UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Doctoral Thesis

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

Emilio Granell Romero

Supervisors:

Dr. Carlos David Martínez Hinarejos
Dra. Verónica Romero Gómez

València, Juny 2017

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Doctoral Thesis

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

Emilio Granell Romero

Supervisors:

Dr. Carlos David Martínez Hinarejos
Dra. Verónica Romero Gómez

València, Juny 2017

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

Emilio Granell Romero

Thesis performed under the supervision of doctors
Carlos David Martínez Hinarejos and Verónica Romero Gómez,
and presented at the *Universitat Politècnica de València*
in partial fulfilment of the requirements for the degree of
Doctor en Informàtica

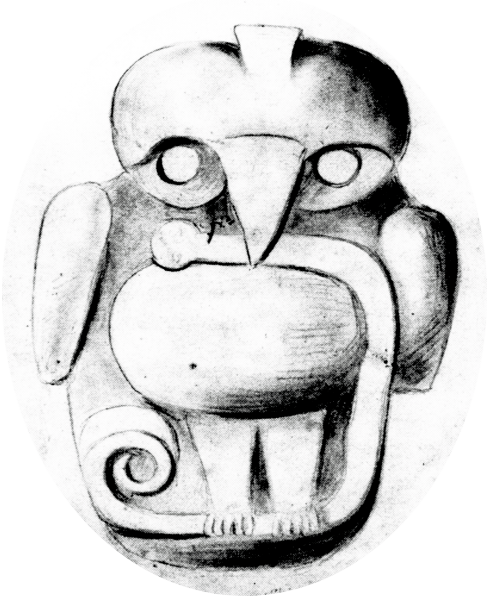
València, Juny 2017

The research leading to the results of this thesis was partially supported by the Spanish Government under the following research projects: Percepción - TSI-020601-2012-50 (MINETUR), SmartWays - RTC-2014-1466-4 (MINECO), STraDA - TIN2012-37475-C02-01 (MINECO), and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO/FEDER).

The mobility in this thesis was partially supported by the Spanish Government through the researchers mobility aids of the R-MIPRVC - MIPRCV-CSD2007-00018 (MINECO), and the European Commission through the European Union programme for education, training, youth and sport Erasmus+.

"Through the sorrow all through our splendour
Don't take offence at my innuendo
You can be anything you want to be
Just turn yourself into anything you think that you could ever be
Be free with your tempo, be free be free
Surrender your ego - be free, be free to yourself"

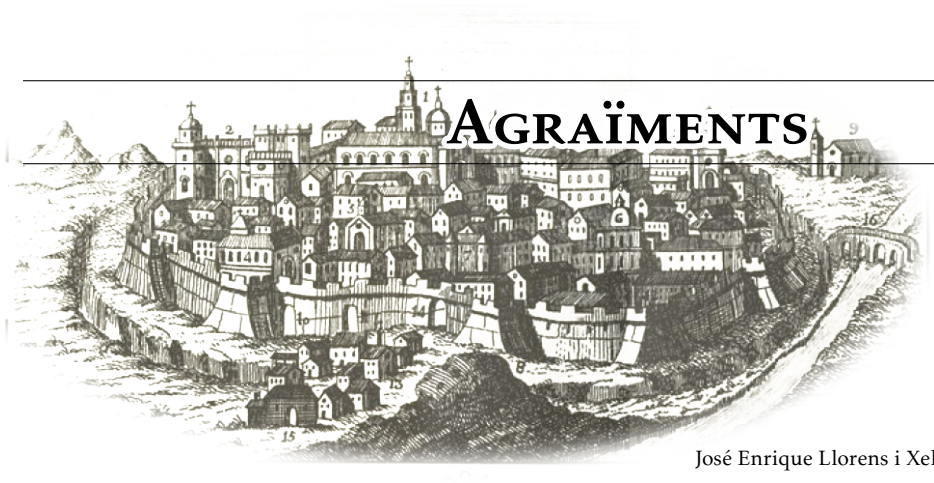
Queen, Innuendo, 1991.



The eagle and the snake as a representation of dual worlds. Sculpture of the table B of the San Agustín Archaeological Park of Colombia. (Drawing of Manuel María Paz, 1857)

Cover illustration info: Artistic montage with the French invasion of Normandy (Jean (*Jean le Bon* -John the Good-), as duke of Normandy, and future king of France as Jean II, Les Grandes chroniques de France, 1332-1350)

Contents



Agraïments

"Coneixement,
coneixement
i coneixement."[1]

José Enrique Llorens i Xelo Peiró, Saviesa valenciana. (∞)

John de Salisbury va escriure en el seu llibre *The Metalogicon* (1159) que Bernard de Chartres (en llatí *Bernardus Carnotensis*) solia comparar-nos (als moderns) amb nans enfilats a les espatlles de gegants, i deia que si veiem més i més lluny que els nostres predecessors, no és perquè tenim una visió més aguda o més altura, sinó perquè som aixecats i portats sobre la seua estatura gegantina.

Ara que per fi veig la llum al final del túnel que representa la realització d'una tesi doctoral. Després d'uns quants anys de dur treball. Crec que per a realitzar una tesi doctoral no només basta amb aconseguir pujar a les espatlles de gegants, sinó que també cal mantindre l'equilibri per a no caure pel camí. Durant aquest recorregut, he tingut la sort de comptar amb l'orientació de dos gegants com són Carlos D. Martínez i Verónica Romero, als quals estic molt agraït. Estic especialment agraït amb Carlos per donar-me l'oportunitat de dedicar-me a la investigació, per la confiança que va dipositar en mi, i per la seua disponibilitat incondicional, tan necessària per als doctorands i a la volta tan extranya per a la majoria d'ells. Moltes gràcies per haver-me alçat sobre les teues espatlles per a veure més enllà i superar tots els obstacles amb els que m'he topat durant la realització d'aquesta tesi.

Sense els amics i els companys del PRHLT i del DSIC aquests anys haurien sigut molt més durs. Durant una temporada, vam compartir tants cafés a l'àgora, que la cambrera començava a servir-los en quant ens veia entrar per la porta de la cafeteria. Moltes gràcies a Ivan Escrivà, Ihab Al-khoury, Gustavo Rovelo, Miriam Luján, Vicent Tamarit, Adrià Giménez, Nicolas Serrano, Vicente Broseta, David Rodriguez, Juan Fernando Martín, Luis A. Leiva, Jesús Alonso, Ahmed Fawzi, Javier Jorge, i d'altres ...

Agraïsc de tot cor les col·laboracions que vaig rebre de totes les persones (amics, companys i fins i tot desconeguts) que van servir per a posar veu a aquesta tesi. Sense les vostres col·laboracions no hauria pogut completar-la. Moltes gràcies a tots, especialment a Vanessa Escrivà.

Vull donar les gràcies a Laurence Likforman pels seus consells i la supervisió del treball que vaig fer durant l'estança d'investigació a *Télécom ParisTech*. *Je vous remercie beaucoup par votre compréhension et sage conseils, lesquels éclairèrent le chemin a suivre après cette thèse, malgré l'obscurité de la Ville Lumière*.

També agraïsc a Steven J. Simske, Enrique Vidal, Antonio Miguel, Alfonso Ortega i Christopher Kermorvant el temps que van dedicar per a revisar i avaluar aquesta tesi. *Your comments helped me to make this thesis brilliant. Thank you very much!*

Però és clar que tampoc hauria estat capaç de fer una tesi doctoral com aquesta sense el suport de la família. He pogut arribar fins ací gràcies a ma mare, que em va ensenyar a ser una persona resilient i a gaudir de la vida. *Muchas gracias mamá*. A les meues germanes per ser, com sempre heu sigut, una

[1] Translation in English: "Knowledge, awareness, and consciousness."

Illustration info: Stamp of the city of Gandia named *"Vista meridional dela civdad de Gandia"* drawn by Juan Fernando Palomino. (Bernardo Espinalt y García, Atlante Español, ó descripción general de todo el reyno de España. Tomo VIII. Descripción del reyno de Valencia. Parte II., 1786)

motivació per continuar endavant. A Judes Moreno, Juan Lucio i Inma Calabuig per motivar-me a entrar al món universitari, sense la vostra influència no seria qui sóc hui en dia, moltes gràcies. Al meus pares i germans bords, per tot l'afecte i saviesa que comparteixen amb mi des de fa més de 20 anys. Quanta raó hi ha en les vostres sàvies paraules! **Coneixement, coneixement i coneixement!**

Finalment, vull donar les gràcies a Carolina per ser una font inesgotable d'estimulació, motivació i estima. *Muchas gracias por animarme a realizar esta tesis a pesar de que significaba volver a tener un océano entre nosotros.* També, al nostre *frijolito* Manel que amb el seu somriure és la principal motivació per al treball actual i futur ...

Sense tots vosaltres no hauria pogut aconseguir-ho. Moltes gràcies!

Resum

El Processament del Llenguatge Natural (PLN, o NLP per les seues sigles en anglès de *Natural Language Processing*) és un camp de recerca interdisciplinar de les Ciències de la Computació, la Lingüística i el Reconeixement de Patrons que estudia, entre d'altres, l'ús del llenguatge natural humà en la interacció Home-Màquina. La majoria de les tasques de recerca del PLN es poden aplicar per resoldre problemes del món real. Aquest és el cas del reconeixement i la traducció del llenguatge natural, que es poden utilitzar per construir sistemes automàtics per a la transcripció i traducció de documents.

Quant als documents manuscrits digitalitzats, la transcripció s'utilitza per facilitar l'accés digital als continguts, ja que la simple digitalització d'imatges només proporciona, en la majoria dels casos, la cerca per imatge i no per continguts lingüístics (paraules clau, expressions, categories sintàctiques o semàntiques). La transcripció és encara més important en el cas dels manuscrits històrics, ja que la majoria d'aquests documents són únics i la preservació del seu contingut és crucial per raons culturals i històriques.

La transcripció de manuscrits històrics sol ser realitzada per paleògrafs, els quals són persones expertes en escriptura i vocabulari antics. Recentment, els sistemes de Reconeixement d'Escriptura (RES, o HTR per les seues sigles en anglès de *Handwritten Text Recognition*) s'han convertit en una eina comuna per ajudar els paleògrafs en la seua tasca, la qual proporciona un esborrany de la transcripció que els paleògrafs poden esmenar amb mètodes més o menys sofisticats. Aquest esborrany de transcripció és útil quan presenta una taxa d'error prou reduïda perquè el procés de correcció siga més còmode que una completa transcripció des de zero. Per tant, l'obtenció d'un esborrany de transcripció amb un baixa taxa d'error és crucial perquè aquesta tecnologia del PLN siga incorporada en el procés de transcripció.

El treball descrit en aquesta tesi se centra en la millora de l'esborrany de la transcripció ofert per un sistema RES, amb l'objectiu de reduir l'esforç realitzat pels paleògrafs per obtenir la transcripció de manuscrits històrics digitalitzats. Aquest problema s'enfronta a partir de tres escenaris diferents, però complementaris:

- **Multimodalitat:** L'ús de sistemes RES permet als paleògrafs accelerar el procés de transcripció manual, ja que són capaços de corregir un esborrany de la transcripció. Una altra alternativa és obtenir l'esborrany de la transcripció dictant el contingut a un sistema de Reconeixement Automàtic de la Parla (RAP, o ASR per les seues sigles en anglès de *Automatic Speech Recognition*). Quan les dues fonts (imatge i parla) estan disponibles, una combinació multimodal és possible i es pot realitzar un procés iteratiu per refinar la hipòtesi final.
- **Interactivitat:** L'ús de tecnologies assistencials en el procés de transcripció permet reduir el temps i l'esforç humà requerits per obtenir la transcripció real, gràcies a la cooperació entre el sistema assistencial i el paleògraf per obtenir la transcripció perfecta. La realimentació (*feedback* en anglès) multimodal es pot utilitzar en el sistema assistencial per proporcionar fonts d'informació addicionals amb senyals que representen la mateixa seqüencia de paraules a transcriure (per exemple, una imatge de text, o el senyal de parla del dictat del contingut d'aquesta imatge de text), o senyals que representen només una paraula o caràcter a corregir (per exemple, una paraula manuscrita mitjançant una pantalla tàctil).
- ***Crowdsourcing*:** La col·laboració distribuïda i oberta (*crowdsourcing*) sorgeix com una poderosa eina per a la transcripció massiva a un cost relativament baix, ja que l'esforç de supervisió dels

paleògrafs pot ser reduït dràsticament. La combinació multimodal permet utilitzar el dictat del contingut de línies de text manuscrit en una plataforma de *crowdsourcing* multimodal, on els col·laboradors poden proporcionar les mostres de parla utilitzant el seu propi dispositiu mòbil en lloc d'utilitzar ordinadors d'escriptori o portàtils, la qual cosa permet ampliar el nombre de col·laboradors reclutables.

Aquests escenaris són la motivació dels principals objectius científics i tecnològics:

- Estudiar les tècniques de combinació unimodal i multimodal, per tal de proposar una nova tècnica de combinació multimodal per millorar la transcripció d'imatges de texts manuscrits històrics utilitzant el dictat dels continguts d'aquestes imatges.
- Estudiar l'ús de tècniques de combinació multimodal en un sistema de transcripció assistida per ordinador per accelerar el procés de transcripció interactiu.
- Desenvolupar una plataforma *crowdsourcing* multimodal per a la transcripció de manuscrits històrics basada en les tècniques de combinació multimodal estudiades.

Les aportacions a l'estat de l'art d'aquesta tesi es poden resumir en: l'avaluació de com combinar les eixides de diferents sistemes de reconeixement de llenguatge natural, la integració de la combinació de diferents senyals en un sistema de transcripció assistida per ordinador, i el desenvolupament d'una plataforma crowdsourcing multimodal per a la transcripció de manuscrits històrics.

Resumen

El Procesamiento del Lenguaje Natural (PLN, o NLP por sus siglas en inglés de *Natural Language Processing*) es un campo de investigación interdisciplinar de las Ciencias de la Computación, Lingüística y Reconocimiento de Patrones que estudia, entre otros, el uso del lenguaje natural humano en la interacción Hombre-Máquina. La mayoría de las tareas de investigación del PLN se pueden aplicar para resolver problemas del mundo real. Este es el caso del reconocimiento y la traducción del lenguaje natural, que se pueden utilizar para construir sistemas automáticos para la transcripción y traducción de documentos.

En cuanto a los documentos manuscritos digitalizados, la transcripción se utiliza para facilitar el acceso digital a los contenidos, ya que la simple digitalización de imágenes sólo proporciona, en la mayoría de los casos, la búsqueda por imagen y no por contenidos lingüísticos (palabras clave, expresiones, categorías sintácticas o semánticas). La transcripción es aún más importante en el caso de los manuscritos históricos, ya que la mayoría de estos documentos son únicos y la preservación de su contenido es crucial por razones culturales e históricas.

La transcripción de manuscritos históricos suele ser realizada por paleógrafos, que son personas expertas en escritura y vocabulario antiguos. Recientemente, los sistemas de Reconocimiento de Escritura (RES, o HTR por sus siglas en inglés de *Handwritten Text Recognition*) se han convertido en una herramienta común para ayudar a los paleógrafos en su tarea, la cual proporciona un borrador de la transcripción que los paleógrafos pueden corregir con métodos más o menos sofisticados. Este borrador de transcripción es útil cuando presenta una tasa de error suficientemente reducida para que el proceso de corrección sea más cómodo que una completa transcripción desde cero. Por lo tanto, la obtención de un borrador de transcripción con una baja tasa de error es crucial para que esta tecnología de PLN sea incorporada en el proceso de transcripción.

El trabajo descrito en esta tesis se centra en la mejora del borrador de transcripción ofrecido por un sistema RES, con el objetivo de reducir el esfuerzo realizado por los paleógrafos para obtener la transcripción de manuscritos históricos digitalizados. Este problema se enfrenta a partir de tres escenarios diferentes, pero complementarios:

- **Multimodalidad:** El uso de sistemas RES permite a los paleógrafos acelerar el proceso de transcripción manual, ya que son capaces de corregir en un borrador de la transcripción. Otra alternativa es obtener el borrador de la transcripción dictando el contenido a un sistema de Reconocimiento Automático de Habla (RAH, o ASR por sus siglas en inglés de *Automatic Speech Recognition*). Cuando ambas fuentes (imagen y habla) están disponibles, una combinación multimodal de las mismas es posible y se puede realizar un proceso iterativo para refinar la hipótesis final.
- **Interactividad:** El uso de tecnologías asistenciales en el proceso de transcripción permite reducir el tiempo y el esfuerzo humano requeridos para obtener la transcripción correcta, gracias a la cooperación entre el sistema asistencial y el paleógrafo para obtener la transcripción perfecta. La realimentación (*feedback* en inglés) multimodal se puede utilizar en el sistema asistencial para proporcionar otras fuentes de información adicionales con señales que representen la misma secuencia de palabras a transcribir (por ejemplo, una imagen de texto, o la señal de habla del dictado del contenido de dicha imagen de texto), o señales que representen sólo una palabra o carácter a corregir (por ejemplo, una palabra manuscrita mediante una pantalla táctil).

- ***Crowdsourcing*:** La colaboración distribuida y abierta (*crowdsourcing*) surge como una poderosa herramienta para la transcripción masiva a un costo relativamente bajo, ya que el esfuerzo de supervisión de los paleógrafos puede ser drásticamente reducido. La combinación multimodal permite utilizar el dictado del contenido de líneas de texto manuscrito en una plataforma de *crowdsourcing* multimodal, donde los colaboradores pueden proporcionar las muestras de habla utilizando su propio dispositivo móvil en lugar de usar ordenadores de escritorio o portátiles, lo cual permite ampliar el número de colaboradores reclutables.

Estos escenarios son la motivación de los principales objetivos científicos y tecnológicos:

- Estudiar las técnicas de combinación unimodal y multimodal, con el fin de proponer una nueva técnica de combinación multimodal para mejorar la transcripción de imágenes de textos manuscritos históricos utilizando el dictado de los contenidos de dichas imágenes.
- Estudiar el uso de técnicas de combinación multimodal en un sistema de transcripción asistida por ordenador para acelerar el proceso de transcripción interactivo.
- Desarrollar una plataforma multimodal de *crowdsourcing* para la transcripción de manuscritos históricos basada en las técnicas de combinación multimodal estudiadas.

Las aportaciones al estado del arte de esta tesis se pueden resumir en: la evaluación de cómo combinar la salida de diferentes sistemas de reconocimiento de lenguaje natural, la integración de la combinación de diferentes señales en un sistema de transcripción asistida por ordenador, y el desarrollo de una plataforma multimodal de *crowdsourcing* para la transcripción de manuscritos históricos.

Abstract

Natural Language Processing (NLP) is an interdisciplinary research field of Computer Science, Linguistics, and Pattern Recognition that studies, among others, the use of human natural languages in Human-Computer Interaction (HCI). Most of NLP research tasks can be applied for solving real-world problems. This is the case of natural language recognition and natural language translation, that can be used for building automatic systems for document transcription and document translation.

Regarding digitalised handwritten text documents, transcription is used to obtain an easy digital access to the contents, since simple image digitalisation only provides, in most cases, search by image and not by linguistic contents (keywords, expressions, syntactic or semantic categories, ...). Transcription is even more important in historical manuscripts, since most of these documents are unique and the preservation of their contents is crucial for cultural and historical reasons.

The transcription of historical manuscripts is usually done by paleographers, who are experts on ancient script and vocabulary. Recently, Handwritten Text Recognition (HTR) has become a common tool for assisting paleographers in their task, by providing a draft transcription that they may amend with more or less sophisticated methods. This draft transcription is useful when it presents an error rate low enough to make the amending process more comfortable than a complete transcription from scratch. Thus, obtaining a draft transcription with an acceptable low error rate is crucial to have this NLP technology incorporated into the transcription process.

The work described in this thesis is focused on the improvement of the draft transcription offered by an HTR system, with the aim of reducing the effort made by paleographers for obtaining the actual transcription on digitalised historical manuscripts. This problem is faced from three different, but complementary, scenarios:

- **Multimodality:** The use of HTR systems allow paleographers to speed up the manual transcription process, since they are able to correct on a draft transcription. Another alternative is to obtain the draft transcription by dictating the contents to an Automatic Speech Recognition (ASR) system. 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 final hypothesis.
- **Interactivity:** The use of assistive technologies in the transcription process allows one to reduce the time and human effort required for obtaining the actual transcription, given that the assistive system and the palaeographer cooperate to generate a perfect transcription. Multimodal feedback can be used to provide the assistive system with additional sources of information by using signals that represent the whole same sequence of words to transcribe (e.g. a text image, and the speech of the dictation of the contents of this text image), or that represent just a word or character to correct (e.g. an *on-line* handwritten word).
- **Crowdsourcing:** Open distributed collaboration emerges as a powerful tool for massive transcription at a relatively low cost, since the paleographer supervision effort may be dramatically reduced. Multimodal combination allows one to use the speech dictation of handwritten text lines in a multimodal crowdsourcing platform, where collaborators may provide their speech by using their own mobile device instead of using desktop or laptop computers, which makes it possible to recruit more collaborators.

These scenarios are the motivation for the main scientific and technological goals:

- To study the unimodal and multimodal combination techniques, in order to propose a new multimodal combination technique for improving the draft transcription of historical text images by using the speech dictation of the contents of the same text images.
- To study the use of multimodal combination techniques in a computer assisted system to accelerate the interactive transcription process.
- To develop a multimodal crowdsourcing platform based on the studied multimodal combination techniques.

The contributions to the state of the art 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.

List of Acronyms

ANN Artificial Neural Network.

ASR Automatic Speech Recognition.

CAD Computer Aided Design.

CAT Computer Assisted Transcription.

CATTI Computer Assisted Transcription of Text Images.

CE Collaboration Effort.

CER Character Error Rate.

CL Computational Linguistics.

CMN Cepstral Mean Normalisation.

CN Confusion Network.

CNC Confusion Network Combination.

DFA Deterministic Finite Automaton.

DNN Deep Neural Network.

EFR Effort Reduction.

EM Expectation-Maximization.

ER Error Rate.

FST Finite State Transducer.

GMM Gaussian Mixture Model.

GSF Grammar Scale Factor.

HCI Human-Computer Interaction.

HMM Hidden Markov Model.

HTR Handwriting/Handwritten Text Recognition.

IPA International Phonetic Alphabet.

IQR Interquartile Range.

LIF Lower Inner Fence.

LM Language Model.

MA Mouse Action.

MAD Median Absolute Deviation.

MAP Maximum *A Posteriori*.

MFCC Mel-Frequency Cepstral Coefficients.

MLLR Maximum Likelihood Linear Regression.

MLP Multi Layer Perceptron.

MM-CATTI Multimodal Computer Assisted Transcription of Text Images.

NLP Natural Language Processing.

OOV Out Of Vocabulary.

PDF Probability Density Function.

PER Phoneme Error Rate.

PMF Probability Mass Function.

PPL Perplexity.

RNN Recurrent Neural Network.

ROVER Recognition Output Voting Error Reduction.

SCMN Segmental Cepstral Mean Normalisation.

UIF Upper Inner Fence.

WER Word Error Rate.

WG Word Graph.

WIP Word Insertion Penalty.

WSR Word Stroke Ratio.

List of Mathematical Symbols

M Number of different observations per state of the HMM.

N Number of states in the HMM.

Σ_n The covariance matrix of the Gaussian component n.

$\hat{B}_{\hat{w}}$ Set of all valid basic unit sequences for a word sequence $\hat{w}$.

$\hat{S}$ Set of suffix word sequences.

$\hat{W}$ Set of word sequences.

$\hat{X}$ Set of feature sequences.

$\hat{\Theta}$ Set of state sequences.

$\hat{\theta}$ State sequence.

$\hat{b}$ Basic linguistic units (usually characters for HTR and phones for ASR) sequence.

$\hat{d}$ Data sequence.

$\hat{o}$ Observation sequence.

$\hat{p}$ Prefix word sequence.

$\hat{s}$ Suffix word sequence.

$\hat{w}$ Word sequence.

$\hat{x}$ Feature sequence.

λ Hidden Markov Model.

$\hat{\mathbf{s}}$ Most likely suffix word sequence.

$\hat{\mathbf{w}}$ Most likely word sequence.

$\mathcal{N}(\mu,\Sigma)$ Gaussian distribution with mean μ and covariance matrix Σ.

$\mathcal{N}(x_t;\mu,\Sigma)$ Likelihood of observation x_t being generated by $\mathcal{N}(\mu,\Sigma)$.

μ_n The mean of the Gaussian component n.

$\mathbf{P}(\hat{x} \mid \hat{w})$ Probability of observing the sequence $\hat{x}$ by assuming that $\hat{w}$ is the underlying word sequence for $\hat{x}$.

$\mathbf{P}(\hat{x} \mid \theta)$ Probability of observing the sequence $\hat{x}$ given the model parameters θ.

$\mathbf{P}(w)$ Probability of w.

$\mathbf{P}(x)$ *A priori* probability of observing x.

c_n The weight of the Gaussian component n.

s_i State i of the HMM.

List of Figures

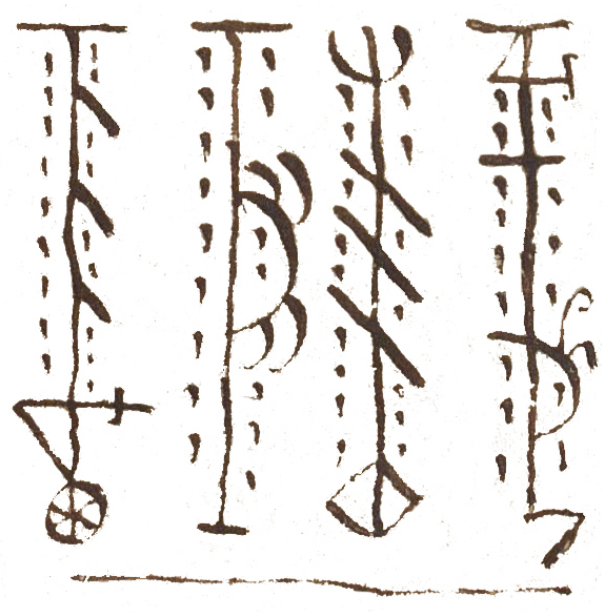
List of Tables

I

Introduction

"Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke, Profiles of the Future: An Inquiry Into the Limits of the Possible, 1962.



Stafur gegn galdri: Icelandic rune to protect against witchcraft from all four corners of the earth. (Galdrakver "Little Book of Magic", 1670)

1 Introduction

"Most of what is unusual about man can be summed up in one word: 'culture' [...] Cultural transmission is analogous to genetic transmission in that, although basically conservative, it can give rise to a form of evolution."

Richard Dawkins, The Selfish Gene, 1976.

Content

In this introductory chapter the motivation for performing this doctoral thesis, the scientific and technological objectives, and the structure of the present document are presented.

1.1 Motivation

Transcription of historical documents is an interesting task for libraries in order to provide efficient information access to text transcription of digitised historical documents. The manual transcription process is done by professionals called paleographers. In the latest years, the use of *off-line* Handwritten Text Recognition (*off-line* HTR) systems (Fischer, 2012) allowed to speed up the manual transcription process. However, state-of-the-art *off-line* HTR systems (Manoj et al., 2016) are far from being perfect, and paleographer supervision is required to really produce a transcription of standard quality. The initial result of automatic recognition may make the paleographers task easier, since they are able to perform corrections on a good draft transcription.

In addition to using *off-line* HTR systems from historical text images, other modalities of natural language recognition systems can be used to help paleographers on the transcription process, such as Automatic Speech Recognition (ASR) (Rabiner and Juang, 1993) from the dictation of the contents, and *on-line* HTR (Plamondon and Srihari, 2000) from touchscreen pen strokes. In this context, a multimodal interactive assistive scenario (Toselli et al., 2011), where the automatic system and the paleographer cooperate to generate the perfect transcription, would reduce the time and the paleographer effort required for obtaining the final result.

The use of multimodal collaborative transcription applications (crowdsourcing) (Fornés et al., 2014), where collaborators can employ speech dictation of text lines as 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.

Illustration info: Alfonso X, Violante de Aragón, and Fernando de la Cerda (Tumbo de Toxosoutos, 13th century).

In this thesis, the reduction of the required effort of the paleographer for obtaining the actual transcription of digitalised historical manuscripts is studied in the following scenarios:

- **Multimodality:** An initial draft transcription of a handwritten text 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 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 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 user feedback and to provide the system with additional sources of information.
- **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 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.

1.2 Scientific and Technological Objectives

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

- **The study of the multimodal combination of Handwriting Text and Speech Recognition systems:** The combination of natural language recognition systems allows the improvement of the recognition accuracy. In most cases, this combination can be performed in three different stages of the recognition process: in the feature extraction stage (feature combination) (Potamianos and Neti, 2001), in the search process (probability combination) (Hernando et al., 1995), and in the decoding output (hypothesis combination) (Fiscus, 1997). However, in this thesis only the hypothesis combination will be studied given the asynchrony between the text image and the speech audio, and the different basic units for words on each modality (characters for HTR and phonemes for ASR).
- **The improvement of an assistive transcription system:** Multimodal combination of natural language recognition systems will allow the incorporation of new sources of information in an interactive transcription system called "Computer Assisted Transcription of Text Images" (CATTI) (Romero et al., 2012; Martín-Albo et al., 2013). We will study the multimodal combination at the CATTI system input, and we will use multimodal combination to integrate the user feedback through touchscreen pen strokes in the CATTI system.
- **The development of a multimodal crowdsourcing platform for the transcription of historical manuscripts:** We will propose a platform with a Client/Server architecture based on multimodal combination of natural language recognition systems where volunteers will be able to collaborate with the transcription of historical manuscripts by using their own mobile devices. In addition,

given that collaborators are a scarce resource, the optimisation of their work load will be studied in order to get the maximum benefit from their efforts.

1.3 Document Structure

This thesis is structured in five parts to facilitate the reading experience. Part I contains this introductory chapter (Chapter 1), and the next one (Chapter 2) introduces the basic concepts. Part II is dedicated to multimodality. Chapter 3 presents the proposed multimodal combination technique, and Chapter 4 shows the performed multimodal experiments. Interactivity is presented in Part III. Chapter 5 details the proposed improvements for the multimodal CATTI framework. Next, the experiments carried out to study the use of speech as an additional source of information at the CATTI input, and the integration of the user feedback (*on-line* HTR) are described in Chapter 6. Crowdsourcing is discussed in Part IV. The suggested multimodal crowdsourcing framework is reviewed in Chapter 7. Then, the experiments performed to assess this multimodal crowdsourcing framework are presented in Chapter 8. Finally, in Chapter 9 (contained in Part V) the conclusions of this thesis, and the future work lines are summarised.

Bibliography

Fischer, A. (2012). *Handwriting Recognition in Historical Documents*. PhD thesis, University of Bern.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 1997)*, pages 347–354.

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

Hernando, J., Ayarte, J., and Monte, E. (1995). Optimization of speech parameter weighting for CDHMM word recognition. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 105–108.

Manoj, A., Borate, P., Jain, P., Sanas, V., and Pashte, R. (2016). A Survey on Offline Handwriting Recognition Systems. *International Journal of Scientific Research in Science, Engineering and Technology*, 2(2):253–257.

Martín-Albo, D., Romero, V., and Vidal, E. (2013). Interactive Off-Line Handwritten Text Transcription Using On-Line Handwritten Text as Feedback. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13)*, pages 1280–1284.

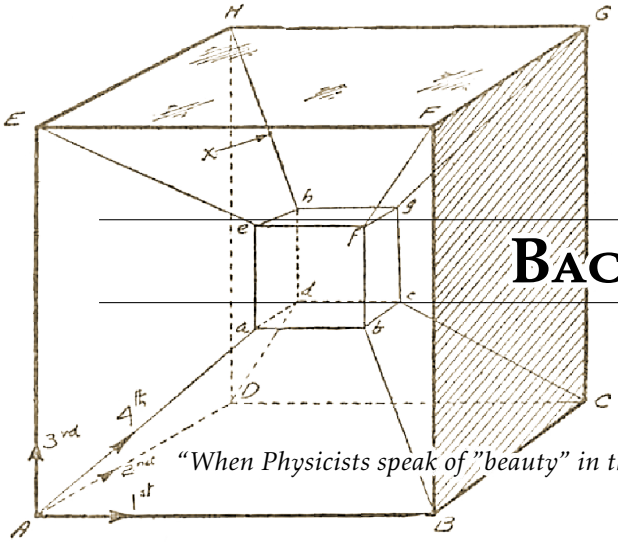
Plamondon, R. and Srihari, S. N. (2000). On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84.

Potamianos, G. and Neti, C. (2001). Automatic Speechreading of Impaired Speech. In *Proceedings of the 2001 International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 177–182.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Romero, V., Toselli, A. H., and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*, volume 80 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing.

Toselli, A. H., Vidal, E., and Casacuberta, F. (2011). Computer Assisted Transcription of Text Images. In *Multimodal Interactive Pattern Recognition and Applications*, chapter 3, pages 61–98. Springer.



2 Background

"When Physicists speak of "beauty" in their theories, they really mean that their theory possesses at least two essential features:
1. A unifying symmetry.
2. The ability to explain vast amounts of experimental data with the most economical mathematical expressions."

Michio Kaku, Hyperspace: A Scientific Odyssey Through Parallel Universes, Time Warps, and The Tenth Dimension, 1994.

Content

This chapter describes the theoretical basis of the work performed in this thesis, together with some important concepts, the evaluation measures, and the datasets used in the experimentation.

Illustration info: Drawing of a tesseract (Alexander Horne, Theosophy and the Fourth Dimension, 1928).

2.1 Statistical Natural Language Recognition

The goal of natural language recognition systems is to obtain the transcription of the input data (a handwritten text image or a speech utterance), usually represented as a determined sequence of features vectors, as a sequence of words given a lexicon (Toselli et al., 2004). The fundamental formulation of statistical natural language recognition is: given a sequence of data $\hat{d} = (d_1, d_2, \ldots, d_{|\hat{d}|})$ (for instance, from a handwritten text image or a speech signal) encoded into the feature vector sequence $\hat{x} = (x_1, x_2, \ldots, x_{|\hat{x}|})$, finding the most likely word sequence $\hat{\mathbf{w}} = (w_1, w_2, \ldots, w_{|\hat{\mathbf{w}}|})$, that is:

$$\hat{\mathbf{w}} = \arg\max_{\hat{w} \in \hat{W}} \mathrm{P}(\hat{w} \mid \hat{x}) = \arg\max_{\hat{w} \in \hat{W}} \frac{\mathrm{P}(\hat{x} \mid \hat{w})\mathrm{P}(\hat{w})}{\mathrm{P}(\hat{x})} = \arg\max_{\hat{w} \in \hat{W}} \mathrm{P}(\hat{x} \mid \hat{w})\mathrm{P}(\hat{w}) \tag{2.1}$$

where $\hat{W}$ denotes the set of all permissible sentences, $\mathrm{P}(\hat{x})$ is the *a priori* probability of observing $\hat{x}$, $\mathrm{P}(\hat{w})$ is the probability of $\hat{w}$, and $\mathrm{P}(\hat{x} \mid \hat{w})$ is the probability of observing $\hat{x}$ by assuming that $\hat{w}$ is the underlying word sequence for $\hat{x}$.

Most current natural language recognition systems handle the recognition process in two steps: a preprocessing step where feature vectors are extracted from the input data, and a recognition step where the system produces the most statistically likely output given the input feature vectors. In Figure 2.1 this scheme is presented.

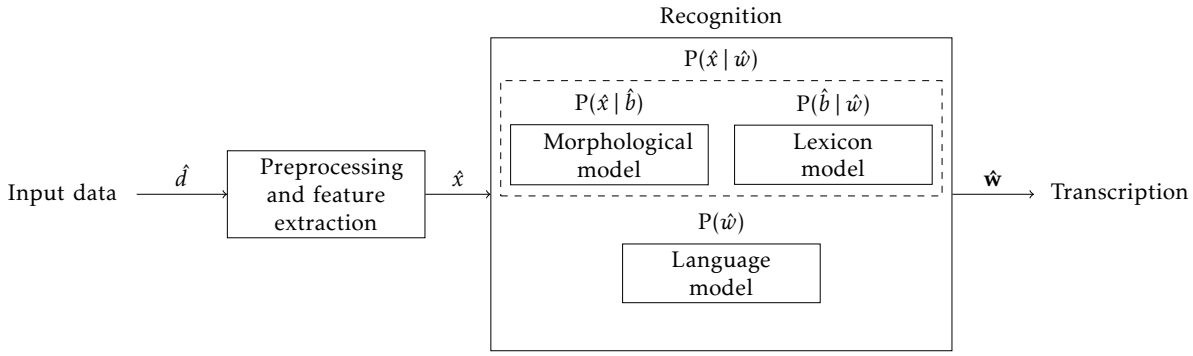


Figure 2.1: Scheme of a two-step process in natural language recognition.

In the recognition step, statistical language properties can be represented by using three statistical models. The morphological model models the relationship between the features of the input signal and the basic units of the language (typically approximated by concatenated Hidden Markov Models -Section 2.3.1-). The lexicon model contains the available vocabulary, where words are modelled as sequences of basic sub-word units $\hat{b} = (b_1, b_2, \ldots, b_{|\hat{b}|})$ (habitually characters for HTR and phones for ASR). Therefore, $\mathrm{P}(\hat{x} \mid \hat{w})$ is evaluated by using the morphological and the lexicon models. Assuming independence between the morphological realisation of words and their transcription (Yun and Oh, 1999), $\mathrm{P}(\hat{x} \mid \hat{w})$ can be represented as:

$$\mathrm{P}(\hat{x} \mid \hat{w}) = \sum_{\hat{b} \in \hat{B}_{\hat{w}}} \mathrm{P}(\hat{x} \mid \hat{b})\mathrm{P}(\hat{b} \mid \hat{w}) \tag{2.2}$$

where $\hat{B}_{\hat{w}}$ represents all possible sequences of basic sub-word units for the word sequence $\hat{w}$, $\mathrm{P}(\hat{x} \mid \hat{b})$ is modelled by the morphological model, and $\mathrm{P}(\hat{b} \mid \hat{w})$ by the lexicon model. Finally, the language model models the permissible word sequences and approximates $\mathrm{P}(\hat{w})$ (usually it is modelled by a n-gram word language model -Section 2.3.2-). In addition of the word order, language models can be used to model the language syntax and the semantic information of the words (Bod, 2000).

This approach can be used to solve different natural language problems depending on the source

of the input data. In this work we have worked with the following three different types of recognition systems:

- **Automatic Speech Recognition (ASR):** The source of the input data sequence is a speech signal (Rabiner and Juang, 1993).
- **Handwriting Text Recognition (HTR):** The input data is related to handwriting. There are two different HTR modes, depending on the handwriting data source:
  - ***Off-line* HTR:** The handwriting input data sequence is originated from text images. This HTR mode allows the transcription of texts originally contained on non-digital formats, such as historical manuscripts (Fischer, 2012). The digitisation can be performed easily using scanners or photographic cameras. However, it presents the drawback that usually the available information is limited to the content of the digital image.
  - ***On-line* HTR:** The handwriting input data is provided by means of touchscreen pen strokes. The data acquisition for this mode requires specialised hardware (such as touch-screens) to acquire the kinematical information of the user handwriting (Martín-Albo Simón, 2016). This kinematical information is richer than that of a text image, since it can be rendered in order to obtain a digital image to be transcribed by an *off-line* HTR system. It also provides data on the sequence of points followed by the user handwriting.

