Multimodal Recognition: Handwriting & Speech

Emilio Granell Romero
PhD program in computer science

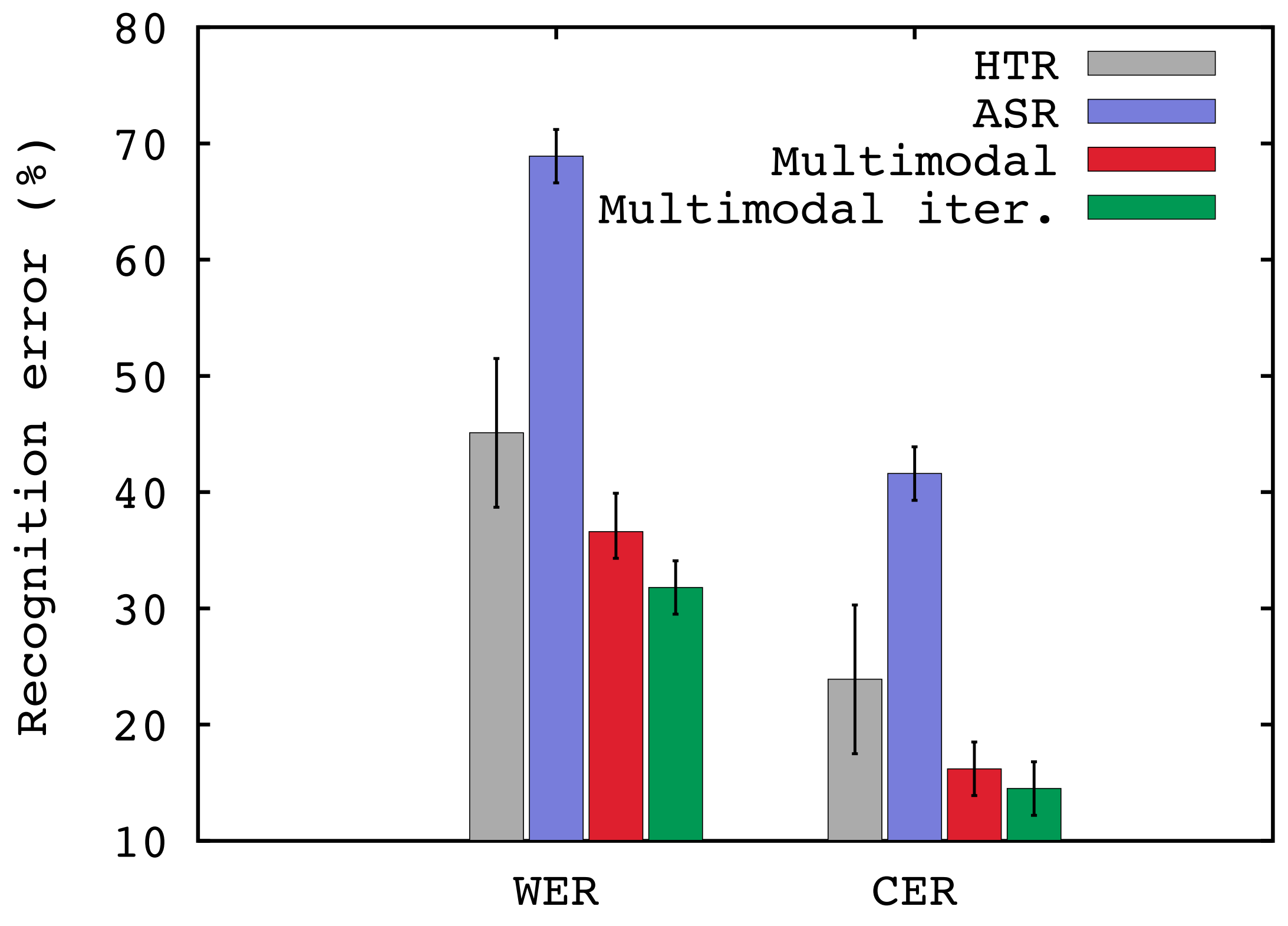
Introduction

To access the information contained in digital historical text documents its transcription is necessary. Transcriptions can be achieved by using Handwritten Text Recognition (HTR) on digitalized pages or using Automatic Speech Recognition (ASR) on the dictation of the contents.

In this work, we will check the effectiveness of a third option, that is using both systems in a multimodal combination.

Expected results

Preliminary results show that our multimodal system improves the HTR baseline despite the huge error in the ASR baseline. However, as we can see to get a significant improvement it is necessary to perform an iterative multimodal recognition.



We expect to obtain highly significant results in the non-iterative multimodal system using robust modeling techniques along with new techniques of multimodal interaction and combination.

Main references

- V. Alabau, C.D. Martínez-Hinarejos, V. Romero and A.L. Lagarda, "An iterative multimodal framework for the transcription of handwritten historical documents", Pattern Recognition Letters, vol. 35, pp. 195-203, 2014.
- V. Alabau, V. Romero, A.L. Lagarda and C.D. Martínez-Hinarejos, "A Multimodal Approach to Dictation of Handwritten Historical Documents", in *Proceedings of Interspeech 2011*, pp. 2245-2248, Florence, Italy, August 27-31, 2011.

