HAL
open science

Multimodal Mathematical Expressions Recognition: Case of Speech and Handwriting

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

To cite this version:

Sofiane Medjkoune, Harold Mouchère, Simon Petitrenaud, Christian Viard-Gaudin. Multimodal Mathematical Expressions Recognition: Case of Speech and Handwriting. 15th International Conference on Human-Computer Interaction, Jul 2013, Las Vegas, United States. pp.77-86, 10.1007/978-3-642-39330-3_9 . hal-00852846

HAL Id: hal-00852846

https://hal.science/hal-00852846v1

Submitted on 21 Aug 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimodal Mathematical Expressions Recognition: Case of Speech and Handwriting

Sofiane Medjkoune[1,2], Harold Mouchere[1],
Simon Petitrenaud[2], and Christian Viard-Gaudin[1]

[1] LUNAM University, University of Nantes, IRCCyN, France
firstname.lastname@univ-nantes.fr
[2] LUNAM University, University of Le Mans, LIUM, France
simon.petit-renaud@lium.univ-lemans.fr

Abstract. In this work, we propose to combine two modalities, handwriting and speech, to build a mathematical expression recognition system. Based on two sub-systems which process each modality, we explore various fusion methods to resolve ambiguities which naturally occur independently. The results that are reported on the HAMEX bimodal database show an improvement with respect to a mono-modal based system.

Keywords: Multimodality, graphical languages, data fusion, handwriting, speech.

1 Introduction and Motivation

Speech and handwriting are very common interaction modalities between humans. With the advances of new devices and of robust recognition algorithms it is possible to extend the usage of such input modalities to Human Computer Interaction (HCI) [1, 2]. In this work, from one hand, we are considering online handwriting produced by interfaces like touch-screens, interactive whiteboards or electronic pens. On the other hand, we suppose also available a speech signal which records the corresponding information as uttered by a speaker. In this regard, the same information is supposed to be available but with the intrinsic capabilities and limitations of each of these two modalities. To take advantage of these two information sources, we have experimented a case where naturally each of them conveys differently the processed information knowledge. The case studies proposed in this work concern the problem of Mathematical Expression (ME) recognition. Of course, some existing tools allow entering MEs in a document. Some are very powerful, as the LaTex language, but they require a high level of expertise. More interactive tools are also available such as the Mathtype equation editor, but, they still suffer from a cumbersome sequence of selections which often delays the ME production. From these observations, it is clear that a more direct way of inputting MEs would be very beneficial. However, this problem is more difficult than text recognition for several reasons. First of all, the mathematical language is composed of a large set of symbols. To cover correctly various domains of sciences, several hundreds of symbols are required. This will

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 77–86, 2013.
© Springer-Verlag Berlin Heidelberg 2013

introduce more confusion between symbols. Second and even more important point, the mathematical language is not a one dimensional (1D) language. Indeed, it is not a left-right sequence of symbols, but a two-dimensional layout where the spatial relations play an important role in the meaning of the expression. The extraction of the layout will be even more difficult from the audio signal, since a spoken language is not specifically adapted to put in plain words spatial relationships.

As Fig.1 shows, speech and handwriting based systems do not have the same drawbacks. Errors committed by each of the two systems may be corrected by the use of the other. So, better performance can be expected by proposing a speech-handwriting system for ME recognition (MER).

The paper is organized in four sections, as follows. In section 2, we describe the global system, by highlighting its main modules. In section 3, we focus on the fusion part. Section 4 is devoted to the experiments: first we check the complementarity of both modalities on isolated mathematical symbols, and then to complete ME. In the last section, we conclude the paper.

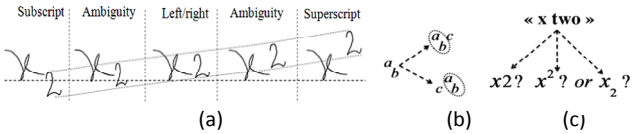

Fig. 1. Some drawbacks that mono-modality based systems encounter, due to the (a) fuzziness nature of the relationships; (b) role of the symbol according to the context; (c) ambiguity of the speech description.

