IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Advances on the Transcription of Historical Manuscripts based on Multimodality, Interactivity and Crowdsourcing

Emilio Granell, Carlos-D. Martínez-Hinarejos, Verónica Romero

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València,
Camí de Vera s/n, 46022, València, Spain
{egranell,cmartine,vromero}@dsic.upv.es

Abstract

The transcription of digitalised documents is useful to ease the digital access to their contents. Natural language technologies, such as Automatic Speech Recognition (ASR) for speech audio signals and Handwritten Text Recognition (HTR) for text images, have become common tools for assisting transcribers, by providing a draft transcription from the digital document that they may amend. This draft is useful when it presents an error rate low enough to make the amending process more comfortable than a complete transcription from scratch.

The work described in this thesis is focused on the improvement of the transcription offered by an HTR system from three scenarios: multimodality, interactivity and crowdsourcing.

The image transcription can be obtained by dictating their textual contents to an ASR system. Besides, when both sources of information (image and speech) are available, a multimodal combination is possible, and this can be used to provide assistive systems with additional sources of information. Moreover, speech dictation can be used in a multimodal crowdsourcing platform, where collaborators may provide their speech by using mobile devices.

Different solutions for each scenario were tested on two Spanish historical manuscripts, obtaining statistically significant improvements

Index Terms: handwritten text recognition, automatic speech recognition, multimodality, combination, interactivity, crowdsourcing

1. Introduction and Motivation

Transcription of digitised historical documents is an interesting task for libraries in order to provide efficient information access to the contents of these documents. The transcription process is done by experts on ancient and historical handwriting called paleographers.

In the latest years, the use of off-line Handwritten Text Recognition (off-line HTR) systems [1] has allowed to speed up the manual transcription process. HTR systems are composed of modules and employ models similar to those of classical speech recognition systems. However, state-of-the-art off-line HTR systems [2] are far from being perfect, and human supervision is required to really produce a transcription of standard quality. The initial result of automatic recognition may make the paleographer task easier, since they are able to perform corrections on a good draft transcription.

In addition to using off-line HTR systems from text line images, other modalities of natural language recognition can be used to help paleographers on the transcription process, such as Automatic Speech Recognition (ASR) [3] from the dictation of the contents, or on-line HTR [4] from touchscreen pen strokes. In this context, a multimodal interactive assistive scenario [5], where the assistive system and the paleographer cooperate to generate the perfect transcription, would reduce the time and the human effort required for obtaining the final result.

The use of multimodal collaborative transcription applications (crowdsourcing) [6], where collaborators can employ speech dictation of text lines as a transcription source from their mobile devices, allows for a wider range of population where volunteers can be recruited, producing a powerful tool for massive transcription at a relatively low cost, since the supervision effort of paleographers may be dramatically reduced.

In this thesis [1] [7], the reduction of the required human effort for obtaining the actual transcription of digitalised historical manuscripts is studied in the following scenarios:

- **Multimodality:** An initial draft transcription of a handwritten text line image can be obtained by using an off-line HTR system. An alternative for obtaining this draft transcription is to dictate the contents of the text line image to an ASR system. Furthermore, when both sources (image and speech) are available, a multimodal combination is possible, and an iterative process can be used in order to refine the draft transcription. Multimodal combination can be used in interactive transcription systems for combining different sources of information at the system input (such as off-line HTR and ASR), as well as to incorporate the user feedback (on-line HTR). At the same time, the multimodal and iterative combination process can be used to improve the initial off-line HTR draft transcription by using the ASR contribution of different speakers in a collaborative scenario.

- **Interactivity:** The use of assistive technologies in the transcription process reduces the time and human effort required for obtaining the actual transcription. The assistive transcription system proposes a hypothesis, usually derived from a recognition process of the handwritten text line image. Then, the paleographer reads it and produces a feedback signal (first error correction, dictation, etc.), and the system uses it to provide an alternative hypothesis, starting a new cycle. This process is repeated until a perfect transcription is obtained. Multimodality can be incorporated to the assistive transcription system, in order to improve the human-computer interaction and to provide the system with additional sources of information.