These recognition systems are composed of two phases:

- **Training:** In the model generation phase the statistical models are estimated by using labeled corpora, i.e. for each entry sequence $\hat{x}$, the expected transcription output $\hat{w}$ is known.
- **Testing:** In the pattern matching phase the system estimates the most likely transcription for the input sequence. In this phase it is possible to measure the error made by the system if a labeled input is available.

2.2 Preprocessing and Feature Extraction

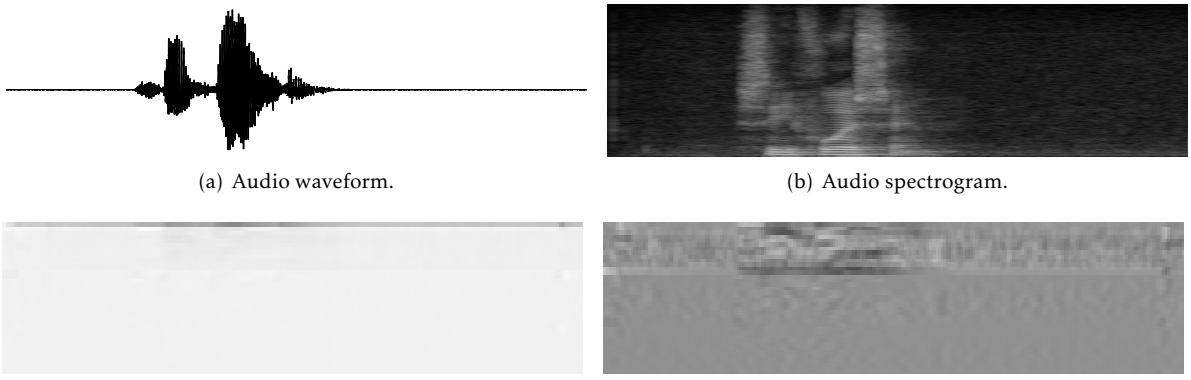
Feature extraction is the process of calculating a compact parametric representation of the input signal features which are relevant for automatic recognition (Rabiner and Juang, 1993). Therefore, the feature extraction is performed after a preprocessing step, which depends on the nature of the input data. On the one hand, ASR preprocessing enhances the audio signal according to the human psychoacoustics. On the other hand, *off-line* HTR preprocessing is aimed at correcting image degradations and geometry distortions, while *on-line* HTR preprocessing involves usually only two simple steps: repeated points elimination and noise reduction.

2.2.1 Automatic Speech Recognition Features

With respect to speech features, computing Mel-Frequency Cepstral Coefficients (MFCC) is the most commonly used feature extraction method from audio signals in ASR. The Fourier transform is calculated over a window of a pre-emphasised signal. Next, a Mel scale filter-bank is applied and the filters outputs are logarithmised. Finally, to obtain the MFCC a discrete cosine transformation is applied (Rabiner and Juang, 1993).

In some applications, ASR systems need to be robust with respect to their acoustical environment, such as ASR systems used from mobile devices. One technique for robust speech recognition is Cepstral Mean Normalisation (CMN). CMN is performed after the MFCC feature extraction, by means of the subtraction of the cepstral mean from all the vectors. This normalisation allows one to compensate the long-term spectral effects caused by different microphones and acoustical environments in the final ASR features (Liu et al., 1993). In robust *on-line* ASR, Segmental Cepstral Mean Normalisation (SCMN) can be applied on buffered segments of the audio signal (Viikki and Laurila, 1998).

In this work, a sampling rate of 16 KHz, a frame length of 25 ms, 10 ms of shift interval, and a filter-bank of 23 equidistant (in Mel frequency domain) triangular filters are used. Then, the first



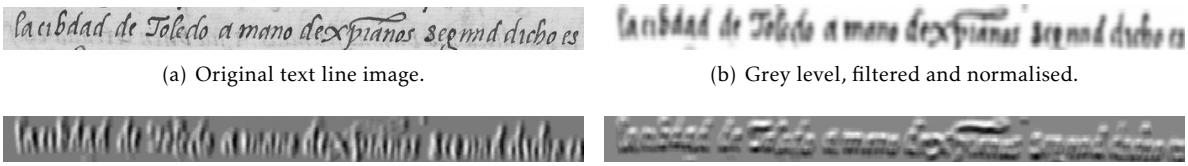
(a) Audio waveform.

(b) Audio spectrogram.

(c) MFCC speech features.

(d) MFCC with CMN speech features.

Figure 2.2: Speech audio signal and ASR feature extraction.



(a) Original text line image.

(b) Grey level, filtered and normalised.

(c) Horizontal derivative.

(d) Vertical derivative.

Figure 2.3: Text line image and *off-line* HTR feature extraction.

12 MFCC and log frame energy with first and second order derivatives are used, resulting in a 39-dimensional feature vector (ETSI, 2003). Figure 2.2 presents an audio signal (waveform and spectrogram) and the graphical representation of the obtained MFCC feature vectors with and without CMN.

2.2.2 *Off-line* Handwriting Text Recognition Features

There are different approaches for obtaining the *off-line* handwritten text features (Bunke et al., 1995; Toselli et al., 2004; Kozielski et al., 2013). In this thesis, the *off-line* handwritten text features are computed in several steps from text line images following the approach presented in (Toselli et al., 2004). In the extraction process, first a bright normalisation is performed. After that, a median filter is applied to the whole image. Next, slant correction is performed by using the maximum variance method (Pastor et al., 2004). Then, a size normalisation is performed and the final image is scaled. Finally, each preprocessed text line image is represented as a sequence of feature vectors. To do this, the text line image is divided into squared cells. From each cell, three features are calculated: normalised grey level, horizontal grey level derivative and vertical grey level derivative. Columns of cells (frames) are processed from left to right and a feature vector is constructed for each frame by stacking the three features computed in its constituent cells.

In this work, a median filter of size 3×3 pixels, a height of 40 pixels, and squared cells of size 20×20 pixels are used. The final feature vectors are of 60 dimensions (20 normalised grey level, 20 horizontal grey level derivative, and 20 vertical grey level derivative) (Toselli et al., 2004). Figure 2.3 presents an example of the feature vectors sequence obtained from one text line image.

2.2.3 *On-line* Handwriting Text Recognition Features

In the *on-line* handwritten text feature extraction a touchscreen coordinates sequence is transformed into a new speed- and size-normalised temporal sequence of feature vectors (Toselli et al., 2007). In

our case, for each point a 6-dimensional feature vector is calculated. First, both coordinates are size-normalised and translated in order to preserve the aspect ratio. Then, the speed-normalised first and second derivatives are calculated (Martín-Albo Simón, 2016).

2.2.4 Tandem Features

In the tandem feature extraction scheme, a neural network is trained to estimate the posterior probabilities of the basic units (usually characters for HTR and phones for ASR) at frame level, which are used as input features in a conventional recogniser (Hermansky et al., 2000). This means that the size of the input layer of the neural network should match the size of the pre-processed feature vectors and the size of the output layer should match the size of the set of basic units. In Figure 2.4, this tandem features extraction scheme is represented. The frame-level labelling required to train this neural network can be generated from a forced alignment decoding by a previously trained recognition system (Hermansky et al., 2000). This forced alignment decoding and the model training must be repeated several times until the convergence of the frame labels. Finally, the tandem features are constituted by the log posterior probabilities of the neural network output.

In this thesis, tandem features were extracted for *off-line* HTR and ASR. Therefore, different Multi-Layer Perceptrons (MLP) with a softmax transfer function at the output layer were trained by backpropagation with a mean-squared error criterion.

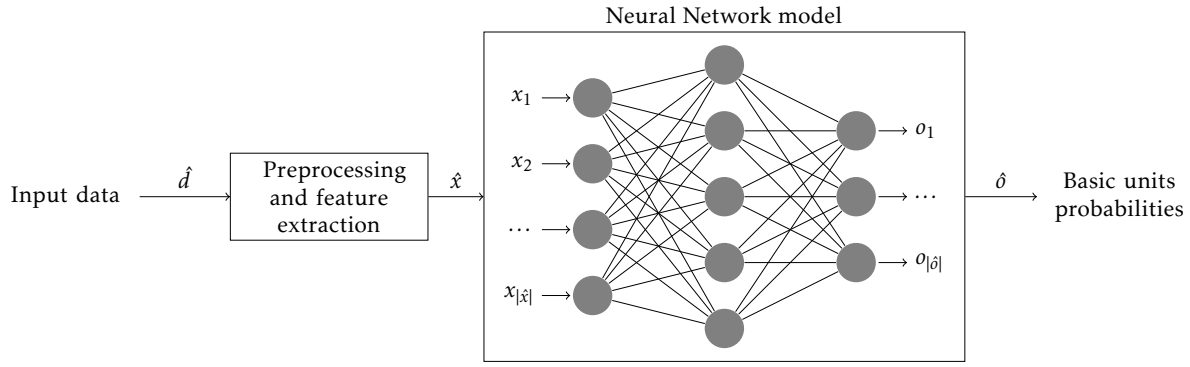


Figure 2.4: Block diagram of the tandem features extraction scheme.

The use of the tandem approach (Hermansky et al., 2000) allowed us to obtain additional recognition systems for the unimodal-multimodal experiment performed in Section 4.4. Although deep learning techniques (Hinton et al., 2012; Goodfellow et al., 2016) allow to train complex neural networks, we chose to use the traditional multilayer perceptron with only one hidden layer due to our limited technical capacities at the time this work was done.

2.3 Statistical Modelling

A statistical model is a non-deterministic class of mathematical model where some variables have probability distributions instead of specific values, i.e. some of the variables are stochastic (McCullagh, 2002). Statistical models describe a set of probability distributions, which are assumed to adequately approximate the generation process of a specific data set distribution in the case of generative models, or to model a direct map to the class labels in the case of discriminative models. Namely, given the inputs x and labels y, generative models learn the joint probability distribution $\mathrm{P}(x, y)$ and the conditional probability distribution $\mathrm{P}(x \mid y)$ can be calculated by using the Bayes rule, whilst discriminative models learn directly $\mathrm{P}(x \mid y)$ (Ng and Jordan, 2001).

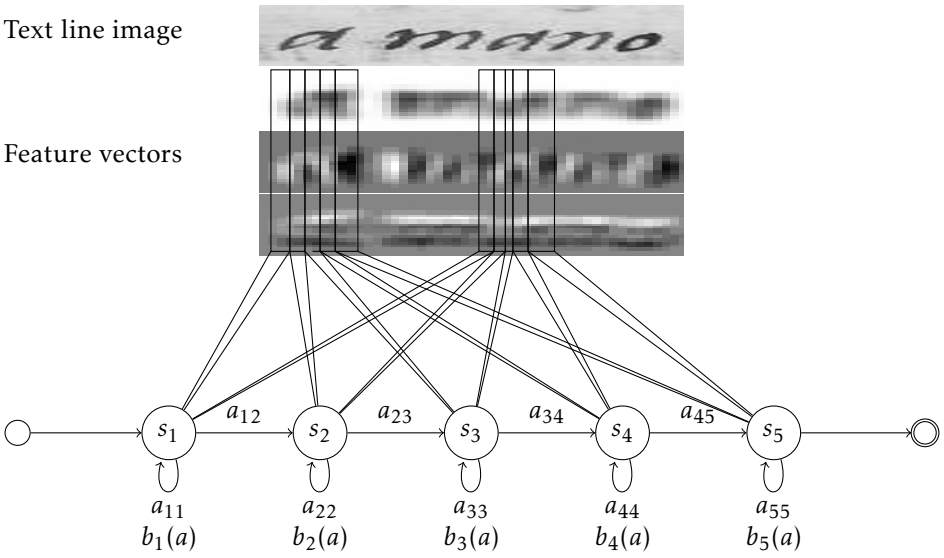


Figure 2.5: A left-to-right HMM with 5 states for HTR. This HMM models the character *a*. a_{ij} is the probability to transit from state s_i to state s_j. $b_i(a)$ is the probability to emit the features related to character *a* in state s_i.

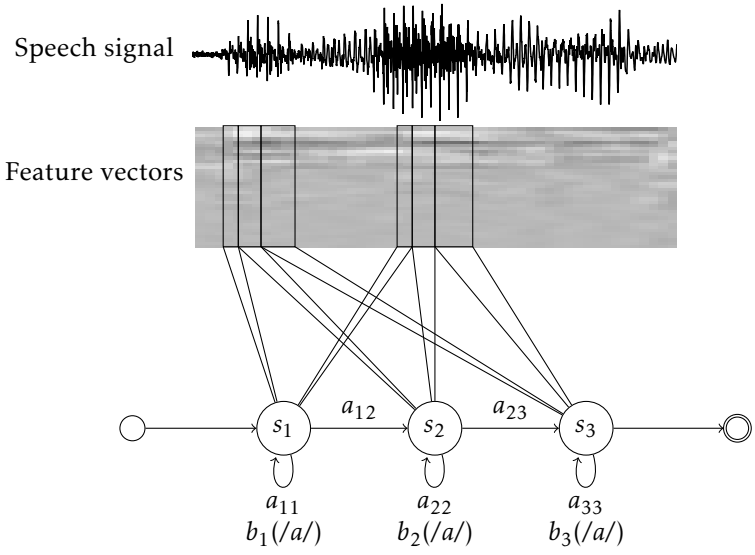


Figure 2.6: A left-to-right HMM with 3 states for ASR. This HMM models the phoneme */a/*. a_{ij} is the probability to transit from state s_i to state s_j. $b_i(/a/)$ is the probability to emit the features related to phoneme */a/* in state s_i.

As previously said, usually, in statistical natural language recognition three different models are used: morphological model, language model, and lexicon model.

2.3.1 Morphological Modelling

Morphological models are used in natural language recognition to represent the relationship between the input signal (such as image, pen strokes or audio) and the basic morphological units that make

up the language in the corresponding modality (phones for speech audio, characters for text images, strokes for *on-line* handwriting, etc.) (Manning and Schütze, 1999).

These models can be implemented through Hidden Markov Models (HMM) (Baum and Petrie, 1966; Baum and Eagon, 1967), Artificial Neural Networks (ANN) (Bishop, 1995), Deep Neural Networks (DNN) (Hinton et al., 2012), or even as Hybrid HMM/ANN models (Bourlard and Morgan, 1998). In the last years, the research in deep learning for morphological modelling has produced a breakthrough in the field of natural language processing (Wang et al., 2012; Graves and Jaitly, 2014; Goodfellow et al., 2016). However, HMM are still widely used for morphological modelling in many natural language recognition tasks, such as speech (Bennett et al., 2014), *off-line* handwriting (Fischer et al., 2013; Giménez et al., 2014), *on-line* handwriting (Samanta et al., 2014), and audio-visual speech (Sad et al., 2015). Besides, HMM can be used to recognise human facial expressions and its associated emotions (Kung et al., 2016).

Although a few years ago morphological deep learning based models have proven to be better than those based on HMM (Bluche et al., 2014), the technological requirements for working with this kind of models are not always available. This was our case; therefore, we decided to use morphological models based on HMM, and focus our research on improving the results after the decoding process.

The morphological models used in this thesis are composed of a set of HMM, where each HMM represents a basic unit of the words according to the nature of the input signal. Usually, these basic units are phones for ASR and characters for HTR. The morphological models are called acoustical models, optical models, and kinematical models, for ASR, *off-line* HTR, and *on-line* HTR, respectively.

An HMM is a probabilistic finite state model, whose succinct notation is $\lambda = (A, B, \pi)$, and has the following elements (Rabiner, 1989):

- A number of states N.
- A number of different observations per state M.
- A probability distribution of transition between states s_i and s_j $A = \{a_{ij}\}$.
- An emission probability distribution of each observation x for each state s_i $B = \{b_i(x)\}$.
- An initial state probability distribution $\pi = \{\pi_i\}$.

This definition corresponds to a generic HMM, and yields the following HMM standard conditional independence assumptions (Rabiner, 1989):

- **The Markov assumption:** States are only conditionally dependent upon the previous state.
- **The stationarity assumption:** Transition probabilities are conditionally independent of time.
- **The output independence assumption:** Observations are conditionally independent of all other observations and only depend on the state that generated them.

Depending on the function used to define the emission probabilities, the HMM can be discrete or continuous. Usually, for continuous HMM the emission probabilities are defined by universal approximators of Probability Density Functions (PDF), such as Artificial Neural Networks (ANN) (Morgan and Bourlard, 1995) or Gaussian Mixture Models (GMM) (Bilmes, 1998). Figure 2.5 and Figure 2.6 illustrate two examples of continuous HMM, the first one for HTR and the second one for ASR.

In this thesis we have used HMM with GMM as its emission probability distribution. A k-dimensional multivariate Gaussian PDF can be expressed as:

$$\mathrm{P}(x \mid \theta) = (2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \tag{2.3}$$

where $x \in \mathbb{R}^k$ is the feature vector, and the PDF is parameterised by $\theta = [\mu, \Sigma]$ where $\mu \in \mathbb{R}^k$ is the mean vector and $\Sigma \in \mathbb{R}^{k \times k}$ the covariance matrix. Then, the output distribution for each state s_i of an HMM with a single multivariate Gaussian is:

$$b_i(x) = \mathcal{N}(x; \mu_i, \Sigma_i) = \mathrm{P}(x \mid [\mu_i, \Sigma_i]) \tag{2.4}$$

More complex statistical models can be obtained by using a mixture distribution composed of a weighted linear combination of PDF. This is called Gaussian Mixture Model (GMM) when the PDF is

composed of a set of G Gaussians, and mathematically this is expressed as:

$$b_i(x) = \sum_{g=1}^{G} c_{ig} \mathcal{N}(x; \mu_{ig}, \Sigma_{ig}) \tag{2.5}$$

where c_{ig} is the weight for the gaussian g of the state s_i. In order for this to be a valid Probability Mass Function (PMF), these mixture coefficients or weights must satisfy the following constraints:

$$\sum_{g=1}^{G} c_{ig} = 1, \quad c_{ig} \geq 0 \tag{2.6}$$

The Evaluation Process: Forward and Backward Algorithms

Given an HMM λ and a sequence of features $\hat{x} = (x_1, x_2, \ldots, x_T)$, the corresponding morphological likelihood is given by:

$$\mathrm{P}(\hat{x} \mid \lambda) = \sum_{\hat{\theta} \in \hat{\Theta}} \mathrm{P}(\hat{\theta}, \hat{x} \mid \lambda) \tag{2.7}$$

where $\hat{\Theta}$ is the set of all state sequences that generate the feature sequence $\hat{x}$ through the model λ and $\hat{\theta} = (\theta_0, \ldots, \theta_{T+1})$ is a specific state sequence. For $\hat{\theta}$, its associated generation probability $\mathrm{P}(\hat{\theta}, \hat{x} \mid \lambda)$ can be obtained as:

$$\mathrm{P}(\hat{\theta}, \hat{x} \mid \lambda) = a_{\theta_0 \theta_1} \prod_{t=1}^{T} a_{\theta_t \theta_{t+1}} b_{\theta_t}(x_t) \tag{2.8}$$

where θ_0 and θ_{T+1} are the non-emitting initial and final states shown in Figure 2.5 and Figure 2.6.

The *forward* and *backward* algorithms permit to efficiently compute $\mathrm{P}(\hat{x} \mid \lambda)$. As their names indicate, the *forward* algorithm goes forward in temporal axis while the *backward* algorithm goes backward in temporal axis.

The forward function $\alpha_j(t) = \mathrm{P}(\hat{x}, \theta_t = s_j; \lambda)$ represents the probability of the partial observation sequence $\hat{x} = (x_1, x_2, \ldots, x_t)$, when it terminates in an intermediary state s_j, with $1 \leq j \leq N$. Correspondently, the backward function $\beta_i(t) = \mathrm{P}(\hat{x} \mid \theta_t = s_i; \lambda)$ can be defined as the probability of the partial observation sequence $\hat{x} = (x_{t+1}, x_{t+2}, \ldots, x_T)$, given that the current state is s_i, with $1 \leq i \leq N$. These probabilities can be calculated via the following recursions (Baum et al., 1970):

$$\alpha_j(t) = \begin{cases} a_{0j} b_j(x_1) & t = 1 \\ \left(\sum_{i=1}^{N} \alpha_i(t-1) a_{ij} \right) b_j(x_t) & 1 < t \leq T \end{cases} \tag{2.9}$$

with state s_j and the initial condition of $\alpha_0(1) = 1$.

$$\beta_i(t) = \begin{cases} a_{iN} & t = T \\ \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_j(t+1) & 1 \leq t < T \end{cases} \tag{2.10}$$

with state s_i and the initial condition of $\beta_N(T) = 1$.

Then, using the *forward* regression, $\mathrm{P}(\hat{x} \mid \lambda)$ can be calculated as:

$$\mathrm{P}(\hat{x} \mid \lambda) = \sum_{i=1}^{N} \alpha_i(T) a_{iN} = \alpha_N(T) \tag{2.11}$$

Similarly, using the *backward* regression, $P(\hat{x} \mid \lambda)$ can be calculated as:

$$P(\hat{x} \mid \lambda) = \sum_{j=1}^{N} a_{0j} b_j(x_1) \beta_j(1) = \beta_0(1) \tag{2.12}$$

Moreover, the probability $\gamma_i(t)$ of being in a state s_i at time t for any observation sequence $\hat{x} = (x_1, x_2, \ldots, x_t)$ can be obtained easily from the *forward* and *backward* probabilities.

$$\gamma_i(t) = P(\theta_t = s_i \mid \hat{x}; \lambda) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{2.13}$$

The Learning Process: Baum-Welch Algorithm

The parameters of an HMM $\lambda = (A = \{a_{ij}\}, B = \{b_i(x)\})$ can be efficiently estimated from a corpus of training observations using the Baum-Welch algorithm, also called the *forward-backward* algorithm (Welch, 2003). This algorithm is an example of an Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Juang, 1985).

Given a training set of R feature sequences $\hat{X}_R = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_R)$, where each feature sequence $\hat{x}_r = (x_{r1}, x_{r2}, \ldots, x_{rT_r})$ contains T_r features, the HMM parameters can be adjusted following the next steps until convergence:

1. Initialise all the parameters of all HMMs.
2. Get the next training features sequence $\hat{x}_r$ of length T_r.
3. Construct a composite HMM by concatenating the corresponding HMMs to the transcription of $\hat{x}_r$.
4. Calculate the *forward* -Equation (2.9)- and *backward* -Equation (2.10)- probabilities for the composite HMM.
5. Calculate the occupation probabilities -Equation (2.13)- for the state s_j at each time frame t and update the corresponding mean μ_i and variance Σ_i.
6. Repeat the steps [2-5] for all the training set R.
7. Use the obtained means μ and variances Σ to estimate the final parameters for all of the HMMs.

Then, the state-transition probabilities can be re-estimated with the following equation:

$$\hat{a}_{ij} = \frac{\sum_{r=1}^{R} P(\hat{x}_r \mid \lambda)^{-1} \sum_{t=1}^{T_r} \alpha_i^r(t) a_{ij} b_j(x_{r(t+1)}) \beta_j^r(t+1)}{\sum_{r=1}^{R} P(\hat{x}_r \mid \lambda)^{-1} \sum_{t=1}^{T_r} \alpha_i^r(t) \beta_i^r(t)} \tag{2.14}$$

where $1 \leq i \leq N$ and $1 \leq j \leq N$.

When the emission probability distribution of each state s_i of the HMM is modelled as a sum of G Gaussian distributions $b_i(x) = (c_{ig}, \mu_{ig}, \Sigma_{ig})$, the probability that the feature sequence $\hat{x}_r \in \mathbb{R}^{T_r}$ be generated by the Gaussian component g in the state s_i can be defined as:

$$\gamma_{ig}^r(t) = \gamma_i(t) \left[\frac{c_{ig} \mathcal{N}(x_{rt}; \mu_{ig}, \Sigma_{ig})}{P(x_{rt} \mid \lambda)} \right] \tag{2.15}$$

Then, the new set of Gaussian parameters ($\hat{c}_{ig}$, $\hat{\mu}_{ig}$, and $\hat{\Sigma}_{ig}$) can be re-estimated by:

$$\hat{c}_{jg} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jg}^r(t)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{l=1}^{G} \gamma_{jl}^r(t)} \tag{2.16}$$

$$\hat{\mu}_{jg} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{jg}^{r}(t)x_{rt}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{jg}^{r}(t)} \tag{2.17}$$

$$\hat{\Sigma}_{jg} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{jg}^{r}(t)(x_{rt}-\hat{\mu}_{jg})(x_{rt}-\hat{\mu}_{jg})^{\top}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{ig}^{r}(t)} \tag{2.18}$$

Morphological Model Adaptation

Training the morphological models requires a large amount of training samples. Depending on the relation between the speakers/writers used to obtain the samples for training the morphological models and the speakers/writers used for decoding, morphological models can be classified into the following three categories (Woodland, 2001):

- **Speaker/writer independent:** These models are trained using training samples obtained from a great variety of speakers/writers; in this way, the charge of producing the training samples is shared, and the obtained models can generalise the speech/handwriting. These models can be used to recognise the speech/handwriting of new users with good performance.
- **Speaker/writer dependent:** These models are trained to recognise the speech/handwriting of a specific user, i.e. all the training samples are produced by the objective user. Given that these models are tailored to a particular person, they give better results than speaker/writer independent models for this particular person. However, such morphological models usually have a poor performance for recognising the speech/handwriting of new users and require a big acquisition effort from the final user.
- **Speaker/writer adaptive:** These models are obtained tuning speaker/writer independent models to fit the speech/handwriting of a specific person using relatively few samples. This approach allows to obtain significant improvements over the speaker/writer independent models, reaching similar results to those offered by speaker/writer dependent models for a particular person, but only using a small amount of training samples of this specific person.

The adaptation of morphological models based on continuous density HMM (Woodland, 2001) can be performed by means of, among others, Maximum *A Posteriori* (MAP) adaptation (Gauvain and Lee, 1994), linear transformation of model parameters such as Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995), and speaker space methods, such as Cluster Adaptive Training (CAT) (Gales, 2000) and Eigenvoice (Kuhn et al., 2000).

In this thesis, writer dependent optical models were trained for *off-line* HTR given that the historical manuscripts (Serrano et al., 2010; Alabau et al., 2011) used in the experimentation are mono-writer, whilst as the speech and the *on-line* handwriting by pen strokes are used as user feedback modalities, the ASR acoustical and *on-line* HTR kinematical models were trained as speaker/writer independent. Besides, for some experiments, these speaker independent acoustical models were adapted to each one of the test speakers by using the Maximum Likelihood Linear Regression (MLLR) technique (Leggetter and Woodland, 1995).

MLLR reduces the mismatch between the morphological model and the adaptation data by computing a set of linear transformations to shift the means and alter the variances in the morphological model (Gales and Woodland, 1996). Given a Gaussian distribution g characterised by a mean vector μ_g, and a covariance matrix Σ_g, the adapted mean vector $\hat{\mu}_g$ is obtained by:

$$\hat{\mu}_g = A\mu_g + b = W_g\xi_g \tag{2.19}$$

where W_g is the $n \times (n+1)$ transformation matrix which maximises the likelihood of the adaptation data of dimensionality n that can be decomposed in:

$$W_g = [b, A] \tag{2.20}$$

where A represents an $n \times n$ transformation matrix, b represents a bias vector, and ξ_g is the extended mean vector, defined as:

$$\xi_g = \left[w, \mu_{g1}, \mu_{g2}, \ldots, \mu_{gn}\right]^\top \tag{2.21}$$

where w is the offset term for the regression (w=1 to include and w=0 to ignore offsets in the regression).

The adapted covariance matrix $\hat{\Sigma}_g$ may be obtained using either the *normalized-full* approach (Gales, 1998), as:

$$\hat{\Sigma}_g = LHL^\top \tag{2.22}$$

where L is the Choleski factor of the original covariance matrix Σ_g, or the *efficient-full* approach (Gales, 1998) as:

$$\hat{\Sigma}_g = H\Sigma_g H^\top \tag{2.23}$$

In both cases, H represents a transformation matrix that can be obtained after the adapted mean vectors $\hat{\mu}_g$ have been estimated (Gales, 1998), as following:

$$H = \frac{\sum_{g=1}^{G} \left\{ L_{g-1}^\top \left[\sum_{\tau=1}^{T} \gamma_g(\tau)(o(\tau) - \hat{\mu}_g)(o(\tau) - \hat{\mu}_g)^\top \right] L_{g-1} \right\}}{\sum_{g=1}^{G} \sum_{\tau=1}^{T} \gamma_g(\tau)} \tag{2.24}$$

where, $\gamma_g(\tau)$ is the *a posteriori* probability determined by the Gaussian component g of the original model at time τ, and $O_T = (o(1), o(2), \ldots, o(T))$ represents the adaptation data.

The complexity of the transformation matrices W_g and H can be determined as full, block–diagonal or diagonal. These transformation matrices are estimated to maximise the likelihood of the adapted models generating the adaptation data by using the Expectation-Maximisation (EM) algorithm (Juang, 1985). Therefore, in addition to the complexity of the transformation matrices, the number of transforms or regression classes may be estimated in order to obtain the best adaptation (Leggetter and Woodland, 1995).

Finally, the main difference between the *normalized-full* and the *efficient-full* variance adaptation approaches is that, while the computation of the *normalized-full* variance transform requires a considerably lower computational cost than the *efficient-full* approach, the likelihood calculation during recognition is faster for *efficient–full* variance transform than for the *normalized-full* one (Gales, 1998).

2.3.2 Language Modelling

In Language Models (LM), the text properties are modelled independently from the morphological models (Marti and Bunke, 2001), i.e. the language model defines all the possible sentences that can be recognised by the system. In addition of the language syntax, the semantic information of the words can be modelled (Romero Gómez, 2010). Language models can be implemented through probabilistic grammars (Suppes, 1970), n-grams (Manning and Schütze, 1999), Finite State Transducers (FST) (Mohri, 1997), and Recurrent Neural Networks (RNN) (Mikolov et al., 2010), among others.

Language modelling based on RNN allows one to model short- and long-term contextual information. This is one of the qualities by which its use (with great success) has become popular in recent research (Arisoy et al., 2015; Zuo et al., 2016). However, in spite of the fact that smoothed n-grams only model short term dependencies, they allow one to obtain a good performance (Romero Gómez, 2010),

and nowadays they are still one of the most wide-spread language models used in natural language processing tasks (Al-Khoury, 2015). In this thesis, language models were implemented as smoothed n-grams.

N-gram Language Models

The *a priori* probability $\mathrm{P}(\hat{w})$ of a word sequence $\hat{w} = w_1, \ldots, w_T$ required in Equation (2.1) is computed as:

$$\mathrm{P}(\hat{w}) = \prod_{t=1}^{T} \mathrm{P}(w_t \mid w_1, \ldots, w_{t-1}) \tag{2.25}$$

where $\mathrm{P}(w_t \mid w_1, \ldots, w_{t-1})$ is the probability of the word w_t after seeing the prior word sequence $w_1, \ldots, w_{t-1}$, which is called the history.

In n-gram models for large vocabulary recognition, the probability $\mathrm{P}(\hat{w})$ of observing the word sequence $\hat{w}$ is approximated by truncating the conditioning word history to $n-1$ words:

$$\mathrm{P}(\hat{w}) \approx \prod_{t=1}^{T} \mathrm{P}(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1}) \tag{2.26}$$

Although the best length n for n-gram language models can be estimated by cross-validation (Kohavi, 1995), in this work, bigrams ($n = 2$) were used based on previous works related with the *Rodrigo* and *Cristo Salvador* corpora (Serrano et al., 2010; Alabau et al., 2011).

The n-gram probabilities estimation can be performed from training text by using the Maximum Likelihood estimation (Scholz, 1985). The conditional probability of a word w_t given a history $h_t = w_{t-(n-1)}, \ldots, w_{t-1}$ can be calculated from the n-gram frequency counts as:

$$\mathrm{P}(w_t \mid h_t) \approx \frac{C(h_t, w_t)}{C(h_t)} \tag{2.27}$$

where $C(h_t, w_t)$ and $C(h_t)$ represent the number of occurrences in the training set of $w_{t-(n-1)}, \ldots, w_{t-1}, w_t$ and $w_{t-(n-1)}, \ldots, w_{t-1}$ respectively.

This method assigns null probability to all unseen events. This data sparsity problem can be mitigated by smoothing the model by a combination of discounting and backing-off. There exist several approaches, such as Laplacian smoothing (Zhai and Lafferty, 2004), which assigns the counts to 1 for unseen n-grams, and techniques based on the Good-Turing discounting (Good, 1953). Good-Turing discounting coefficients can be obtained as follows:

$$\lambda(h, w) = (r+1)\frac{N_{r+1}}{rN_r} \tag{2.28}$$

where N_r is the number of n-grams that have appeared $r = C(h, w)$ times in the training data. This discounting factor is used, for example, in the back-off method. In this method, the unseen n-grams are approximated by a weighted version of the corresponding n-gram, and the discounted probability is distributed over the unseen n-grams as follows:

$$\tilde{\mathrm{P}}(w \mid h) = \begin{cases} \lambda(h, w)\mathrm{P}(w \mid h) & \text{if } C(h, w) > 0 \\ \Gamma(h)\beta(w \mid h) & \text{if } C(h, w) = 0 \end{cases} \tag{2.29}$$

where $\beta(w \mid h)$ is some less specific distribution (usually a *(n-1)*-gram), and $\Gamma(h)$ is the normalised total amount of discounted probability given the history h, that is:

$$\Gamma(h) = \frac{1 - \sum_{w:C(h,w)>0} \lambda(h, w)\mathrm{P}(w \mid h)}{\sum_{w:C(h,w)=0} \beta(w \mid h)} \tag{2.30}$$

There are different methods based on the Good-Turing discounting (Chen and Goodman, 1999). One of the most effective is the Kneser-Ney back-off smoothing method (Kneser and Ney, 1995). In this method, the concept of absolute-discounting interpolation is used incorporating information from the higher- and the lower-order n-grams. When the counts of the higher-order n-gram is near zero, the lower-order n-gram adds more weight to the probability and vice versa.

2.3.3 Lexicon Modelling

The lexicon model indicates to the recognition system the basic units (usually, characters for HTR and phonemes for ASR) of each one of the words that make up the language model. In this way, the lexicon model links the morphological level representation with the word sequence output (Adda-Decker and Lamel, 2000). In the lexicon model the transcription (pronunciation or spelling) of each word w is modelled as a sequence $\hat{b} = (b_1, b_2, \ldots, b_n)$ of n basic sub-word units. In this model multiple transcriptions are allowed (Wooters and Stolcke, 1994; Yun and Oh, 1999); if it is the case, the probability $\mathrm{P}(\hat{x} \mid \hat{w})$ of Equation (2.1) can be computed over the different transcriptions.

$$\mathrm{P}(\hat{x} \mid \hat{w}) = \sum_{\hat{b} \in \hat{B}_{\hat{w}}} \mathrm{P}(\hat{x} \mid \hat{b}) \mathrm{P}(\hat{b} \mid \hat{w}) \tag{2.31}$$

where $\hat{B}_{\hat{w}}$ denotes the set of all valid transcriptions $\hat{b}$ for the word sequence $\hat{w}$.

Usually, these models are implemented as Deterministic Finite Automaton (DFA) (Lucchesi and Kowaltowski, 1993; Ciura and Deorowicz, 2001). For example, the Spanish word *Frijolito* would be generated by the finite state model for ASR presented in Figure 2.7.

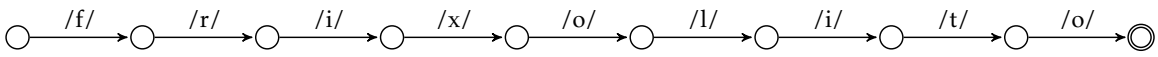


Figure 2.7: Finite state automata for the Spanish word *Frijolito* [fri xo 'li to] in a lexicon for ASR.

Lexicon models for HTR can be build easily by splitting the words into their constituent character sequences. However, given that words are composed of phonemes in spoken language, expert knowledge is required to build the lexicon models for ASR. These lexicon models can be made manually by annotating the phonetic transcription of each word (Riley et al., 1999), or automatically by using systems based on rules (Cremelie and Martens, 1997). As *automatic phonotisation* is treated in Text-to-Speech synthesis (Dutoit and Stylianou, 2003; Taylor, 2009), some tools developed for speech synthesis can be useful to build ASR lexicon models. One example is eSpeak (Duddington, 1995) which is an open source speech synthesiser that allows one to obtain phonetic transcriptions in the International Phonetic Alphabet (IPA) (International Phonetic Association, 1999) for several languages.

Nevertheless, in the last years some works on lexicon free ASR (Harwath and Glass, 2014; Maas et al., 2015) have appeared. Eliminating the need for a lexicon, the required expert knowledge is considerable reduced, and the speech recognition capabilities are expanded; for instance, it allows to transcribe new words, word fragments, and disfluencies.

In this thesis, the lexicon models for HTR were obtained by splitting the words into their constituent character sequences. Similarly, for the lexicon models for ASR an initial phonetical transcription for each word was obtained automatically based on rules for current Spanish. Then, all phonetical transcriptions were supervised in order to adjust their accuracy to the actual word, in spite of the medieval scripting where heterographs (one sound represented by several spellings) are frequently used (Llamas Pombo, 2012). For instance, this is the case of the word *cristianos*, which could be written also as *xp̃ianos* and *christianos*, while the phonetical transcription for the three forms is [kris 'tja nos].

2.4 Decoding

As previously said, the goal of natural language recognition systems is to obtain the transcription of a determined sequence of features vectors $\hat{x}$ as a sequence of words $\hat{\mathbf{w}}$ given a lexicon (Toselli et al., 2004). Most natural language recognition systems are composed of three statistical models (see Section 2.3). Therefore, the fundamental formulation of statistical natural language decoding or recognition (Equation (2.1)) can be represented as follows:

$$\hat{\mathbf{w}} = \arg\max_{\hat{w}\in\hat{W}} \max_{\hat{b}\in\hat{B}_{\hat{w}}} \mathrm{P}(\hat{x} \mid \hat{b})\mathrm{P}(\hat{b} \mid \hat{w})\mathrm{P}(\hat{w}) \tag{2.32}$$

where $\hat{W}$ denotes the set of available word sequences, $\hat{B}_{\hat{w}}$ the set of different transcriptions of the word sequence $\hat{w}$, $\mathrm{P}(\hat{x} \mid \hat{b})$ is modelled by the morphological model, $\mathrm{P}(\hat{b} \mid \hat{w})$ by the lexicon model, and $\mathrm{P}(\hat{w})$ by the language model.

2.4.1 The Viterbi Algorithm

For natural language recognition systems based on HMM, the efficient recursive *forward* algorithm for computing the forward probabilities also allows to obtain the total morphological likelihood $\mathrm{P}(\hat{x} \mid \hat{b})$. Thus, if the sum in Equation (2.9) is replaced by the dominating term, this algorithm could also be used to find the sequence of states which yields the maximum value of $\mathrm{P}(\hat{x} \mid \hat{b})$, i.e. finding the most likely state sequence $\hat{\theta} = (\theta_1, \theta_2, \ldots, \theta_T)$ within the finite-state search space of the morphological model (conditioned by the language and lexicon models) for a given sequence of observations $\hat{x} = (x_1, x_2, \ldots, x_T)$. Hence, this approximation could be used for recognition, and it is commonly called the Viterbi algorithm (Viterbi, 1967; Jelinek, 1998).

The Viterbi algorithm consists of the search of the best path in a trellis of states and time (Soong and Huang, 1991), where the search space consists of a finite-state network of HMM states (obtained from a network of basic sub-word units and a network of words), given a time sequence of feature vectors. Figure 2.8 presents an example of state-time trellis for an ASR system. Each dot represents the observation probability for each state given a temporal feature vector, and the transition probabilities are represented by the edges between dots. The transitions allowed by the morphological model are highlighted in black, the transitions allowed by the lexical model in red, and the transitions allowed by the language model in blue.