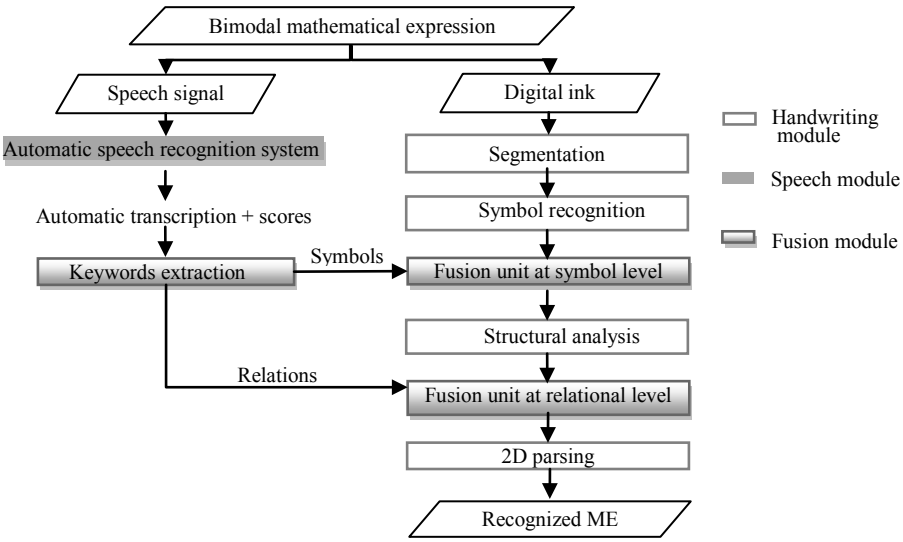

Fig. 2. The collaborative architecture for bimodal mathematical expression recognition

2 Global Overview of the Proposed Method

We propose in this work a combined system composed of two specialized ones: an online handwritten ME system and a speech recognition one. The system in charge of handwritten MER receives as input a set of elementary strokes, and gives a formatted ME as an output. Concerning the system in charge of the audio signal processing, it takes as input the audio signal and provides as a result an automatic transcription which is a textual description of the ME as uttered by the speaker. The information coming from both modalities are merged through the fusion module. This module uses the textual description issued from the speech module and extracts two kinds of information that will be supplied to the handwriting module. The combination process is done using classical data fusion techniques [3] as in Fig. 2. These three modules (handwriting, speech and fusion) will be briefly presented in the remaining of this section. The next section reports a deeper presentation of the main module in this work: the fusion module.

2.1 The Handwriting Recognition Module

The handwriting module has to make the complete interpretation of the handwritten signal and propose the final interpreted version of the ME. This is mainly done at two levels: symbol identification (segmentation and recognition) and relationships discovering (through which the identified symbols are spatially arranged). Thus, recognizing a handwritten ME includes three sequential but interdependent steps [5, 6]: segmentation, symbol recognition and spatial relations interpretation. The aim of the segmentation process is to form the symbol hypotheses *"h_s"* from the set of strokes. Each *"h_s"* has to be labeled; this is the role of the recognition stage, where a list of the most probable symbols with confidence scores is assigned to *"h_s"*. The structural analysis of the global layout including the identified symbol hypotheses is the third step. Finally, the results of these three steps are used to deduce the final ME layout thanks to a bi-dimensional grammatical parsing. Optimizing separately each step has a major drawback since the failure of one step can lead to the failure of the next one. To alleviate this problem, the simultaneous optimization of the segmentation and recognition steps is reported in various works as in [6, 7]. The handwritten MER subsystem used in the architecture of Fig. 2 is largely based on Awal and al.'s system [6].

2.2 The Speech Recognition Module

Using speech for mathematical expression recognition is usually done by means of two successive processes [7, 8]. The first one is a classical automatic speech recognition (ASR) system which provides a textual description of the ME according to the speech describing the ME. The second one is a syntactical-grammatical one. It analyzes the text given by the ASR (1D) to deduce the corresponding ME written in a mathematical language (2D). Thus, even if there are no automatic transcription errors, the relative (un)-clarity of the description might result in ambiguous interpretations. Furthermore, even if both ASR system and speaker are hundred percent accurate, the

bidimensional aspect of the ME is hard to retrieve (*ref.* Fig.1). In the rare existing systems [8, 9], the parsing (1D to 2D) is most of the time assisted by either introducing some dictation rules or using an additional source of information (such as using a mouse to point the position where to place the different elements). This makes the editing process less natural and far from what is expected from this kind of systems.