Aim and specific objectives

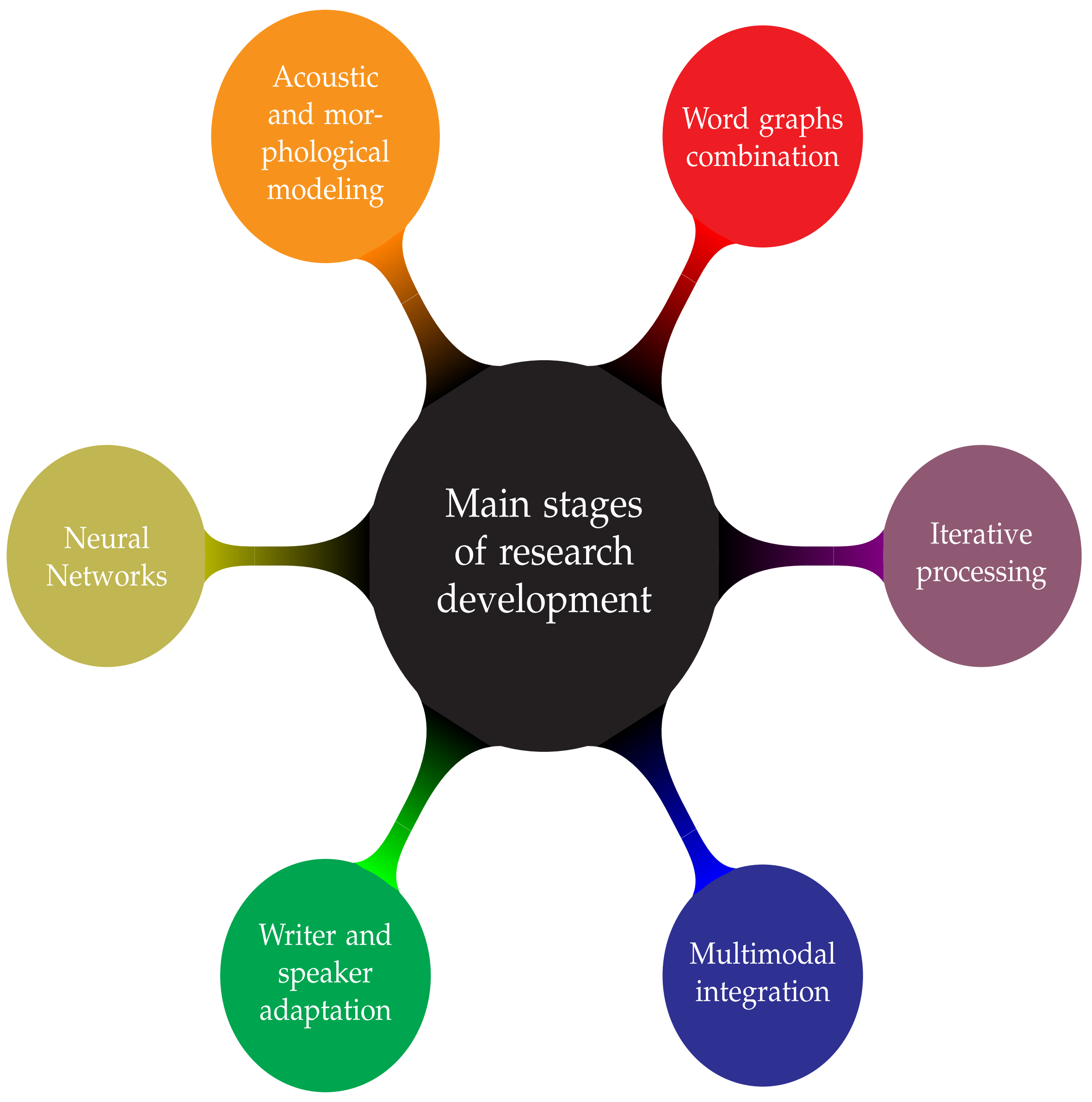
Aim

To reduce the error in the recognition of handwritten text at both word (WER) and character (CER) level.

Objectives

1. To study the techniques of morphological and acoustic modeling.
2. To study the use of Recurrent Neural Networks (RNN) on the morphological and acoustic modeling with Long Short-Term Memory (LSTM) features.
3. To study the techniques of writer and speaker adaptation.
4. To study the iterative and non-iterative multimodal interaction.
5. To study the combination of word graphs.