[1] Publicly available in the UPV institutional repository: http://hdl.handle.net/10251/86137

10.21437/IberSPEECH.2018-35

- **Crowdsourcing:** Open distributed collaboration to obtain initial transcriptions is another option for improving the draft transcription to be amended by the paleographer. However, current transcription crowdsourcing platforms are mainly limited to the use of non-mobile devices, since the use of keyboards in mobile devices is not friendly enough for most users. An alternative, is the use of speech dictation of handwritten text lines as a transcription source in a crowdsourcing platform where collaborators may provide their speech by using their own mobile device. Multimodal combination allows the improvement of the initial handwritten text recognition hypothesis by using the contribution of speech recognition from several speakers, providing as a final result a better draft transcription to be amended by a paleographer with less effort. In this framework, since collaborators are usually a scarce resource, their acquisition effort should be optimised with respect to the quality of the draft transcriptions.

The rest of this paper is structured as follows: Section 2 offers the main scientific and technological goals; Section 3 summarises the contents of this thesis; Section 4 contains the main conclusions; Section 5 draws the current work derived from this thesis and the future work lines; Finally, Section 6 presents the achievements, and the scientific contributions.

2. Scientific and Technological Goals

The main scientific and technological goals of this thesis are the following:

- To study the unimodal and multimodal combination techniques, in order to propose a new multimodal combination technique for improving the transcription of digitalised historical manuscripts by using the speech dictation of their contents.
- To study the use of multimodal combination techniques in a computer assisted system to improve the computer-human interaction and to accelerate the interactive transcription process.
- To develop a multimodal crowdsourcing platform based on the studied multimodal combination techniques to ease and widespread the transcription of digitalised historical manuscripts.

3. Thesis Overview

The thesis document [7] is structured in five parts to facilitate the reading experience. It starts with a first introductory part, followed by a part for each one of the three studied scenarios, and it finishes with a part which presents the general conclusions and future work lines. This section presents an overview of the contents of the three central parts (multimodality, interactivity, and crowdsourcing).

3.1. Multimodality

The integration of knowledge given by off-line HTR and ASR processes presents two limitations: both signals are asynchronous and each modality uses different basic linguistic units (usually, characters for off-line HTR and phonemes for ASR). An initial approach for solving this limitation was proposed in previous works [8, 9], where the output of the recognition process of one modality, in form of word-graph lattice, is used to modify the general language model in order to make more likely the decoded sentences; this modified language model is employed in the decoding for the other modality. This procedure can be used iteratively. This approach presents a few drawbacks: there is not a single hypothesis given that each modality provides its own, and it is not known beforehand which one is more accurate, and the initial modality must be chosen arbitrarily.

Chapter 3 of the thesis (*Combining Handwriting and Speech*) presents a new proposal based on the use of Confusion Networks for obtaining a single hypothesis from the combination of the hypotheses obtained from an off-line HTR and an ASR recognisers for decoding a text line image and the dictation of its contents. In the next chapter (Chapter 4), our multimodal proposal is tested and compared with other combination methods.

The experiments were performed on two different Spanish historical manuscripts. *Cristo Salvador*, which is a single writer book from the 19th century provided by *Biblioteca Valenciana Digital*, and *Rodrigo* [10], that corresponds to the digitisation of the book *Historia de España del arçobispo Don Rodrigo*, which was written in old Castilian (Spanish) in 1545. Both corpora are publicly available for research purposes on the website of the Pattern Recognition and Human Language Technology (PRHLT) research center [2]. Acoustics models were trained by using the Spanish phonetic corpus Albayzin [11].