In our framework, the speech module role is limited to the task of automatic speech transcription providing the textual description. This textual description is then used within the fusion. This ASR task is carried out by a system based on the one developed at the LIUM [10], which is based on the CMU-Sphinx transcription system [11].

2.3 The Fusion Module

The fusion module ensures the connection between the specialized systems (handwriting and speech modules) in order to benefit from the existing complementarity between both modalities. This module is the main contribution of the current work and it is inspired from the data fusion field. Let us first present in the following section this concept and after discuss its use for our purpose: automatic MER.

3 The Fusion Module Description

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. This makes the conversation 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. Generally, data fusion methods are divided in three main categories [3, 4]: 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. Rules based approaches represent the first category and include methods using simple operators such as max, (weighted) mean or product. The second category is based on classification techniques and the last one is based on estimation.

In order to accomplish its task (combination of the information coming from both modalities for MER), the fusion module uses the textual description given by the ASR system to assist the handwriting module at two levels: symbol and relation. This why this module can be broken down into three distinct parts: **the keyword extraction unit, the fusion unit at symbol level and finally the fusion unit at relational level.** Since the signals coming from both modalities are heterogeneous and with the objective of using suitable recognition techniques for each modality, we chose to use a late fusion strategy. We give in the following the complete description of each unit.

3.1 Keyword Extraction Unit

The purpose of this unit is to analyze the text describing the ME provided by the ASR 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', 'two', 'parentheses'*), either relations (*'subscript', 'over'*) or both (*'integral', 'square root'*). The second category of words includes all the other words. These words are only used to make sense from the 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. For example if the word *'squared'* exists in the transcription, the ME under process could contain the symbol *'2'* and the relation *'superscript'*. This dictionary is used within the fusion units to identify the included symbols/relations in the current ME according to speech.

3.2 Fusion Unit at Symbol Level

Besides of considering the late fusion strategy, in this work, we explored some rule based methods to perform the fusion at symbol level. Let us define some notations that we will use to describe the fusion methods we explored. Let us denote by C = $\{c_1 \ldots c_N\}$, the set of the N possible symbol classes we consider. If an hypothesis *'x'* has to be classified with respect to the modality *'i'* ($i \in \{s,h\}$, where *'s'* represents speech and *'h'* is for handwriting), let us define the score of this symbol class *'c_j'* assigned to this hypothesis as: $d_{i,j}(x)$. The decision score after fusion is denoted as $d_j(x)$. Now, we focus on the methods used to obtain the scores after fusion.

1. **Weighted summation:** in this case, the fusion score $d_j(x)$ is given by equation 1.

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

where the $w_{i,j}$ are some weights that can be defined in several ways. These weights can be the same for both modalities (simple **mean**) if we trust in the same way both modalities: $w_{h,j}=w_{s,j}=0.5, \forall j$. They can depend on the global performances (**meanWGR**), using the global recognition rates R_h and R_s with respect to each modality: $w_{h,j}= R_h/(R_h + R_s)$; $w_{s,j}= R_s/(R_h + R_s), \forall j$. They can also be related to the local performances (**meanWCR**) using the class recognition rates $R_{h,j}$ and $R_{s,j}$ with respect to each modality: $w_{h,j}= R_{hj}/(R_{hj} + R_{sj})$; $w_{s,j}= R_{sj}/(R_{hj} + R_{sj}), \forall j$.