In a forward pass, the most likely path is found by keeping the most probable predecessor that reaches each state-time point. Then, the obtained most likely path is traced backwards from the final state for obtaining the most likely sequence of words. For instance, the most probable path is highlighted by dashed backward arrows in the example presented in Figure 2.8. The total score of each path or hypothesis $\hat{w}$ is composed of three parts, the probability of the lexical-morphological models, the probability obtained by the language model scaled by a Grammar Scale Factor (GSF), and a Word Insertion Penalty (WIP) factor to avoid bias towards large or small words. Therefore, in the logarithmic domain the total likelihood for each path is obtained as follows (Gales and Young, 2008):

$$\log \mathrm{P}(\hat{w} \mid \hat{x}) = \log \mathrm{P}(\hat{x} \mid \hat{w}) + \mathrm{GSF} \cdot \log \mathrm{P}(\hat{w}) + \mathrm{WIP} \cdot |\hat{w}| \tag{2.33}$$

where $|\hat{w}|$ represents the number of words in $\hat{w}$.

The performance of the decoding process can be improved if it is not restricted to the single best hypothesis offered by the Viterbi approach. Instead, it is used to obtain a set of the n-best hypotheses. However, the language and the lexicon models greatly expand the search space (Gales and Young, 2008). To deal with this problem, the search space can be expanded dynamically by beam-search techniques as the search progresses (Haeb-Umbach and Ney, 1994).

The global beam pruning is a heuristic state-space search algorithm, such as breadth-first search, best-first search or depth-first search (Zhang, 1999). This pruning technique builds a search tree, where

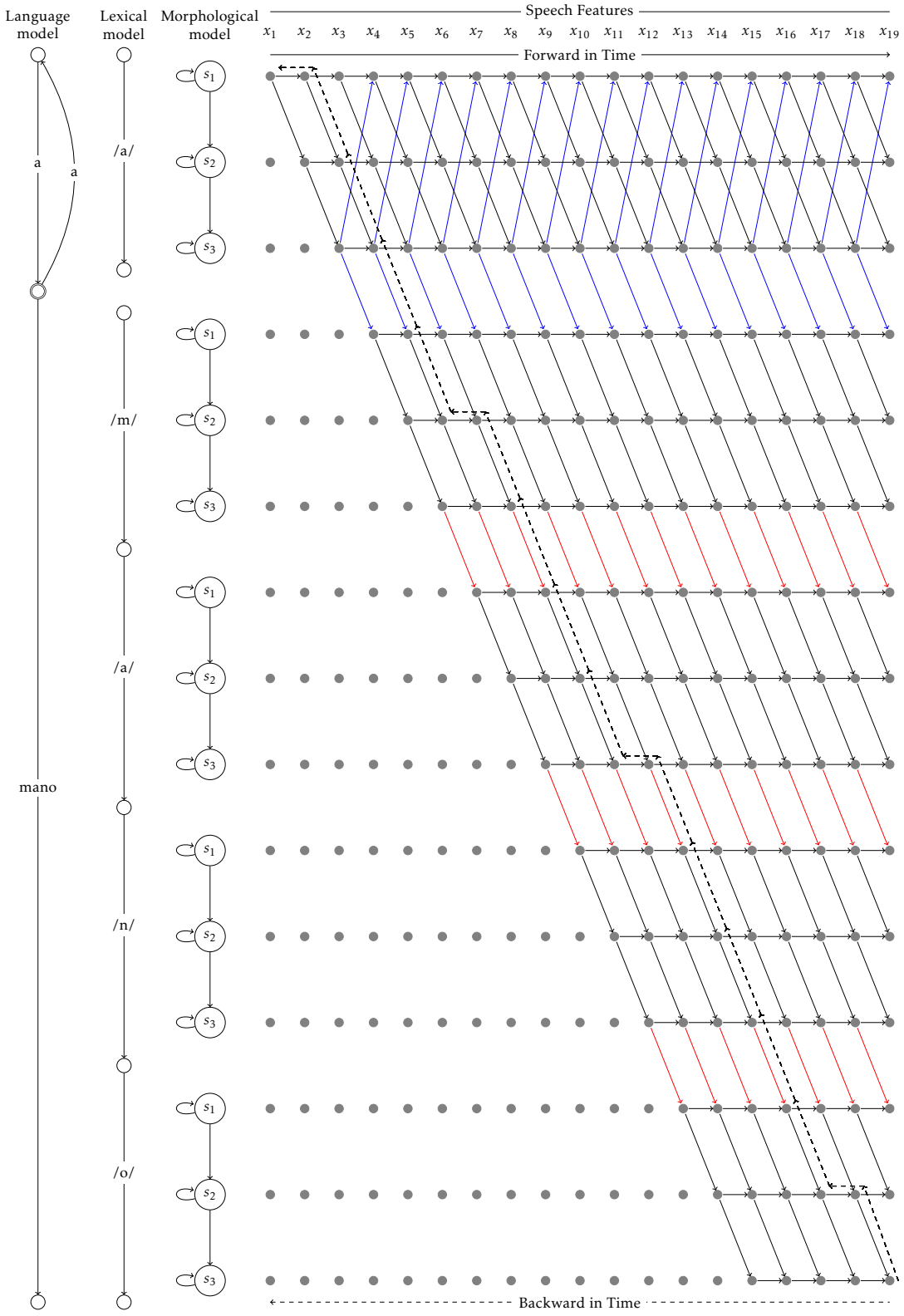


Figure 2.8: State-time trellis of the observation sequence for an ASR system. In this trellis, in black are presented the transitions allowed by the morphological model, in red the transition of the lexical model, and in blue the transitions of the language model. The most likely path is highlighted by the dashed backward arrows.

at each state, the search is limited to the paths with a likelihood score close to the best partial path hypothesis.

Formally, the likelihood of the best partial path hypothesis that ends at time t in state s with word history $\hat{w}$ can be defined as (Ortmanns et al., 1997; Pylkkönen, 2005):

$$\mathrm{P}_{GB}(t) = \max_{(\hat{w}\in\hat{W}, s\in\lambda)} \mathrm{P}(\hat{w} \mid t, s) \tag{2.34}$$

where $\hat{W}$ denotes the set of available word sequences and λ the HMM parameters.

Then, at each time t and state s, in the logarithmic domain, the paths with a log likelihood lower than a determined threshold f_{GB} are pruned:

$$\log \mathrm{P}(\hat{w} \mid t, s) < \log \mathrm{P}_{GB}(t) - f_{GB} \tag{2.35}$$

f_{GB} controls the beam width that limits the search space, given that at each tree level, only the states included in the beam are expanded. In the rest of the document, f_{GB} is denominated as *beam factor*.

Given the fact that the GSF, WIP and *beam factor* parameters have a significant effect on the decoding performance, they usually have to be adjusted by using some validation data.

In any case, the set of n-best hypotheses obtained from the recognition process can be represented into different formats. For example, this output can be compactly represented as lattices (Soong and Huang, 1991; Oerder and Ney, 1993). Section 2.4.2 gives details on the output formats used in this thesis.

2.4.2 Recognition Output Formats

Natural language decoding n-best solutions can be compactly represented into lattices, structured as Word Graphs or as Confusion Networks (Jurafsky and Martin, 2009).

A Word Graph (WG) (Figure 2.9) is a directed, acyclic and weighted graph that represents a huge set of hypotheses in a very efficient way (Ljolje et al., 1999). The nodes in a WG correspond to horizontal positions for *off-line* HTR, and discrete time points for ASR and *on-line* HTR. The edges are labelled with words and weighted with the morphological, lexical, and language likelihoods of the word that appears in the signal delimited between the starting and ending nodes of the edge. The likelihoods are derived from the morphological and language models during the decoding process (Bahl et al., 1983; Soong and Huang, 1991).

On the other hand, a Confusion Network (CN) (Figure 2.10) is also a directed, acyclic and weighted graph that shows at each point which word hypotheses are competing or confusable. Each hypothesis goes through all the nodes. The words and their probabilities are stored in the edges, and the total probability of the words contained in a subnetwork (all edges between two consecutive nodes) sum up to 1 (Mangu et al., 2000). On each subnetwork, one special word (*DELETE*) can be inserted to allow hypotheses having different lengths.

Confusion Networks reduce the complexity of Word Graphs losing the segmentation information (Xue and Zhao, 2005). However, a CN contains all hypotheses of the original WG which is originated from together with new hypotheses. These new hypotheses are originated due to the CN structure and the special word (*DELETE*). The properties of CN allow their use for many tasks, such as lattice compression (Mangu and Brill, 1999), word spotting (Mangu et al., 2014), machine translation (Bertoldi et al., 2007), confidence annotation, and system combination (Evermann and Woodland, 2000).

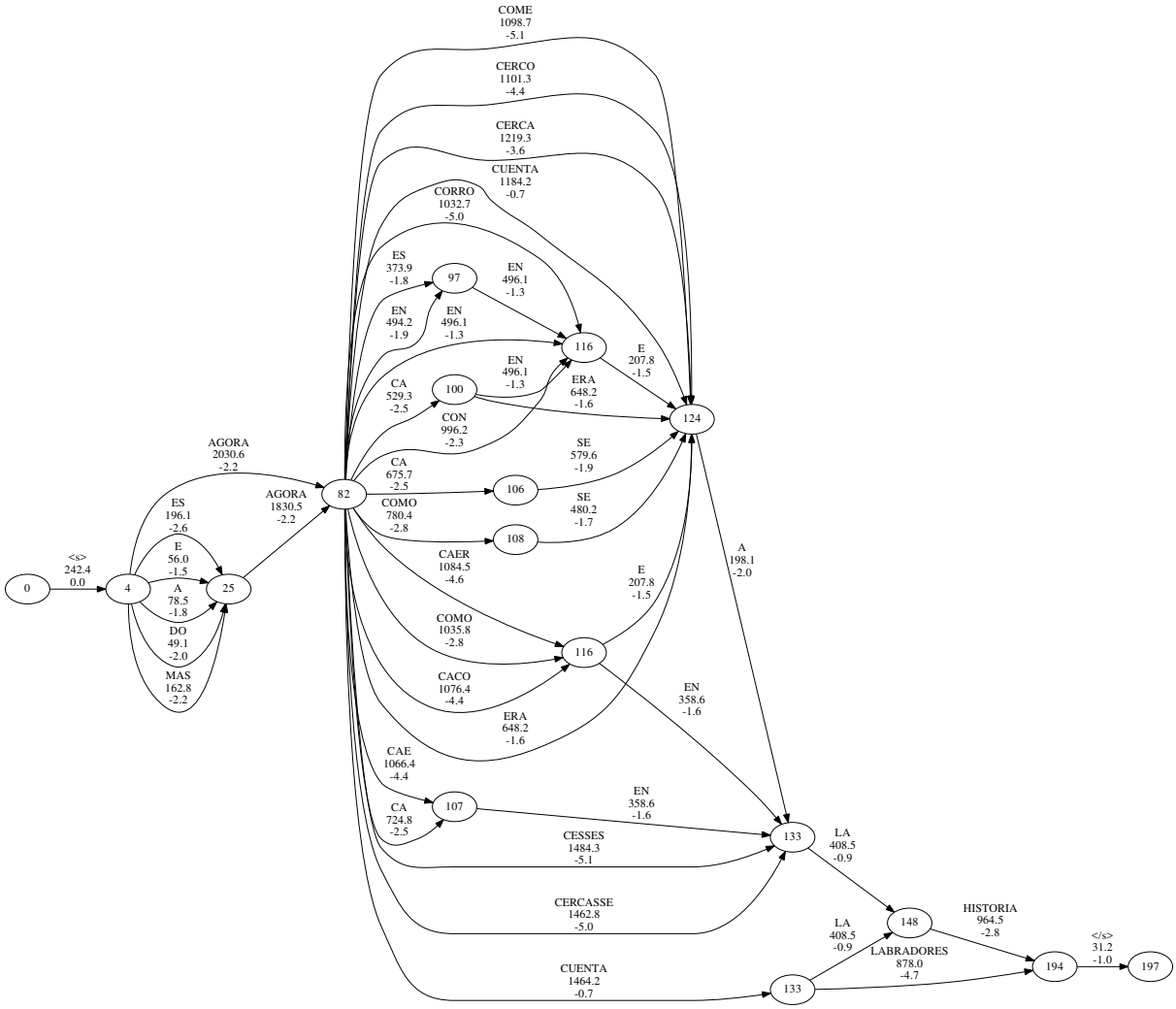


Figure 2.9: Lattice as a Word Graph. Nodes correspond with frame positions, and each edge contains a word and its corresponding likelihoods (morphological likelihood at top, and language likelihood at bottom).

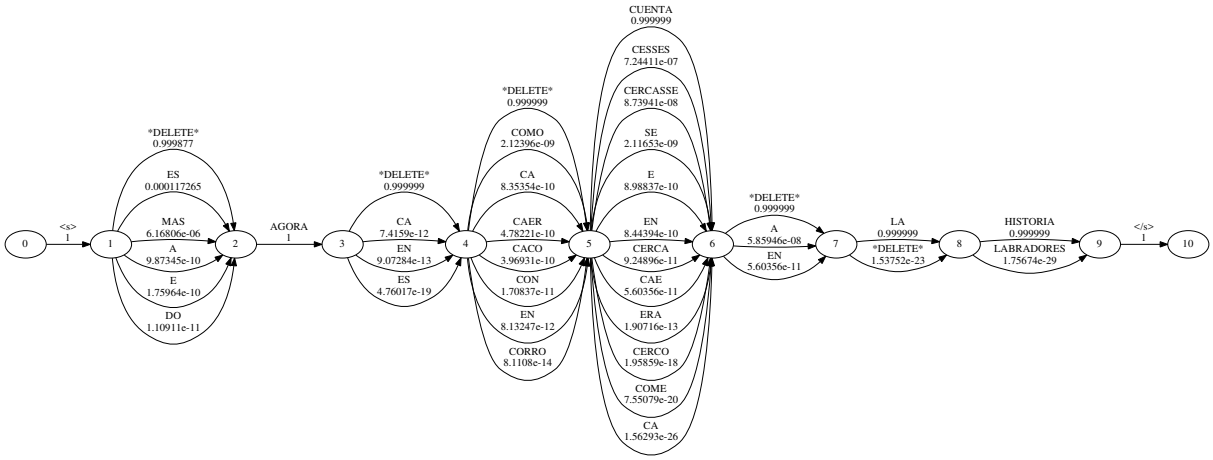


Figure 2.10: Lattice as a Confusion Network. Nodes represent the connections in a sequence of words, and each edge represents a word with the likelihood of this word being in this position of the sequence.

2.5 Assistive Transcription of Historical Manuscripts

HTR systems can be seen as part of an assistive technology. Assistive technologies have been of traditional use in many fields of computer applications, such as Computer Aided Design (CAD) field (Machover, 1995), medical diagnosis (Doi, 2007), automatic driving (Malit, 2009), Computational Linguistics/Natural Language Processing (CL/NLP) (Barrachina et al., 2009; Revuelta-Martínez et al., 2012; Silvestre-Cerdà et al., 2013), and Pattern Recognition and Image Processing (Romero et al., 2012).

In these tasks, the computer allows the human user to have easier and faster work, providing the final user with a series of tools that allow the user to speed-up the process. Among these tools appear some automatic processing elements, such as medical data analysers, processors for data from driving sensors, speech and handwritten text recognisers, image feature extractors, etc. Apart from that, these systems need an interface that allows the user to amend the possible errors obtained by the automatic process. The interface tool could provide, by using underlying systems that employ the results of the automatic process and user feedback, autonomous actions that avoid the user to perform some of the corrections.

Consequently, the main objective in these systems is not obtaining the most accurate result from the automatic system, but achieving the lowest effort for the human user (although both facts could be correlated). This requires new evaluation measures and frameworks that follow this criterion (minimising user effort) and are adapted to the corresponding task. For example, for the transcription of speech or handwritten text, the number of corrective actions that the user has to perform (taking into account automatic corrections given by the system) is a good measure of the effort.

In this assistive context, the multimodal paradigm arose as a new form of improving these systems by reducing the final user effort. The multimodal paradigm has experimented a spectacular growth in the latest years because of the development of mobile devices (Di Fabbrizio et al., 2009), where different modalities (speech and touch mainly) are employed for the device management. In the case of Image or Natural Language Processing tasks, multimodality has been applied to problems where signals of different nature that represent the same final object are available (Mihalcea, 2012; Potamianos et al., 2003; Sebe et al., 2005; Granell and Martínez-Hinarejos, 2015). In any case, multimodality is strongly linked to Human-Computer Interaction (HCI), since the user may employ different modalities to obtain a more ergonomic or faster interaction to achieve an objective.

One interesting computer assisted application where multimodality can provide productivity improvements is the transcription of handwritten documents (Gordo et al., 2008). In this case, the assistive system provides the final users an initial draft transcription of the handwritten image. Then, the system can supply alternative transcriptions every time the user makes an amendment, with the final aim of reducing the user effort to obtain a perfect transcription. An example of assistive framework that presents these features is the Computer Assisted Transcription of Text Images (CATTI) system presented in (Romero et al., 2012).

The CATTI system takes as input the text image to be transcribed. This data is employed to offer the user a first hypothesis and to search for alternatives when the user makes a correction, usually by employing an HTR system. The multimodality can be incorporated in CATTI by providing another signal that represents the same sequence of words, e.g., a speech dictation of the text that can be processed by an Automatic Speech Recognition (ASR) engine and gives as a result different alternatives. Since both HTR and ASR systems employ similar models (Hidden Markov Models - HMM - and n-grams) and can obtain results in a similar format, its combination seems feasible despite the different nature of the signals and their asynchrony.

Additionally, the user must provide feedback several times to the CATTI system, independently of the initial transcription given by the available data sources. Although the number of interactions may change depending on these initial sources, making the interaction process comfortable to the user is crucial for the success of an interactive system. Since paleographers usually employ touchscreen tablets for their task, using touchscreen pen strokes to provide feedback appears an appropriate interactive option (Romero et al., 2012; Martín-Albo et al., 2013).

2.6 Crowdsourcing for Natural Language Processing Tasks

In (Estellés-Arolas and González-Ladrón-De-Guevara, 2012) crowdsourcing is defined as follows: "Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organisation, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilise to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."

Crowdsourcing can be used in different tasks related to natural language processing (Parent and Eskenazi, 2011), such as: speech acquisition, document transcription and annotation, and the assessment of natural language processing technology. Regarding the document transcription task, the appearing of crowdsourcing platforms (Doan et al., 2011) has had a strong impact on the transcriber's task. In these platforms, many volunteers provide a transcription of a digital document at a very small (or even null) cost; the inherent difficulties of those documents make necessary the posterior revision of the professional transcriber, but the workload is considerably lower than that of scratch transcription. There are several generic crowdsourcing platforms available, such as Mechanical Turk[1] or CrowdFlower[2], but for handwritten text transcription (and in particular for historical text) several platforms have been developed in the last years (such as AnnoTate[3], Transcribe Bentham[4], or Transkribus[5]).

Crowdsourcing approaches to the acquisition of speech data have become really popular in the last decade. In (Parent and Eskenazi, 2011), a review on different works based on crowdsourcing reveal a high number of research articles (29) and experiments (37) in the topic. Works such as that of (Caines et al., 2016) reveal the feasibility of the acquisition of speech corpora by using mobile devices and the capacity of the crowdsourcing framework to obtain annotated speech corpora at several levels.

2.7 Evaluation Measures

In all the experimental work presented in this thesis the text reference was available in order to automatically evaluate the proposed solutions. There are different types of evaluation measures depending on the task: classification, recognition, assistive transcription, etc. In the following, the different evaluation measures used in the experiments of this thesis are presented.

2.7.1 Natural Language Recognition Evaluation

Classification Error Rate

Classification Error Rate (ER) permits one to measure the percentage of samples incorrectly classified, and it is calculated as:

$$\mathrm{ER} = \frac{f}{n} \cdot 100 \tag{2.36}$$

[1] https://www.mturk.com/
[2] https://www.crowdflower.com/
[3] https://anno.tate.org.uk/
[4] http://blogs.ucl.ac.uk/transcribe-bentham/
[5] https://transkribus.eu/Transkribus/

where, f is the number of samples incorrectly classified, and n is the total number of samples. This measure is usually used in tasks such as isolated word recognition, where each word is treated as a class, and it is classified as correct or incorrect.

Word Error Rate

The quality of the obtained transcriptions in the decoding process of natural language recognition systems (speech and handwriting) can be evaluated using the edition distance with respect to a reference text. This edition distance is derived from the Levenshtein distance (Levenshtein, 1966), and it is defined as the minimum number of substitutions, deletions and insertions needed to transform the transcription into the reference, divided by the total number of words in the reference. Therefore, Word Error Rate (WER) is this edition distance at word level, and can be calculated as follows:

$$\mathrm{WER} = \frac{s+d+i}{n} \cdot 100 \tag{2.37}$$

where, s is the number of substitutions, d the number of deletions, i the number of insertions, and n is the total number of words in the reference.

Additionally, when the decoding output is not limited to the most likely hypothesis, such as a n-best list, the best transcription that can be obtained from this output is measured, at word level, by the oracle WER.

This measure can be used to evaluate the transcription error at different levels, depending on the type of linguistic unit taken into consideration. Words are a common unit in speech and handwriting recognition systems. However, words can be composed of phonemes (speech) or of characters (handwriting), depending on the recognition system. Therefore, this measure can be also used at phoneme level -Phoneme Error Rate (PER)- and at character level -Character Error Rate (CER)-.

2.7.2 Language Model Evaluation

Perplexity

As language models are probability distributions over entire sentences or texts, Perplexity (PPL) (Brown et al., 1992) can be used for evaluating their performance over a reference text. Perplexity is a measure in standard use from the field of information theory, and it is defined as the exponential of the entropy.

Given a sequence of words $\hat{w} = w_1, w_2, \ldots, w_N$, the entropy (H) for this reference text, can be approximated with:

$$\mathrm{H} = -\frac{1}{N} \log_2 \mathrm{P}(\hat{w}) \tag{2.38}$$

where N is the number of words in the sequence $\hat{w}$ and $\mathrm{P}(\hat{w})$ is the probability assigned to the word sequence $\hat{w}$ by the evaluated language model. Then, perplexity can be estimated as follows (Rabiner and Juang, 1993):

$$\mathrm{PPL} = 2^{\mathrm{H}} = \mathrm{P}(\hat{w})^{-\frac{1}{N}} \tag{2.39}$$

Specifically, perplexity can be estimated for n-gram language models as follows:

$$\mathrm{PPL} = \prod_{i=1}^{N} \mathrm{P}(w_i \mid w_{i-n}, \ldots, w_{i-1})^{-\frac{1}{N}} \tag{2.40}$$

where n is the order of the n-gram model.

Perplexity can be considered to be a measure of, on average, the number of equally most probable words that can follow any given word. Therefore, lower perplexities represent better language models when comparing the performance of different language models over the same reference text.

2.7.3 Computer Assisted Transcription Evaluation

Given that on interactive assistive approaches for transcription the user and the system work together to obtain the perfect transcription, the following measures were used to assess the performance of the interactive system, and to estimate the human effort reduction.

Word Stroke Ratio

The performance of computer assisted transcription systems can be measured by the Word Stroke Ratio (WSR), which can be computed using the reference transcription. After each user interaction, the longest common prefix between the hypothesis and the reference is obtained and the first unmatching word from the hypothesis is replaced by the corresponding reference word. This process is iterated until a full match is achieved. Therefore, the WSR can be defined as the number of user interactions that are necessary to produce correct transcriptions using the assistive system, divided by the total number of reference words.

$$\mathrm{WSR} = \frac{i}{n} \cdot 100 \tag{2.41}$$

where i is the number of user interactions and n is the total number of words in the reference.

Effort Reduction

The main objective of computer assisted transcription systems is to reduce the required human effort for obtaining the actual transcription. Given the definition of WER and WSR, the relative difference between them can be used to estimate the human Effort Reduction (EFR) that can be achieved by using the assistive system with respect to using a conventional handwriting text recognition system followed by human post-editing.

$$\mathrm{EFR} = \frac{\mathrm{WER} - \mathrm{WSR}}{\mathrm{WER}} \cdot 100 \tag{2.42}$$

2.7.4 Multimodal Crowdsourcing

In order to avoid using collaborations that could worsen the transcriptions, the multimodal crowdsourcing framework proposed in Part IV must verify the reliability of the hypotheses obtained from the decoding processes. The reliability verification is also useful to optimise the collaboration effort. Therefore, in this multimodal crowdsourcing framework the following measures were used to assess the decoding reliability, and the collaboration effort.

Decoding Reliability

The statistical formulation of the natural language recognition problem -see Equation (2.1)-, allows one to take the posterior probability $\mathrm{P}(\hat{w} \mid \hat{x})$ as a good confidence measure for the recognition reliability. However, generative classifiers (such as classifiers based on HMM) provide joint probabilities $\mathrm{P}(\hat{x}, \hat{w})$ (Ng and Jordan, 2001) that are inadequate to obtain this reliability.

Nevertheless, when the recognition scores of a fairly large n-best list can be re-normalised to sum up to 1, the joint probability $\mathrm{P}(\hat{x}, \hat{w})$ obtained for the best hypothesis can be used as a good confidence measure, since it is a measure of the match between $\hat{x}$ and $\hat{w}$ (Rueber, 1997). Therefore, the decoding reliability R is measured by the re-normalised 1-best joint probability:

$$R = \frac{\max\limits_{\hat{w} \in \hat{W}} \mathrm{P}(\hat{x}, \hat{w})}{\sum\limits_{\hat{w} \in \hat{W}} \mathrm{P}(\hat{x}, \hat{w})} \tag{2.43}$$

where $\hat{W}$ denotes the set of all permissible sentences in the evaluated decoding output.

Collaboration Effort

We define the Collaboration Effort (CE) as the number of speech utterances used in the crowdsourcing platform for obtaining a determined output, i.e., the CE corresponds with the product between the number of lines (batch size B) that the system asks the collaborators to read, and the actual number n of collaborators involved in the obtainment of a determined output.

$$\mathrm{CE} = n \cdot B \tag{2.44}$$

2.7.5 Statistical Significance

Statistical significance of experimental results can be estimated by means of confidence intervals. In this work, confidence intervals of probability 95% ($\alpha = 0.025$) were calculated by using the bootstrapping method with 10,000 repetitions (Bisani and Ney, 2004).

For calculating the confidence intervals with probability $100(1 - 2\alpha)\%$ in the evaluation of a set of s sentences, an iterative process is repeated β times (in this work we used $\beta = 10,000$). During this iterative process, β bootstrap samples are generated by selecting randomly with replacement s pairs (number of elements in the reference - number of errors) from the original set of sentences. These samples will contain several of the original sentences multiple times, while others are missing. Then, each bootstrap sample is evaluated, and the obtained results are sorted. Finally, for a chosen error threshold α, the extreme points of the confidence interval are represented by the $\alpha\beta^{\text{th}}$ smallest and the $\alpha\beta^{\text{th}}$ largest values of the obtained results. Only if the confidence intervals of two systems do not overlap, we can say that the difference is statistically significant.

In addition to the confidence intervals, the statistical significance can be confirmed by means of p-values. In this work, p-values were calculated by means of the Welch's t-test (Welch, 1947) by using the statistical computing tool R (R Core Team, 2017). The significance threshold was set at $\alpha = 0.025$.

Welch's t-test is an adaptation of Student's t-test, which can be used to test the null hypothesis that two observed results represent equal means μ. In other words, the Welch's t-test is useful for obtaining the p-values to reject the null hypothesis H_0, and to confirm the alternative hypothesis H_a:

$$H_0 : \mu_1 = \mu_2 \tag{2.45}$$

$$H_a : \mu_1 \neq \mu_2 \tag{2.46}$$

Assuming the truth of the null hypothesis H_0, the p-value is the probability that the statistical summary for the two compared results would be the same. Therefore, the smaller the p-value, the larger the statistical significance, and only when the obtained p-value is less than or equal to a set significance threshold, the null hypothesis H_0 is rejected and we can say that the difference is statistically significant.

2.8 Datasets

The *off-line* HTR experiments were performed on two different historical manuscripts *Cristo Salvador* and *Rodrigo*. Although both manuscripts were written in Spanish by a single writer, they present some characteristics that give them different degrees of difficulty. One of the most important is the language variation (Llamas Pombo, 2012), *Cristo Salvador* was written in Modern Spanish (19^{th} century), whilst *Rodrigo* was written in an ancient Spanish called Early Modern Spanish (15^{th} - 17^{th} century), been much more challenging the transcription for *Rodrigo*. Given that the additional multimodal sources of information used in this thesis provide from collaborators, only few pages were selected as representative of the whole *off-line* HTR test sets.

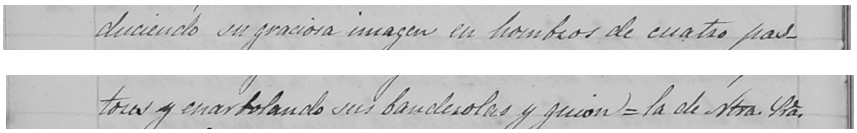


Figure 2.11: Some lines of the *Cristo Salvador* corpus.

The phonetical Spanish corpus *Albayzin* (Moreno et al., 1993) was used for training the acoustical models for the ASR experimentation. However, the speech samples used in the multimodal experiments were acquired from the dictation of the contents of the HTR test sets of *Cristo Salvador* and *Rodrigo*.

The *on-line* HTR feedback was simulated by using the touch screen handwriting dataset UNIPEN (Guyon et al., 1994). To increase realism, different writers were used for training the kinematical models and testing the *on-line* HTR interaction.

2.8.1 Historical Manuscript Corpora (*Off-line* Handwriting)

Cristo Salvador

Cristo Salvador is a handwritten book of the 19th century provided by *Biblioteca Valenciana Digital* (BiValDi). It is a single writer book with different image features that cause some problems, such as smear, background variations, differences in bright, and bleed-through (ink that trespasses to the other surface of the sheet). It is composed of 53 pages that were automatically divided into lines (such as those shown in Figure 2.11).

This corpus presents a total number of 1,172 lines, with a vocabulary of 3,287 different words. For training the optical models for *off-line* HTR, a partition with the first 33 pages (675 lines) was used. Test data for *off-line* HTR is composed of the 24 lines of page 41 (Figure 2.12), which contains 222 words. This page was selected for being, according to preliminary error recognition results (Alabau et al., 2014), a representative page of the whole test set (the remaining 20 pages, 497 lines, not used on the training).

Rodrigo

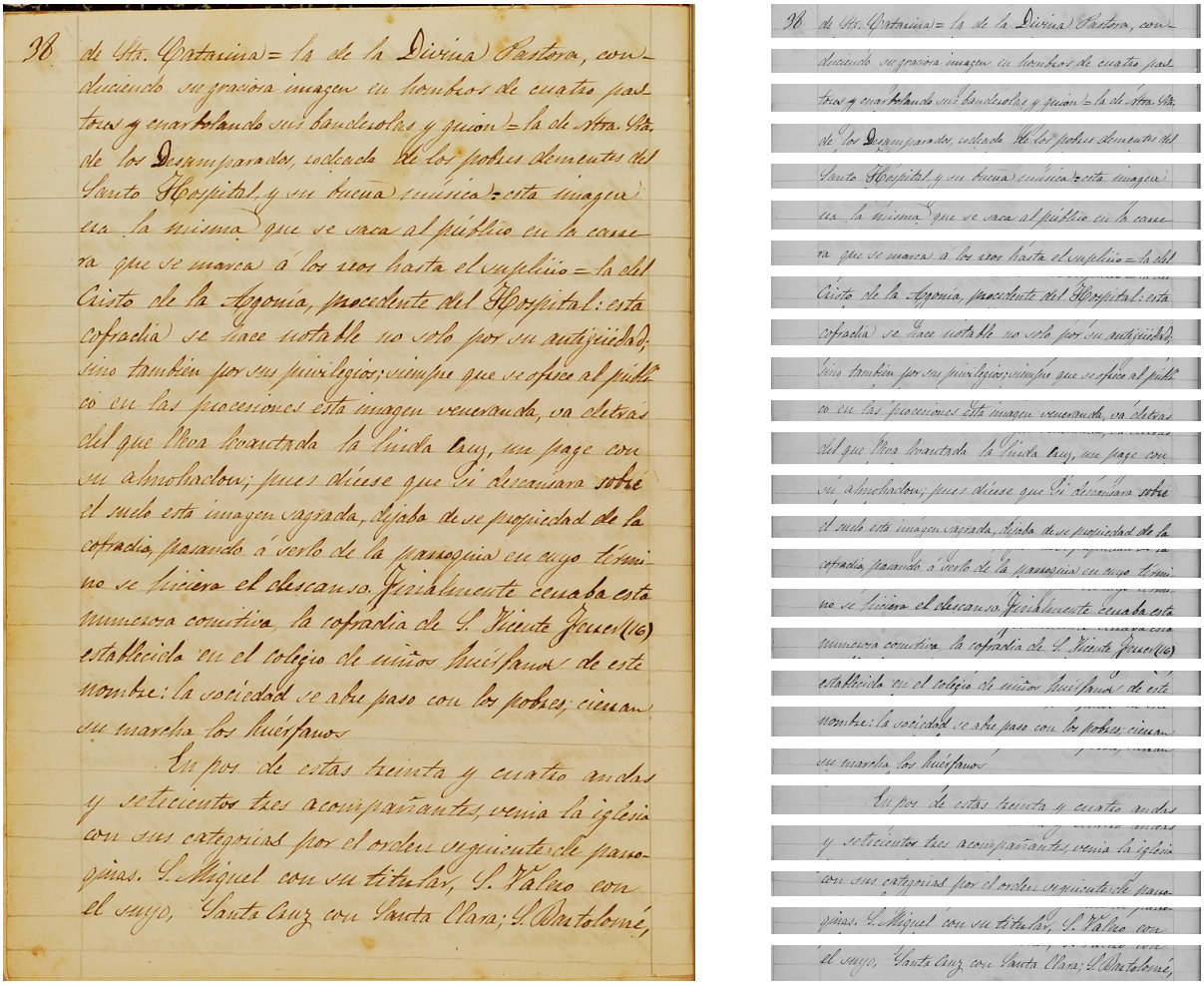
Rodrigo (Serrano et al., 2010) is a corpus obtained from the digitalisation of the book "Historia de España del arçobispo Don Rodrigo", written in ancient Spanish in 1545. It is a single writer book where most pages consist of a single block of well separated lines of calligraphical text. It is composed of 853 pages that were automatically divided into lines (see example in Figure 2.13), giving a total number of 20,356 lines.

The vocabulary size is of about 11,000 words. For training the optical models, a standard partition with a total number of 5,000 lines (about 205 pages) was used. Test data for *off-line* HTR was composed of two pages that were not included in the training part (pages 515 -Figure 2.14- and 579 -Figure 2.15-) and that were representative of the average error of the standard test set (of about 5,000 lines). These two pages contain 50 lines and 514 words.

2.8.2 Touch Screen Handwriting Corpus (*On-line* Handwriting): UNIPEN

The production of the touchscreen feedback data in the assistive transcription experiments has been simulated using the UNIPEN `Train-R01/V07` dataset (Guyon et al., 1994)[6]. It comes organised into several categories such as lower and upper-case letters, digits, symbols, isolated words and full sentences. Unfortunately, the UNIPEN isolated words category does not contain the required word instances to be handwritten by the user in the interactive processes with the text images of the *off-line* HTR corpora

[6]For a detailed description of this dataset, see http://www.unipen.org.



(a) Page 41.

(b) Text lines extracted from the page 41.

Figure 2.12: Page 41 of the *Cristo Salvador* corpus.

used in this thesis (*Cristo Salvador* and *Rodrigo*). Therefore, the process for generating synthetic samples used in (Romero et al., 2012) was followed here. The samples were generated by concatenating random character instances from three UNIPEN categories: `1a` (digits), `1c` (lowercase letters) and `1d` (symbols).

To increase realism, the generation of each of these test words was carried out employing characters belonging to a same writer. Three different writers were randomly chosen, taking care that sufficient samples of all the characters needed for the generation of the required word instances were available from each writer. Each character needed to generate a given word was plainly aligned along a common word baseline, except if it had a descender, in which case the character baseline was raised 1/3 of its height. The horizontal separation between characters was randomly selected from *one* to *three* trajectory points. The selected writers are identified by their name initials as BS, BH and BR. At bottom of Figure 2.16, three examples of the word "*historia*" for the three different writers generated in this way are shown.

Training data were produced in a similar way using 17 different UNIPEN writers. For each of these writers, a sample of each of the 42 symbols and digits needed was randomly selected and one sample of each of the 1,000 most frequent Spanish and English words was generated, resulting in 34,714 training tokens (714 isolated characters plus 34,000 generated words). To generate these tokens, 186,881 UNIPEN character instances were used, using as many repetitions as required out of the 17,177 unique character samples available. Table 2.1 summarises the amount of UNIPEN training and

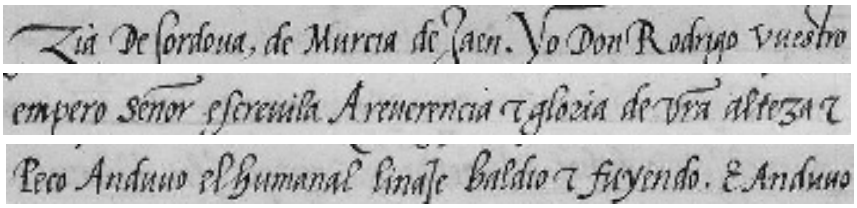


Figure 2.13: Some lines of the *Rodrigo* corpus.

Table 2.1: Basic statistics of the UNIPEN corpus.

Number of different:	Train	Test	Lexicon
writers	17	3	-
digits (1a)	1,301	234	10
letters (1c)	12,298	2,771	26
symbols (1d)	3,578	3,317	32
total characters	17,177	6,322	68

test data used.

It should be mentioned here that, even though the kinematical models (*on-line* HTR HMM) were trained from artificially built words, the accuracy in real operation with real users performed in (Romero et al., 2012) is observed to be similar to that shown in the laboratory results reported with synthetic samples.

2.8.3 Training Speech Corpus: *Albayzin*

For training the ASR acoustical models we used a partition of the Spanish phonetic corpus *Albayzin* (Moreno et al., 1993). This corpus consists of a set of three sub-corpus recorded by 304 speakers using a sampling rate of 16 kHz and a 16-bit quantisation. The training partition used in this work includes 4,800 phonetically balanced utterances. Specifically, 200 utterances were read by four speakers and 25 utterances were read by 160 speakers, with a total length of about 4 hours.