The transcriptions quality is assessed using the Word Error Rate (WER) value, which allows us to obtain a good estimation for the paleographer post-edition effort, and the lattices quality by the oracle WER, which represents the WER of the best hypotheses contained in the word lattices (more details about corpora and evaluation metrics can be found in Chapter 2 of the thesis).

Table 1: *Summary of the multimodal experimental results.*

Experiment	*Cristo Salvador*		*Rodrigo*	
	WER	Oracle WER	WER	Oracle WER
Off-line HTR	32.9%	27.5%	39.3%	28.2%
ASR	43.3%	27.4%	62.9%	29.5%
Multimodal	29.3%	13.4%	35.9%	14.8%

Table 1 summarises the results of the multimodal experiments. As it can be observed, the behaviour is similar for both corpora. The use of the ASR does not improve the WER of the draft offered by the off-line HTR system, although the word-graphs generated offer similar values of oracle WER for both modalities. However, combining both modalities by using our proposal, not only the WER is improved, but the oracle WER value of the multimodal word-graph lattices is substantially reduced. Given that the oracle WER value is related to the quality of the alternatives offered by our interactive and assistive system, an outstanding effect on interactive transcription can be expected.

3.2. Interactivity

The result of combining the knowledge given by off-line HTR and ASR processes may make the paleographer task easier, since they are able to correct on an improved draft transcription. However, given that paleographer revision is required to

[2] https://prhlt.upv.es/

produce a transcription of standard quality, an interactive assistive scenario, where the automatic system and the paleographer cooperate to generate the perfect transcription, would provide an additional reduction of the human effort and time required for obtaining the final result.

Chapter 5 of the thesis (*Assistive Transcription*) presents a multimodal interactive transcription system where the paleographer feedback is provided by means of touchscreen pen strokes, traditional keyboard, and mouse operations. The combination of the different sources of information is based on the use of Confusion Networks derived from the decoding output of three recognition systems: two HTR systems (off-line and on-line), and an ASR system. Off-line HTR and ASR are used to derive (by themselves or by combining their recognition results) the initial hypothesis, and on-line HTR is used to provide feedback. In the next chapter (Chapter 6 of the thesis), our multimodal and interactive proposals are tested.

In this case, the interactive performance is given by Word Stroke Ratio (WSR), the definition of which makes it comparable with the WER. The relative difference between them gives us the effort reduction (EFR), which is an estimation of the transcription effort reduction that can be achieved by using the interactive system (see Chapter 2 of the thesis for more details).

Table 2: *Summary of the multimodal interactivity experimental results.*

Experiment	*Cristo Salvador*		*Rodrigo*	
	WSR	EFR	WSR	EFR
Off-line HTR	30.2%	8.2%	36.2%	7.9%
ASR	35.1%	−6.7%	47.2%	−20.1%
Multimodal	14.1%	57.1%	27.0%	31.3%

Table 2 summarises the results of the assistive and interactive experiments. As it can be observed, the estimated interactive human effort (WSR) required for obtaining the perfect transcription from the off-line HTR decoding represents about 8% of relative effort reduction (EFR) over the off-line HTR WER for both corpora (see Table 1). However, in the case of ASR no effort reduction can be considered. Regarding multimodality, as expected, the use of the proposed multimodal approach allows the interactive system to achieve more than 30% of relative effort reduction over the off-line HTR WER for both corpora.

Table 3: *Summary of the multimodal feedback and interactivity experimental results for* Cristo Salvador.

Experiment	WSR				EFR
	Deletions	TS	KBD	Global	
Off-line HTR	5.5%	26.0%	6.7%	32.7%	0.6%
ASR	5.1%	31.6%	3.5%	35.1%	−6.7%
Multimodal	1.9%	10.7%	1.3%	12.0%	63.5%

In Table 3 a summary of the multimodal feedback (i.e. on-line handwriting feedback) and interactivity experimental results for *Cristo Salvador* are presented. In this case, the WSR is calculated under the assumptions that the deletion of words have no cost, and that the cost of keyboard-correcting an erroneous on-line feedback word is similar to another on-line HTR interaction. Therefore, the WSR correspond with the percentage of words written with the on-line HTR feedback (TS) and the percentage of words corrected by means of the keyboard (KBD). As it can be observed, the multimodal combination of the on-line feedback with the input hypotheses reduces significantly the amount of words that are required to be corrected by using the keyboard, and most of the paleographer effort is concentrated in the more ergonomic touchscreen feedback.