2. **Belief functions based fusion (Belief F):** the belief functions theory aim's is to determine the belief concerning different propositions from some available information [12, 13]. It is based on two ideas: 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. Various combination operators are defined in literature. In this work, we focus on the most used and optimal one [13]. It is the Dempster's combination rule. For two belief functions m_1 and m_2, we obtain $\tilde{m}$ using the conjunctive binary operator:

$$\forall A \in \Omega, \tilde{m}(A) = \sum_{B\cap C=A} m_1(B)\, m_2(C) \tag{2}$$

In our experiment, the belief functions are deduced from the recognition scores of symbols assigned by the specialized systems. 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 combine. The recognition processes in both modalities give recognition lists (symbol label and score *s*). The associated masses (beliefs) can be:

Recognized labels list (with scores)

$$\text{for } S_{hyp} = \begin{cases} s(x) = 0.62 \\ s(s) = 0.10 \end{cases}$$

$$\text{for } H_{hyp} = \begin{cases} s(n) = 0.52 \\ s(x) = 0.46 \end{cases}$$

$\Longrightarrow$

Example of associated beliefs (masses)

$$\text{for } S_{hyp} = \begin{cases} m(x) = 0.62 \\ m(s) = 0.10 \\ m(\Omega) = 0.28 \end{cases}$$

$$\text{for } H_{hyp} = \begin{cases} m(n) = 0.52 \\ m(x) = 0.46 \\ m(\Omega) = 0.02 \end{cases}$$

The score $d_j(x)$ for a hypothesis *'x'* to be the class *'j'* is then equals to $\widetilde{m}$(x) obtained from equation 2 in which m_1 represent handwriting masses and m_2 speech masses.

3. Fusion classification based: a support vector machine classifier (SVM) with a Gaussian kernel is used to perform this task. We use the scores from each of the up-stream systems as input features of an SVM classifier. Thus, this classifier knows the two score lists provided by each independent specialized classifier (2xN features) and computes a new score to every classes (N outputs).

3.3 Fusion Unit at Relational Level

At relational level, the fusion 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 symbol hypotheses proposed by the symbol recognition module. 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*. For each explored relation a cost is associated [6]. 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 extracted keyword list. If an explored relation exists in this keywords list, its cost is decreased, otherwise it is increased. This is expressed in equation 3, whith $RC(R_i)$ and $RC_{new}(R_i)$ are respectively the relational costs before and after fusion for the relation R_i and α_e ($\alpha_e<1$) and α_p ($\alpha_p>1$) are respectively parameters to enhance relations present in both modalities and penalize those missing in the speech modality:

$$RC_{new}(R_i) = \begin{cases} \alpha_e \, RC(R_i) & \textit{if the relation } R_i \textit{ is in the keyword list} \\ \alpha_p \, RC(R_i) & \textit{otherwise} \end{cases} \quad (3)$$

4 Results and Discussions

In this section we present two kinds of results. The first one is to validate the hypothesis of the existing complementarity between speech and handwriting modalities.

We propose a first experiment considering only the recognition of isolated symbols. In this case, in both modalities, the symbols are already segmented and no relations between symbols exist. Thus, this experiment is a very simplified scenario which does not completely match with the real-life application. Then a more complete experiment is presented including all the steps of the MER process.

4.1 Database Description

The data used to perform the experiment is from the HAMEX [14] database. This database includes a set of approximately 4350 ME, each of them available in the spoken and the handwritten modalities. The vocabulary covered by HAMEX contains 74 mathematical symbols, including all the Latin alphabet letters, the ten digits, six letters from the Greek alphabet and various mathematical symbols (integral, summation…).

4.2 Case of Isolated Mathematical Symbol (Gain at Symbol Recognition Level)

The on-line handwriting recognition is performed by the symbol recognizer used in the global MER system of Fig.2. It is globally based on a Time Delay Neural Network (TDNN) classifier [6]. The corresponding output is a list of Nbest classes with their scores which are normalized in the range [0, 1]. The isolated spoken words recognition is performed using a system based on MFCC coefficients and template matching using a DTW algorithm [15]. Here again, a list of most probable symbols with scores in the range [0, 1] is given.