2.8.4 Multimodal (Text - Speech) Corpora

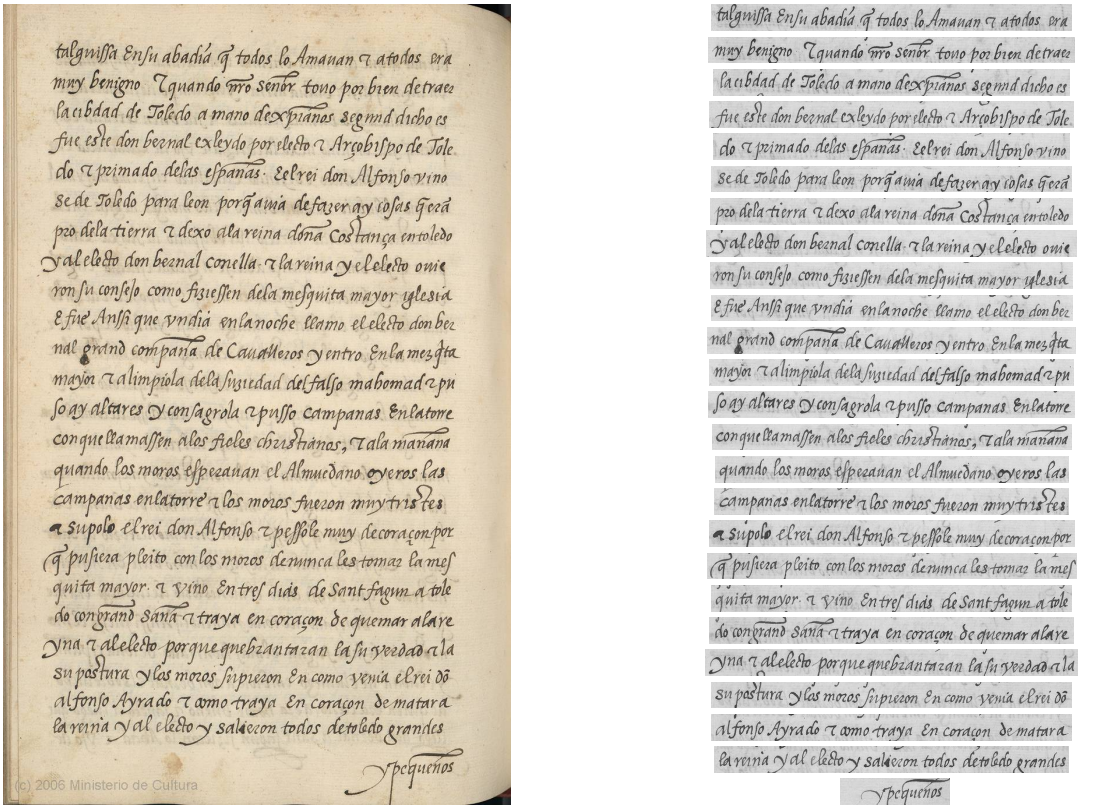
For the multimodal test (*off-line* HTR and ASR), the test data for ASR was the product of the acquisition of the dictation of the contents of the handwritten text line images that compose the HTR test set of both historical manuscripts by different native Spanish speakers, using a sample rate of 16 kHz and an encoding of 16 bits (to match the conditions of *Albayzin* data).

In the ASR system we are dealing with two major sources of errors, i.e. on the one side we have the differences between the training and test audio samples (speakers, devices and environment), and on the other side speakers can make mistakes while reading the manuscript. In order to alleviate these sources of errors, speakers were provided with a text guide of reading along with text images during the speech acquisitions.

Speech Acquisition on a Controlled Environment

In the first speech acquisition, performed for both historical manuscripts, the speech samples were registered by using a desktop computer in a controlled laboratory environment.

In the case of *Cristo Salvador*, the ASR test data was composed of the acquisition of the dictation



(a) Page 515.

(b) Text lines extracted from the page 515.

Figure 2.14: Page 515 of the *Rodrigo* corpus.

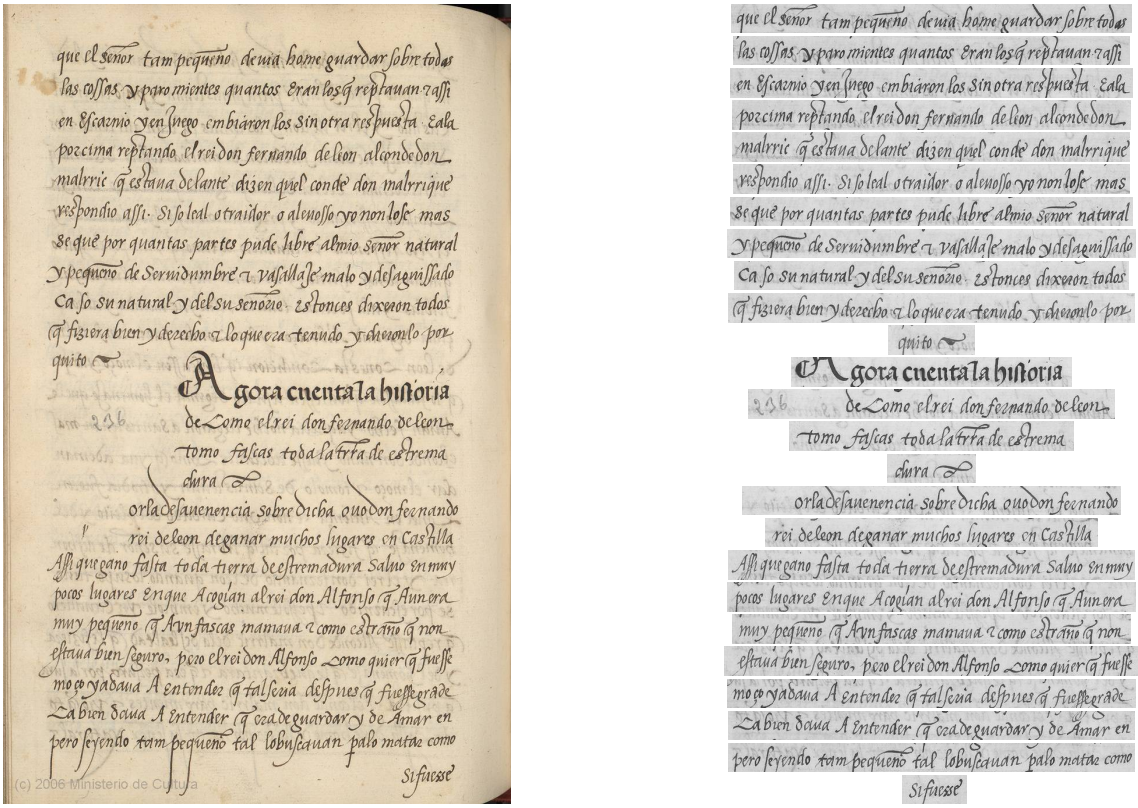
of the contents of the 24 test lines (those of page 41) by five different speakers (i.e., a total set of 120 utterances, with a total length of about 9 minutes), while in the case of *Rodrigo* seven different speakers read the 50 handwritten test lines (those of pages 515 and 579), giving a total set of 350 utterances (about 15 minutes).

Speech Acquisition on a Real Crowdsourcing Environment

A second speech acquisition was performed only for the *Rodrigo* corpus on a real use scenario. The mobile application *Read4SpeechExperiments* (Granell and Martínez-Hinarejos, 2016) (see Figure 8.2) was used for acquiring the speech samples, and the mailing list of our research group for collaboration demand. *Read4SpeechExperiments* is an Android free software application designed to facilitate the speech acquisition from mobile devices. The source code is available on GitLab[7], and it can be installed from the Google Play[8] and the F-Droid[9] platforms.

None of the received contributions was rejected, given that we intentionally wanted a rather broad and real sample. We obtained the collaboration of 27 different speakers who installed the application on their own mobile devices, and read the 50 handwritten text lines without any control from our side, i.e. the speakers read the text lines where and when they wanted, giving a total set of 1,350 utterances (about 1 hour and 50 minutes).

[7] https://gitlab.com/egranell/Read4SpeechExperiments
[8] https://play.google.com/store/apps/details?id=com.prhlt.aemus.Read4SpeechExperiments
[9] https://f-droid.org/wiki/page/com.prhlt.aemus.Read4SpeechExperiments



(a) Page 579.

(b) Text lines extracted from the page 579.

Figure 2.15: Page 579 of the *Rodrigo* corpus.

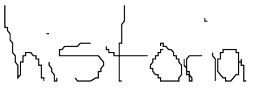
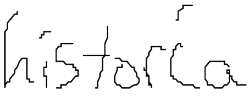
Writter:	BH	BR	BS
Sample:			

Figure 2.16: Samples of the word *"historia"* generated using the characters of three writers of the UNIPEN corpus.

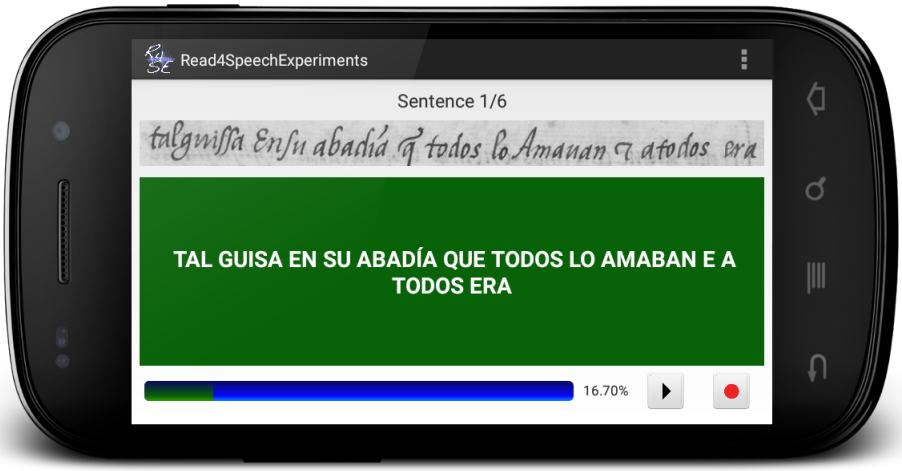


Figure 2.17: Screenshot of the application *Read4SpeechExperiments*.

Bibliography

Adda-Decker, M. and Lamel, L. (2000). The Use of Lexica in Automatic Speech Recognition. In *Lexicon Development for Speech and Language Processing*, Text, Speech and Language Technology, pages 235–266. Springer.

Al-Khoury, I. (2015). *Arabic Text Recognition and Machine Translation*. PhD thesis, Universitat Politècnica de València.

Alabau, V., Martínez-Hinarejos, C.-D., Romero, V., and Lagarda, A. L. (2014). An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195–203. Frontiers in Handwriting Processing.

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

Arisoy, E., Sethy, A., Ramabhadran, B., and Chen, S. (2015). Bidirectional recurrent neural network language models for automatic speech recognition. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*, pages 5421–5425.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J. M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of American Mathematical Society*, 73(3):360–363.

Baum, L. E. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Bennett, I. M., Babu, B. R., Morkhandikar, K., and Gururaj, P. (2014). Speech recognition system interactive agent. US Patent App. 14/256,648.

Bertoldi, N., Zens, R., and Federico, M. (2007). Speech Translation by Confusion Network Decoding. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 4, pages 1297–1300.

Bilmes, J. A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report TR-97-021, International Computer Science Institute.

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, volume 1, pages 409–412.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bluche, T., Ney, H., and Kermorvant, C. (2014). A Comparison of Sequence-Trained Deep Neural Networks and Recurrent Neural Networks Optical Modeling for Handwriting Recognition. In *Proceedings of the 2nd International Conference on Statistical Language and Speech Processing (SLSP 2014)*, pages 199–210.

Bod, R. (2000). Combining Semantic and Syntactic Structure for Language Modeling . In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 3, pages 106–109.

Bourlard, H. and Morgan, N. (1998). Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions. In *Adaptive Processing of Sequences and Data Structures*, volume 1387 of *Lecture Notes in Artificial Intelligence*, pages 389–417. Springer.

Brown, P. F., Della Pietra, V. J., Mercer, R. L., Della Pietra, S. A., and Lai, J. C. (1992). An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, 18(1):31–40.

Bunke, H., Roth, M., and Schukat-Talamazzini, E. G. (1995). Off-line cursive handwriting recognition using hidden markov models. *Pattern recognition*, 28(9):1399–1413.

Caines, A., Bentz, C., Graham, C., Polzehl, T., and Buttery, P. (2016). Crowdsourcing a Multi-lingual Speech Corpus: Recording, Transcription and Annotation of the CrowdIS Corpora. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2145–2152.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Ciura, M. G. and Deorowicz, S. (2001). How to squeeze a lexicon. *Software: Practice and Experience*, 31(11):1077–1090.

Cremelie, N. and Martens, J.-P. (1997). Automatic rule-based generation of word pronunciation networks. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2459–2462.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Di Fabbrizio, G., Okken, T., and Wilpon, J. G. (2009). A Speech Mashup Framework for Multimodal Mobile Services. In *Proceedings of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '09)*, pages 71–78.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM*, 54(4):86–96.

Doi, K. (2007). Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. *Computerized Medical Imaging and Graphics*, 31(4–5):198 – 211.

Duddington, J. (1995). eSpeak: text to speech. `http://espeak.sourceforge.net`. Last access: February 2017.

Dutoit, T. and Stylianou, Y. (2003). Text-to-speech synthesis. In *The Oxford Handbook of Computational Linguistics*, chapter 17, pages 323–338. Oxford University Press.

Estellés-Arolas, E. and González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2):189–200.

ETSI (2003). Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Frontend feature extraction algorithm; Compression algorithms. Standard ETSI-ES-201-108, European Telecommunications Standards Institute (ETSI).

Evermann, G. and Woodland, P. C. (2000). Posterior Probability Decoding, Confidence Estimation and System Combination. In *Proceedings of the NIST 2000 Speech Transcription Workshop*.

Fischer, A. (2012). *Handwriting Recognition in Historical Documents*. PhD thesis, University of Bern.

Fischer, A., Frinken, V., and Bunke, H. (2013). Hidden Markov Models for Off-Line Cursive Handwriting Recognition. In *Handbook of Statistics: Machine Learning: Theory and Applications*, volume 31, chapter 17, pages 421–442. Elsevier.

Gales, M. and Young, S. (2008). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98.

Gales, M. J. F. (2000). Cluster Adaptive Training of Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 8(4):417–428.

Gales, M. J. F. and Woodland, P. C. (1996). Mean and Variance Adaptation within the MLLR framework. *Computer Speech & Language*, 10(4):249–264.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Giménez, A., Khoury, I., Andrés-Ferrer, J., and Juan, A. (2014). Handwriting word recognition using windowed Bernoulli HMMs. *Pattern Recognition Letters*, 35:149–156.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gordo, A., Llorens, D., Marzal, A., Prat, F., and Vilar, J. M. (2008). State: A Multimodal Assisted Text-Transcription System for Ancient Documents. In *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems (DAS '08)*, pages 135–142.

Granell, E. and Martínez-Hinarejos, C.-D. (2015). Multimodal Output Combination for Transcribing Historical Handwritten Documents. In *Proceedings of the 16th International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 246–260.

Granell, E. and Martínez-Hinarejos, C.-D. (2016). 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)*, pages 411 – 417.

Graves, A. and Jaitly, N. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1764–1772.

Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., and Janet, S. (1994). UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR)*, volume 2 - Conference B: Computer Vision & Image Processing, pages 29–33.

Haeb-Umbach, R. and Ney, H. (1994). Improvements in Beam Search for 10000-Word Continuous-Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 2(2):353–356.

Harwath, D. and Glass, J. (2014). Speech Recognition without a Lexicon - Bridging the Gap between Graphemic and Phonetic Systems. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2655–2659.

Hermansky, H., Ellis, D. P. W., and Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 3, pages 1635–1638.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Juang, B.-H. (1985). Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains. *AT&T Technical Journal*, 64(6):1235–1249.

Jurafsky, D. and Martin, J. H. (2009). *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition*. Artificial Intelligence. Prentice Hall.

Kneser, R. and Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, volume 1, pages 181–184.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 14, pages 1137–1145.

Kozielski, M., Doetsch, P., and Ney, H. (2013). Improvements in RWTH's system for off-line handwriting recognition. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13)*, pages 935–939.

Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N. (2000). Rapid Speaker Adaptation in Eigenvoice Space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707.

Kung, S. H., Zohdy, M. A., and Bouchaffra, D. (2016). 3D HMM-based Facial Expression Recognition using Histogram of Oriented Optical Flow. *Transactions on Machine Learning and Artificial Intelligence*, 3(6):42–69.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Liu, F.-H., Stern, R. M., Huang, X., and Acero, A. (1993). Efficient Cepstral Normalization for Robust Speech Recognition. In *Proceedings of the workshop on Human Language Technology (HLT '93)*, pages 69–74.

Ljolje, A., Pereira, F., and Riley, M. (1999). Efficient General Lattice Generation and Rescoring. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'99)*.

Llamas Pombo, E. (2012). Variation and standardization in the history of Spanish spelling. In *Orthographies in Early Modern Europe*, pages 15–62. Walter de Gruyter.

Lucchesi, C. L. and Kowaltowski, T. (1993). Applications of Finite Automata Representing Large Vocabularies. *Software: Practice and Experience*, 23(1):15–30.

Maas, A. L., Xie, Z., Jurafsky, D., and Ng, A. Y. (2015). Lexicon-Free Conversational Speech Recognition with Neural Networks. In *Proceedings of the Human Language Technologies: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, pages 345–354.

Machover, C. (1995). *The CAD/CAM Handbook*. Visual Technology. McGraw-Hill.

Malit, R. F. (2009). Computer assisted driving of vehicles. US Patent 7,513,508.

Mangu, L. and Brill, E. (1999). Lattice Compression in the Consensual Post-Processing Framework. In *Proceedings of the 3rd World Multiconference on Systemics, Cybernetics and Informatics Joint with the 5th International Conference on Information Systems Analysis and Synthesis (SCI/ISAS'99)*, volume 5, pages 246–252.

Mangu, L., Brill, E., and Stolcke, A. (2000). Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400.

Mangu, L., Kingsbury, B., Soltau, H., Kuo, H.-K., and Picheny, M. (2014). Efficient spoken term detection using confusion networks. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*, pages 7894–7898.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Marti, U.-V. and Bunke, H. (2001). Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition System. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):65–90.

Martín-Albo, D., Romero, V., and Vidal, E. (2013). Interactive Off-Line Handwritten Text Transcription Using On-Line Handwritten Text as Feedback. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13)*, pages 1280–1284.

Martín-Albo Simón, D. (2016). *Contributions to Pen & Touch Human-Computer Interaction*. PhD thesis, Universitat Politècnica de València.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, 30(5):1225–1267.

Mihalcea, R. (2012). Multimodal Sentiment Analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '12)*, pages 1–1.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1045–1048.

Mohri, M. (1997). Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, 23(2):269–311.

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

Morgan, N. and Bourlard, H. (1995). An Introduction to Hybrid HMM/Conectionist Continuous Speech Recognition. *IEEE Signal Processing Magazine*, pages 25–42.

Ng, A. Y. and Jordan, M. I. (2001). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01)*, volume 2, pages 841–848.

Oerder, M. and Ney, H. (1993). Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'93)*, volume 2, pages 119–122.

Ortmanns, S., Eiden, A., Ney, H., and Coenen, N. (1997). Look-ahead techniques for fast beam search. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume 3, pages 1783–1786.

Parent, G. and Eskenazi, M. (2011). Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3037–3040.

Pastor, M., Toselli, A. H., and Vidal, E. (2004). Projection Profile Based Algorithm for Slant Removal. In *Proceedings of the 2nd International Conference on Image Analysis and Recognition (ICIAR'04)*, volume 3212 of *Lecture Notes in Computer Science*, pages 183–190.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. (2003). Recent Advances in the Automatic Recognition of Audiovisual Speech. *Proceedings of the IEEE*, 91(9):1306–1326.

Pylkkönen, J. (2005). New pruning criteria for efficient decoding. In *Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech)*, volume 5, pages 581–584.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Last access: May 2017.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.

Revuelta-Martínez, A., Rodríguez, L., and García-Varea, I. (2012). A Computer Assisted Speech Transcription System. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 41–45.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G. (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29(2):209–224.

Romero, V., Toselli, A. H., and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*, volume 80 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing.

Romero Gómez, V. (2010). *Multimodal Interactive Transcription of Handwritten Text Images*. PhD thesis, Universitat Politècnica de València.

Rueber, B. (1997). Obtaining confidence measures from sentence probabilities. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 739–742.

Sad, G. D., Terissi, L. D., and J, C. G. (2015). Combination of Standard and Complementary Models for Audio-Visual Speech Recognition. In *Proceedings of the 16th Argentine Symposium on Artificial Intelligence (ASAI 2015)-JAIIO 44*, pages 113–120.

Samanta, O., Bhattacharya, U., and Parui, S. K. (2014). Smoothing of HMM parameters for efficient recognition of online handwriting. *Pattern Recognition*, 47(11):3614–3629.

Scholz, F. W. (1985). Maximum Likelihood Estimation. In *Encyclopedia of Statistical Sciences*.

Sebe, N., Cohen, I., and Huang, T. S. (2005). Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*.

Serrano, N., Castro, F., and Juan, A. (2010). The RODRIGO Database. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2709–2712.

Silvestre-Cerdà, J. A., Pérez, A., Jiménez, M., Turro, C., Juan, A., and Civera, J. (2013). A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC'13)*, pages 3994–3999.

Soong, F. K. and Huang, E.-F. (1991). A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition. In *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'91)*, pages 705–708.

Suppes, P. (1970). Probabilistic Grammars for Natural Languages. In *Synthese*, volume 22 of *Semantics of Natural Language*, pages 95–116. Springer.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

Toselli, A. H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E., and Casacuberta, F. (2004). Integrated Handwriting Recognition and Interpretation using Finite-State Models. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):519–539.

Toselli, A. H., Pastor, M., and Vidal, E. (2007). On-Line Handwriting Recognition System for Tamil Handwritten Characters. In Martí, J., Mendonça, J. M. B. A. M., and Serrat, J., editors, *Pattern Recognition and Image Analysis (IbPRIA 2007)*, volume 4477 of *Lecture Notes in Computer Science*, pages 370–377. Springer, Berlin, Heidelberg.

Viikki, O. and Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Wang, T., Wu, D. J., Coates, A., and Ng, A. Y. (2012). End-to-End Text Recognition with Convolutional Neural Networks. In *Proceedings of the 21^{tst} International Conference on Pattern Recognition (ICPR 2012)*, pages 3304–3308.

Welch, B. L. (1947). The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35.

Welch, L. R. (2003). Hidden Markov Models and the Baum-Welch Algorithm. *IEEE Information Theory Society Newsletter*, 53(4):10–13.

Woodland, P. C. (2001). Speaker Adaptation for Continuous Density HMMs: A Review. In *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, pages 11–19.

Wooters, C. and Stolcke, A. (1994). Multiple-pronunciation lexical modeling in a speaker independent speech understanding system. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP'94)*, pages 1363–1366.

Xue, J. and Zhao, Y. (2005). Improved confusion network algorithm and shortest path search from word lattice. In *Proceedings of the International Conference in Acoustics, Speech and Signal Processing (ICASSP 2005)*, volume 1, pages 853–856.

Yun, S.-J. and Oh, Y.-H. (1999). Stochastic Lexicon Modeling for Speech Recognition. *IEEE Signal Processing Letters*, 6(2):28–30.

Zhai, C. and Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.

Zhang, W. (1999). *State-Space Search: Algorithms, Complexity, Extensions, and Applications*. Springer Science & Business Media.

Zuo, L., Wan, X., and Liu, J. (2016). Comparison of Various Neural Network Language Models in Speech Recognition. In *Proceedings of the 3rd International Conference on Information Science and Control Engineering (ICISCE 2016)*, pages 894–898.

II

Multimodality

"It is the tension between creativity and skepticism that has produced the stunning and unexpected findings of science."

Carl Sagan, Broca's Brain: Reflections on the Romance of Science, 1979.

Skeletal illustration. (Andreae Vesalii Bruxellensis, scholae medicorum Patauinae professoris, de Humani corporis fabrica Libri septem, 1543)

3

Combining Handwriting and Speech

"Imagination is the only weapon in the war against reality."

Lewis Carroll, Alice in Wonderland, 1865.

Content

Transcription of historical documents is an interesting task for libraries in order to make their content available electronically. In the lasts years, the use of Handwritten Text Recognition (HTR) systems allowed paleographers to speed up the manual transcription process, since they are able to correct on a draft transcription (Fischer, 2012). Another alternative is obtaining the draft transcription by dictating the contents to an Automatic Speech Recognition (ASR) system. When both sources (image and speech) are available, the multimodal combination of both sources of information permits one to obtain better draft transcriptions, allowing for a faster transcription process (Alabau et al., 2011, 2014). In this chapter, a multimodal combination method based on Confusion Networks is presented.

3.1 Introduction

The combination of natural language recognition systems allows one to improve the recognition accuracy. In most cases, this combination can be performed in three different stages of the recognition process (Li, 2005): in the feature extraction stage (feature combination), in the search process (probability combination), and in the decoding output (hypothesis combination).

- **Feature combination:** Feature combination is performed concatenating the different features at feature vector level to form a new feature vector sequence to be used in the recognition process (Potamianos and Neti, 2001). This combination method usually requires synchronous parallel feature streams.

Illustration info: Five triads of concepts drawn inside a circle, used to draw conclusions from comparisons of its basic components (Ramon Llull, Arts demostrativa, 1283).

- **Probability combination:** In probability combination methods, the recognition class probabilities are combined before the final search process. The probability combination can be performed synchronously (Hernando et al., 1995) combining the observation probabilities of the HMM states frame-by-frame, or asynchronously (Dupont and Luettin, 2000) combining the probabilities at a higher-level, such as characters or phonemes. Synchronous probability combination requires synchronous parallel feature streams, while asynchronous probability combination allows one to combine asynchronous parallel feature streams of the same nature (they use the same higher-level unit, such as in audio-visual speech recognition (Gurban and Thiran, 2005)).
- **Hypothesis combination:** The last stage where the combination can be performed is at recogniser output (Fiscus, 1997). In this stage, the hypotheses obtained after the completion of the search process from each recogniser are combined. In hypothesis combination, the parallel feature streams can be synchronous or asynchronous, and the only restriction is that all feature streams must represent the same final sequence of words.

The similar nature of the HTR and ASR processes (common use of HMM as optical or acoustical models, use of n-grams as language models, Viterbi decoding process, etc.) makes this integration of knowledge easier than for other modalities. In fact, it can be seen as the integration of the result of two different recognisers.

Many techniques have been proposed for unimodal system combination (Fiscus, 1997; Evermann and Woodland, 2000; Bertolami et al., 2006; Ishimaru et al., 2011). In the case of bimodal fusion (image and speech), (Woodruff and Dupont, 2005) presents a technique that can be applied for isolated words. However, bimodal combination in continuous decoding for the two modalities is a hard problem because of the time asynchrony between the two signals, i.e., the sequence of feature vectors for each modality differs in length and therefore it is not an easy problem to find the time points where the same elements (words in this case) are synchronised.

An initial approach for solving this limitation was proposed in (Alabau et al., 2011). This technique uses the output of the recognition process of one of the modalities, in form of lattice or word-graph (WG) (Ortmanns et al., 1997), 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. An iterative version of this procedure was presented in (Alabau et al., 2014). The results reported for a historical document in Spanish showed significant improvements in the final results.

Anyway, the approach proposed in (Alabau et al., 2011, 2014) presents a few drawbacks: there is not a single hypothesis for the two modalities (each modality provides its own hypothesis and it is not known beforehand which one is more accurate), and the initial modality must be chosen arbitrarily (which according to the presented results is crucial for the quality of the final decoding).

In this chapter we present a new proposal based on the use of Confusion Networks (CN) (Saldarriaga and Cheriet, 2011; Mangu et al., 2000) for obtaining a single hypothesis from the combination of the outputs (in initial form of WG) of an HTR and an ASR recogniser that refer to the same text. The chapter is structured as follows: Section 3.2 introduces the combination of natural language recognition systems, Section 3.3 draws the details on the proposed combination technique, and Section 3.4 summarises the conclusions on the presented topics.

3.2 Hypothesis Combination on Natural Language Recognition

The multimodal combination of HTR and ASR systems can not be performed easily at any stage of the recognition process given the different nature of these modalities and the asynchrony with respect to each other. The easiest way of performing this multimodal combination is to combine the results of both systems by using a hypothesis combination method.

Many techniques on joining results have been proposed with the idea of reducing the error in the combined output. Some examples are: Recogniser Output Voting Error Reduction (ROVER) (Fiscus, 1997), N-best ROVER (Stolcke et al., 2000), Lattices Rescoring (Stolcke et al., 1997), and Confusion

Network Combination (CNC) (Evermann and Woodland, 2000). These methods can be used to combine the outputs of recognition systems of different modalities that represent the same sentence. They all effectively improve the recognition performance, even though each one presents different characteristics.

3.2.1 Recogniser Output Voting Error Reduction (ROVER)

The widely used ROVER method (Fiscus, 1997) misses part of the information contained in the recognition outputs as it performs the combination by voting (at word level) among the different system outputs using only the 1-best hypothesis.

The ROVER method is implemented in two modules. In the first one, the 1-best decoding outputs are aligned and combined in a word transition network (with a structure similar to a Confusion Network -Figure 2.10-). Then, the second module (the voting search module) evaluates each subnetwork to select the best scoring word (using a voting scheme) for the new transcription.

Voting is performed as follows: for each subnetwork the number of occurrences of each word w in the corresponding subnetwork i is accumulated in an array $N(w,i)$, and normalised by dividing $N(w,i)$ by the number of combined systems (N_s) to scale the frequency of occurrence to the unity. Moreover, depending on the voting scheme, the confidence scores for word w in the subnetwork i are measured and normalised in an array $C(w,i)$. The confidence score of NULL transition arcs can be defined by the $Conf(@)$ parameter.

The balance between using word frequency and confidence scores can be adjusted by means of a parameter α:

$$\text{Score}(w,i) = \alpha\left(\frac{N(w,i)}{N_s}\right) + (1-\alpha)C(w,i) \tag{3.1}$$

where $0 \leq \alpha \leq 1$.

The voting search module offers the following three different voting schemes:

1. **Frequency of occurrence.** In the voting by frequency of occurrence scheme all confidence scoring information is ignored, i.e. the α parameter is set to 1.
2. **Frequency of occurrence and average word confidence.** In this voting method, the confidence score of each word w in the array $C(w,i)$ is set to the average value of the appearance of this word w in the subnetwork i. Both parameters α and $Conf(@)$ must be trained *a priori*.
3. **Frequency of occurrence and maximum confidence.** In the last voting scheme, the confidence score of each word w in the array $C(w,i)$ is set to the maximum value of the appearance of this word w in the subnetwork i. In this case, both parameters α and $Conf(@)$ must be also trained *a priori*.

3.2.2 N-best ROVER

The combination of multiple hypotheses can produce an output more accurate than combining only the 1-best hypothesis. This is the idea behind the N-best ROVER method (Stolcke et al., 2000), which uses n-best outputs to perform the combination.

This method works in three steps. In a first step, the n-best h hypotheses from the decoding of a feature vector sequence x by using different systems S_i are aligned like in the ROVER method. Then, in a second step the normalised and weighted log-linear word posteriors are estimated for each system. In the last step, the combined word posterior is computed as a linear combination.

The word posteriors for each word w and system i are computed for each subnetwork j by log-linear score weighting, followed by a normalisation over all hypotheses.

$$\mathrm{P}_i(w \mid x) = \frac{\sum\limits_{h:w\in h} \exp\left(\sum\limits_j \lambda_{ij} s_{ij}(h \mid x)\right)}{\sum\limits_{\forall h} \exp\left(\sum\limits_j \lambda_{ij} s_{ij}(h \mid x)\right)} \tag{3.2}$$

where $s_{ij}(h \mid x)$ is the log-score, and λ_{ij} are the log-score combination weights for the subnetwork j of the hypothesis h of the system i. Then, the combined posterior can be computed as a linear combination:

$$\mathrm{P}(w \mid x) = \sum_i \mu_i \mathrm{P}_i(w \mid x) \tag{3.3}$$

where μ_i represents the system weight.

Finally, the combined hypothesis is formed by the concatenation of the most probable word hypotheses at each position in the alignment. Therefore, like ROVER, this method presents the following constraint: the result of the combination is composed of a single hypothesis.

3.2.3 Lattices Rescoring

Combining multiple lattices on a new lattice not only may improve the most likely hypothesis, but also this new lattice may contain better hypotheses than the most likely. The N-best List and Lattices Rescoring method (Ostendorf et al., 1991; Stolcke et al., 1997) optimises the word-level recognition scores and constructs a word lattice from all information contained in the lattices to combine.

This algorithm has two components. In the first one, the scores of the hypotheses contained in the lattices to combine are weighted by using a parameter, and then all these hypotheses are aligned and merged in one n-best list. In the second one, the optimisation of the word-level recognition scores is made by means of the substitution of the normalisation term $\mathrm{P}(\hat{x})$ of Equation (2.1) by a finite sum over the set $\hat{W}$ of all the hypotheses in the joint n-best list:

$$\mathrm{P}(\hat{x}) = \sum_{\hat{w}\in\hat{W}} \mathrm{P}(\hat{w} \mid \hat{x}) \tag{3.4}$$

Finally, a new combined lattice is built from the rescored n-best hypotheses.

3.3 Our proposal: Bimodal Confusion Network Combination

As an output format for handwriting and speech recognisers, Confusion Networks (CN) reduce the complexity of the Word Graph (WG) without losing important information (Xue and Zhao, 2005).

The bimodal CN combination technique presented here is based on the unimodal CN combination technique shown in (Ishimaru et al., 2011). Our technique combines the CN derived from the outputs of two recognition systems of different modality, one for HTR and the other for ASR, and it acts in two steps. In the first step, the subnetworks of both CN are aligned by similarity, marking the selected subnetworks as immovable anchors. In the second step, a new CN is composed from the base of the first CN, by using combination, insertion and deletion of subnetworks.

During the development of this bimodal combination technique, the operation was verified by using CN obtained from several samples of the multimodal test set of the *Rodrigo* corpus. From this experience, the value of some variables was defined. Specifically, 10^{-4} for the smoothing factor Θ, 0.75 for the delete threshold δ, and 0.25 for the insertion threshold γ.

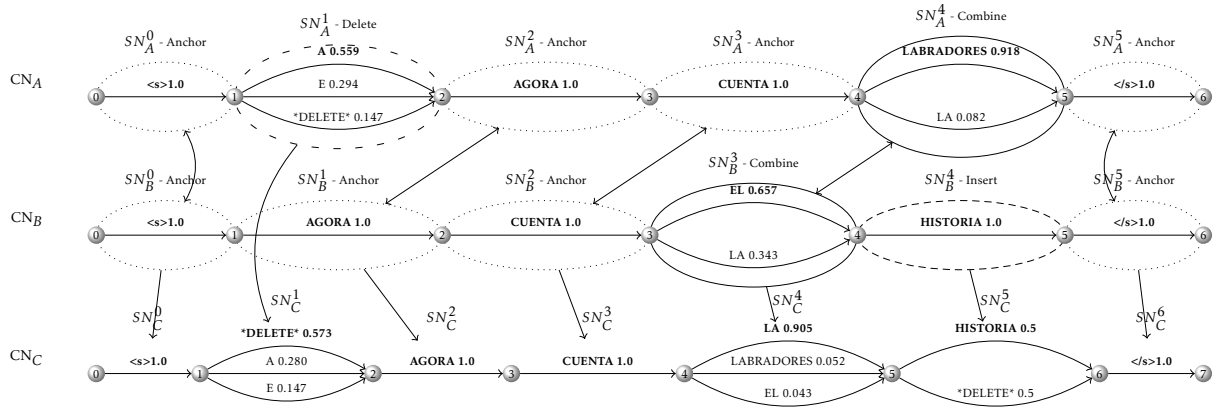


Figure 3.1: Bimodal combination example, reference: *<s>AGORA CUENTA LA HISTORIA </s>*.

3.3.1 Subnetworks Based Alignment

Due to the fact that outputs of different recognition systems may have different errors, it is necessary to find some reference subnetworks that serve as anchors between the CN we wish to combine. The search of anchor subnetworks is performed in both directions (from left to right and vice versa) simultaneously, taking as anchor subnetworks only those where the search on both directions coincide. This search can be adjusted by several parameters, such as:

- Searching unigrams, bigrams or skip-bigrams.
- Searching only on the most probable word or on all the words contained in the subnetworks.
- Setting a gram matching error threshold ϵ between words.

As words can be decomposed on characters (basic unit for HTR) and phonemes (basic unit for ASR), the quadratic mean of the Character Error Rate (CER) and the Phoneme Error Rate (PER) is used to assess the gram matching error between words of both CN:

$$\mathrm{E}(w_A, w_B) = \sqrt{\frac{\mathrm{CER}(w_A, w_B)^2 + \mathrm{PER}(w_A, w_B)^2}{2}} \tag{3.5}$$

where E represents the gram matching error, and the words of the first and the second CN are represented by w_A and w_B respectively. CER and PER are defined on Section 2.7.1.

In Figure 3.1 a complete example of the performance of our technique is shown. In this example, CN$_A$, CN$_B$ represent the two CN to combine, and CN$_C$ the resulting CN. When searching for anchor subnetworks on bigrams and unigrams with $\epsilon = 0$, it would find the following anchor subnetwork pairs: $SN_A^0 - SN_B^0$, $SN_A^2 - SN_B^1$, $SN_A^3 - SN_B^2$, and $SN_A^5 - SN_B^5$.

3.3.2 Composing a New Confusion Network

The final goal of the CN combination is to compose a new CN with higher accuracy than the two original CN. The edit operations used to compose the new CN are: combination, insertion and deletion of subnetworks.

Combination

Combination of subnetworks allows to maximise the probability of the correct word, if it is present on both subnetworks (SN_A and SN_B). Based on the Bayes theorem and assuming a strong independence

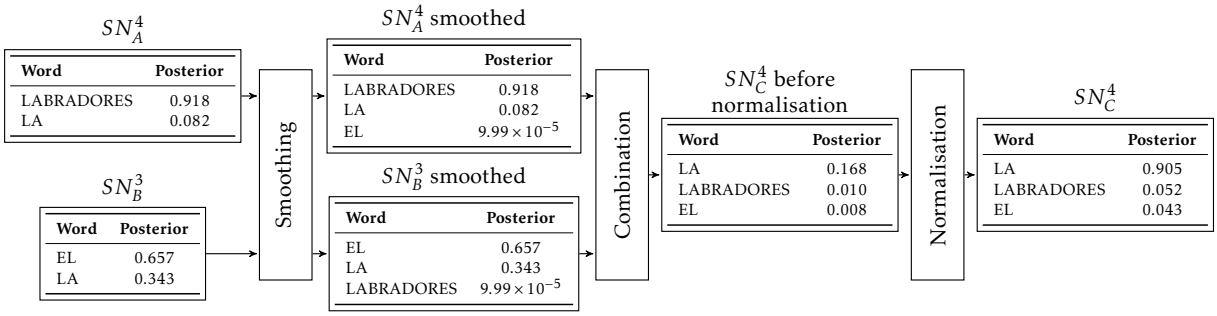


Figure 3.2: Example of subnetwork combination with $\alpha = 0.5$ and $\Theta = 10^{-4}$. The Smoothing block represents the use of Equation (3.8), and the Combination block the use of Equation (3.7).

between SN_A and SN_B, we get:

$$\mathrm{P}(w \mid SN_A, SN_B) \simeq \mathrm{P}(w \mid SN_A)\mathrm{P}(w \mid SN_B) \tag{3.6}$$

However, in practice it is usual to employ a weighted version of Equation (3.6) in which the weight factor α permits to balance the relative reliability between the different probability distributions models (as happens in HTR and ASR systems to balance the influence of the optical/acoustical models and the language model). Thus, in practice we use the following:

$$\mathrm{P}(w \mid SN_A, SN_B) \simeq \mathrm{P}(w \mid SN_A)^{\alpha}\mathrm{P}(w \mid SN_B)^{1-\alpha} \tag{3.7}$$

Before combining two subnetworks with Equation (3.7), it is necessary to smooth all the word probabilities. Otherwise, uncommon words would have null probabilities. Equation (3.8) permits one to smooth the word probabilities of all words ($n = common + uncommon$) in both subnetworks. This equation is based on Laplacian smoothing (Zhai and Lafferty, 2004). However, here the word counts are obtained by dividing the word probabilities by a defined granularity Θ.

$$\mathrm{P}_s(w \mid SN) = \frac{\mathrm{P}(w \mid SN) + \Theta}{1 + n\Theta} \tag{3.8}$$

Finally, the word probabilities on the resulting subnetwork are normalised.