3.3. Crowdsourcing

As an alternative to the keyboard, volunteers could employ voice as input for transcription. Nearly all mobile devices provide this modality, which widens the range of population and situations where collaboration can be performed. The main drawback is that the audio transcription, usually obtained by ASR systems [3], presents an ambiguity not present in typed input. Even the state-of-the-art techniques [12], although more accurate than a few years ago, produce a considerable amount of errors in the recognition process, which makes it necessary to obtain a balance between the amount of collaborations and the quality they provide.

In any case, the need for final supervision by a paleographer enables the possibility that, although not perfect, voice inputs combined with off-line HTR provide an initial draft transcription more accurate than that given only by off-line HTR. This fact was confirmed with the statistically significantly improvements obtained in the experiments performed for the previous parts of this thesis, multimodal, and interactive transcription. Thus, the employment of speech collaborations will allow us to significantly reduce the final transcription effort.

Chapter 7 of the thesis (*Collective Collaboration*) explores how a crowdsourcing framework that allows for text line dictations acquisition could decrease the transcription effort. The framework is based on the use of multimodal recognition, both employing and combining off-line HTR and ASR results, to improve the final transcription that is going to be offered to the paleographer. The multimodal recognition approach is based on language model interpolation and Confusion Network combination techniques. The crowdsourcing platform was implemented by using a client-server architecture. The client is a mobile application [13] that allows speech acquisition and the server part performs the recognition and combination operations. In the next chapter (Chapter 8), our multimodal crowdsourcing proposal is tested in a supervised and in an unsupervised mode for the *Rodrigo* [10] corpus.

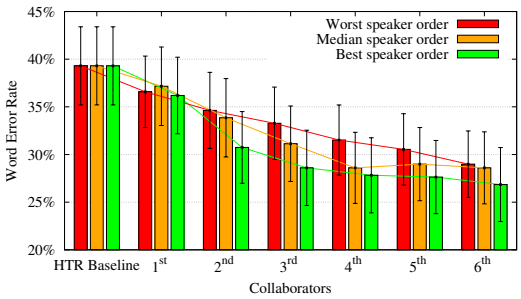

Figure 1: *Results of the speaker ordering in the supervised crowdsourcing experiments. Best, worst and the median of 11 different random orders.*

In the supervised experiments, the collaborators were randomly sorted 11 times giving 11 different order lists. Figure 1

shows the evolution, from the initial off-line HTR baseline until the process of the speech of the last collaborator, for the lists that obtained the worst, the median and the best final results. As it can be observed, the worst and the best final results do not represent any statistically significant differences. These results show that, in the best case, only two speakers are needed to obtain significant improvements. Meanwhile, in the worst case at least four speakers are needed.

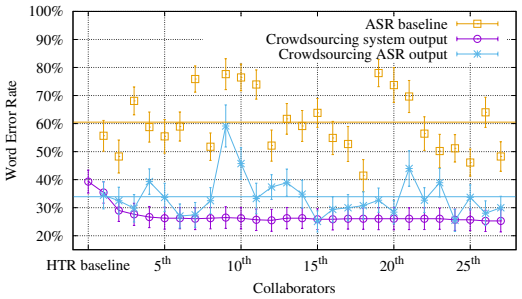

Figure 2: *Baseline values and the evolution of the system and ASR outputs for the whole unsupervised collaborations. The horizontal lines represent the corresponding average ASR WER.*

From the unsupervised experiments, Figure 2 draws the baseline values for both modalities and the evolution of the system and ASR outputs. As it can be observed, the language model interpolation permits to reduce the error level in the next speech decoding process [8], and the combination with the speech decoding results allows the system output to converge to a better hypothesis with less errors to correct [16]. Besides, the ASR performance is considerably improved, reducing the average WER baseline value. Finally, after processing the speech of the last collaborator, the system outputs presented 25.3% of WER that represents 35.6% of relative statistically significant improvement over the off-line HTR baseline, and an estimated time reduction for the paleographer revision of about 5 minutes per page [10].