Handwritten Text & Automatic Speech Recognition


Main stages of research development

Acoustic and morphological modeling

Word graphs combination

Iterative processing

Multimodal integration

Writer and speaker adaptation

Neural Networks

The HTR and ASR problems are formulated in a very similar way that allows integration into a multimodal system. The unimodal formulation is: given a handwritten text image or a speech signal encoded into the feature vector sequence $\vec{x} = (x_1, x_2, \ldots, x_T)$, finding the most likely word sequence $\vec{w} = (w_1 w_2 \ldots w_l)$, that is:

$$\hat{\vec{w}} = \arg\max_{\vec{w}} \Pr(\vec{w}|\vec{x})$$

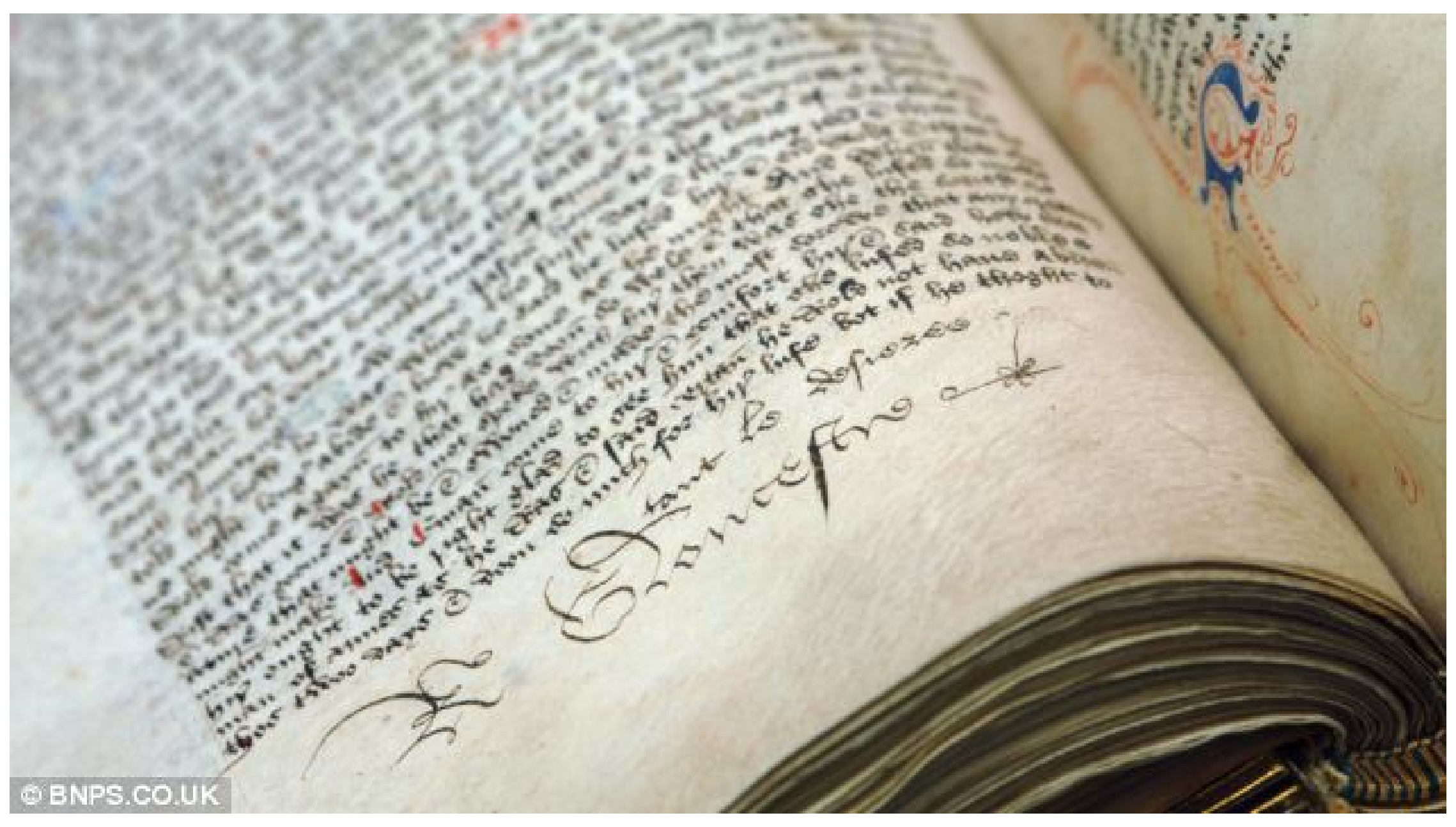
$$\hat{\vec{w}} = \arg\max_{\vec{w}} \Pr(\vec{x}|\vec{w}) \Pr(\vec{w})$$

Where, $\Pr(\vec{x}|\vec{w})$ is the morphologic or acoustic model and $\Pr(\vec{w})$ is the language model.

Possible uses

The results of this research could be used to create assistance systems that combine handwriting with speech. Some utilities might be:

1. As assistance in transcription of historical text documents.



Richard III book
`http://www.dailymail.co.uk`

2. As a complement during the drafting of documents with graphics tablets or touch screens.

Intuos pen
`http://www.wacom.com`

Contact info

Author:
Emilio Granell Romero
Email: *egranell@dsic.upv.es*
Phone: 963877000 (Ext. 73565)

Supervisor:
Dr. Carlos D. Martínez Hinarejos
Email: *cmartine@dsic.upv.es*
Phone: 963877000 (Ext. 73529)

Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València
Camino de Vera, s/n, 46022, Valencia, Spain