In the example presented in Figure 3.1, SN_A^4 and SN_B^3 are selected for combination. In this case, the correct word (LA) is not the most probable word in either subnetwork. However, it becomes the most probable word when combining both subnetworks with $\alpha = 0.5$ and $\Theta = 10^{-4}$, as it can be seen in SN_C^4 (in Figure 3.2 this combination process is shown in detail).

Insertion and Deletion

Insertion and deletion editing actions allow one to reach a compromise when there is a disagreement between the CN as to whether a particular position between two words should or should not be another word. Specifically, insertion occurs when the second CN subnetwork presents a word with a probability greater than a threshold γ and which is not present in the same position of the first CN subnetwork. Deletion occurs similarly, when the second CN considers that a word is not necessary in a specific position of the first CN, and the most probable word of the first CN subnetwork to delete does not reach a threshold δ.

Both operations use the same procedure: the subnetwork to insert or to delete is combined with a subnetwork with an only *DELETE* arc with probability 1.0. As an example, the subnetworks SN_B^4 and SN_A^1 (see Figure 3.1) are inserted and deleted, respectively.

Composition of the New Confusion Network

The first step on the composition of the new CN is to combine the subnetworks labeled as anchor. Thereby between consecutive anchored subnetworks, a series of aligned fragments appear in both CN. Each fragment can contain from none to several subnetworks. In the example in Figure 3.1, two fragments appear, the first one between anchors $SN_A^0 - SN_B^0$ and $SN_A^2 - SN_B^1$, and the second one between anchors $SN_A^3 - SN_B^2$, and $SN_A^5 - SN_B^5$. We use these fragments sizes to decide what to do with each subnetwork. Comparing the sizes of two aligned fragments, we can find the following cases, when both fragment sizes are not null:

1. If both fragment sizes match, all subnetworks are combined one by one.
2. If the fragment size of only one CN is null, we must choose whether to insert or to delete (as explained above) for all the subnetworks contained in the fragment of the other CN. This is the case of the first fragment in the example of Figure 3.1, which is composed only by SN_A^1. It is deleted, since the probability of the first word does not reach the δ threshold (in this case, $\delta = 0.75$ was chosen).
3. If both fragment sizes are different and none is null, we must find every additional anchor subnetworks in a relaxed search, and decide whether to insert or delete for the rest of subnetworks. This is the case of the second fragment in the example of Figure 3.1, which is formed by SN_A^4 on a side, and by SN_B^3 and SN_B^4 on the other. When searching for unigrams on the whole subnetworks, it is found that SN_A^4 can be combined with SN_B^3, and given that SN_B^4 exceeds the threshold γ (in this case, $\gamma = 0.25$ was taken), it is inserted.

Finally, a new CN is obtained as a result of this process. In Figure 3.1, CN_C is the resulting CN; it can be seen that several errors have been corrected, and the correct sentence (*<s>AGORA CUENTA LA HISTORIA </s>*) has the highest probability.

3.4 Conclusions

In this chapter, a multimodal combination technique for improving the transcription of historical handwritten documents has been presented. This technique takes advantage of the fact that different natural language recognition systems make different errors; thus, editing operations can correct errors. Insertion and deletion create new word sequences that enrich the resulting CN, and the combination can maximise the probability of the correct word. This can occur when both subnetworks contain the correct word, and even when this word has a low probability in both subnetworks. Conversely, if only one subnetwork contains the correct word and both subnetworks contain the same erroneous word, this error will be maximised at the expense of the correct word. Despite of this fact, we will see in the next chapter (Chapter 4) how the performed experiments confirm the strengths of this combination technique.

The initial experimentation detailed in next chapter (Chapter 4) includes iterative and non-iterative combination experiments (in the iterative fashion proposed in (Alabau et al., 2014)), and experiments where the outputs of more than two recognition systems (of the same or different modality) are combined by using a hierarchical combination. Moreover, the quality of the hypotheses contained in the resulting CN are studied, and the proposed combination method is compared with the other three combination techniques presented in Section 3.2.

This combination method was used to provide the multimodal input and to improve the integration of the multimodal user feedback in an assistive environment for transcribing text images (Part III). Finally, a multimodal crowdsourcing system for transcribing historical handwritten documents (Part IV) was designed using this multimodal combination method and the idea of iterative language model interpolation presented in (Alabau et al., 2014).

Bibliography

Alabau, V., Martínez-Hinarejos, C.-D., Romero, V., and Lagarda, A. L. (2014). An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195–203. Frontiers in Handwriting Processing.

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

Bertolami, R., Halter, B., and Bunke, H. (2006). Combination of Multiple Handwritten Text Line Recognition Systems with a Recursive Approach. In *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 61–65.

Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151.

Evermann, G. and Woodland, P. C. (2000). Posterior Probability Decoding, Confidence Estimation and System Combination. In *Proceedings of the NIST 2000 Speech Transcription Workshop*.

Fischer, A. (2012). *Handwriting Recognition in Historical Documents*. PhD thesis, University of Bern.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 1997)*, pages 347–354.

Gurban, M. and Thiran, J.-P. (2005). Audio-visual speech recognition with a hybrid SVM-HMM system. In *Proceedings of the 13th European Signal Processing Conference (EUROSIPCO 2005)*.

Hernando, J., Ayarte, J., and Monte, E. (1995). Optimization of speech parameter weighting for CDHMM word recognition. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 105–108.

Ishimaru, S., Nishizaki, H., and Sekiguchi, Y. (2011). Effect of Confusion Network Combination on Speech Recognition System for Editing. In *Proceedings of the 3rd Asia-Pacific Signal and Information Processing Asocition Annual Summit and Conference (APSIPA ASC 2011)*, volume 4, pages 1–4.

Li, X. (2005). *Combination and Generation of Parallel Feature Streams for Improved Speech Recognition*. PhD thesis, Carnegie Mellon University.

Mangu, L., Brill, E., and Stolcke, A. (2000). Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400.

Ortmanns, S., Ney, H., and Aubert, X. (1997). A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition. *Computer Speech & Language*, 11(1):43–72.

Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R. M., and Rohlicek, J. R. (1991). Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses. In *Proceedings of the Workshop on Speech and Natural Language (HLT '91)*, pages 83–87.

Potamianos, G. and Neti, C. (2001). Automatic Speechreading of Impaired Speech. In *Proceedings of the 2001 International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 177–182.

Saldarriaga, S. P. and Cheriet, M. (2011). Indexing On-line Handwritten Texts Using Word Confusion Networks. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR'11)*, pages 197–201.

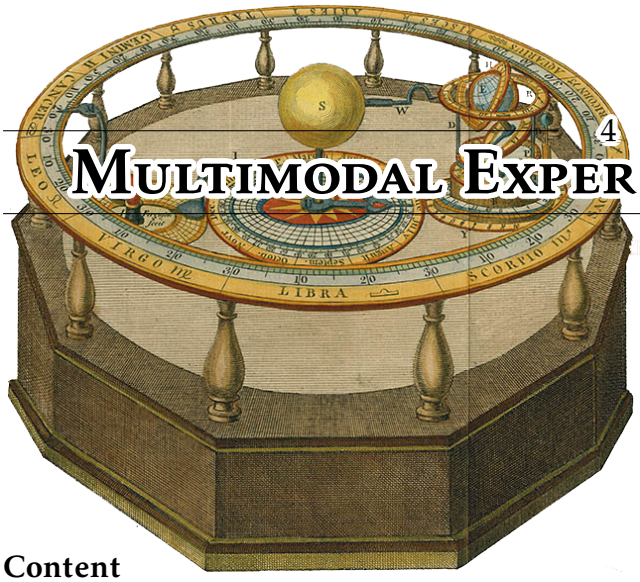
Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Gadde, V. R. R., Plauché, M., Richey, C., Shriberg, E., Sönmez, K., Weng, F., and Zheng, J. (2000). The SRI March 2000 Hub-5 conversational speech transcription system. In *Proceedings of the NIST Speech Transcription Workshop*.

Stolcke, A., Konig, Y., and Weintraub, M. (1997). Explicit Word Error Minimization in N-best List Rescoring. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, volume 1, pages 163–166.

Woodruff, P. and Dupont, S. (2005). Bimodal combination of speech and handwriting for improved word recognition. In *Proceedings of the 13th European Signal Processing Conference (EUROSIPCO 2005)*.

Xue, J. and Zhao, Y. (2005). Improved confusion network algorithm and shortest path search from word lattice. In *Proceedings of the International Conference in Acoustics, Speech and Signal Processing (ICASSP 2005)*, volume 1, pages 853–856.

Zhai, C. and Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.



4 Multimodal Experimental Results

"No importa
-decía José Arcadio Buendía-.
Lo esencial es no perder la orientación."[1]

Gabriel García Márquez, 100 años de soledad, 1967.

Content

Handwritten Text Recognition (HTR) allows us to speed up the manual transcription process of digitalised historical manuscripts. However, recent research shows that as other modalities of natural human language can be useful for improving the obtained draft transcription to be corrected by paleographers (Alabau et al., 2011, 2014). For instance, this is the case of Automatic Speech Recognition (ASR) on the dictation of the contents of the historical manuscript that can be used as an additional source of information in a multimodal combination. Apart from that, an iterative process can be used in order to refine the final hypothesis. Moreover, more than one recognition system per modality can be combined to obtain a draft transcription, given that combining the outputs of different recognition systems will generally improve the recognition accuracy.

This chapter describes the performed experiments to test the effectiveness of the combination method presented in the previous chapter (Chapter 3). In an initial experiment, this proposal is tested on two different Spanish historical books with different difficulty level. The obtained results show that

[1] "It's all right -José Arcadio Buendía said-. The main thing is not to lose our bearings."
Illustration info: The calculator (James Ferguson, Astronomy explained upon Sir Isaac Newton's Principles, and made easy to those who have not studied mathematics, 1806).

the proposed technique provides similar or better draft transcriptions than a previously proposed approach (Alabau et al., 2011, 2014), allowing for an easier transcription process. In a second experiment, its effectiveness when combining several recognisers of both modalities is tested for transcribing a Spanish historical book. Results present improvements on both unimodal combination (with different optical -for HTR- and acoustical -for ASR- models), and on multimodal combination, where the improvement is statistically significant. In this second experiment, the improvements produced in the set of hypotheses contained in the resulting Confusion Network is studied, and the difficulty of reaching the best hypothesis (oracle) is estimated. Finally, in a last experiment, the performance of our proposal is compared with other well established combination techniques on both Spanish historical books.

The rest of the chapter is structured as follows: Section 4.1 presents the experimental framework, Section 4.2 explains the experimental setup, Section 4.3 shows the results of the iterative and non-iterative combination experiments, Section 4.4 draws the results of the unimodal and multimodal combination experiments, Section 4.5 describes the results of the comparison with other combination techniques, and Section 4.6 summarises the conclusions.

4.1 Experimental Framework

This section introduces the datasets, features, models, and evaluation metrics used on the experiments.

4.1.1 Datasets

Historical Manuscript Corpora

The following two different historical manuscripts were employed in the experimentation:

- *Cristo Salvador* is a single writer handwritten book of the 19th century provided by *Biblioteca Valenciana Digital* (BiValDi). This corpus represents a total number of 1,172 lines, with a vocabulary of 3,287 different words. For training the optical models for HTR, a partition with the first 32 pages (675 lines) was used. Test data for HTR is composed of the lines of page 41 (24 lines, 222 words). This corpus is the same employed in (Alabau et al., 2011, 2014) and allows for an initial comparison with our proposal.
- *Rodrigo* (Serrano et al., 2010) is a corpus obtained from the digitalisation of the single writer book "Historia de España del arçobispo Don Rodrigo", written in Early Modern Spanish in 1545. It is composed of 853 pages that were automatically divided into lines (see example in Figure 2.11), giving a total number of 20,356 lines. The vocabulary size is of about 11,000 words. For training the optical models, a standard partition with a total number of 5,000 lines (about 205 pages) was used. Test data for HTR was composed of two pages that were not included in the training part (pages 515 and 579). These two pages contain 50 lines and 514 words.

Figure 4.1 presents two text line samples, one for each one of the historical manuscripts used in the experiments. More information about these datasets can be found in Section 2.8.1.

Speech Training Corpus: *Albayzin*

The acoustical models for ASR were trained by using a partition of the *Albayzin* Spanish database (Moreno et al., 1993) presented in Section 2.8.3. The training partition used includes 4800 phonetically balanced utterances, with a total length of about 4 hours.

Multimodal (Text - Speech) Corpora

Test data for ASR was the product of the controlled acquisition (see Section 2.8.4 for more information) of the dictation of the contents of the lines contained in the test set of both historical manuscripts.

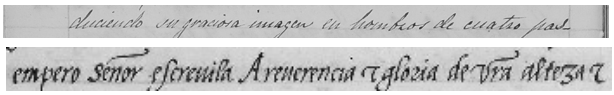


Figure 4.1: Sample lines for *Cristo Salvador* (top) and for *Rodrigo* (bottom).

Specifically, in the case of *Cristo Salvador*, the ASR test data was composed of the acquisition of the dictation of the contents of the 24 test lines by five different speakers (i.e., a total set of 120 utterances, with a total length of about 9 minutes), while in the case of *Rodrigo* seven different speakers read the 50 handwritten test lines, giving a total set of 350 utterances (about 15 minutes).

4.1.2 Features

Off-line handwritten text features were computed in several steps as explained in Section 2.2.2. Final regular feature vectors were composed of 60 dimensions.

Mel-Frequency Cepstral Coefficients (MFCC) were used as speech features. These regular ASR features were obtained following the procedure presented in Section 2.2.1, resulting in 39 dimensional features vectors.

In the tandem feature extraction scheme (see Section 2.2.4 for more information), two Multi-Layer Perceptrons (MLP) with 2,000 neurones at the hidden layer and a softmax transfer function at the output layer were trained to estimate symbol-phoneme posterior probabilities. On the one hand, for HTR, a MLP with 60 neurones at the input layer and 106 neurones at the output layer was trained with Torch (Collobert et al., 2002). On the other hand, for ASR, the MLP had 39 neurones at the input layer and 25 neurones at the output layer and it was trained by using QuickNet (Johnson, 2004). Both MLP were trained by backpropagation with a mean-squared error criterion. The final tandem features are constituted by the log posteriors probabilities of the MLP.

4.1.3 Models

Optical and acoustical models were trained by using HTK (Young et al., 2006). On the one hand, symbols on the optical models are modelled by a continuous density left-to-right HMM with 12 and 4 states for *Cristo Salvador* and *Rodrigo*, respectively, and 32 Gaussians per state. On the other hand, phonemes on the acoustical model are modelled as a left-to-right HMM with 3 states and 64 Gaussians per state.

In order to test the influence of the speaker adaptation in the acoustical models for the *Rodrigo* corpus, the speaker independent acoustical models were adapted to each speaker and page using HTK's Maximum Likelihood Linear Regression global adaptation (Young et al., 2006). For each independent acoustical model, two adapted models were obtained per speaker, since we got the adapted models for decoding one page by using the audio samples of the other page, and vice-versa.

The lexicon models were estimated in the HTK lexicon format, where each word is modelled as a concatenation of symbols for HTR or phonemes for ASR.

The baseline language models (LM) were estimated as a 2-gram with Kneser-Ney back-off smoothing (Kneser and Ney, 1995) directly from the transcriptions of the pages included on the HTR training sets (32 pages for *Cristo Salvador*, and about 205 pages for *Rodrigo*). The *Cristo Salvador* LM was interpolated with the whole lexicon in order to avoid out-of-vocabulary (OOV) words. In contrast, the *Rodrigo* LM presents a 6.2% of OOV words. All processes on language models (inference, interpolation, ...) were done by using the SRILM toolkit (Stolcke, 2002).

4.1.4 Evaluation Metrics

The measures used to evaluate the obtained experimental results are the following: WER, oracle WER, CER, and oracle CER, with their corresponding confidence intervals at 95%. The statistical significance was confirmed by using p-values and a significance threshold set at $\alpha = 0.025$. On the other hand, perplexity was used to measure the performance of language models. Section 2.7 explains the details of these evaluation measures.

The statistical dispersion of the position (in a n-best list) of the best hypothesis (the best but maybe not the most likely) that allow one to obtain the oracle value permits to estimate the difficulty of reaching this oracle value. The Interquartile Range (IQR), the median, and the Median Absolute Deviation (MAD) were used to measure this statistical dispersion.

4.2 Experimental Setup

The recognition systems were implemented by using the iATROS (Luján-Mares et al., 2008) recogniser, and the SRILM *lattice-tool* (Stolcke, 2002) utility was used to obtain CN from the WG obtained in the decoding processes.

The values of the main decoding variables (GSF, WIP, and *beam factor*) were tuned with respect to the test set so as to optimise the experimental baseline results. In this way, the main research challenge of this thesis was presented, that is, to improve, after the recognition process, the best transcription that recognisers are able to offer. As explained in Section 2.4 these parameters have a significant effect on the decoding performance. In an initial tuning, the best results were obtained by using the values of the upper half of Table 4.1 for both modalities and each data set. These values were used only on the first experiment (Section 4.3). Afterwards, a more extensive search was performed, and new combinations of decoding parameters that allowed us to improve the decoding results were found. These final decoding values are presented on the bottom half of Table 4.1. These final values were used for obtaining the decoding baseline for the rest of the experiments performed in this thesis.

Table 4.1: Tuning of the main decoding variables.

Tunning	Modality	*Cristo Salvador*			*Rodrigo*		
		GSF	WIP	*beam factor*	GSF	WIP	*beam factor*
Initial	HTR	80	160	3000	20	-20	600
	ASR	5	7	1000	20	-20	400
Final	HTR	60	200	3000	30	-20	3500
	ASR	5	7	1000	15	-5	300

In the experiments, the anchor subnetworks search was performed several times, looking for skip-bigrams (allowing only one wrong word in the gap) and unigrams, throughout the whole hypothesis in the subnetworks. Specifically, we started with a perfect matching search of skip-bigrams, followed by a perfect matching search of unigrams. Next, we made a relaxed matching search of skip-bigrams, and a relaxed matching search of unigrams, setting the matching error threshold to $\epsilon = 2^{-1/2}$. A relaxed search with this threshold would allow to align words like *3* and *TRES* that coincide in their phonetical transcription (['tres]), since it is the same word written differently. We used a weight factor of $\alpha = 0.5$ (despite the fact that in the baseline experiments the HTR is more reliable than the ASR), a granularity for smoothing of $\Theta = 10^{-4}$, and $\gamma = 0.25$ and $\delta = 0.75$ as thresholds for insertion and deletion respectively.

4.3 Experiment 1: Iterative and Non-Iterative Combination

In order to test the performance of the multimodal combination method presented in Chapter 3, we have experimented with two HTR corpus (*Cristo Salvador* and *Rodrigo*, for more information see Section 2.8) with different degrees of complexity. For both corpora, we started obtaining the baseline values for both modalities. Next, we reproduced the technique of (Alabau et al., 2011) to get the multimodal WG baseline, and the technique of (Alabau et al., 2014) to get the iterative multimodal WG baseline during 10 iterations. Finally, we made the multimodal CN combination described in Section 3.3, and we got the iterative multimodal CN results.

At each iteration of the iterative experiments, the original LM (the LM obtained on the training) is interpolated with the decoding results so as to obtain an improved LM for the next decoding process (Alabau et al., 2014).

4.3.1 Experiments with *Cristo Salvador*

The obtained results are shown in Table 4.2. The baseline results are similar to those presented in (Alabau et al., 2011). In our experiments, the best results on the multimodal WG were obtained in the direction $HTR \rightarrow ASR$. Compared with the HTR baseline there is an improvement, but the confidence intervals overlap ($p = .076$). However, with the iterative multimodal WG a significantly ($p = .020$) relative improvement of 30% of WER is achieved in the same direction.

Table 4.2: *Cristo Salvador* experiment results. Best results in **boldface**.

Experiment	WER	CER	LM Perplexity
Baseline HTR	33.3% ± 7.0	15.8% ± 3.9	742.8
Baseline ASR	43.2% ± 3.2	20.3% ± 1.8	742.8
Multimodal WG	25.5% ± 2.9	11.2% ± 1.4	102.3
Iterative Multimodal WG	**23.3% ± 2.9**	**11.1% ± 1.5**	**27.4**
Multimodal CN	30.6% ± 2.9	14.2% ± 1.7	46.9
Iterative Multimodal CN	26.0% ± 2.7	12.0% ± 1.5	31.6

Furthermore, the improvement produced by our CN technique achieves similar results to those of the WG technique, in both multimodal and iterative multimodal experiments. However, these results are not significant when compared to the HTR baseline.

Perplexity measures the usefulness of a language model for decoding a reference text. As can be observed in Table 4.2, the original LM (used in the HTR and ASR baseline decodings) presents a huge perplexity value (742.8) that gives an idea of the difficulty of the task. After the first interpolation in the multimodal approaches, the perplexity of the obtained LM drops considerably (102.3 and 46.9 for WG and CN, respectively). Finally, after 10 iterations, the perplexity of the final interpolated LM drops to a value close to 30 for both approaches.

4.3.2 Experiments with *Rodrigo*

The same experimental procedure was used with this corpus. Nevertheless, the results are completely different. Although the original LM for *Rodrigo* presents a lower perplexity (298.4) compared with the observed in the previous experiment, some differences as the language (*Rodrigo* was written in Early Modern Spanish and *Cristo Salvador* in Modern Spanish), and the presence of hyphenated and OOV words make *Rodrigo* a much more challenging corpus. Baseline values in both HTR and ASR are very high, due to the difficulty of the corpus, and to the fact that the decoding parameters (upper half of Table 4.1) were not optimal, as we observed posteriorly.

Table 4.3 shows the results obtained. On the multimodal WG, not only the ASR is ineffective

Table 4.3: *Rodrigo* experiment results. Best results in **boldface**.

Experiment	WER	CER	LM Perplexity
Baseline HTR	45.1% ± 5.4	23.9% ± 5.2	298.4
Baseline ASR	68.9% ± 2.3	41.6% ± 1.6	298.4
Multimodal WG	45.1% ± 2.2	24.7% ± 2.3	74.5
Iterative Multimodal WG	42.6% ± 2.0	24.0% ± 1.9	73.8
Multimodal CN	38.8% ± 1.7	19.3% ± 0.9	81.0
Iterative Multimodal CN	**38.0**% ± **1.6**	**17.4**% ± **0.9**	**56.9**

correcting the HTR, but also worsens (although, not statistically significant, $p = .497$) the CER level on the best results obtained in the direction *ASR* → *HTR*. The iterative multimodal WG produces a slight and not significant ($p = .414$) WER improvement in the direction *HTR* → *ASR*.

On the other hand, our Multimodal CN technique produces improvements on both WER and CER level, from the first combination. The iterative multimodal CN achieves significant improvements, with a relative improvement of 15.7% of WER ($p = .023$) and 27.2% of CER ($p < .001$).

Regarding the LM perplexities, as in the previous experiment, after the first interpolation in the multimodal approaches, the perplexity of the obtained LM drops considerably (from 298.4 to 74.5 for WG and 81.0 for CN). However, after 10 iterations, the perplexity of the final interpolated LM drops slightly for the WG approach to 73.8, while for our CN approach achieves a value of 56.9.

4.4 Experiment 2: Unimodal and Multimodal Combination

Hierarchical combination was tested with the *Rodrigo* corpus. In this experiment, we used regular HTR and ASR features (see Section 2.2), and the tandem procedure (Hermansky et al., 2000) to obtain tandem HTR and ASR features (see Section 2.2.4). Moreover, the MLLR speaker adaptation technique (Leggetter and Woodland, 1995) was used to improve the performance of the acoustical models (more information about morphological models and its adaptation can be found in Section 2.3.1). In this way, six different recognition systems were obtained: two HTR systems (regular and tandem), and four ASR systems (regular and tandem, with and without speaker adaptation). As a first step, the reference values for each recognition system were obtained. As a second step, the unimodal combination was performed. Finally, the multimodal combination was conducted. With regard to the order of the combinations, the best results were obtained with the order represented in Figure 4.2. However, the differences were not significant with respect to the rest of possible combination orders.

4.4.1 Baseline Experiments

The reference values shown in Table 4.4 and Table 4.5 were obtained by using the selected final values for the decoding parameters presented on the bottom half of Table 4.1. As can be observed, the HTR reference values are better than the ASR reference values. As to the baseline HTR results (Table 4.4), the tandem system produces lower error rates. Therefore, we take these values as the baseline reference for this experiment; namely, 32.9% for WER and 15.7% for CER. Regarding the oracle baseline results, the lower bounds were obtained in the HTR modality; specifically, 24.9% for WER and 11.6% for CER. Both are also from the tandem system.

The baseline ASR results (Table 4.5) are quite poor because of the difficulty of the corpus. In *Rodrigo* were faced with text images containing hyphenated words (e.g., *REYNA*, where a part of the word *RE* is at the end of a line and the second part *YNA* is at the beginning of the following line), abbreviations (e.g., *NRÕ*) that are pronounced as the whole word (*NUESTRO* [’nwes tro]), and words written in multiple forms (e.g., *XPIÃNOS* and *CHRISTIANOS*, or numbers as *5* and *V*) but that are pronounced in the same way ([kris ’tja nos], [’θiŋ ko]). In spite of these facts, speaker adaptation and

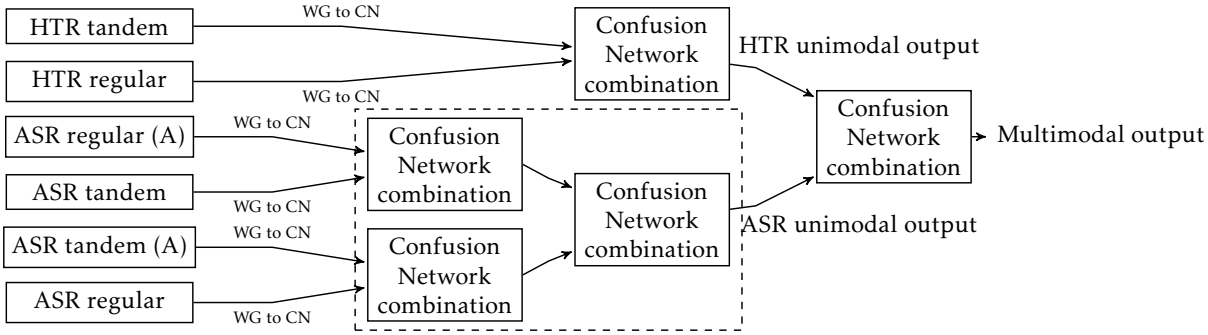


Figure 4.2: Unimodal - Multimodal combination diagram. (A) denotes speaker adapted models

Table 4.4: HTR baseline results for the *Rodrigo* corpus. Best results are highlighted in **boldface**.

System	WER	Oracle WER	CER	Oracle CER
Regular	39.3% ± 4.1	28.0% ± 3.1	20.6% ± 2.5	14.0% ± 2.1
Tandem	**32.9% ± 5.3**	**24.9% ± 5.6**	**15.7% ± 4.9**	**11.6% ± 4.9**

Table 4.5: ASR baseline results for the *Rodrigo* corpus. (A) denotes speaker-adapted models and the best results are highlighted in **boldface**.

System	WER	Oracle WER	CER	Oracle CER
Regular	62.9% ± 2.2	44.9% ± 2.1	35.4% ± 1.4	25.1% ± 1.4
Regular (A)	**51.0% ± 2.2**	**33.7% ± 2.0**	**26.4% ± 1.3**	**17.2% ± 1.2**
Tandem	58.6% ± 2.2	41.1% ± 2.0	31.3% ± 1.3	21.9% ± 1.3
Tandem (A)	55.5% ± 2.2	38.4% ± 2.1	28.9% ± 1.3	19.7% ± 1.1

tandem features provide improvements for ASR when compared to the regular baseline. Specifically, the best results were obtained with the regular model with speaker adaptation (A). Regarding the WER, a value of 51.0% was obtained, while the value of CER reached 26.4%. In this experiment, the search of oracle values was limited to the 2000-best. As can be observed, the ASR oracle values are worse than the HTR baseline values.

4.4.2 Unimodal Combination Experiments

In these experiments, the output of the different systems of the same modality were combined. The input signals to each recognition system are the same. Therefore, the combination of these outputs does not represent the difficulty of asynchrony.

On the one hand, in the case of the HTR unimodal combination (Figure 4.2), once the image signals were processed through each HTR system, the WG outputs were transformed into CN. Then, these CN were processed by our CN combination technique, which returned a new CN by combining the information from both HTR systems.

On the other hand, for the ASR unimodal combination, as our technique allows us to combine only two CN and we used four ASR systems, it was necessary to use the CN combination technique three times as described in Figure 4.2. First, the voice signals were processed through each ASR system, and their WG outputs were transformed in CN. Secondly, these CN were processed by our CN combination technique by pairs, in order to obtain two combined CN. Finally, the two combined CN were processed by our CN combination technique. Thereby, we got a new CN that combines the information from the four ASR systems.

Table 4.6: Combination results for the *Rodrigo* corpus.

Combination	WER	Oracle WER	CER	Oracle CER
HTR unimodal	31.1% ± 3.7	20.6% ± 3.2	13.3% ± 1.9	7.8% ± 1.4
ASR unimodal	50.4% ± 2.0	34.9% ± 1.9	28.6% ± 1.3	18.5% ± 1.2
Multimodal	**28.2% ± 1.3**	**16.6% ± 1.3**	**13.1% ± 0.7**	**6.2% ± 0.5**

As shown in Table 4.6, the HTR unimodal combination reduced the error with respect to the baseline HTR system at both the WER and CER level: 1.8% and 2.4% respectively. Meanwhile, the combination introduced new information in the new CN that reduced the oracle levels, being of special interest the case of the oracle CER since it presents 32.8% of relative reduction over the HTR oracle CER baseline reference. In the case of the ASR unimodal combination, only a small improvement is produced at the WER level when compared to the ASR baseline.

4.4.3 Multimodal Combination Experiment

In the multimodal combination experiment, the unimodal CN were combined with the aim of obtaining a multimodal CN (Figure 4.2). The multimodal CN combination produced improvements compared with the results obtained by the unimodal combinations, despite the high error values obtained by the ASR unimodal combination. As can be seen in Table 4.6, a relative WER improvement of 9.3% is achieved when compared to the HTR unimodal combination, and this relative improvement increases to 14.3% when compared to the baseline reference as well. Furthermore, in terms of the CER, a relative improvement of 1.5% is produced when compared to the HTR unimodal combination. The relative improvement over the baseline reached 16.6% for this approach.

The oracle WER level presents a statistically significant ($p = .001$) relative improvement of 33.3% when compared to the oracle WER baseline, while the statistically significant ($p < .001$) relative improvement presented by the oracle CER level when compared with the oracle CER baseline attained 46.6%.

4.4.4 Difficulty of Reaching the Oracle Values

The difficulty of reaching the oracle values from the information contained in each Confusion Network was estimated by means of a statistical study of the dispersion of the n-best positions that allowed to achieve those oracle values. In this experiment, the search of oracle values was limited to the 2000-best.

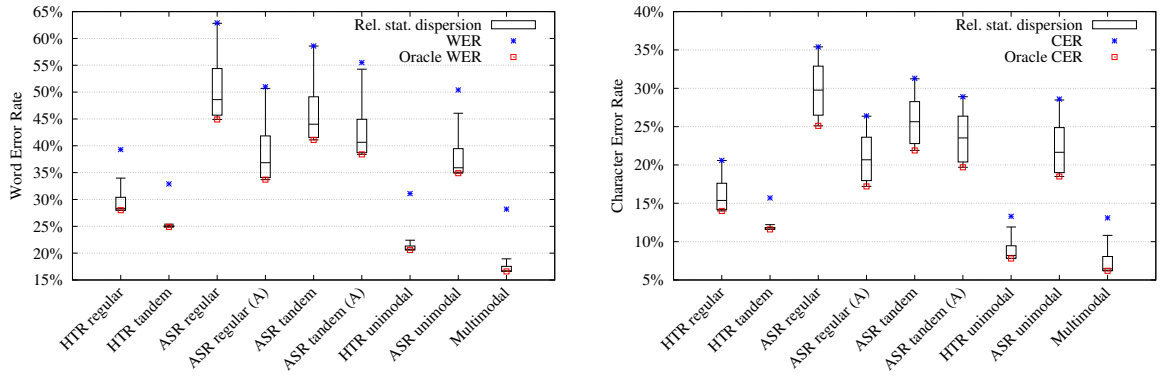
In Table 4.7, the statistical dispersion obtained for each CN is outlined for oracle WER and oracle CER. Regarding the CN obtained from the recognition systems, the CN obtained from the tandem model of the HTR offered the narrowest interquartile range (IQR), side by side with the lowest median and median absolute deviation (MAD) values. On the opposite side, all the CN of the ASR presented a wide IQR with high median and MAD values. As to the CN acquired from the combinations, the depth of the search is reduced because of the increase in the amount of information and the word probability correction resulting from the combination.

To show the importance of these statistical dispersions, it is necessary to represent them with their related error values, as in Figure 4.3. In Figures 4.3(a) and 4.3(b), the statistical dispersions are plotted as box plots, where the positions are normalised, the range of values is set to the difference between the 1-best error value and the oracle error value, and the minimum value is set to the oracle error value. In the box plots, the IQR, the median, the minimum and the maximum values of the statistical dispersions are represented. The smaller the box plot, the easier it will be to reach the oracle error value. Therefore, the better systems have the lower values of oracle error, WER, etc.

Regarding the oracle WER (Figure 4.3(a)), in the CN of the reference system (HTR tandem) it is very easy to reach the oracle WER value (24.9%) from the WER value (32.9%), whereas in the CN from the different ASR systems it is not easy to reach the oracle WER value. With the combination, the

Table 4.7: Statistical dispersion of the positions in the n-best list of the hypothesis that allows one to obtain the oracle values. (A) denotes speaker-adapted models.

Confusion network	Oracle WER			Oracle CER		
	IQR	Median	MAD	IQR	Median	MAD
HTR regular	423	62	61	1041	415	409
HTR tandem	50	2	2	119	17	16
ASR regular	967	412	398	1242	906	617
ASR regular (A)	897	365	359	1230	756	601
ASR tandem	866	334	329	1164	796	591
ASR tandem (A)	728	263	260	1300	831	658
HTR unimodal combination	137	26	25	594	144	142
ASR unimodal combination	567	129	127	1161	625	573
Multimodal combination	166	29	28	510	100	99



(a) Statistical dispersion relative to WER and oracle WER.

(b) Statistical dispersion relative to CER and oracle CER.

Figure 4.3: Relative statistical dispersion in the set of the positions in the n-best list of the hypothesis that obtain the oracle values. (A) denotes speaker-adapted models.

difficulty of reaching the oracle values is reduced, especially in the CN obtained from the multimodal combination where it is quite easy to reach the oracle WER value (16.6%) from the WER value (28.2%). In the oracle CER (Figure 4.3(b)), a similar behaviour is observed.

4.5 Experiment 3: Multimodal Combination Comparative

From the previous experiments, we can say that the multimodal combination of different sources of information (in this case HTR and ASR) performed by using the proposed method provides a better draft transcription than the HTR baseline to be offered to the paleographer in a *post-edition* correction approach. However, one advantage of this multimodal combination is the fact that this method also enriches the set of hypotheses contained in the resulting lattice formatted as CN. This additional enrichment can be very useful for improving the performance of interactive and assistive transcription systems, and for developing further ideas such as using multimodality and crowdsourcing for document transcription.

The next experiment was performed in order to compare the performance of our multimodal CN combination proposal on two different handwritten text datasets (*Cristo Salvador* and *Rodrigo*) with other three well stablished combination techniques (ROVER, N-best ROVER, and Lattices Rescoring: see Section 3.2 for more information about combination of natural language recognition systems). Given that usually on interactive transcription systems the corrections and measures are made at word level,

in the following experiment the obtained results are compared at word level. Moreover, in this experiment the best WER values contained in the lattices, i.e. the oracle WER values, were obtained without limitation by using the SRILM toolkit (Stolcke, 2002).

In order to optimise the experimental results, the values of the main decoding parameters (GSF, WIP, and *beam factor*) were set to the values presented on the bottom half of Table 4.1. In the multimodal combination both modalities were equally weighted. Therefore, the different combination parameters were: frequency of occurrence voting scheme for ROVER (since the multimodal combination is performed without training), uniform weights ($\lambda = 1$ and $\mu = 0.5$) for N-best ROVER, and combination weight of 0.5 for the Lattices Rescoring and CN Combination techniques.

For both historical manuscripts, the unimodal lattices were obtained from the decoding processes of both modalities, and then, the obtained lattices of both modalities were combined by using the different combination techniques.

Table 4.8: Multimodal combination comparative. Best results are highlighted in **boldface**.

Experiment		*Cristo Salvador*		*Rodrigo*	
		1-best WER	Oracle WER	1-best WER	Oracle WER
Unimodal	Baseline HTR	32.9% ± 6.8	27.5% ± 6.4	39.3% ± 4.1	28.2% ± 3.0
	Baseline ASR	43.3% ± 3.4	27.4% ± 2.2	62.9% ± 2.2	29.5% ± 1.6
Multimodal	ROVER	32.7% ± 2.9	32.7% ± 2.9	44.9% ± 1.8	44.9% ± 1.8
	N-best ROVER	33.3% ± 2.9	33.3% ± 2.9	38.4% ± 1.8	38.4% ± 1.8
	Lattices Rescoring	31.3% ± 2.6	**10.3**% ± **1.7**	37.2% ± 1.7	**10.6**% ± **0.9**
	Proposed CNC	**29.3**% ± **2.5**	13.4% ± 2.1	**35.9**% ± **1.6**	14.8% ± 1.0

Similar behaviour can be observed for both data sets in the final results presented in Table 4.8. Regarding the unimodal results, speech recognition does not seem to be a good substitute for handwriting recognition for transcribing historical manuscripts. However, the lattices obtained from the ASR decoding processes present similar oracle WER values to those obtained from the HTR decoding processes.

As could be expected, in the multimodal experiments the use of all hypotheses (compactly contained in lattices) in the combination (used by Lattices Rescoring and our CN Combination proposal) allows one to obtain better draft transcriptions to be corrected by a paleographer on a *post-edition* procedure. Concretely, the best WER results were obtained by using our proposed CN Combination method. In the case of *Cristo Salvador*, 29.3% ± 2.5 of WER was obtained, representing a relative improvement over the HTR baseline WER (32.9%±6.8) of 10.9%. For *Rodrigo* a WER value of 35.9%±1.6 was obtained, with a relative improvement over the HTR baseline WER (39.3% ± 4.1) of 8.7%. Concerning the oracle WER, the best values were obtained by using the Lattices Rescoring method. Specifically, oracle WER values of 10.3% and 10.6% were obtained for *Cristo Salvador* and *Rodrigo*, respectively.

Results show that WER improvements are not significant with respect to the HTR baseline WER, but the oracle WER values are statistically significant lower ($p < .001$) in the case of lattice combination methods. Therefore, an outstanding effect of multimodal lattices combination in interactive transcription systems can be expected, since this low oracle WER is related to the amount and quality of the alternatives offered by the combination technique (the lower the oracle WER, the more and better alternatives).