4. Main Conclusions

Regarding multimodality, the benefits of multimodal combination of the results obtained from off-line HTR with additional sources of information for the transcription of historical manuscripts have been confirmed.

With respect to interactivity, multimodality was applied on an interactive tool for transcribing historical handwritten documents. On the one hand, the multimodal hypotheses combination allows to reduce the human time and workload required for transcribing historical books, due to the increased recognition accuracy and the better quality of the alternatives contained in the multimodal lattice. On the other hand, the use of multimodal combination allows to improve the human-computer interaction (by using on-line touch-screen handwritten pen strokes), given that the multimodal combination allows to correct errors on the interactive system hypothesis by using the information provided by the on-line handwritten text introduced by the user.

Finally, the proposed multimodal crowdsourcing framework is based on the iterative refinement of the language model and hypotheses combination. This framework uses a client / server architecture in order to allow collaborators to decide when and where to collaborate. The mobile application used for speech acquisition is publicly available [13].

The experiments showed that, in this framework, the number of collaborators is more important than the order in which their speech is processed. Through this experimentation, it has been shown that the use of speech is a good additional source of information for improving the transcription of historical manuscripts, and that this modality allows people to collaborate in this task using their own mobile device.

5. Current and Future Work

Currently, we are testing the performance of our assistive and interactive system with more robust modelling methods based on deep learning.

Regarding multimodality, we propose for future studies the use of whole sentences instead of lines of the handwritten text document because it might make multimodality more natural from the point of view of the paleographer or speaker who has to dictate the contents of the handwritten text images to the ASR system.

In the case of interactive transcription, we have already tested the use of speech not only as an additional source of information of the handwritten text line image to transcribe in the interactive and assistive system, but as an additional modality for human-computer interaction [14]. Furthermore, our future works aim also at taking advantage of the real samples that are produced while the system is used for adapting the feedback natural language recognisers to the user.

Finally, the proposed multimodal crowdsourcing framework and the multimodal interactive transcription system were integrated [15], and in the near future, we are planning to test it with other datasets.

6. Scientific Contributions

The main contributions of this thesis can be summarised in: the evaluation on how to combine the decoding output of different natural language recognition systems, the integration of the combination of different signals in a computer assisted transcription system, and the development of a multimodal crowdsourcing platform for the transcription of historical manuscripts.

The scientific impact of this thesis was supported by eight publications at the time of the dissertation presentation. Concretely, the multimodality part was supported by two articles presented in two international conferences (*ICDAR 2015* [16], and *CAIP 2015* [17]), the interactivity part by two publications, one in an international conference and the other in a book chapter (*DAS 2016* [18], *Handwriting, Nova 2017* [19]), and the crowdsourcing part by four publications, two in international conferences, one in a book chapter, and one in a JCR international journal (*DocEng 2016* [20], *IberSPEECH 2016* [21, 13], *IEEE/ACM TASLP* [22]).

Moreover, in the time of writing this paper, an additional publication on an international conference is supporting the interactivity part (*DAS 2018* [14]), and another in a JCR international journal the crowdsourcing part (*COIN* [15]).

7. Acknowledgments.

Work partially supported by: Percepción - TSI-020601-2012-50 (MINETUR), SmartWays - RTC-2014-1466-4 (MINECO), STraDA - TIN2012-37475-C02-01 (MINECO), and CoMUNHaT - TIN2015-70924-C2-1-R (MINECO/FEDER).