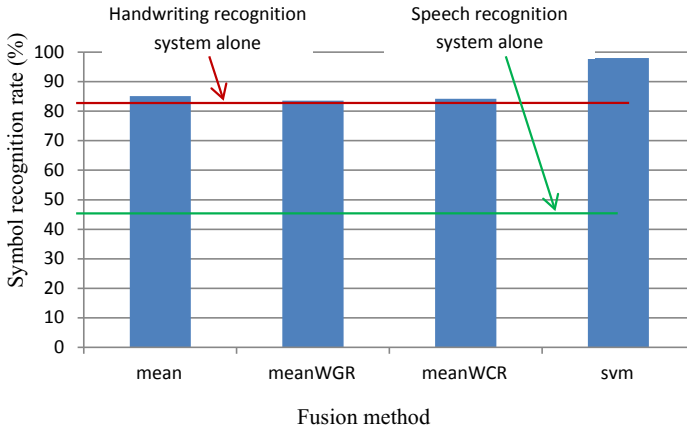

Fig. 3. Recognition rates before and after fusion at symbol recognition level

As we can see on Fig.3, the bimodal based recognition outperforms the mono-modality based systems regardless of the used fusion method. The classification based approach appears to be the best fusion method (recognition rate of 98.04% against the highest recognition rate in the mono modality mode, 81.55%). This classifier takes

clearly advantage of the strengths and weaknesses of each individual classifier because of its training stage, while other combination methods are simpler, they rely on more heuristic functions.

4.3 Case of Complete Mathematical Expression

In the case of a complete ME, the architecture in Fig.2 is used. The fusion process, here, is more complicated (cf. section 3.3). The handwriting recognition task is accomplished with the online handwritten MER system we participated with for CROHME2012[1] competition [16]. A set of 500 ME from the HAMEX train part is used to tune the fusion parameters (cf. equations 1, 2 and 3). The results reported here concern a set of 519 ME of the HAMEX test part selected in such a way to satisfy to the CROHME grammar (task 2). Finally, the models of the ASR system are trained on the whole speech data of the HAMEX train part. Concerning the fusion process itself, the selection of the speech segment to combine with the handwriting group of strokes 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 a common label in the 3best recognition lists from both modalities exists. Since at the moment of running this first experiment the alignment at the ground truth level of the handwriting and speech streams is still not available, the classification based fusion is not explored. We report on Fig.4, the recognition rates at the expression level for the different fusion methods explored.

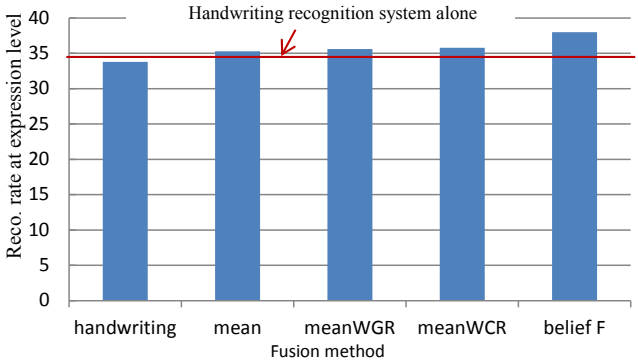

Fig. 4. Recognition rates at expression level before and after fusion

Similarly to the case of isolated symbols recognition, the fusion process improves the performances compared to a purely mono-modality based system. The recognition rates at expression level show that whatever the fusion strategy used, the performance is better. The best fusion configuration is the one based on the belief function theory. The exploration of the various mean weighted methods showed that a good weighting of the scores coming from both modalities is important, since it allows dealing with

[1] http://www.isical.ac.in/~crohme/

the problem of score normalization in both modalities. This support the hypothesis that using a classification based approach can fix the score normalization problem.

Deeper analysis comparing the best fusion and the handwriting systems, reported in table 1, show the fusion gain at lower levels (segmentation and recognition).

Table 1. Performances comparison of handwriting and belief functions fusion based systems

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

The improvement brought by the fusion process concerns both low (strokes and symbols) and high (complete expression) levels. Another important remark is that when allowing only one error (symbol or relation), we gain around 30% of ME (from 38.34% to 50.10%); this suggests that there is still scope for additional contribution of the fusion process, especially by exploring classification fusion methods. We give in Fig.5 a real example of results, where the handwriting system fails to provide the right solution when the fusion one, thanks to this bimodal processing, succeeds on this task.