4.6 Conclusions and Future Work

In this chapter, the benefits of combining additional sources of information for the transcription of historical manuscripts has been confirmed. The proposed combination method takes advantage of the fact that different systems make different errors; thus, editing operations can correct errors. Insertion and deletion create new word sequences than enrich the resulting CN, and the combination can maximise

the probability of the correct word, when both subnetworks contain the correct word, even when this word has a low probability in both subnetworks. Conversely, if only one subnetwork contains the correct word and both subnetworks contain the same erroneous word, this error will be maximised at the expense of the correct word. Despite this, the experiments performed confirm the strengths of this CN combination technique. In the iterative fashion, an additional advantage is that the new word sequences generated by the proposed method enrich the language model, reducing their perplexity for the next iteration, even with the existence of Out-Of-Vocabulary words.

The results observed lead us to believe that there is still room for improvement. We propose for future studies the possibility of using not lines but whole sentences of the handwritten text corpus in order to 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.

Eventually, integrating multimodality in an interactive and assistive transcription system will reduce the time and the workload of paleographers for transcribing historical books, due to the increased recognition accuracy and the quality of the alternatives contained in the multimodal lattice. The work realised for integrating multimodality on an interactive tool for transcribing historical handwritten documents is detailed in the next part of this thesis (Part III).

Bibliography

Alabau, V., Martínez-Hinarejos, C.-D., Romero, V., and Lagarda, A. L. (2014). An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195–203. Frontiers in Handwriting Processing.

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

Collobert, R., Bengio, S., and Mariéthoz, J. (2002). Torch: a modular machine learning software library. Reseach Report IDIAP-RR-02-46, IDIAP Research Institute.

Hermansky, H., Ellis, D. P. W., and Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 3, pages 1635–1638.

Johnson, D. (2004). ICSI Quicknet software package. `http://www1.icsi.berkeley.edu/Speech/qn.html`. Last access: May 2015.

Kneser, R. and Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, volume 1, pages 181–184.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.-D., Pastor, M., Sanchis, A., and Toselli, A. (2008). iATROS: A speech and handwritting recognition system. In *Proceedings of the V Jornadas en Tecnologías del Habla (VJTH'2008)*, pages 75–78.

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

Serrano, N., Castro, F., and Juan, A. (2010). The RODRIGO Database. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2709–2712.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the 3rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 901–904.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

III

Interactivity

"The laws of chess are as beautiful as those governing the universe - and as deadly."

Katherine Neville, A Calculated Risk, 1992.

Illustration of coexistence between two great cultures. (Juegos diversos de Axedrez, dados, y tablas, con sus explicaciones, ordenados por mandado del Rey Don Alonso X el Sabio, 1251-1283)

5 Assistive Transcription

"Il ne peut y avoir de langage plus universel et plus simple,
plus exempt d'erreurs et d'obscurités,
c'est-à-dire plus digne d'exprimer les rapports invariables des êtres naturels. [...]
l'analyse mathématique [...]
semble être une faculté de la raison humaine
destinée à suppléer à la brièveté de la vie et à la imperfection des sens."[1]

Jean-Baptiste Joseph Fourier, Théorie Analytique de la Chaleur, 1822.

Content

The initial result of automatic recognition may make the paleographer's task easier, since they are able to correct on a draft transcription. However, given that paleographer revision is required to produce a transcription of standard quality, an interactive assistive scenario, where the automatic system and the paleographer cooperate to generate the perfect transcription, would reduce the time and the paleographer effort required for obtaining the final result.

In this context, the assistive transcription system proposes a hypothesis, usually derived from a recognition process. The recognition can be unimodal (e.g., from a handwritten text image or an audio utterance with its dictation) or multimodal (two or more signals which represent the same sequence of words). Then, the paleographer reads it and produces a feedback signal (first error correction, positioning, 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.

In this chapter, we present a multimodal interactive transcription system where user feedback is provided by means of touchscreen pen strokes, traditional keyboard, and mouse operations. The combination of the main and the feedback data streams is based on the use of Confusion Networks derived from the output of three recognition systems: two Handwritten Text Recognition systems (*off-line* and *on-line*), and an Automatic Speech Recognition system. *Off-line* text recognition and speech recognition are used to derive (by themselves or by combining their recognition results) the initial hypothesis, and *on-line* text recognition is used to provide feedback. The use of the proposed multimodal interactive assistive system not only reduces the required transcription effort, but it also helps to optimise the overall performance and usability, allowing for a faster and more comfortable transcription process.

[1] "There cannot be a language more universal and more simple, more free from errors and obscurities, i.e. more worthy to express the invariable relations of all natural things. [...] mathematical analysis [...] seems to be a faculty of human reason to complement the brevity of life and the imperfection of the senses."
Illustration info: Buddha offers fruit to the devil, (Rashid-al-Din Hamadani, Jāmi' al-tawārīkh, 14th century).

The rest of the chapter is organised as follows: Section 5.1 introduces the CATTI framework; Section 5.2 specifies the particulars of our multimodal CATTI assistive transcription proposal; and Section 5.3 offers the final conclusions.

5.1 Computer Assisted Transcription Overview

In the last few years, the use of natural language recognition systems has allowed us to speed up the manual transcription of digitised documents, usually done by professional transcribers. However, state-of-the-art natural language recognition systems are far from being perfect, and human revision is required to produce a transcription of standard quality. Therefore, once the full recognition process of one document has finished, heavy human expert revision is required to really produce a transcription of standard quality. Such a post-editing solution is rather inefficient and uncomfortable for the human corrector.

In order to reduce the time and human effort required for obtaining the perfect transcription of digitised documents, transcribers can use interactive and assistive approaches, where the transcriber and the computer work together to obtain the perfect transcription. This is the case of Computer Assisted Transcription (CAT) of speech (Rodríguez et al., 2007) or handwritten text documents (Toselli et al., 2007). For instance, the assistive framework called Computer Assisted Transcription of Text Images (CATTI) (Romero et al., 2012) provides transcribers an initial draft transcription of the handwritten text image or speech utterance. Then, at each interaction step, the CATTI system uses the information obtained from automatic recognition processes, and the part of the transcription (prefix) corrected and validated by the transcriber to propose a new, hopefully better, continuation.

In the CATTI framework, the user is directly involved in the transcription process, since he/she is responsible for validating and/or correcting the system hypothesis during the transcription process. The system takes into account the handwritten text image and the feedback of the user in order to improve these proposed hypotheses. The process starts when the system proposes a full transcription $\hat{\mathbf{s}}$ of a text line image. Then, the user reads this transcription until finding a mistake and makes a Mouse Action (MA) m, or equivalent pointer-positioning keystrokes, to position the cursor at this point. By doing so, the user is already providing some very useful information to the system: he is validating a prefix $\hat{p}$ of the transcription, which is error-free and, in addition, he is signalling that the following word e located after the cursor is incorrect. Hence, the system can already take advantage of this fact and directly propose a new suitable suffix, i.e. a new $\hat{\mathbf{s}}$ in which the first word is different from the first wrong word of the previous suffix. In this way, many explicit user corrections are avoided (Romero et al., 2009). If the new suffix $\hat{\mathbf{s}}$ corrects the erroneous word, a new cycle starts. However, if the new suffix has an error in the same position than the previous one, the user can make a new MA or can enter a word v to correct the erroneous one. This last action produces a new prefix $\hat{p}$ (the previously validated prefix followed by the new word v). Then, the system takes into account the new prefix to suggest a new suffix and a new cycle starts. This process is repeated until a correct transcription is accepted by the user.

Figure 5.1 illustrates an example of the CATTI process. In this example, without interaction with a CATTI system, a user should have to correct about three errors from the original recognised hypothesis (*abadia*, *segun* and *el*). Using CATTI only one explicit user-correction is necessary to get the final error-free transcription in two CATTI iterations: the iteration 1 only needs one MA to find the correct word, but in the iteration 2 a single MA does not succeed and the correct word needs to be typed.

The CATTI framework can be defined as a traditional natural language recognition problem - Equation (2.1)-. In this case, in addition to the given feature sequence $\hat{x}$, a prefix $\hat{p}$ of the transcription is available, depending on the editing operation that the user performed to correct the erroneous text. The editing operations considered are substitution, insertion, deletion, and rejection (Romero et al., 2012). Therefore, the CATTI system should try to complete the transcription from this prefix $\hat{p}$ by searching for a most likely suffix $\hat{\mathbf{s}}$:

$$\hat{\mathbf{s}} = \arg\max_{\hat{s}\in\hat{S}} \mathrm{P}(\hat{s} \mid \hat{x}, \hat{p}) \approx \arg\max_{\hat{s}\in\hat{S}} \mathrm{P}(\hat{x} \mid \hat{p}, \hat{s})\mathrm{P}(\hat{s} \mid \hat{p}) \tag{5.1}$$

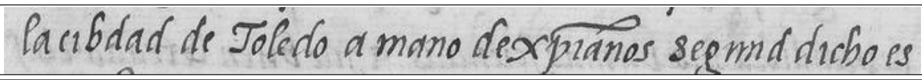
Text line image		
ITER-0	$\hat{p}$	
ITER-1	$\hat{\mathbf{s}}$	la abadia de Toledo a mano de xpiānos segun el dicho es
	m	⇑
	$\hat{p}$	la
ITER-2	$\hat{\mathbf{s}}$	cibdad de Toledo a mano de xpiānos segun el dicho es
	m	⇑
	$\hat{p}$	la cibdad de Toledo a mano de xpiānos
	$\hat{\mathbf{s}}$	sigue el dicho es
	v	**segund**
	$\hat{p}$	la cibdad de Toledo a mano de xpiānos segund
FINAL	$\hat{\mathbf{s}}$	dicho es
	v	#
	$\hat{p} \equiv T$	la cibdad de Toledo a mano de xpiānos **segund** dicho es

Figure 5.1: Example of CATTI operation using Mouse Actions. Starting with an initial recognised hypothesis $\hat{\mathbf{s}}$ from the text line image, the user validates its longest well-recognised prefix $\hat{p}$, making a Mouse Action (MA) m, and the system emits a new recognised hypothesis $\hat{\mathbf{s}}$. As the new hypothesis corrects the erroneous word, a new cycle starts. Now, the user validates the new longest prefix $\hat{p}$, which is error-free, making another MA m. The system provides a new suffix $\hat{\mathbf{s}}$ taking into account this information. As the new suffix does not correct the mistake, the user types the correct word v, generating a new validated prefix $\hat{p}$. Taking into account the new prefix, the system suggests a new hypothesis $\hat{\mathbf{s}}$. As the new hypothesis corrects the erroneous word, a new cycle starts. This process is repeated until the final error-free transcription T is obtained. The underlined boldface word in the final transcription is the only one which was corrected by the user. Note that in the iteration 2 it is needed two user interactions (a MA and then, to type the correct word). However, the iteration 1 only needs one user interaction (one MA).

where $\hat{S}$ represents the set of all possibles suffixes $\hat{s}$ of $\hat{p}$.

Equation (5.1) is very similar to Equation (2.1), being $\hat{w}$ the concatenation of $\hat{p}$ and $\hat{s}$. The difference is that now a part of the transcription $\hat{p}$ is given. As shown in (Romero et al., 2012), this search can be efficiently carried out using the lattices obtained during the Viterbi decoding of the whole input signal representation $\hat{x}$.

5.2 Multimodal Computer Assisted Transcription

Multimodal combination can be used to add additional sources of information to the CATTI process. This is the case of the dictation of the text contents in the transcription of historical text images, where taking into account both the handwritten text image and the speech utterance the CATTI system might propose a better hypothesis in each interaction step. In this way, many user corrections could be avoided.

Since *off-line* HTR and ASR systems share most part of the recognition process (see Section 2.1), the possibility of using the results of both systems in CATTI arises immediately. Through this **multimodal hypotheses combination** (Chapter 3), the CATTI system would take advantage from two different data sources. Moreover, the **multimodal hypotheses correction** approach allows the integration of the more ergonomic *on-line* HTR feedback (based on the combination technique presented in Chapter 3), which provides CATTI with an additional source of information for correcting specific errors.

Figure 5.2 illustrates an example of the integration of the *on-line* HTR feedback on the CATTI process. In this example, the initial recognised hypothesis can be derived from two different input signals, a text line image and a speech utterance of the dictations of the contents of the text line image to transcribe. Besides, in this case, when the MA does not correct the erroneous word, the user can

Figure 5.2: Example of CATTI operation using *on-line* HTR feedback. Starting with an initial recognised hypothesis $\hat{\mathbf{s}}$ from the the input signal (a text line image, a speech utterance, or a multimodal combination of both), the user validates its longest well-recognised prefix $\hat{p}$, making a Mouse Action (MA) m, and the system emits a new recognised hypothesis $\hat{\mathbf{s}}$. As the new suffix does not correct the mistake, the user writes in an *on-line* mode the correct word t by using pen strokes, and the system emits a new recognised hypothesis $\hat{\mathbf{s}}$. As the new hypothesis corrects the erroneous word, a new cycle starts. Now, the user validates the new longest prefix $\hat{p}$, which is error-free, making another MA m. The system provides a new suffix $\hat{\mathbf{s}}$ taking into account this information. Given that the new suffix does not correct the mistake, the user writes (in *on-line* mode) the correct word t, and the system emits a new recognised hypothesis $\hat{\mathbf{s}}$. As the new suffix does not correct the mistake, the user types the correct word v, generating a new validated prefix $\hat{p}$. Taking into account the new prefix, the system suggests a new hypothesis $\hat{\mathbf{s}}$. As the new hypothesis corrects the erroneous word, a new cycle starts. This process is repeated until the final error-free transcription T is obtained.

provide the correct word by using pen-strokes, and if this *on-line* handwriting feedback fails, the user can type the correct word by using a keyboard.

5.2.1 Multimodal Hypotheses Combination in CATTI

Formally, in the traditional CATTI framework (Romero et al., 2012), the system uses a given feature sequence, $\hat{x}_{htr}$, representing a text image and a user validated prefix $\hat{p}$ of the transcription (Equation (5.1)). In our multimodal proposal, in addition to $\hat{x}_{htr}$, a sequence of feature vectors $\hat{x}_{asr}$, which represents the speech dictation of the text image contents, is used to improve the system performance. Therefore, the CATTI system should try to complete the validated prefix $\hat{p}$ by searching for a most likely suffix $\hat{s}$ taking into account both sequences of feature vectors:

$$\hat{\mathbf{s}} = \arg\max_{\hat{s}\in\hat{S}} \mathrm{P}(\hat{s} \mid \hat{x}_{htr}, \hat{x}_{asr}, \hat{p}) \tag{5.2}$$

Making the naive assumption that $\hat{x}_{htr}$ does not depend on $\hat{x}_{asr}$, and applying the Bayes' rule, we can rewrite Equation (5.2) as:

$$\hat{\mathbf{s}} = \arg\max_{\hat{s}\in\hat{S}} \mathrm{P}(\hat{x}_{htr} \mid \hat{p}, \hat{s}) \cdot \mathrm{P}(\hat{x}_{asr} \mid \hat{p}, \hat{s}) \cdot \mathrm{P}(\hat{s} \mid \hat{p}) \tag{5.3}$$

where the concatenation of $\hat{p}$ and $\hat{s}$ is $\hat{w}$. As in conventional HTR and ASR, $\mathrm{P}(\hat{x}_{htr} \mid \hat{p}, \hat{s})$ and $\mathrm{P}(\hat{x}_{asr} \mid \hat{p}, \hat{s})$ can be approximated by HMMs and $\mathrm{P}(\hat{s} \mid \hat{p})$ by an n-gram language model conditioned by $\hat{p}$. Therefore, the search must be performed over all possible suffixes of $\hat{p}$ (Romero et al., 2012).

This suffix search can be efficiently carried out by using Word Graphs (WG) (Romero et al., 2012) or Confusion Networks (CN) (Granell et al., 2016) obtained from the combination of the HTR and ASR recognition outputs, i.e., the terms in Equation (5.3) are derived from a WG or CN obtained from both modalities (more information about multimodality can be found on Chapter 3). In each interaction step, the decoder parses the validated prefix p over the WG or CN, and then continues searching for a suffix which maximises the posterior probability according to Equation (5.3). This process is repeated until a complete and correct transcription of the input text image is obtained.

5.2.2 Multimodal Hypotheses Correction in CATTI

In the CATTI framework, users are repeatedly interacting with the system. Therefore, the quality and ergonomics of the interaction process is crucial for the success of the system. Traditional peripherals like keyboard and mouse can be used to unambiguously provide the feedback associated with the validation and correction of the successive system predictions. Nevertheless, using more ergonomic multimodal interfaces should result in an easier and more comfortable Human-Computer Interaction (HCI), at the expense of a less deterministic feedback. It is important to note that the use of this more ergonomic user interaction will produce new errors coming from the decoding of the feedback signals. Here, we will focus on touchscreen communication, which is perhaps the most natural feedback modality for CATTI. In this way, the user corrective feedback can be quite naturally provided by means of *on-line* text or pen strokes exactly registered over the text produced by the system.

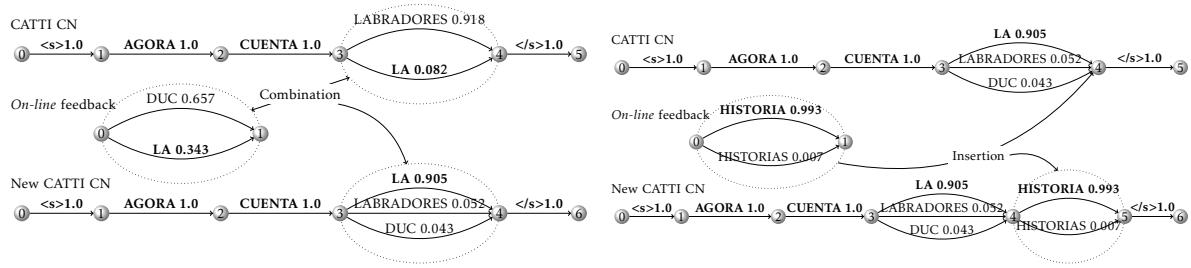
In (Romero et al., 2012), the multimodal interaction process is formulated into two steps. In the first step, a CATTI system solves the problem presented in Equation (5.1). In the second step, the user enters some pen-strokes, t, typically aimed at accepting or correcting parts of the suffix suggested by the system in the previous interaction step, $\hat{s}$, validating a prefix which is error free, $\hat{p}'$. Then, an *on-line* HTR feedback subsystem is used to decode t into a word d, taking into account $\hat{s}$ and $\hat{p}'$.

The multimodal interaction process presented in this thesis, that we call multimodal CATTI (MM-CATTI), differs from the previous one into two main points. On the one hand, the process is formulated in only one step. On the other hand, both the input data and the *on-line* HTR feedback help each-other to optimise the system accuracy.

Formally speaking, let $\hat{x}$ be the input feature sequence and t the *on-line* touchscreen pen strokes that the user introduces to insert or substitute a word. Let $\hat{p}'$ be the user-validated prefix of the previously suggested transcription which is error-free and e the wrong word that the user tries to correct. Using this information, the system has to suggest a new suffix, $\hat{\mathbf{s}}$, as a continuation of the validated prefix $\hat{p}'$, conditioned by the *on-line* touchscreen strokes t and the erroneous word e. Therefore, the problem is to find $\hat{\mathbf{s}}$ given $\hat{x}$ and a feedback information composed of $\hat{p}'$, e and t. By further considering the decoding d as a hidden variable, we can write:

$$\begin{aligned} \hat{\mathbf{s}} &= \arg\max_{\hat{s}\in\hat{S}} \sum_{d} \mathrm{P}(\hat{s}, d \mid \hat{x}, \hat{p}', t, e) \\ &\approx \arg\max_{\hat{s}\in\hat{S}} \sum_{d} \mathrm{P}(t \mid \hat{p}', e, \hat{s}, d, \hat{x}) \cdot \mathrm{P}(\hat{x} \mid \hat{p}', e, \hat{s}, d) \cdot \mathrm{P}(\hat{s} \mid \hat{p}', e, d) \cdot \mathrm{P}(d \mid \hat{p}', e) \end{aligned} \tag{5.4}$$

We can now make the reasonable assumption that t only depends on d and, that $\hat{x}$ and $\hat{s}$ do not depend on e and, approximating the sum over all the possible decodings d of t by the dominating term,



(a) Subnetwork combination example.

(b) Subnetwork insertion example.

Figure 5.3: MM-CATTI editing actions. Ref.: <s>AGORA CUENTA LA HISTORIA </s>

Equation (5.4) can be rewritten as:

$$\hat{\mathbf{s}} \approx \arg\max_{\hat{s}\in\hat{S}} \max_{d} \mathrm{P}(\hat{x} \mid \hat{p}', \hat{s}, d) \cdot \mathrm{P}(\hat{s} \mid \hat{p}', d) \cdot \mathrm{P}(t \mid d) \cdot \mathrm{P}(d \mid \hat{p}', e) \tag{5.5}$$

The first two terms of Equation (5.5) are very similar to Equation (5.1), being $\hat{p}$ the concatenation of $\hat{p}'$ and d. The main difference is that now d is unknown. On the other hand, the last two terms correspond to the HTR decoding of the *on-line* feedback, conditioned by the previously validated prefix $\hat{p}'$ and the erroneous word e. As in conventional CATTI, the probabilities $\mathrm{P}(\hat{x} \mid \hat{p}', \hat{s}, d)$ and $\mathrm{P}(t \mid d)$ are modelled by morphological models, whereas, $\mathrm{P}(\hat{s} \mid \hat{p}', d)$ and $\mathrm{P}(d \mid \hat{p}', e)$ are modelled by using conditioned n-gram language models.

In order to cope with the erroneous word e that follows the validated prefix, and given that this word only affects to the decoding of t, $\mathrm{P}(d \mid \hat{p}', e)$ can be formulated as follows:

$$\mathrm{P}(d \mid \hat{p}', e) = \frac{\bar{\delta}(d, e) \cdot \mathrm{P}(d \mid \hat{p}')}{1 - \mathrm{P}(e \mid \hat{p}')} \tag{5.6}$$

where $\bar{\delta}(i, j)$ is a negative Kronecker delta (Lai et al., 2010) that is 0 when $i = j$ and 1 otherwise.

In conventional CATTI, this decoding can be implemented easily using WG or CN. However, given that the feedback integration in our MM-CATTI proposal is based on the combination of Confusion Networks, the decoding on this proposal must be implemented using CN. In each interaction step, the validated prefix $\hat{p}'$ is parsed over the CN obtained from the input feature sequence (CATTI CN). This parsing procedure will end defining a node q of the CN whose associated word sequence is $\hat{p}'$. Then, a CN is obtained from the *on-line* handwriting feedback recogniser. Assuming that the user corrects only one word in each interaction, this *on-line* CN is composed by a list of words that corresponds with the different decodings of t. This *on-line* CN is combined with the CATTI CN after the node q. Then, the system continues searching for the most probable suffix, according to Equation (5.5), using this new combined CN.

As the *on-line* HTR feedback is limited to one word, the *on-line* CN obtained is composed of only two nodes, like a subnetwork. This *on-line* CN is combined or inserted into the CATTI CN at the point that the previous parsing of the user validated prefix has defined. Therefore, two different editing operations can be carried out to generate the new CATTI CN: combination and insertion of subnetworks:

Combination: Given two subnetworks, one from the CATTI hypotheses SN_{CATTI} and the other from the *on-line* HTR feedback SN_{FB}, the word posterior probabilities of the combined CATTI subnetwork SN_{Comb} are obtained applying a normalisation on the logarithmic interpolation of the smoothed word posterior probabilities of both SN (SN_{CATTI} and SN_{FB}):

$$\mathrm{P}(w \mid SN_{Comb}) = \mathrm{P}_s(w \mid SN_{CATTI})^{\alpha} \mathrm{P}_s(w \mid SN_{FB})^{1-\alpha} \tag{5.7}$$

where the weight factor α allows us to balance the reliability between modalities, and the smoothing of the word posterior probabilities is calculated according to the following equation which is based on Laplacian smoothing:

$$P_s(w \mid SN) = \frac{P(w \mid SN) + \Theta}{1 + n\Theta} \tag{5.8}$$

where Θ is a defined granularity that represents the minimum probability for a word and n is the number of different words in the final CATTI SN. This subnetwork combination process was previously presented in Chapter 3.
In the example presented in Figure 5.3(a), the marked subnetwork of the CATTI CN (subnetwork between the 3rd and 4th nodes) is selected for combination. In this case, the correct word (*LA*) is not the most probable word, either in the *on-line* feedback subnetwork. However, it becomes the most probable word when combining both subnetworks with $\alpha = 0.5$ and $\Theta = 10^{-4}$, as can be seen in the appointed subnetwork of the new CN (this combination process is explained in detail in Figure 3.2).

Insertion: The subnetwork insertion allows us to add a word into the CATTI CN on a particular position. This position is determined by the parsing of the validated prefix p' that precedes the *on-line* word inserted by the user in the CATTI interaction.
As an example, the *on-line* SN (see Figure 5.3(b)) is inserted just after the 4th node of the CATTI CN.

5.3 Conclusions

This chapter presents how multimodality can be very useful on assistive transcription. Concretely, in this work multimodal combination is applied for improving an assistive transcription tool called Computer Assisted Transcription of Text Images (CATTI).

The main advantage of the presented approach is that the error reduction produced by multimodal combination (see Part II for more information about multimodal combination) allows us to reduce the human effort significantly when using an assistive transcription system.

The next chapter (Chapter 6) presents the experiments performed with the multimodal CATTI approach presented in this chapter. These experiments include multimodal hypotheses combination and multimodal hypotheses correction.

Bibliography

Granell, E., Romero, V., and Martínez-Hinarejos, C.-D. (2016). 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)*, pages 269–274.

Lai, W. M., Rubin, D. H., Rubin, D., and Krempl, E. (2010). *Introduction to Continuum Mechanics*. Butterworth-Heinemann.

Rodríguez, L., Casacuberta, F., and Vidal, E. (2007). Computer Assisted Transcription of Speech. In Martí, J., Mendonça, J. M. B. A. M., and Serrat, J., editors, *Pattern Recognition and Image Analysis (IbPRIA 2007)*, volume 4477 of *Lecture Notes in Computer Science*, pages 241–248. Springer, Berlin, Heidelberg.

Romero, V., Toselli, A. H., and Vidal, E. (2009). Using Mouse Feedback in Computer Assisted Transcription of Handwritten Text Images. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 96–100.

Romero, V., Toselli, A. H., and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*, volume 80 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing.

Toselli, A., Romero, V., Rodríguez, L., and Vidal, E. (2007). Computer Assisted Transcription of Handwritten Text Images. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*, volume 2, pages 944–948.

6

Interactivity Experimental Results

"Ea res libera dicitur, quae ex solâ suae naturae necessitate existit, et à se solâ ad agendum determinatur : necessaria autem, vel potiùs coacta, quae ab alio determinatur ad existendum, et operandum certâ, ac determinatâ ratione."[1]

Baruch Spinoza, Ethica ordine geometrico demonstrata, 1677.

Content

A multimodal interactive assistive tool, where the automatic system and the paleographer cooperate to generate the perfect transcription, would reduce the time and the paleographer effort required for transcribing historical handwritten documents.

In this chapter, we present the experimentation performed in a multimodal interactive transcription system wherein user feedback can be provided by means of touchscreen pen strokes, traditional keyboard, and mouse operations. As seen in the previous chapter (Chapter 5), the combination of the

[1] "A thing is free when it exists by the sole necessity of its nature and it is determined to act only by itself. A thing is necessary or rather constrained when it is determined by another thing to exist and to act according to a certain and determined law."

Illustration info: Amulets to protect women in childbirth and newborn infants from the spirit of Lilith. (Bi-defus Stanislaus Augustus Melekh Polin, Sefer Raziel (The Book of Raziel), 1793).

main and the feedback data streams is based on the use of Confusion Networks derived from the output of three recognition systems: two Handwritten Text Recognition systems (*off-line* and *on-line*), and an Automatic Speech Recognition system. *Off-line* text recognition and speech recognition are used to derive (by themselves or by combining their recognition results) the initial hypothesis, and *on-line* text is used to provide feedback. The use of the proposed multimodal interactive assistive system not only reduces the required transcription effort, but it also helps to optimise the overall performance and usability. This allows for a faster and more comfortable transcription process.

Section 6.1 presents the experimental framework (data, conditions, and assessment measures); Section 6.2 offers the results of the multimodal hypotheses combination experiments; Section 6.3 describes the results of the multimodal hypotheses correction experiments; Section 6.4 shows the results of the multimodal hypotheses combination and multimodal hypotheses correction experiments; and Section 6.5 draws the final conclusions and outlines the future work lines.

6.1 Experimental Framework

This section introduces the data sets, features, models, and evaluation metrics used on the experiments. More details about these topics can be found in Chapter 2.

6.1.1 Datasets

Off-line Handwritten Text: *Cristo Salvador*, and *Rodrigo*

The two historical manuscripts (*Cristo Salvador* and *Rodrigo* (Serrano et al., 2010)) employed in the experimentation of the previous part (see Section 4.1.1) were also used in the following *off-line* HTR experiments. Section 2.8.1 presents the details about these datasets. Figure 6.1 presents two text line samples, one for each one of these historical manuscripts.

Speech: *Albayzin*, *Cristo Salvador*, and *Rodrigo*

The acoustical models were trained by using a partition of the *Albayzin* Spanish database (Moreno et al., 1993) presented in Section 2.8.3. Test data for ASR was the product of a controlled acquisition (see Section 2.8.4 for more information) of the dictation of the contents of the lines contained in the test set of both historical manuscripts.

On-line Handwritten Text: UNIPEN, *Cristo Salvador*, and *Rodrigo*

The touchscreen feedback was simulated following the process for generating synthetic samples used in (Romero et al., 2012) with the UNIPEN `Train-R01/V07` dataset. The kinematical models were trained by using samples from 17 different UNIPEN writers. For the *on-line* HTR test, three different writers were randomly chosen for testing. Character samples for each writer were selected in a number enough to fulfil the data requirements. Data amount depends on the number of word instances to be handwritten by the user in the multimodal CATTI process. This includes *Cristo Salvador* and *Rodrigo* corpora. The selected writers are identified by their name initials as BS, BH and BR, and Figure 6.2 presents an example of the word "*historia*" for each one of them. In Section 2.8.2 more details about this corpus can be found.

6.1.2 Features

Feature extraction in the *off-line* HTR case transforms a preprocessed text line image into a sequence of 60-dimensional feature vectors (Section 2.2.2), whereas a touchscreen coordinates sequence (*on-line*

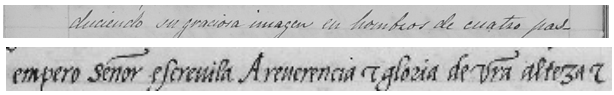


Figure 6.1: Sample lines for *Cristo Salvador* (top) and for *Rodrigo* (bottom).

Figure 6.2: Examples of the word *"historia"* generated by using characters from the three selected UNIPEN test writers (BH, BR, BS).

feedback) is transformed into a new speed- and size-normalised temporal sequence of 6-dimensional real-valued feature vectors (Section 2.2.3). In the ASR case, from each speech utterance a sequence of 39-dimensional feature vectors is extracted (Section 2.2.1).

6.1.3 Models

Optical, acoustical, and kinematical models were trained by using HTK (Young et al., 2006). In the first place, symbols on the optical models are modelled by a continuous density left-to-right HMM with 12 and 4 states for *Cristo Salvador* and *Rodrigo*, respectively, and 32 Gaussians per state. Secondly, phonemes on the acoustical model are modelled as a left-to-right HMM with 3 states and 64 Gaussians per state. Finally, a variable number of states for the different *on-line* characters was used with 16 Gaussians per state on the kinematical model.

The lexicon models for both systems are in HTK lexicon format, where each word is modelled as a concatenation of symbols for HTR or phonemes for ASR.

The language models (LM) were estimated directly from the transcriptions of the pages included on the *off-line* HTR training sets (32 pages for *Cristo Salvador*, and about 205 pages for *Rodrigo*) by using the SRILM *ngram-count* tool (Stolcke, 2002). The same language models used in the experiments of the previous part (see Section 4.1.3) were used in the following experiments. Nevertheless, in addition to the open vocabulary language model for *Rodrigo*, for the second experiment (Section 6.3) this language model was interpolated with the whole lexicon for obtaining a closed vocabulary language model.

6.1.4 Evaluation Metrics

Different evaluation measures have been adopted. On the one hand, the quality of the recognition (post-edition approach) is given by the well known Word Error Rate (WER), and oracle WER to assess the quality of the obtained lattices. On the other hand, the CATTI performance is given by the Word Stroke Ratio (WSR), and the relative difference between them gives us a good estimation of the reduction in human effort (EFR). Finally, the classification Error Rate (ER) was used to assess the recognition quality of the *on-line* HTR feedback. For each measure, confidence intervals of 95% were calculated, and p-values were obtained to confirm the statistical significance with the significance threshold set at $\alpha = 0.025$. In Section 2.7 the details of these evaluation measures can be found.

6.1.5 Experimental Setup

The three recognition systems were implemented by using the iATROS recogniser (Luján-Mares et al., 2008). The SRILM toolkit (Stolcke, 2002) was used for all processes on language models, and for obtaining CN from the WG of the decoding outputs.

In order to optimise the experiments results, the values of the main decoding variables were tuned, and set to the values presented on the bottom half of Table 4.1.

6.2 Experiment 1: Multimodal Hypotheses Combination

Multimodal hypotheses combination allows us to enrich the CATTI hypotheses from different sources of information (in this case, *off-line* HTR and ASR decoding results). Several experiments were performed in order to test our multimodal proposal by using two different handwritten text datasets (*Cristo Salvador* and *Rodrigo*). Moreover, the performance of this multimodal proposal was tested by using four different combination techniques (ROVER, N-best ROVER, Lattices Rescoring, and our CN Combination proposal). In Chapter 3, more information about combination of natural language recognition systems can be found.

For both *off-line* HTR corpora, the unimodal post-edition baseline values were obtained. Next, the classical unimodal CATTI was tested. Then, both modalities (*off-line* HTR and ASR) were combined by using the different combination techniques. Post-edition values in all these cases are those presented in Section 4.5. Finally, the new multimodal CATTI proposal was tested.

In the CATTI experiments, the limit of mouse actions was set to 3. In the multimodal combination both modalities were equally weighted. Therefore, the different combination parameters were: frequency of occurrence voting scheme for ROVER (since the multimodal combination is performed without training), uniform weights ($\lambda = 1$ and $\mu = 0.5$) for N-best ROVER, and combination weight of 0.5 for the Lattices Rescoring and CN Combination techniques.

6.2.1 Experiments with *Cristo Salvador*

Table 6.1: *Cristo Salvador* experimental results. The relative human effort reduction (EFR) represents the relative difference between the obtained CATTI WSR over the unimodal *off-line* HTR post-edition WER value (32.9% ± 6.8). Best results are highlighted in **boldface**.

Experiment		Post-edition		CATTI	
		WER	Oracle WER	WSR	EFR
Unimodal	Baseline *off-line* HTR	32.9% ± 6.8	27.5% ± 6.4	30.2% ± 6.4	8.2%
	Baseline ASR	43.3% ± 3.4	27.4% ± 2.2	35.1% ± 3.5	−6.7%
Multimodal	ROVER	32.7% ± 2.9	32.7% ± 2.9	32.8% ± 2.6	0.3%
	N-best ROVER	33.3% ± 2.9	33.3% ± 2.9	35.9% ± 2.6	−9.1%
	Lattices Rescoring	31.3% ± 2.6	**10.3% ± 1.7**	**13.7% ± 2.0**	**58.4%**
	Proposed CNC	**29.3% ± 2.5**	13.4% ± 2.1	14.1% ± 2.2	57.1%

Table 6.1 presents the obtained experimental results for the *Cristo Salvador* corpus. As can be observed in the unimodal post-edition results, speech recognition does not seem to be a good substitute for handwriting recognition in this task. However, the ASR oracle WER value is similar to the HTR oracle WER.

Regarding the unimodal CATTI results presented in the top-right part of the table, the estimated interactive human effort (WSR) required for obtaining the perfect transcription from the *off-line* HTR decoding represents 8.2% of relative effort reduction (EFR) over the HTR baseline WER. However, no effort reduction can be considered when only ASR is used at the input of the CATTI system.

As expected, in the multimodal experiments the use of lattices in the combination (used by Lattices Rescoring and our proposed CN Combination) allows us to obtain better post-edition results. The best result was obtained by using the CN Combination method with a 29.3% ± 2.5 of WER, which represents a relative improvement over the *off-line* HTR baseline WER (32.9% ± 6.8) of 10.9%. The best

oracle WER (10.3%) was obtained by using the Lattices Rescoring method. Results show that WER improvements are not significant ($p = .413$) with respect to *off-line* HTR baseline WER, but the oracle WER values are statistically significant lower ($p < .001$) in the case of lattice combination methods. Therefore, an outstanding effect of multimodal lattices combination in interactive transcription systems can be expected, since this low oracle WER is related to the amount and quality of the alternatives offered by the combination technique (the lower the oracle WER, the more and better alternatives).

Regarding the obtained results in the multimodal CATTI experiments, the use of the ROVER and N-best ROVER combination methods produce worse results -although differences are not statistically significant ($p > .900$)- when comparing with the unimodal *off-line* HTR CATTI baseline WSR ($30.2\% \pm 6.4$). In contrast, as expected because of their low oracle WER, the values obtained by the CN Combination and Lattices Rescoring methods not only represent improvements, but these improvements are also statistically significant ($p < .001$). Concretely, the overall best result ($13.7\% \pm 2.0$) was achieved by using the Lattices Rescoring method and it represents a relative improvement of 54.6% over the unimodal *off-line* HTR CATTI WSR, and an EFR of 58.4% over the unimodal *off-line* HTR baseline WER.

6.2.2 Experiments with *Rodrigo*

Table 6.2: *Rodrigo* experimental results. The relative human effort reduction (EFR) represents the relative difference between the obtained CATTI WSR over the unimodal *off-line* HTR post-edition WER value ($39.3\% \pm 4.1$). Best results are highlighted in **boldface**.

Experiment		Post-edition		CATTI	
		WER	Oracle WER	WSR	EFR
Unimodal	Baseline *off-line* HTR	$39.3\% \pm 4.1$	$28.0\% \pm 3.0$	$36.2\% \pm 3.6$	7.9%
	Baseline ASR	$62.9\% \pm 2.2$	$29.5\% \pm 1.6$	$47.2\% \pm 2.3$	−20.1%
Multimodal	ROVER	$44.9\% \pm 1.8$	$44.9\% \pm 1.8$	$44.8\% \pm 1.7$	−14.0%
	N-best ROVER	$38.4\% \pm 1.8$	$38.4\% \pm 1.8$	$41.0\% \pm 1.8$	−4.3%
	Lattices Rescoring	$37.2\% \pm 1.7$	$\mathbf{10.6\% \pm 0.9}$	$\mathbf{25.2\% \pm 1.6}$	**35.9%**
	Proposed CNC	$\mathbf{35.9\% \pm 1.6}$	$14.8\% \pm 1.0$	$27.0\% \pm 1.8$	31.3%

The same procedure was followed with this corpus. In Table 6.2 the obtained experimental results are shown. With the post-edition experiments, we confirmed that speech recognition is not a good substitute for historical handwriting recognition. However, as happened with *Cristo Salvador*, for this corpus both modalities also present similar oracle WER values.

The obtained WSR value ($36.2\% \pm 3.6$) in the unimodal *off-line* HTR CATTI experiment represents a relative effort reduction of 7.9% over the *Rodrigo* unimodal *off-line* HTR baseline ($39.3\% \pm 4.1$). Nevertheless, in the unimodal ASR CATTI experiment, no effort reduction can be considered from the obtained results.

Regarding the multimodal results, in post-edition all techniques present similar performances, except ROVER, which yields a statistically significantly worse result. However, in the CATTI experiments, Lattices Rescoring and CN Combination methods presented significantly better results than ROVER and N-best ROVER, which is in consonance with their oracle WER results.

