7">
  <annotation type="truth">Ground truth</annotation>
  <traceGroup xml:id="8">
   <annotation type="truth">a</annotation>
   <annotationXML href="A"/>
   <traceView traceDataRef="1"/>
   <traceView traceDataRef="2"/>
  </traceGroup>
       ...
 </traceGroup>
</ink>
```

Listing 1. Example of an INKML file for the expression $a < \frac{b}{c}$.

B. Audio data format

As stated in sub-section III-B, the mathematical expressions uttered by a given speaker are originally available as a unique WAV audio file. Then, the file needs to be segmented into each of its expressions, and the transcriptions need to be done. The speech transcription consists in associating the corresponding spoken text to each pronounced expression or word. This corresponds to the ground truth for audio signals. Examples of some French transcriptions are given in Table V (with the corresponding English translation given in brackets).

Table V
EXAMPLES OF AUDIO TRANSCRIPTIONS

Mathematical expression	Audio transcription
α	"alpha"
$y = 2x$	"y égale deux x" (y equals two x)
$(a + b) \times c$	"a plus b fois c" (a plus b times c)

The segmentation and transcription of a whole file into its expressions is done using TRANSCRIBER[5], a tool to assist the manual annotation of speech signals. An example of an audio file with two expressions is given in Figure 6. TRANSCRIBER also enables to add some information about

[5]http://trans.sourceforge.net/en/presentation.php

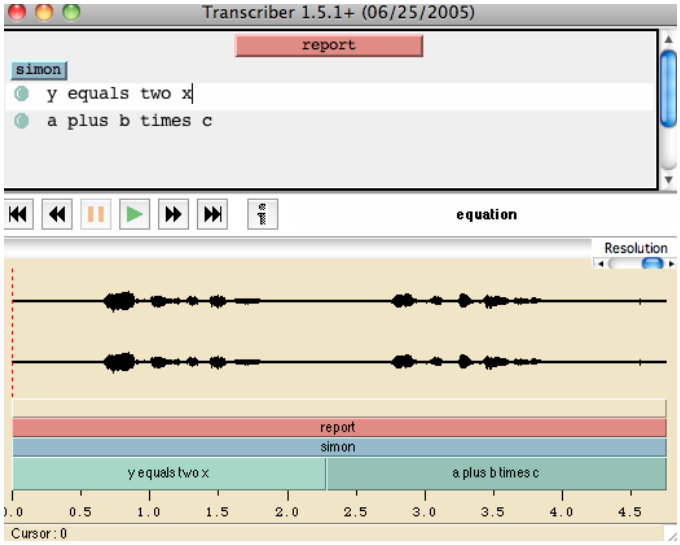


Figure 6. Example of the segmentation and transcription of an audio file with 2 expressions, using TRANSCRIBER.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Simon␣Petitrenaud"
     audio_filename="equation" version="1"
     version_date="110307">
 <Speakers>
  <Speaker id="spk1" name="simon" check="no" type="male"
       dialect="native" accent="" scope="local"/>
 </Speakers>

 <Episode>
  <Section type="report" startTime="0"
       endTime="4.76009082794">
   <Turn startTime="0" endTime="4.76009082794"
        speaker="spk1" fidelity="high" channel="studio">
    <Sync time="0"/>
    y equals two x
    <Sync time="2.285"/>
    a plus b times c
   </Turn>
  </Section>
 </Episode>
</Trans>
```

Listing 2. Example of a transcription file with 2 expressions.

the speaker (given at the beginning of the audio file): their name, gender, age, mother tongue, and also the fidelity of the pronunciation and the channel (studio here). Several steps are required to obtain the transcript of each of the mathematical expressions of a form uttered by a speaker. After opening a form audio file in TRANSCRIBER, we perform the manual segmentation of the mathematical expressions and we enter the corresponding transcription of each expression. Then, speaker information is entered, and the audio segment of each expression is saved in the WAV format. The transcriptions, along with speaker information, and segment offsets and duration, are stored in an XML file. An example of an XML file containing two expressions is given in Listing 2.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented HAMEX, a new multi-modal dataset. This dataset is freely available and contains about 4 350 mathematical expressions for the two most common interaction modalities for human beings: handwriting and speech. We have shown how this dataset has been drawn up, from the choice of the mathematical expression corpora to the segmentation and transcription of the collected data, via the data collection process itself. At the end, the handwritten and audio mathematical expressions are provided with their ground truth.

This dataset will help in bridging the evaluation gap between diverse systems dedicated to recognizing on-line handwritten mathematical expressions that use their own dataset. Moreover, this dataset is the cornerstone for systems looking to combine on-line handwritten signals and speech signals. In this way, we hope our dataset will pave the way towards the new generation of multimodal recognition systems.

ACKNOWLEDGMENT

This work is supported by the French *Région Pays de la Loire* in the context of the DEPART project http://www.projet-depart.org/. We would also like to express our sincere thanks to all those who have contributed to the data collection process.

REFERENCES

[1] A. Humm, J. Hennebert, and R. Ingold, "Combined handwriting and speech modalities for user authentication," *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 39, no. 1, pp. 25–35, 2009.

[2] K. Kurihara, M. Goto, J. Ogata, and T. Igarashi, "Speech pen: predictive handwriting based on ambient multimodal recognition," in *Proc. of CHI*, 2006, pp. 851–860.

[3] R. Anderson, C. Hoyer, P. Criag, J. Su, F. Videon, and S. Wolfman, "Speech ink and slides: the interaction on content channels," in *Proc. of ACM Multimedia*, 2006, pp. 796–803.

[4] J. Anthony, J. Yang, and K. Koedinger, "Evaluation of multimodal input for entering mathematical equation on the computer," in *Proc. of CHI*, 2005, pp. 1184–1187.

[5] S. Vemulapalli and M. H. III, "Using audio based disambiguation for improving handwritten mathematical content recognition in classroom videos," in *Proc. of DAS*, 2010.

[6] S. Medjkoune, H. Mouchere, S. Petitrenaud, and C. Viard-Gaudin, "Handwritten and audio information fusion for mathematical symbol recognition," in *Submitted to ICDAR 2011*.

[7] S. MacLean, G. Labahn, E. Lank, M. Marzouk, and D. Tausky, "Grammar-based techniques for creating ground-truthed sketch corpora," *International Journal of Document Analysis and Recognition*, vol. 14, no. 1, pp. 65–74, 2011.

[8] A. Nijenhuis and H. S. Wilf, *Combinatorial algorithms for computers and calculators*, 2nd ed. New York: Academic Press, 1978.