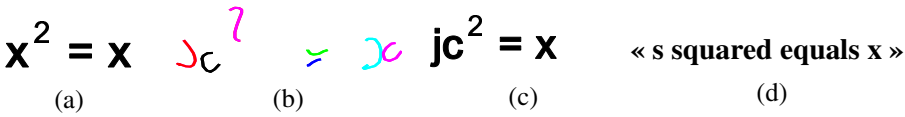

Fig. 5. Real example of a contribution of the bimodal processing(misrecognized in handwriting and recognized in fusion); (a) ME ground-truth, (b) its handwritten version, (c) the recognized result without fusion, (d) the automatic transcription of its spoken description

In this example, the first two strokes (going from the left, in Fig.5-b) should belong to the same symbol. However during the handwritten recognition, combining both of these strokes into the same symbol hypothesis leads to its misclassification. Indeed, the classifier suggests that this segmentation is not valid and assign a high score for rejection label 0.84 and answers, in a second rank, that it can be an *'x'* with a score of 0.15. When fusing, this segmentation hypothesis is combined with the audio segment containing also an 'x' label as a recognition hypothesis. Unfortunately, apart from the belief functions fusion method, 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 classes 's' (0.48) and 'x' (0.45). The belief functions method, by modeling a part of ignorance (equation 2), makes the 'x' label score high enough to rank it as a first hypothesis and to include it during the structural analysis process.

5 Conclusion and Future Work

After a first experiment on isolated symbols recognition to prove the existing complementarity between speech and handwriting, we proposed a new architecture for complete MER based on bimodal processing. The obtained results are quiet satisfying since the performances are improved compared to a mono-modal system.

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 the 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 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.: Human-Computer Interaction: Overview on State of the Art. IJSSIS, 137–159 (2008)
2. Jaimes, L., Sebe, N.: Multimodal human computer interaction: A survey. Computer Vision and Image Understanding 108, 116–134 (2007)
3. Thiran, J.-P., Marquès, F., Bourlard, H.: Multimodal Signal Processing - Theory and Applications for Human-Computer Interaction. Elsevier (2010)
4. Atrey, P.K., Hossain, M., AEl Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems, 345–379 (2010)
5. Zanibbi, R., Blostein, D.: Recognition and retrieval of mathematical expressions. IJDAR 15, 331–357 (2012)
6. Awal, A.-M., Mouchère, H., Viard-Gaudin, C.: A global learning approach for an online handwritten mathematical expression recognition system. In: PRL, pp. 1046–1050 (2012)
7. Rhee, T.H., Kim, J.H.: Robust recognition of handwritten mathematical expressions using search-based structure analysis. In: ICFHR, pp. 19–24 (2008)
8. Fateman, R.: How can we speak math? University of California, Tech. report (2012)
9. 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. JCMST 28, 177–189 (2009)
10. Deléglise, P., Estève, Y., Meignier, S., Merlin, T.: Improvements to the LIUM French ASR system based on CMU Sphinx: what helps tosignificantly reduce the word error rate? Interspeech (2009)
11. Cmu sphinx system, `http://cmusphinx.sourceforge.net/html/cmusphinx.php`
12. Denoeux, T.: Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. AI 172, 234–264 (2007)
13. Smets, P., Kennes, R.: The transferable belief model. AI 66, 191–234 (1994)
14. Quiniou, S., Mouchere, H., Peña Saldarriaga, S., Viard-Gaudin, C., Morin, E., Petitrenaud, S., Medjkoune, S.: HAMEX – a Handwritten and Audio Dataset of Mathematical Expressions. In: ICDAR, pp. 452–456 (2011)
15. Muda, L., Begam, M., Elamvazuthi, I.: Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic Time Warping. TJC 2 (2010)
16. Mouchère, H., Viard-Gaudin, C., Kim, D.H., Kim, J.H., Garain, U.: ICFHR2012: Competition on recognition of online handwritten mathematical expressions (crohme 2012). In: ICFHR (2012)