In the post-edition experiments, the best result ($35.9\% \pm 1.6$) was obtained by using the CN Combination method, and it represents 8.7% relative improvement ($p = .094$) over the *off-line* HTR baseline WER ($39.3\% \pm 4.1$). Meanwhile, the Lattices Rescoring method allowed to obtain the best oracle WER ($10.6\% \pm 0.9$), representing 62.1% of statistically significant ($p < .001$) improvement over the oracle WER baseline ($28.0\% \pm 3.0$).

On the other hand, the use of the ROVER and N-best ROVER combination methods on the multimodal CATTI does not improve the unimodal *off-line* HTR baseline WSR ($36.2\% \pm 3.6$). However, the CN Combination and Lattices Rescoring combination methods allow us to obtain statistically significant

($p < .001$) improvements with an EFR higher than 30% over the *off-line* HTR baseline WER. The Lattices Rescoring combination method allowed us to obtain the overall best WSR result, specifically 25.2% ± 1.6 of WSR, which represents a statistically significant ($p < .001$) relative improvement of 30.4% over the unimodal *off-line* HTR baseline WSR.

6.3 Experiment 2: Multimodal Hypotheses Correction

Several experiments were performed to assess the multimodal CATTI (MM-CATTI) approach presented in Section 5.2.2. Multimodal interaction offers ergonomy and increased usability at the expense of the system having to deal with non-deterministic feedback signals. Therefore, the main concern here is how much *on-line* HTR feedback decoding accuracy can be boosted by the combination of both *off-line* and *on-line* handwritten text recognition.

This experiment was performed using the *Rodrigo* corpus with open and closed vocabulary language models. The limit of mouse actions was set to 3, and for the multimodal subnetwork combination, a weight factor of $\alpha = 0.5$ and a granularity factor of $\Theta = 10^{-4}$ were used. As a first step, the non interactive *off-line* HTR baseline was obtained by using open and closed vocabulary. As a second step, the CATTI approach was applied. Next, the *on-line* HTR baseline result was obtained by decoding only the words that the user must introduce during the CATTI process (154 in the open vocabulary task and 146 in the closed vocabulary). Finally, the new MM-CATTI proposal was tested.

6.3.1 *Off-line* and *On-line* HTR Results

Table 6.3: *Off-line* HTR baseline results for *Rodrigo*.

Vocabulary	WER	Oracle WER
Open	39.3% ± 4.1	28.0% ± 3.1
Closed	37.0% ± 3.8	27.4% ± 3.6

Table 6.3 presents the WER and the oracle WER obtained during the conventional, non-interactive *off-line* HTR experiments performed on the *Rodrigo* dataset.

Table 6.4: *On-line* HTR baseline results for *Rodrigo*.

Vocabulary	Words	ER_b	ER_m
Open	154	28.8% ± 3.4	24.9% ± 3.3
Closed	146	12.3% ± 2.7	8.4% ± 2.2

Table 6.4 shows the writer average *on-line* HTR feedback decoding error rates for closed and open vocabulary. The column ER_b is the baseline defined in Section 5.2.2, which corresponds with the two-steps approach presented in (Romero et al., 2012). The column ER_m corresponds with the accuracy of the multimodal approach presented here (Equation (5.5)), wherein both modalities are combined using CN. As can be observed, feedback decoding accuracy increases when both the main and the feedback data help each other, although differences are not statistically significant ($p > .300$).

6.3.2 CATTI and Multimodal CATTI Results

Table 6.5 shows the estimated interactive human effort (WSR) required for obtaining the perfect transcription and the corresponding estimated effort reduction (EFR) when compared with post-edition

Table 6.5: CATTI results for *Rodrigo*.

Vocabulary	WSR	EFR
Open	36.2% ± 3.5	7.9%
Closed	33.5% ± 3.8	9.5%

effort (*off-line* HTR baseline WER). Note that the WSR obtained by CATTI is limited by the *off-line* HTR oracle WER (Table 6.3). The obtained WSR in both experiments (open and closed vocabulary), represents a minimum EFR of 7.9%.

Table 6.6: Multimodal CATTI results for *Rodrigo*.

Vocabulary	WSR				EFR
	Deletions	TS	KBD	Global	
Open	1.8% ± 0.5	29.9% ± 2.0	7.4% ± 1.2	37.3% ± 2.6	5.1%
Closed	1.8% ± 0.5	28.4% ± 2.1	2.3% ± 0.6	30.7% ± 2.4	17.0%

In Table 6.6 the MM-CATTI results are presented. In this case, the estimated interactive human effort (WSR) is decomposed into the percentage of deleted words, the percentage of words written with the *on-line* HTR feedback -TouchScreen (TS)-, and the percentage of those words for which the correction with the *on-line* HTR feedback failed and the corrections had to be entered by means of the keyboard (KBD), i.e., in MM-CATTI 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. This is a pessimistic assumption, since interaction through touchscreen is more ergonomic than through keyboard. Despite the presence of 6.2% of OOV words, 5.1% of EFR was obtained. On the other hand, without OOV words the EFR reached 17.0%. According to these results, the expected user effort for the more ergonomic and user preferred touchscreen based MM-CATTI is only moderately higher than that of CATTI for the open vocabulary experiments.

6.4 Experiment 3: Multimodal Hypotheses Combination and Correction

The next experiments were performed to assess our multimodal proposals for improving the assistive transcription system presented in the previous chapter (Chapter 5) on the *Cristo Salvador* corpus. In the CATTI and MM-CATTI experiments, the limit of mouse actions was set to 5, and for the multimodal combination, a weight factor of $\alpha = 0.5$ and a granularity factor of $\Theta = 10^{-4}$ were used.

We started obtaining the non interactive post-edition baseline, for the *off-line* HTR, for the ASR, and for the multimodal combination of both unimodal recognition systems by using our CN combination proposal (see Chapter 3). Then, the CATTI and the multimodal CATTI (MM-CATTI) approaches were applied to the three input possibilities formatted as CN, two unimodal (*off-line* HTR and ASR), and one multimodal (*off-line* HTR combined with ASR).

6.4.1 Post-Edition Baseline Results

In Table 6.7 the baseline results are presented. This table shows the WER and the oracle WER presented by the CN obtained during the conventional, non-interactive experiments performed on the *Cristo Salvador* dataset. This values are similar to those presented in Section 4.5 and Section 6.2.1. As can be observed in the post-edition results, the *off-line* HTR decoding output presents 32.9% ± 6.4 of WER with

Table 6.7: Post-edition experimental results for *Cristo Salvador*.

Modality	WER	Oracle WER
Off-line HTR	32.9% ± 6.4	27.5% ± 6.4
ASR	43.7% ± 3.3	27.4% ± 2.2
Multimodal	29.3% ± 2.5	13.4% ± 2.1

an oracle WER value of 27.5% ± 6.4. Regarding the ASR obtained results, speech recognition does not seems to be a good substitute for handwriting recognition in this task, although both modalities present similar oracle WER values. Given that these unimodal oracle WER values are not significantly better than the *off-line* HTR baseline WER value, a significant effort reduction produced by the CATTI and MM-CATTI systems can not be expected.

However, the multimodal combination of both sources allows us to reduce the WER value to 29.3% ± 2.5, which represents a relative improvement ($p = .413$) of 10.9% over the *off-line* HTR baseline, and 33.0% over the ASR baseline ($p = .002$). One of the best advantages of the multimodal combination is that not only the best hypothesis is improved, but also the rest of hypotheses. This fact can be observed through the oracle WER of this multimodal combination (13.4% ± 2.1), which is significantly ($p < .001$) reduced given the two unimodal sources. Given that the oracle WER represents the WER of the best hypothesis that can be achieved, a significant beneficial effect on interactive systems can be expected.

6.4.2 CATTI Results

Table 6.8: CATTI experimental results for *Cristo Salvador*.

CATTI Input	WSR	EFR
Off-line HTR	31.1% ± 6.0	5.5%
ASR	31.6% ± 3.1	4.0%
Multimodal	12.9% ± 2.1	60.8%

Table 6.8 presents the estimated interactive human effort (WSR) required for obtaining the perfect transcription using the interactive CATTI approach for the three different input possibilities. Notice that these results differ from those presented in Table 6.1, because here the input lattices were formatted as CN (and not as WG) since this format is necessary for the hypothesis correction (see Section 5.2.2). As expected, the obtained WSR for the unimodal inputs represents a slight effort reduction (EFR) of around 5% with respect to the *off-line* HTR baseline. However, in the case of the multimodal input the WSR reaches 12.9% ± 2.1, which represents a significant ($p < .001$) effort reduction of 60.8% over the *off-line* HTR baseline (32.9% ± 6.4). Notice that, in the multimodal case, the obtained WSR value is a bit lower than the oracle WER value (13.4% ± 2.1); this is possible because the presented CATTI approach, by means of mouse actions, allows to reduce the number of words explicitly corrected by the user. Therefore, in this case the CATTI approach not only offers the best hypothesis contained in the multimodal lattices, but it improves the oracle WER value deleting several erroneous words of this hypothesis.

6.4.3 Multimodal CATTI Results

In MM-CATTI, the *on-line* HTR feedback results (see Table 6.9) were obtained by decoding only the words that the user must introduce during the MM-CATTI process (on the mean, 57.7 words when the input was the unimodal *off-line* HTR, 70.4 words when the input was the unimodal ASR, and 23.9 words when the input was multimodal).

The *on-line* HTR feedback presented moderated writer average decoding error rates (ER_b). As

Table 6.9: *On-line* HTR feedback results for *Cristo Salvador*.

MM-CATTI Input	*On-line* HTR		
	Words	ER_b	ER_m
Off-line HTR	57.7	7.0% ± 3.2	25.8% ± 5.4
ASR	70.4	3.6% ± 0.9	11.0% ± 0.7
Multimodal	23.9	8.7% ± 2.4	12.1% ± 1.4

presented in Table 6.9, the more words are decoded, the better the decoding error rates. This is due to the fact that the MM-CATTI input with worst hypotheses (ASR) needs the *on-line* HTR feedback to correct easier words (so there is less error), while the input with better hypotheses (Multimodal) needs the *on-line* HTR feedback to correct more difficult words (biggest mistake in proportion, although the overall number of words to correct is lower). In this case, the multimodal integration of the *on-line* feedback in the MM-CATTI (ER_m) did not produce any improvement with respect to ER_b.

Table 6.10: Multimodal CATTI experimental results.

MM-CATTI Input	WSR				EFR
	Deletions	TS	KBD	Global	
Off-line HTR	5.5% ± 2.6	26.0% ± 5.5	6.7% ± 3.2	32.7% ± 7.2	0.6%
ASR	5.1% ± 1.0	31.6% ± 3.4	3.5% ± 1.1	35.1% ± 4.0	−6.7%
Multimodal	1.9% ± 0.8	10.7% ± 1.8	1.3% ± 0.6	12.0% ± 2.1	63.5%

In Table 6.10 the MM-CATTI results 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). Despite the observed ER_m results in the previous table (Table 6.9), the multimodal combination of the *on-line* feedback with the MM-CATTI hypotheses allowed us to reduce significantly the amount of words that are required to be corrected by using the keyboard. In the unimodal input experiments, only 6.7% of words for *off-line* HTR and 3.5% for ASR were corrected by using the keyboard, only a slight EFR was obtained for *off-line* HTR, and none for the ASR case. However, with the multimodal input a 12.0% of WSR was obtained, which represents a significant ($p < .001$) EFR of 63.5% with respect to the *off-line* HTR baseline (32.9% ± 6.4).

According to these results, in MM-CATTI most of the user effort is concentrated in the more ergonomic and user preferred touchscreen feedback. Moreover, the overall user effort in MM-CATTI can be lower than that of CATTI when the input presents a low oracle WER value (EFR of 63.5% instead of 60.8%).

6.5 Conclusions and Future Work

In this chapter, we have presented how the use of Confusion Networks Combination allows to improve the interaction (by using *on-line* touch-screen handwritten pen strokes) in an multimodal interactive transcription system (MM-CATTI) presented in previous works (Romero et al., 2012). The main advantage of the presented approach is that the multimodal combination allows us to correct errors on the MM-CATTI hypothesis by using the information provided by the *on-line* handwritten text introduced by the user.

The obtained results show the benefits of using speech as an additional source of information for the transcription of historical manuscripts. Moreover, the use of the more ergonomic feedback (*on-line* HTR) modality usually comes at the cost of only a reasonably small number of additional interaction

steps needed to correct the few feedback decoding errors. In fact, when the input presents a low oracle WER, the use of the *on-line* HTR feedback modality can even produce an additional reduction of the required human effort for obtaining the actual transcription.

Our future works aim at using speech also as a feedback modality. Besides, the MM-CATTI system can be improved by taking advantage of the real samples that are produced while the system is used for adapting the natural language recognition systems (*on-line* HTR and ASR) to the user. Eventually, this approach will be opened to be tested with other corpora.

Finally, until now we have considered that the acquisition of the speech samples does not require any effort for the palaeographer. However, the speech acquisition of a historical manuscript can represent an important extra cost if a single palaeographer should do so. This effort can be avoided, or rather distributed among different collaborators, thanks to an external speech acquisition through a multimodal crowdsourcing platform as the proposed in the next part (Part IV).

Bibliography

Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.-D., Pastor, M., Sanchis, A., and Toselli, A. (2008). iATROS: A speech and handwritting recognition system. In *Proceedings of the V Jornadas en Tecnologías del Habla (VJTH'2008)*, pages 75–78.

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

Romero, V., Toselli, A. H., and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*, volume 80 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing.

Serrano, N., Castro, F., and Juan, A. (2010). The RODRIGO Database. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2709–2712.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the 3rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 901–904.

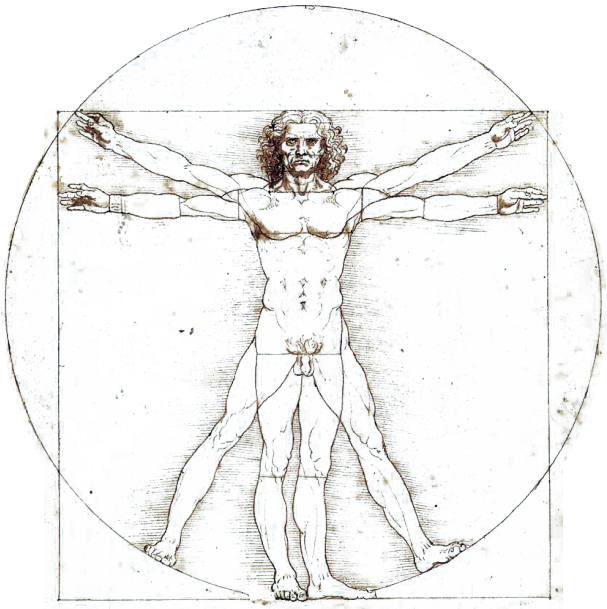
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

IV

Crowdsourcing

"- What are your thoughts on constant connectivity and its effect on mankind?
- We are all now connected by the Internet, like neurones in a giant brain. [. . .]
Now anyone anywhere in the world can react immediately to new work.
Science has become more inclusive."

Stephen Hawking, USA Today (2014-12-02).



The Vitruvian Man. (Leonardo da Vinci, Le proporzioni del corpo umano secondo Vitruvio, 1487)

7

Collective Collaboration

"- Agnes: Will you read us a bedtime story?
- Gru: [reluctantly] No.
- Agnes: But we can't go to sleep without a bedtime story.
- Gru: Well then it's going to be a long night for you, isn't it? So good night, sleep tight and don't let the bed bugs bite. Because... there are literally thousands of them. [beats, then whispers sinisterly] Oh, and there's probably something in your closet. [closes the door and chuckles]
- Margo: [referring to Gru] He's just kidding, Agnes." [1]

Dialogue from the film Despicable Me (2010).

Content

Most crowdsourcing platforms for document transcription (such as the presented in Section 2.6) make users employ the keyboard for providing the transcription. This poses a severe limitation on the kind of devices that can be used in the collaboration: only desktop or laptop computers seem suitable for those platforms. Although mobile devices (tablets and smartphones) admit keyboard input by using their virtual keyboard, the lack of ergonomics makes the transcription task a frustrating experience. Consequently, the range of volunteers gets constrained by this limitation.

As an alternative, 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 (Rabiner and Juang, 1993), presents an ambiguity not present in typed input. Even the state-of-the-art techniques (Hinton et al., 2012), although more accurate than a few years ago, produce a

Illustration info: Drawing of a bed bug infestation (Johann Prüss, Ortus sanitatis, 1499).

[1] This is a funny situation, until the day that you feel the effects of their bite on your skin, and you wake up surrounded of bed bugs. It seems that this insect has been fed by human blood from the prehistory (Potter, 2011). This bug was an important nuisance in the middle ages, and nowadays in the times of Deep Learning and Mars exploration this little bug not only is still here but also is becoming more resistant to insecticides (Morand, 2014). So, remember: Don't let the bed bugs bite. Because, there are LITERALLY thousands of them!

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 HTR provide an initial draft transcription more accurate than that given only by HTR. This fact was confirmed with the statistically significantly improvements obtained in the experiments performed for the previous parts of this thesis, multimodal transcription (Chapter 4), and multimodal interactive transcription (Chapter 6). Thus, the employment of speech collaborations will allow us to significantly reduce the final transcription effort.

This chapter 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 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 (Bellegarda, 2004) and Confusion Network combination (Xue and Zhao, 2005) techniques. The crowdsourcing platform was implemented by using a client-server architecture. The client is a mobile application that allows speech acquisition and the server part performs the recognition and combination operations.

The rest of this chapter is structured as follows: Section 7.1 presents the details on the proposed multimodal crowdsourcing framework, and Section 7.2 summarises the conclusions.

7.1 Multimodal Crowdsourcing Framework

In the proposed crowdsourcing framework, the main objective is, given a text image and different dictations (usually from different speakers) of that text, to obtain a final transcription with the lowest number of errors. This transcription will be provided to a paleographer to obtain the final quality transcription with the lowest effort.

The framework is mainly based on two ideas: using the current system output to obtain an adapted language model that can be employed in the next decoding step (Alabau et al., 2011), and combining the decoding outputs of the two modalities (more information about multimodal combination can be found on Chapter 3) to obtain a final output with less errors (Granell and Martínez-Hinarejos, 2015).

Apart from that, the framework includes a speech reliability verification module that may exclude utterances that are considered of insufficient quality. This takes into account that volunteers may experience difficulties when dictating historical text (hesitations in some ancient words, word misses, inconvenient pauses, etc.). Using a similar idea, and with the aim of reducing the collaborator's effort, a line selection module is incorporated to select lines whose transcription have a low reliability. The aim is to obtain more speech samples for those lines than for other lines. It is supposed that this strategy would allow us to improve the global results on the whole set of lines to be transcribed.

Figure 7.1 presents the working diagram of this multimodal crowdsourcing system. The operation is as follows:

1. The initial system output is given by the HTR decoding.
2. When a collaborator offers to help, the crowdsourcing loop starts:
   (a) In the language model (LM) interpolation module, the previous system output is interpolated with the original LM, giving an improved language model for the next ASR decoding.
   (b) The reliability of the system output is evaluated and the lines are selected by its reliability (in increasing order); thus, the collaborator is asked to read only a subset of lines with the lowest reliability.
   (c) The collaborator speech is decoded in the ASR module using the improved language model.
   (d) The reliability of the obtained ASR output is verified and filtered, i.e., only those utterances which reach a minimum reliability value are given as output by the reliability verification module.

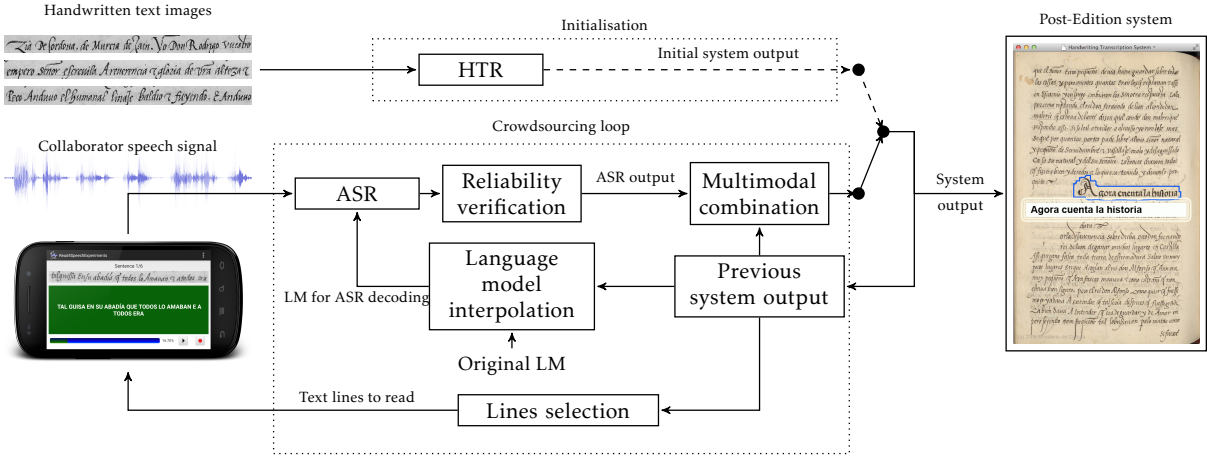


Figure 7.1: Multimodal crowdsourcing transcription framework.

(e) The multimodal combination module produces the new system output by combining the previous system output and this verified ASR output.

3. Every time a new collaborator offers to help, the crowdsourcing loop is executed and the system output is improved by using the new audio samples.

The following subsections describe in detail, including the presentation of some examples, the different modules of the framework.

7.1.1 Language Model Interpolation

Decoding outputs from HTR and ASR processes can be obtained in rich formats that provide several alternatives in the form of lattices. Two usual forms of representing lattices are Word Graphs (WG) and Confusion Networks (CN). In Section 2.4.2 more information about the recognition output formats can be found.

The language model interpolation module builds a statistical language model conditioned on a sample $\hat{x}$ as follows (Alabau et al., 2011):

1. The decoding lattices for $\hat{x}$ are formatted as WG.
2. The posterior probabilities for each WG node ($\mathrm{P}(q \mid \hat{x})$) and link ($\mathrm{P}(l \mid \hat{x})$) are computed by using the forward $\alpha(q)$ and backward $\beta(q)$ probabilities of the nodes (Wessel et al., 2001).
3. The counts for a word sequence $\hat{w}_{i-n+1}^{i} = (w_{i-n+1}, \ldots, w_i)$ are estimated as:

$$C^*(\hat{w}_{i-n+1}^{i} \mid \hat{x}) = \sum_{l_1^n \in N(\hat{w}_{i-n+1}^{i})} \frac{\prod_k \mathrm{P}(l_k \mid \hat{x})}{\prod_k \mathrm{P}(s(l_k) \mid \hat{x})} \tag{7.1}$$

where $N(\hat{w}_{i-n+1}^{i})$ are all the sequences of concatenated links that generate $\hat{w}_{i-n+1}^{i}$. Figure 7.2 presents some of the word sequences (n-grams) and weighted counts that could be obtained from a WG as the one presented in Figure 2.9.

4. The word posterior probabilities associated to the current input $\hat{x}$ can be calculated from these counts. Prior to that, a discount method (for back-off estimation), a smoothing method -to avoid the Out Of Vocabulary (OOV) problem-, and a proper normalisation are applied. The final estimation follows:

$$\mathrm{P}^{\hat{x}}(\hat{w}) = \prod_i \frac{C^*(\hat{w}_{i-n+1}^{i} \mid \hat{x})}{C^*(\hat{w}_{i-n+1}^{i-1} \mid \hat{x})} \tag{7.2}$$

N-gram	$C^*(\hat{w}_{i-n+1}^{i} \mid \hat{x})$
<s>AGORA	0.999797
AGORA	0.999797
AGORA CUENTA	0.999797
AGORA CUĒTA	4.75237e-197
CUENTA EL	1.25754e-32
CUENTA LA	0.999797
HISTO </s>	1.25754e-32
HISTORIA	0.999797
HISTORIA </s>	0.999797
LA HISTORIA	0.999797
LABRADORES </s>	4.75104e-197

Figure 7.2: Some weighted word sequence counts $C^*(\hat{w}_{i-n+1}^{i} \mid \hat{x})$ estimated from the Word Graph in Figure 2.9.

N-gram	$\log \mathrm{P}^{\hat{x}}(\hat{w})$	$\log \mathrm{P}(\hat{w})$	$\log \mathrm{P}_{\lambda}^{\hat{x}}(\hat{w})$
...			
AGORA	-0.6989701	-2.864318	-0.9970424
<s>AGORA	-0.3010741	-2.427961	-0.5988735
AGORA ABRAÇAN	–	-2.835677	-3.136707
AGORA CUENTA	-0.3010741	-0.6985934	-0.4558558
AGORA CUĒTA	–	-2.835677	-3.136707
...			
...			
HISTORIA	-0.6989701	-3.969872	-0.9997674
HISTORIAS	-10.74464	-4.056476	-4.357506
DE HISTORIAS	–	-4.235704	-4.536734
HISTORIA </s>	-0.3010741	-0.571207	-0.4154676
HISTORIA A	–	-1.18019	-1.48122
HISTORIA DE	–	-0.4723006	-0.7733306
LA HISTORIA	-0.3010741	-1.468681	-0.5735402
LAS HISTORIAS	–	-1.983436	-2.284466
MUCHAS HISTORIAS	–	-2.842337	-3.143367
QUE HISTORIAS	–	-3.912265	-4.213295
...			

Figure 7.3: Example of language model interpolation from the counts in Figure 7.2 by using $\lambda = 0.5$ and a smoothing factor of 1^{-10}. The probabilities are in log domain.

5. The new conditioned language model $\mathrm{P}^{\hat{x}}(\hat{w})$ is linearly interpolated with the original language model $\mathrm{P}(\hat{w})$ by using a weight factor λ:

$$\mathrm{P}_{\lambda}^{\hat{x}}(\hat{w}) = \lambda \mathrm{P}^{\hat{x}}(\hat{w}) + (1-\lambda)\mathrm{P}(\hat{w}) \tag{7.3}$$

The weight factor λ balances the reliability in the interpolation between the language model estimated from the previous system output and the original one. Figure 7.3 presents an example in which a general language model is refined according to the n-grams presented in Figure 7.2 (by using $\lambda = 0.5$ and a smoothing factor of 10^{-10}). As can be observed, the probability of the n-grams that allow to obtain the correct transcription is increased in the new language model. This shows that through this interpolation the knowledge acquired in form of lattices can be used for the next decoding processes.

7.1.2 Multimodal Combination

The multimodal combination employs Confusion Networks (CN) to combine the ASR decoding output with the previous system output. Specifically, this framework employs the bimodal Confusion Network combination method defined in Section 3.3.

7.1.3 Reliability Verification

As seen in Section 2.7.4, when the recognition scores of a fairly large n-best list can be re-normalised to sum up to 1, the re-normalised joint probability $\mathrm{P}(\hat{x}, \hat{w})$ of the obtained best hypothesis can be used as a good confidence measure, since it is a measure of the match between $\hat{x}$ and $\hat{w}$ (Rueber, 1997; Wessel et al., 2001).

N-best	$P(\hat{x}, \hat{w})$
<s>Y PEQUEÑOS </s>	75.1%
<s>Y NUEUE AÑOS </s>	25.8%
<s>Y VEINTE AÑOS </s>	12.5%
<s>Y SIETE AÑOS </s>	12.5%
<s>Y DE DUEÑAS </s>	3.4%

Figure 7.4: Example of n-best list with their corresponding joint probabilities. This n-best list presents a reliability factor of $R = 58.1\%$.

Therefore, the reliability verification module employs the re-normalised 1-best joint probability:

$$R = \frac{\max_{\hat{w} \in \hat{W}} P(\hat{x}, \hat{w})}{\sum_{\hat{w} \in \hat{W}} P(\hat{x}, \hat{w})} \tag{7.4}$$

where $\hat{W}$ denotes the set of all permissible sentences in the evaluated decoding output. As an example of this confidence measure calculation, the small n-best list showed in Figure 7.4 presents a reliability factor of $R = 58.1\%$.

For every ASR decoding of a collaborator utterance, this module is applied in order to assess if the utterance is incorporated into the combination process. Only when the value of R is higher than a threshold value τ, the decoding of the utterance is used in the multimodal combination and a new system output is computed.

7.1.4 Lines Selection

Given that collaborators are a scarce resource, their efforts must be optimised. This can be seen as obtaining the maximum benefit, i.e., the highest possible number of lines improved by their collaboration for a given amount of collaborations.

Consequently, since there are lines where the current system output presents more reliability than the other, it can be supposed that those low reliability lines are more susceptible to be improved by collaborators utterances than the other.

Therefore, it is necessary to select the subset of lines that would be offered to the collaborator according to their current reliability. This is the role of the lines selection module, that acts as follows:

1. The current system output (total set of lines to be transcribed) is evaluated by using the re-normalised 1-best joint probability (R, Equation (7.4) and example in Figure 7.4), giving an estimation of the confidence of the current transcription for each one of the text line images to transcribe.
2. The lines are ranked according to their estimated confidence value R.
3. The system selects the subset of B (batch size) lines with the lowest confidence.
4. The collaborator is asked to read only the selected lines.

With this policy, each collaborator would dictate the subset of lines that, according to their reliability, would potentially have improvement when included with speech dictation. The number of lines given by the batch size B is important as well, since it determines the effort of a collaborator for an acquisition session.

7.1.5 Client Application for Speech Acquisition

In the proposed multimodal crowdsourcing framework, collaborators interact with the system through a client application installed on their own mobile devices. In this way, when a collaborator offers to help, the subset of B lines with the lowest confidence is loaded into the client application. Given that speech acquisition is a very expensive and time consuming process (Hughes et al., 2010), the client

application allows collaborators to dictate the contents of the text images in an *off-line* mode. Therefore, collaborators are free to decide when and where to collaborate.

Although there are commercial applications and platforms to collect speech utterances, such as Mechanical-Turk (Lane et al., 2010), at the moment this work was performed, we have not found any useful free software application. Therefore, we developed a client application called *Read4SpeechExperiments* (Granell and Martínez-Hinarejos, 2016) for acquiring speech samples from mobile devices (See Section 2.8.4 for more details about the performed speech acquisition from mobile devices).

The main features of *Read4SpeechExperiments* are the following:

- The text to read can be presented as plain text, as images, or as images and a text guide of reading.
- The speech utterances can be recorded either by using the internal or an external microphone.
- The recorded speech utterances can be shared through any communication application available on the mobile device (such as traditional e-mail).
- A handsfree mode allows the speech acquisition in those environments where speakers can not have the mobile device on their hands (for instance, driving a vehicle).

Read4SpeechExperiments is developed for mobile devices with the Android operating system, it can be installed easily from the *Google Play*[2] and *F-Droid*[3] platforms. Moreover, the source code is publicly available in a *GitLab*[4] repository with a GPLv3 license. More information about this application can be found in (Granell and Martínez-Hinarejos, 2016).

7.2 Conclusions

This chapter presents a multimodal crowdsourcing framework for the transcription of historical handwritten documents, wherein volunteers may employ voice as input for transcription. This framework is based on the iterative refinement of the language model, and on the combination of decoding outputs. The client application permits collaborators to decide when and where to collaborate. On the other hand, the lines selection module on the server application analyses the transcription reliability of the handwritten text lines to transcribe at the system output, and selects the set of lines with lower reliability to be presented to the collaborators. This two characteristics allow one to obtain more collaborations and, at the same time, to focus the collaboration effort to the lines whose transcription need more refinement.

The performed experiments on the behaviour of the proposed framework are presented in the next chapter (Chapter 8). These experiments include an initial experimentation using speech samples acquired on the laboratory under our supervision, and a real volunteers-based experiment using speech samples acquired by using mobiles devices without our supervision. Moreover, some experiments were performed to study how to optimise the collaborators effort in terms of number of collaborations, including how many lines and which lines should be selected for the speech dictation.

Bibliography

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

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93–108.

[2] https://play.google.com/store/apps/details?id=com.prhlt.aemus.Read4SpeechExperiments
[3] https://f-droid.org/repository/browse/?fdid=com.prhlt.aemus.Read4SpeechExperiments
[4] https://gitlab.com/egranell/Read4SpeechExperiments.git

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

Granell, E. and Martínez-Hinarejos, C.-D. (2016). 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)*, pages 411 – 417.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. J., and LeBeau, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1914–1917.

Lane, I., Waibel, A., Eck, M., and Rottmann, K. (2010). Tools for Collecting Speech Corpora via Mechanical-Turk. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (NAACL HLT 2010)*, pages 184–187.

Morand, C. (2014). *La punaise de lit (Cimex Lectularius): résurgence d'un nuisible*. PhD thesis, École nationale vétérinaire d'Alfort.

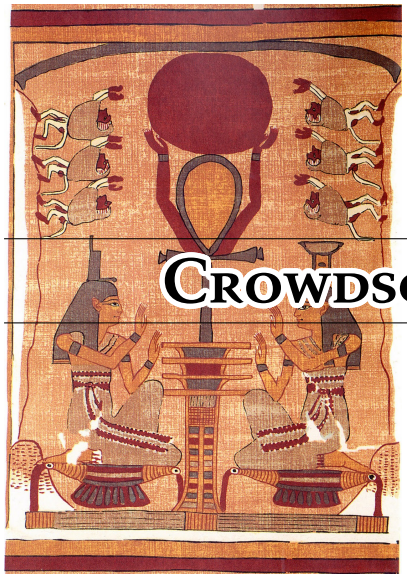
Potter, M. F. (2011). The History of Bed Bug Management - With Lessons from the Past. *American Entomologist*, 57(1):14–25.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rueber, B. (1997). Obtaining confidence measures from sentence probabilities. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 739–742.

Wessel, F., Schlüter, R., Macherey, K., and Ney, H. (2001). Confidence Measures for Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298.

Xue, J. and Zhao, Y. (2005). Improved confusion network algorithm and shortest path search from word lattice. In *Proceedings of the International Conference in Acoustics, Speech and Signal Processing (ICASSP 2005)*, volume 1, pages 853–856.



8 Crowdsourcing Experiments

"It's still magic even if you know how it's done."

Terry Pratchett, A Hat Full of Sky, 2004.

Content

In crowdsourcing platforms, users generally employ keyboard input to provide transcription. This limits the use of crowdsourcing platforms to desktop or laptop computers, losing the potential transcription capability that could be provided by the use of mobile devices (tablets and smartphones), where keyboard input is not ergonomic enough to make its intensive use attractive. As an alternative to that, volunteers could employ voice as input for transcription. This modality is available in nearly all mobile devices, and would allow a researcher to obtain a larger number of volunteers.

This chapter studies how to employ multimodal recognition (combining HTR and ASR) in the crowdsourcing platform presented in the previous chapter (Chapter 7) where volunteer speakers dictate the transcription of a historical handwritten text image. The framework is based on techniques of language model interpolation (Alabau et al., 2014) and Confusion Network combination (Granell and Martínez-Hinarejos, 2015), that allow the fusion of multimodal natural language recognition decoding outputs in a single transcription hypothesis.

In an initial experimentation using speech utterances acquired on the laboratory under our supervision, the influence of the order of the volunteers was examined, in order to check its robustness

Illustration info: The Adoration of Re (The Papyrus of Ani -The Egyptian Book of the Dead-, 1240 BC.

against different sequences of contributors. As the framework includes a reliability verification module, different configurations for this module were analysed. Then, the robustness of the proposed platform against lost contributions (e.g., when volunteers avoid their contribution for a text line because of its difficulty) was tested. Finally, given that volunteers are a scarce resource, optimisation of the work load on the side of collaborators was studied.

The final evaluation of this initial set of experiments provided clues on the feasibility of using this type of platforms for handwritten historical text transcription. Therefore, speech acquisition was made by using the mobile application *Read4SpeechExperiments* (Granell and Martínez-Hinarejos, 2016) for acquiring a rather broad and real sample of collaborators' speech, where the collaborators read the text lines where and when they wanted, i.e. without any supervision from our part (see Section 2.8.4). On this data, a second set of experiments was performed.

The chapter is structured as follows: Section 8.1 presents the details of the data acquisition and experimental conditions, Section 8.2 shows the supervised experiments, Section 8.3 presents the unsupervised experiments, Section 8.4 summarises the conclusions and the outline of future work lines.

8.1 Experimental Conditions

This section introduces the dataset, speech acquisition, features, models, and evaluation metrics used on the experiments. More details about these topics can be found on Chapter 2.

8.1.1 Datasets

Historical Manuscript Corpus: *Rodrigo*

The *Rodrigo* corpus (Serrano et al., 2010) was the historical manuscript employed in the experiments with the same partitions used in the previous parts of this thesis (see Section 4.1.1). This corpus presents several difficulties, such as the following examples, that are present in the first 5 lines of the page 515 (Figure 8.1):

- Text images containing abbreviations (e.g., *nrõ* in the second line) that must be pronounced as the whole word (*nuestro* ['nwes tro]).
- Archaic words (e.g., *Amauan*, *touo*, and *cibdad* in the first, second, and third lines, respectively) that are not used or have a different spelling in modern Spanish (*Amaban*, *tuvo*, and *ciudad*).
- Words written in multiple forms (e.g., *xpiãnos* -in the third line- and *christianos*, or numbers as *5* and *V*) but that are pronounced in the same way ([kris 'tja nos], ['θiŋ ko]).
- Hyphenated words (e.g., *Toledo* in the fourth and fifth lines, where a part of the word -*Tole*- is at the end of a line and the second part -*do*- is at the beginning of the following line).

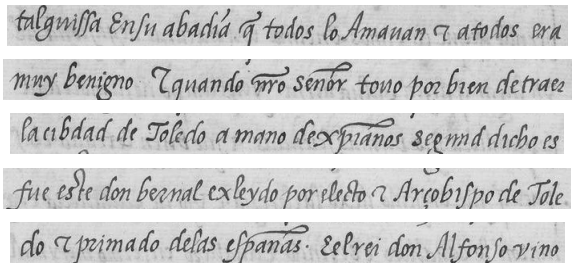


Figure 8.1: The 5 first lines of the page 515 of *Rodrigo*.

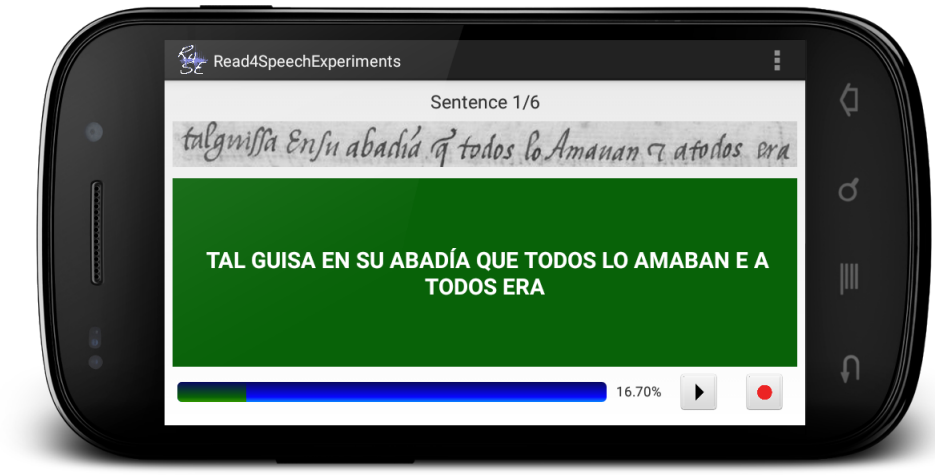


Figure 8.2: Screenshot of the application *Read4SpeechExperiments*.

Speech: *Albayzin* and *Rodrigo*

For the training of the acoustical models we used a partition of the Spanish phonetic corpus *Albayzin* (Moreno et al., 1993). In Section 2.8.3, more information about this corpus is presented.