8. References

[1] A. Fischer, "Handwriting Recognition in Historical Documents," Ph.D. dissertation, University of Bern, 2012.

[2] A. Manoj, P. Borate, P. Jain, V. Sanas, and R. Pashte, "A Survey on Offline Handwriting Recognition Systems," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 2, no. 2, pp. 253–257, 2016.

[3] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[4] R. Plamondon and S. N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[5] A. H. Toselli, E. Vidal, and F. Casacuberta, "Computer Assisted Transcription of Text Images," in *Multimodal Interactive Pattern Recognition and Applications*. Springer, 2011, ch. 3, pp. 61–98.

[6] A. Fornés, J. Lladós, J. Mas, J. M. Pujades, and A. Cabré, "A Bimodal Crowdsourcing Platform for Demographic Historical Manuscripts," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*, 2014, pp. 103–108.

[7] E. Granell, "Advances on the Transcription of Historical Manuscripts based on Multimodality, Interactivity and Crowdsourcing," Ph.D. dissertation, Universitat Politècnica de València, 2017, supervisors: C.-D. Martínez-Hinarejos and V. Romero, Available: http://hdl.handle.net/10251/86137.

[8] V. Alabau, V. Romero, A. L. Lagarda, and C.-D. Martínez-Hinarejos, "A Multimodal Approach to Dictation of Handwritten Historical Documents," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 2245–2248.

[9] 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, frontiers in Handwriting Processing.

[10] N. Serrano, F. Castro, and A. Juan, "The RODRIGO Database," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 2010, pp. 2709–2712. [Online]. Available: http://aclweb.org/anthology/L10-1330

[11] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, 1993, pp. 175–178.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] E. Granell and C.-D. Martínez-Hinarejos, "Read4SpeechExperiments: A Tool for Speech Acquisition from Mobile Devices," in *Proceedings of the IX Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop (IberSPEECH'2016)*, 2016, pp. 411 – 417.

[14] C.-D. Martínez-Hinarejos, E. Granell, and V. Romero, "Comparing different feedback modalities in assisted transcription of manuscripts," in *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS '18)*, 2018, pp. 115–120.

[15] E. Granell, V. Romero, and C.-D. Martínez-Hinarejos, "Multimodality, interactivity, and crowdsourcing for document transcription," *Computational Intelligence*, vol. 34, no. 2, pp. 398–419, 2018.

[16] E. Granell and C.-D. Martínez-Hinarejos, "Combining Handwriting and Speech Recognition for Transcribing Historical Handwritten Documents," in *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR'15)*, 2015, pp. 126–130.

[17] ——, "Multimodal Output Combination for Transcribing Historical Handwritten Documents," in *Proceedings of the 16th International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2015, pp. 246–260.

[18] E. Granell, V. Romero, and C.-D. Martínez-Hinarejos, "An Interactive Approach with *Off-line* and *On-line* Handwritten Text Recognition Combination for Transcribing Historical Documents," in *Proceedings of the 12th IAPR International Workshop on Document Analysis Systems (DAS '16)*, 2016, pp. 269–274.

[19] ——, "Using Speech and Handwriting in an Interactive Approach for Transcribing Historical Documents," in *Handwriting: Recognition, Development and Analysis*. Nova Science, 2017.

[20] E. Granell and C.-D. Martínez-Hinarejos, "A Multimodal Crowdsourcing Framework for Transcribing Historical Handwritten Documents," in *Proceedings of the 16th ACM Symposium on Document Engineering (DocEng)*, 2016, pp. 157–163.

[21] ——, "Collaborator Effort Optimisation in Multimodal Crowdsourcing for Transcribing Historical Manuscripts," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 234–244.

[22] ——, "Multimodal Crowdsourcing for Transcribing Handwritten Documents," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 409–419, 2017.