For the initial ASR test, we used the speech samples acquired in a controlled environment from 7 different native Spanish speakers who read the 50 handwritten test lines of *Rodrigo* (those of pages 515 and 579), giving a total set of 350 utterances (about 15 minutes) acquired by using a computer in the laboratory (see Section 2.8.4 for more information). These speech utterances were the same used in the previous parts of this thesis.

For testing the framework in a real scenario, we used the mobile application *Read4SpeechExperiments* (Granell and Martínez-Hinarejos, 2016) (see Figure 8.2) for acquiring the collaborators speech, and the mailing list of our research group for collaboration demand. None of the received contributions was rejected, given that we intentionally wanted a rather broad and real sample. We obtained the collaboration of 27 different speakers who installed the application on their own mobile devices, and read the 50 handwritten text lines (those of pages 515 and 579) without any control from our side. The collaborators read the text lines where and when they wanted, giving a total set of 1,350 utterances (about 1 hour and 50 minutes, see Section 2.8.4).

8.1.2 Features

Handwritten text features were computed in several steps as explained in Section 2.2.2. In the case of speech, Mel-Frequency Cepstral Coefficients (MFCC) were used (Section 2.2.1). Specifically, in the experiments with the samples obtained from mobile devices, we used MFCC with Cepstral Mean Normalisation (CMN) as ASR features. This normalisation allows us to compensate the long-term spectral effects caused by different microphones and audio environments in the final features.

8.1.3 Models

Optical and acoustical models were trained as continuous density Gaussian mixtures left-to-right HMM by using HTK (Young et al., 2006). The lexicon models were modelled as a concatenation of symbols for HTR or phonemes for ASR. Language models were estimated as a 2-gram with Kneser-Ney back-off smoothing (Kneser and Ney, 1995). The same models for the *Rodrigo* corpus as used in the earlier experiments (see Section 4.1.3, and Section 6.1.3) were used in the following experiments. Only the new acoustical models, with MFCC-CMN features as employed in Section 8.3 were used in this part of the work.

Table 8.1: Baseline results for supervised multimodal crowdsourcing.

Modality	WER
HTR	39.3% ± 4.1
ASR	62.9% ± 2.2

8.1.4 Evaluation Metrics

Different measures were used to assess the performance of our multimodal crowdsourcing framework proposal. The quality of the transcription is given by the well known WER with confidence intervals of 95% (Bisani and Ney, 2004). The statistical significance was confirmed by using p-values obtained by the Welch's t-test (Welch, 1947). The significance threshold was set at $\alpha = 0.025$. The decoding reliability R was verified by using the re-normalised 1-best joint probability -Equation (2.43)-, which is a good estimation of the decoding confidence. Finally, the Collaboration Effort (CE) applied to a determined draft transcription was measured as the product between the number of lines (batch size B) that the system asks the collaborators to read, and the actual number of collaborators involved in the obtainment of the measured draft. More information about these evaluation measures can be found in Section 2.7.

8.1.5 Experimental Setup

The HTR and the ASR systems were implemented by using the iATROS recogniser (Luján-Mares et al., 2008). All processes on language models (inference, interpolation, ...), the decoding output evaluation, and the transformation from Word Graph to Confusion Network were done by using the SRILM toolkit (Stolcke, 2002).

8.2 Experiment 1: Supervised Multimodal Crowdsourcing

To check the performance of the multimodal crowdsourcing framework proposed in the previous chapter (Chapter 7), we have experimented with the 50 text-line images of the *Rodrigo* corpus, and the 350 speech utterances recorded in a controlled environment from 7 different collaborators as described in Section 2.8.4. We started obtaining the baseline values for both modalities. Next, we selected the speaker who best represented the average error rate of the speech set for adjusting the values of the LM interpolation factor λ and the CN combination factor α. Finally, with the other 6 speakers we tested the effects of the speakers ordering, the ASR reliability verification, the absence of speech utterances, and the collaborator effort optimisation.

8.2.1 Baseline and Framework Adjustment

The baseline values were obtained by using the original LM in the decoding process of both modalities. As can be observed in Table 8.1 (that coincide with the results presented in Section 4.4.1), the HTR and ASR WER values are quite high due to the difficulty of the task.

The values of the α and λ parameters must be adjusted in order to obtain the best result. We tested the multimodal crowdsourcing framework adjusting the α and λ parameters with the values $\{0.4, 0.5, 0.6\}$, by using only the speech of the selected speaker. We measured the average reliability $\langle R \rangle$ of the speech decoding output, and the same average reliability but weighted by the number of words contained in the 1-best $\langle R_w \rangle$. Table 8.2 presents the results obtained for the adjustment. Both measures, $\langle R \rangle$ and $\langle R_w \rangle$, present the same tendency. The system presents the highest reliability when the multimodal combination is a bit balanced to the speech output ($\alpha = 0.6$), and the LM interpolation to the original LM ($\lambda = 0.4$).

Table 8.2: Framework adjustment reliability results. Best results are highlighted in **boldface**.

α	λ	$\langle R \rangle$	$\langle R_w \rangle$
0.4	0.4	53.3%	45.8%
	0.5	51.7%	44.1%
	0.6	50.3%	42.7%
0.5	0.4	45.5%	38.4%
	0.5	44.8%	37.7%
	0.6	43.4%	36.3%
0.6	0.4	**62.4%**	**54.5%**
	0.5	61.3%	53.4%
	0.6	61.1%	53.3%

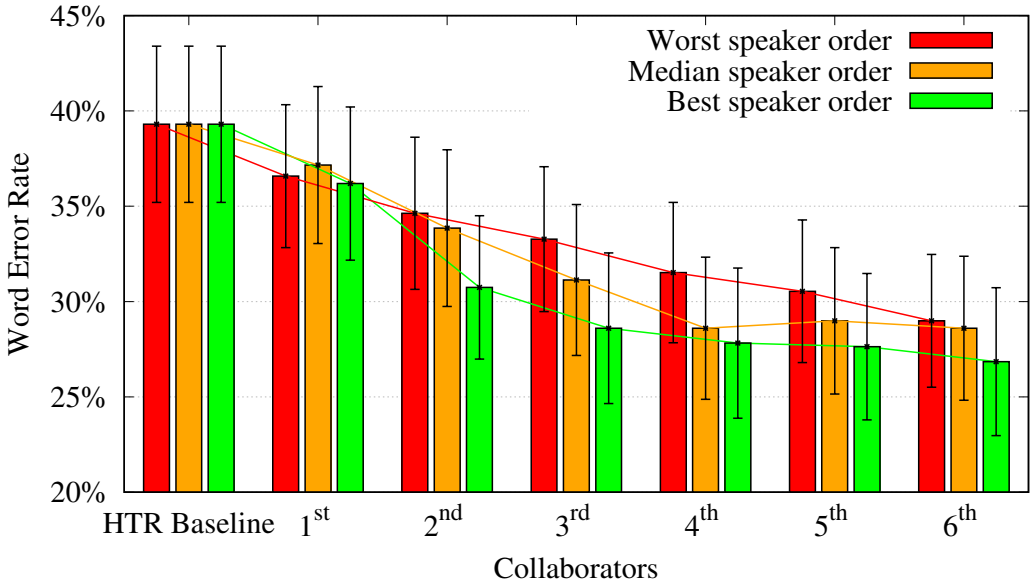


Figure 8.3: Results of the speaker ordering experiments. Best, worst and the median of 11 different random orders.

Table 8.3: Ordering experiments final results.

Order	WER
Worst	29.0% ± 3.5
Median	28.6% ± 3.8
Best	26.9% ± 3.9

8.2.2 Speaker Ordering

The 6 speakers not used in the framework adjustment were randomly sorted 11 times giving 11 different order lists. Figure 8.3 shows the evolution in the system output, from the initial 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 (see Table 8.3). As can be observed in Table 8.3, the worst and the best final results do not represent any statistically significant differences ($p = .445$).

Regarding the ordering of speakers, the obtained results show that in the best case, only two speakers are needed to obtain significant improvements (30.7% ± 3.8). Meanwhile, in the worst case

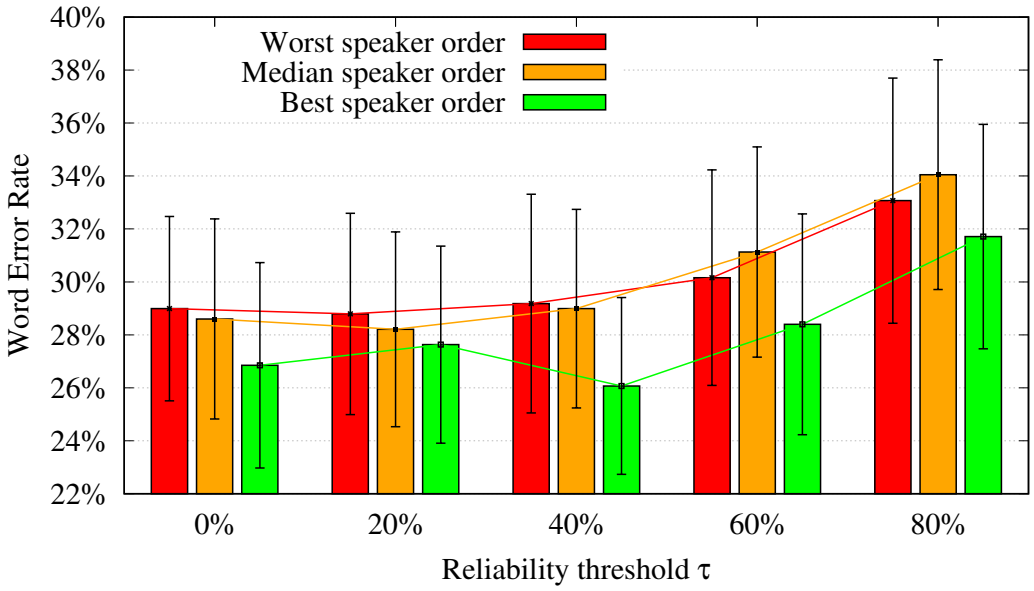


Figure 8.4: Results of the reliability verification experiments on the best, the worst and the median speaker orders. $\tau = 0\%$ means no rejected samples.

Table 8.4: Reliability experiments final results.

Order	Threshold τ	WER
Worst	20%	$28.8\% \pm 3.8$
Median	20%	$28.2\% \pm 3.7$
Best	40%	$26.1\% \pm 3.3$

at least four speakers are needed ($31.5\% \pm 3.6$). As can be seen in the final results from Table 8.3, the presented framework reached, in the worst order, a relative statistically significant improvement ($p < .001$) higher than 26% when compared with the HTR baseline ($39.3\% \pm 4.1$). From these results, we conclude that speaker order is not important for obtaining significant improvements.

8.2.3 ASR Reliability Verification

In the presented framework, if the dictations were made only by expert speakers in historical manuscripts with good pronunciation of ancient words, the output error could be reduced significantly with the collaboration of less people. However, the aim of this framework is to distribute the effort among a larger group of non-experts. Therefore, in order to ensure that the speech of the collaborator enriches the final system output it is necessary to set a minimum reliability threshold.

Figure 8.4 presents the obtained final results when varying the reliability threshold on 3 of the 11 lists (which presented the worst, the median, and the best final results). As can be seen, in all cases there exists a threshold where the rejection of several speech utterances improves the final results.

In Table 8.4 the summary of the best obtained results is shown. Although these improvements are not statistically significant when compared with the results obtained without reliability verification (see Table 8.3), it highlights the importance of verifying the reliability of the speech recognition for obtaining the best results in a crowdsourcing framework as the one presented in this part of the thesis.

Moreover, we observed that the corrections were made in most cases at the beginning and/or

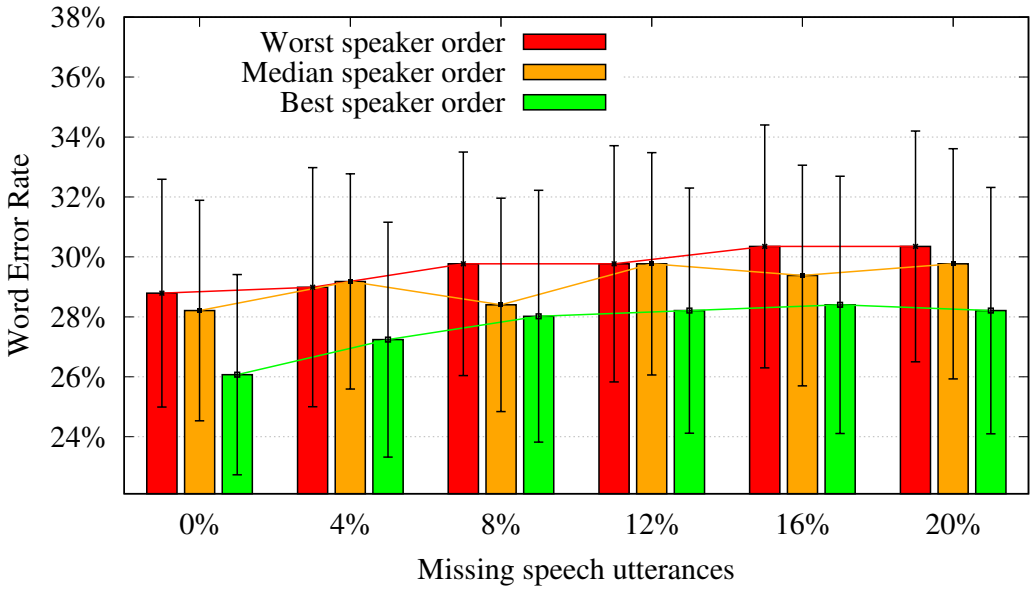


Figure 8.5: Results of the speech missing experiments on the best, the worst and the median speaker orders. 0% means no speech utterances were missing.

at the end of the text lines. This is due to the fact that this selection permits a better refinement of the language model, making it more reliable for the parts of the lines where its estimation is more complicated in the initial training.

8.2.4 Absence of Speech Utterances

In this experiment, we tested the strength of this crowdsourcing framework against the absence of speech utterances. The absence of speech utterances can appear because some collaborators did not read part of the sentences, or because some speech samples got lost in the communication process.

For each of the 6 speakers used for the test experiments, 20% of their speech samples were randomly selected as missing utterances. Then, we tested the performance of this crowdsourcing framework against the loss of speech samples from 4% to 20% in an incremental way, i.e., the missing sentences set of the 8% contains the missing sentences of the previous 4%, etc.

As can be seen in Figure 8.5, the performance of the presented framework decays when some speech utterances are missing. Nevertheless, even losing 20% of the speech utterances of each speaker, in the worst order the final result obtained still achieved a statistically significant ($p = .003$) value of $30.4\% \pm 3.8$, representing a relative improvement of 22.6% over the HTR baseline ($39.3\% \pm 4.1$).

8.2.5 Collaborator Effort Optimisation

In previous experiments, the speaker order and the reliability verification did not show a significant impact on the results. Therefore, the configuration with best results was used in the next experiment, i.e. the best order with a speech decoding reliability threshold of $\tau = 40\%$.

In this experiment we tested the influence of the number of collaborators and the number of lines to read with the aim of optimising the effort made by the collaborators. Figure 8.6 and Table 8.5 present the obtained results. As can be observed, the best results are obtained when all people collaborate with a full effort, i.e., giving the speech transcription of the whole set of text lines. In this case, the best result is $26.1\% \pm 3.3$ of WER, by using the 50 speech utterances of the 6 speakers. This result

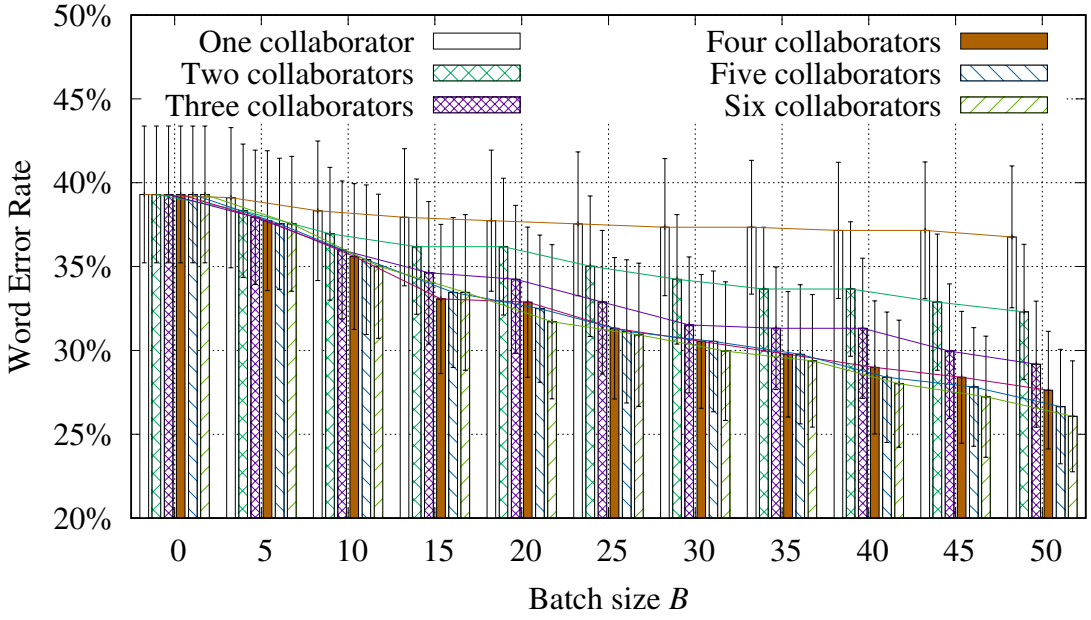


Figure 8.6: Results of the collaborator effort optimisation experiments.

represents a statistically significant ($p < .001$) relative improvement of 33.6% over the HTR baseline, which represents an important effort reduction for obtaining the final transcription.

However, similar statistically significant improvements can be achieved optimising the effort of the collaborators. As can be observed in Figure 8.6 and Table 8.5, the optimal collaborator load is 30 lines (30.0% ± 4.1 of WER for 6 collaborators). Besides, with this load, the obtained improvements are statistically significant ($p < .001$) after the collaboration of the 4^{th} volunteer. In this case, the system output presented a WER of 30.5% ± 4.2, which represents a relative improvement of 22.4% over the baseline. Furthermore, the difference between this value and that obtained by using the 50 speech utterances of the 6 speakers is not statistically significant ($p = .128$). However, it represents a collaborator effort reduction of 33.3% in the number of collaborators and 40.0% in the number of lines read by each collaborator.

The collaborators' load optimisation allowed us to reduce the number of global collaborations, given that with a CE = 120 collaborations (30 lines read by 4 collaborators) this crowdsourcing system improved significantly the transcription of 50 lines. Therefore, it can be expected that with 300 collaborations (50 lines read by 6 collaborators) this system could improve significantly the transcription of a test set of 125 lines.

Table 8.5: Results of the collaborator effort (CE) optimisation experiments (WER). The WER result obtained with the best CE (CE = 120, 4 collaborators with $B = 30$) is highlighted in **boldface**.

Batch size B	Number of Collaborators					
	1	2	3	4	5	6
5	39.1% ± 4.2	38.3% ± 4.0	37.9% ± 4.0	37.7% ± 4.2	37.6% ± 3.9	37.6% ± 4.0
10	38.3% ± 4.2	37.0% ± 4.0	36.0% ± 4.1	35.6% ± 4.4	35.4% ± 4.5	35.0% ± 4.3
15	37.9% ± 4.1	36.2% ± 4.0	34.6% ± 4.3	33.1% ± 4.5	33.5% ± 4.5	33.5% ± 4.6
20	37.7% ± 4.2	36.2% ± 4.1	34.2% ± 4.4	32.9% ± 4.5	32.5% ± 4.4	31.7% ± 4.6
25	37.6% ± 4.3	35.0% ± 4.2	32.9% ± 4.3	31.3% ± 4.2	31.1% ± 4.3	30.9% ± 4.3
30	37.4% ± 4.1	34.2% ± 3.9	31.5% ± 4.1	**30.5% ± 4.0**	30.5% ± 4.2	30.0% ± 4.1
35	37.4% ± 4.0	33.7% ± 3.7	31.3% ± 3.7	29.8% ± 3.7	29.8% ± 4.2	29.4% ± 4.0
40	37.2% ± 4.1	33.7% ± 4.0	31.3% ± 4.2	29.0% ± 4.0	28.4% ± 3.9	28.0% ± 3.8
45	37.2% ± 4.1	32.9% ± 4.1	30.0% ± 4.0	28.4% ± 3.9	27.8% ± 3.5	27.2% ± 3.6
50	36.8% ± 4.2	32.3% ± 4.0	29.2% ± 3.8	27.6% ± 3.5	26.7% ± 3.4	26.1% ± 3.3

Table 8.6: Baseline results for unsupervised multimodal crowdsourcing.

Modality	WER
HTR	39.3% ± 4.1
ASR	60.5% ± 1.3

8.3 Experiment 2: Unsupervised Multimodal Crowdsourcing

To check the performance of the presented multimodal crowdsourcing framework in a real scenario, we have experimented with the 50 text line images of the *Rodrigo* corpus, and the 1,350 speech utterances recorded from 27 different collaborators in a real crowdsourcing environment as described in Section 2.8.4. We started obtaining the baseline values for both modalities; after that, we performed some preliminary experiments, and then we tested the effects of the ASR reliability verification and the optimisation of the collaborators work load. Finally, the Collaboration Effort (CE) per line was studied.

8.3.1 Baseline and Framework Adjustment

The baseline values presented in Table 8.6 were obtained by using the original language model in the decoding process of both modalities. The HTR baseline correspond with the best result obtained in the HTR decoding process (same than in Table 8.1). However, in the ASR system of this experiment we are dealing with an additional source of errors due to the differences between the training and test audio samples (speakers, devices, and environment). In order to alleviate this source of errors, we normalised the cepstral features. Doing this, we obtained a similar ASR baseline WER value to the obtained in the previous experiment, in spite of the fact that here the speech utterances were acquired using different mobile devices without supervision.

In the previous experiment, we observed that this crowdsourcing framework presents the highest reliability (for this corpus) when the multimodal combination is a bit balanced to the speech output ($\alpha = 0.6$, with $\Theta = 10^{-4}$), and the language model interpolation to the original model ($\lambda = 0.4$). We also noted that the speaker ordering did not show a significant impact on the results. Therefore, in this experiment, the speaker ordering was defined by the order of reception of the speech collaborations.

8.3.2 Preliminary Experiments

We started evaluating the performance of the multimodal crowdsourcing platform presented in the previous chapter by using all the collaboration utterances without reliability verification. Figure 8.7

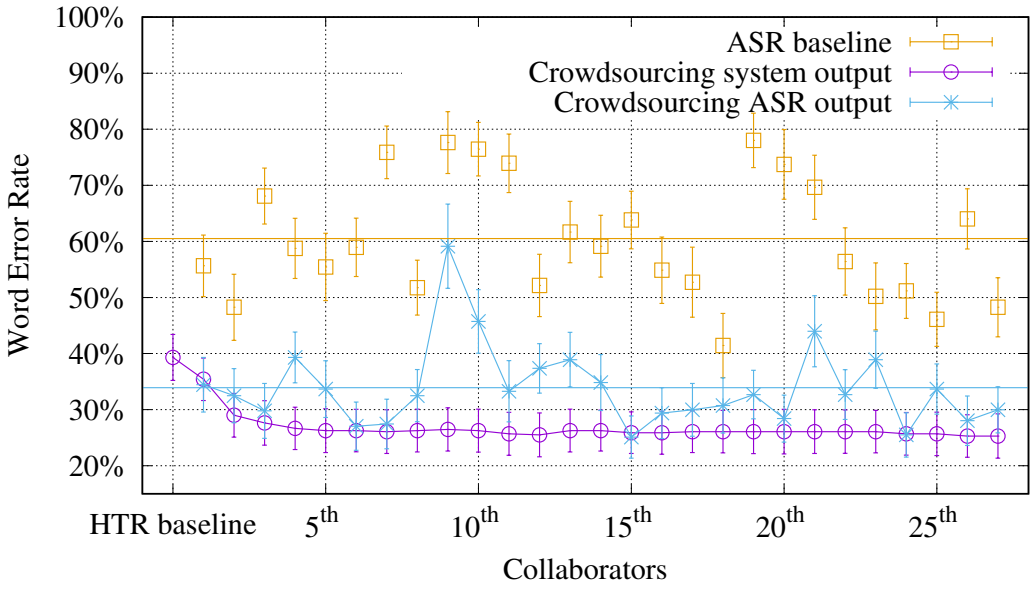


Figure 8.7: Baseline values and the evolution of the system and ASR outputs for the whole test speech corpus without reliability verification nor lines selection. The horizontal lines represent the corresponding average ASR WER values.

draws the baseline values for both modalities and the evolution of the system and ASR outputs for the whole test ASR corpus (CE = 1,350) without reliability verification. As can be observed, the language model interpolation permits one to reduce the error level in the next speech decoding process (Alabau et al., 2011), and the combination with the speech decoding results allows the system output to converge to a better hypothesis with less errors to correct (Granell and Martínez-Hinarejos, 2015). Besides, the ASR performance is considerably improved, reducing the average WER baseline value (60.5% ± 1.3) to 33.9% ± 4.8. Finally, after processing the speech of the last collaborator, the ASR and the system outputs presented 30.0% ± 4.1 and 25.3% ± 3.9 of WER, respectively. The 25.3% ± 3.9 of WER present in this final system output represents 35.6% of relative statistically significant ($p < .001$) improvement over the HTR baseline, and an estimated time reduction for the paleographer revision of about 5 minutes per page (Serrano et al., 2010).

Additionally, in order to test the unimodal performance of this framework, we conducted an experiment in the same conditions without HTR initialisation, i.e., only the speech of the collaborators was processed. As can be observed in Figure 8.8, the behaviour of the system is similar to which was obtained in the previous experiment. In this case, the ASR decoding output presented an average WER of 44.2%±1.2. The WER at the system output decreased to 35.2%±4.6 from an initial value of 55.6%±5.5. In spite of the fact that this is a remarkable improvement over the initialisation, this improvement is not statistically significant ($p = .156$) over the HTR baseline.

8.3.3 ASR Reliability Verification and Collaboration Effort

In order to analyse the behaviour of our multimodal crowdsourcing platform, we tested setting different speech reliability thresholds (τ), and different amount of lines -batches (B)- to be read by the collaborators.

Figure 8.9 presents the effect of the batch size B and the threshold τ on the WER level at the system output after processing the speech of the last collaborator (the 27th collaborator). We can observe that a minimum batch size of $B = 20$ is required to obtain a significant improvement (see details in the final output column of Table 8.7) over the HTR baseline (39.3% ± 4.1). On the other hand, the ASR

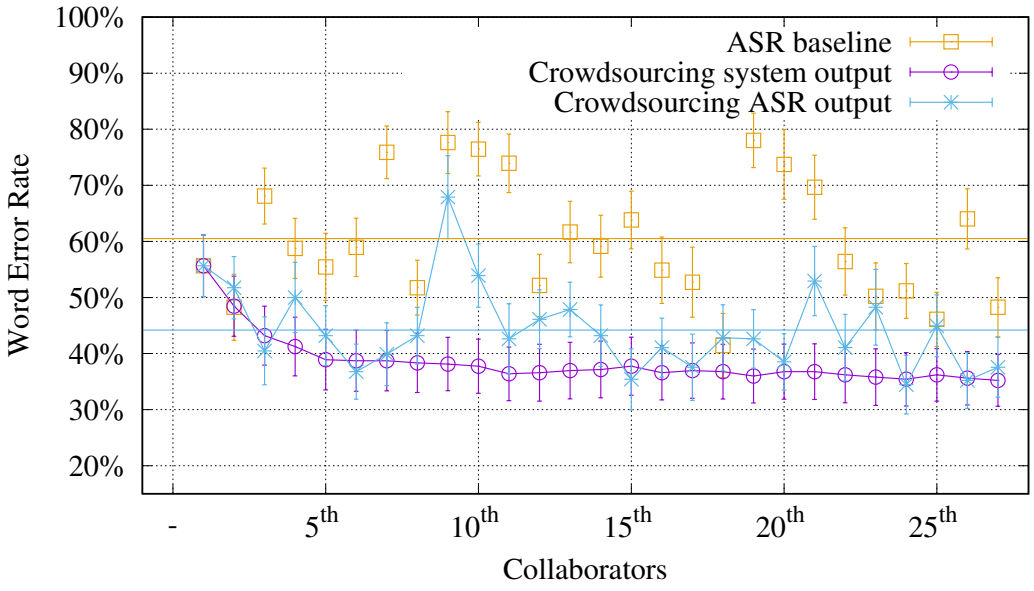


Figure 8.8: ASR baseline values and the evolution of the system and ASR outputs processing only the speech, without HTR initialisation nor reliability verification nor lines selection. The horizontal lines represent the corresponding average ASR WER values.

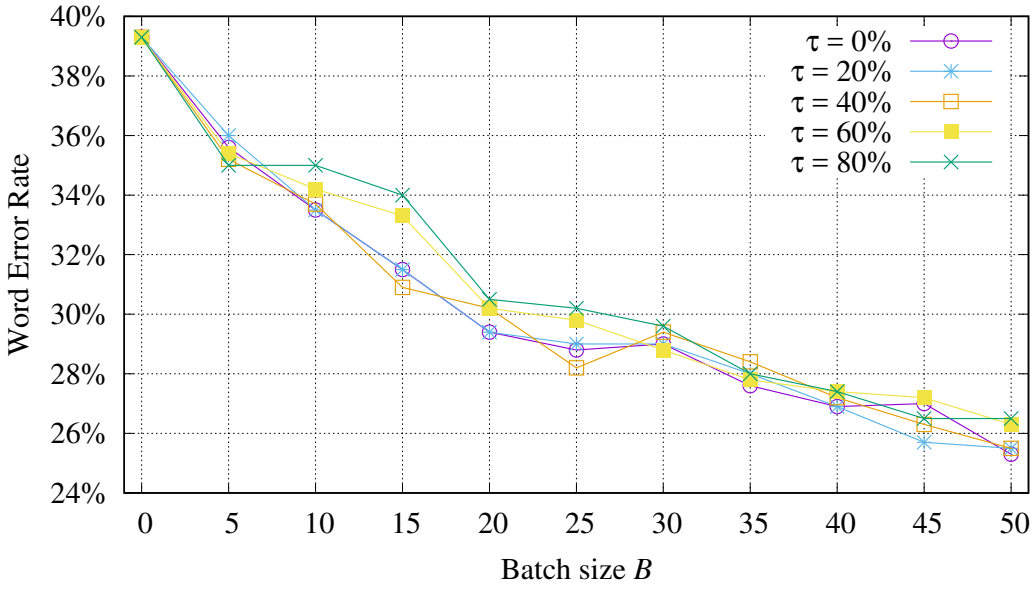


Figure 8.9: Effect of the batch size B and the threshold τ on the WER of the final output.

reliability verification allows one to filter the utterances that can worsen the system output. As we can observe, high values of τ remove too many utterances; therefore, the best performance is obtained when the value of τ is lower than or equal to 40%.

Figure 8.10 presents the effect of the batch size B and the threshold τ on the minimum number of collaborators for improving the system output significantly, i.e. the minimum number of collaborators that allow one to obtain a WER value at the system output lower than 31.2% (which represents

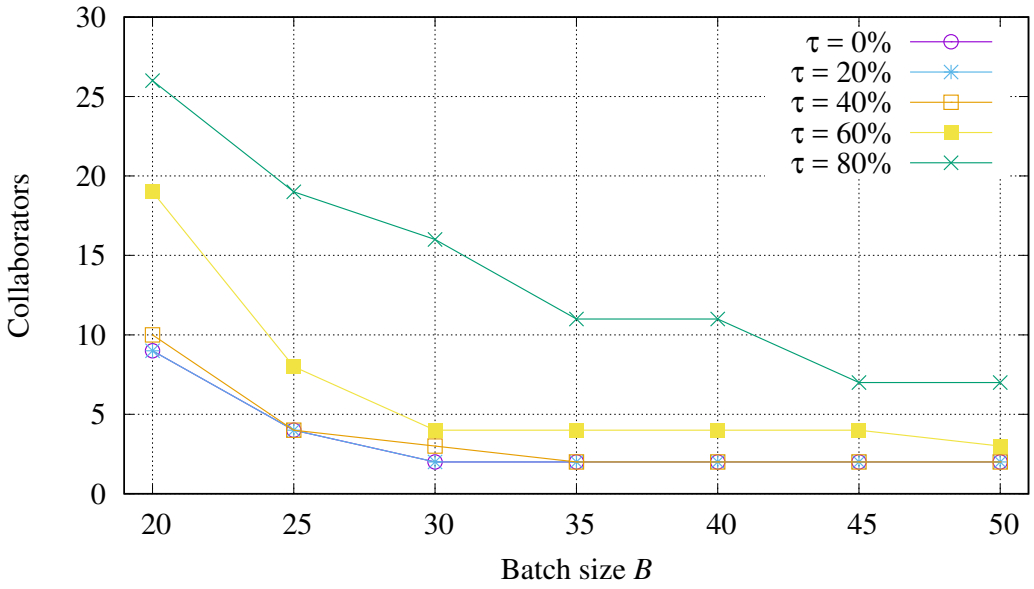


Figure 8.10: Effect of the batch size B and the threshold τ on the minimum number of collaborators for improving the output significantly.

a minimum relative improvement of 20.6%). Therefore, Figure 8.10 only shows results for batch sizes $B \geq 20$, where statistically significant improvements appear. The main conclusion that we can extract from Figure 8.10 is that high values of τ require more collaborators to significantly refine the system output, and that for τ in 0% – 40% the system presents a similar behaviour.

Table 8.7 summarises the obtained results for the B and τ ranges that present significant improvements with respect to baseline results. As can be observed, the overall best result in terms of collaboration effort (CE) was obtained with $B = 30$ and $\tau = 0\%$. In this case, the system output presented a statistically significant ($p = .005$) improvement ($31.1\% \pm 3.8$ of WER) after processing the speech of the second collaborator, i.e., with a CE of only 60 utterances. This WER value represents a relative improvement of 20.9% over the HTR baseline ($39.3\% \pm 4.1$), and an estimated time reduction for the paleographer revision of about 3 minutes per page. Moreover, differences with the overall best result ($25.3\% \pm 3.9$ obtained with a CE of 1,350 utterances) are not statistically significant ($p = .038$). Supposing a similar behaviour on other lines of the corpus, this means that with the whole collaboration effort (1,350 utterances), 1,125 lines would obtain transcription improvements.

8.3.4 Collaboration Effort per Line

We observed as some lines needed more refinement than others. Thus, we analysed the collaboration distribution over the set of lines. Figure 8.11 presents a histogram with the collaboration distribution on the experiments without ASR reliability verification ($\tau = 0$, in order to avoid its influence), and the selective batches ($B = [5,..,45]$) in order to observe the presence of lines that were never refined. This distribution presents the characteristics described in Table 8.8.

As can be observed, several lines, such as the lines number 7, 12, 46, and 48 can be considered as upper mild outliers, while other lines, such as the lines number 3, 22, 33, 37, and 45 can be considered as lower mild outliers. There are several lines of special interest, such as the lines number 7 and 48 that required full collaboration, and the lines number 37 and 45 that were never refined. These lines are presented in Figure 8.12. When comparing their linguistics and visual features, no differences were appreciated, which led us to verify their features in terms of reliability.

In consequence, we studied the relation between the reliability R obtained in the HTR baseline

Table 8.7: Collaboration Effort (CE) experiment results summary. The best CE result is highlighted in **boldface**.

B	τ	First significant improvement			Final output
		Collaborators	CE	WER	WER
20	0%	9	180	30.9% ± 4.0	29.4% ± 3.6
	20%	9	180	30.9% ± 4.2	29.4% ± 4.0
	40%	10	200	30.7% ± 4.2	30.2% ± 4.2
25	0%	4	100	30.5% ± 4.2	28.8% ± 4.0
	20%	4	100	30.5% ± 4.3	29.0% ± 4.1
	40%	4	100	30.9% ± 4.2	28.2% ± 3.8
30	0%	2	**60**	31.1% ± 3.8	29.0% ± 3.9
	20%	4	120	30.5% ± 4.3	29.0% ± 4.1
	40%	3	90	30.9% ± 3.9	29.4% ± 4.0
35	0%	2	70	30.2% ± 3.7	27.6% ± 3.5
	20%	2	70	30.4% ± 3.9	28.0% ± 3.6
	40%	2	70	31.1% ± 3.8	28.4% ± 3.8
40	0%	2	80	30.0% ± 3.8	26.9% ± 3.6
	20%	2	80	30.2% ± 3.7	26.9% ± 3.8
	40%	2	80	30.7% ± 4.0	27.2% ± 3.8
45	0%	2	90	29.6% ± 4.1	27.0% ± 4.1
	20%	2	90	29.8% ± 4.0	25.7% ± 3.7
	40%	2	90	30.9% ± 4.2	26.3% ± 3.7
50	0%	2	100	29.0% ± 3.9	25.3% ± 3.9
	20%	2	100	29.2% ± 3.8	25.5% ± 3.9
	40%	2	100	30.2% ± 4.1	25.5% ± 3.9

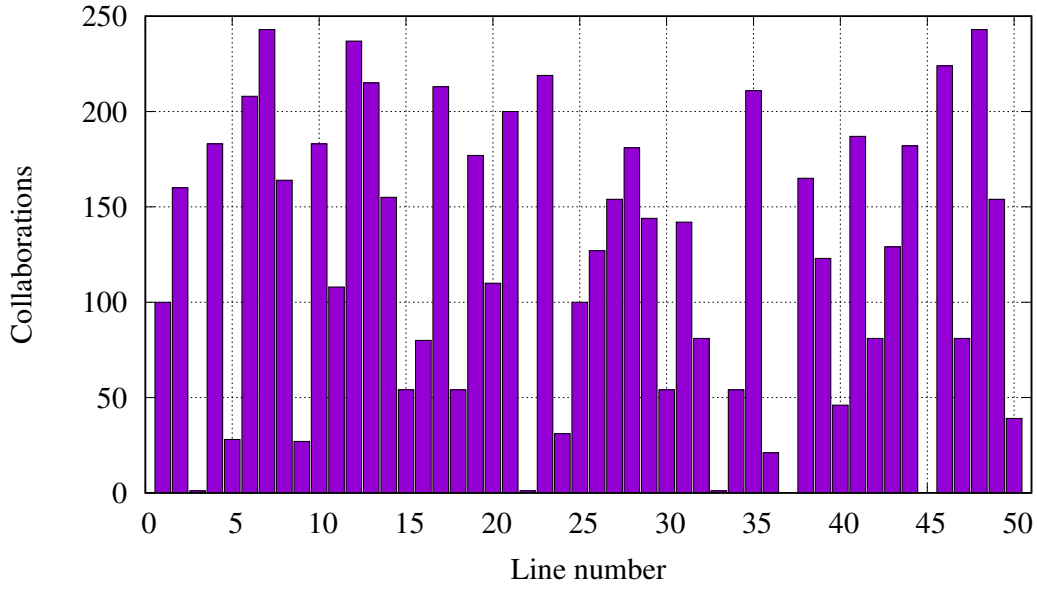


Figure 8.11: Histogram representing the number of collaborations (times read) for each text line in the experiments for $\tau = 0$ and $B = [5,\ldots,45]$.

with the collaboration effort per line. This relation is presented in Figure 8.13 and, as can be observed, the lines with lower R require a higher amount of collaboration. Specifically, the 50% of lines with lower

Table 8.8: Features of the collaborations per line distribution. Q_1, Q_2, and Q_3 are respectively the 1st, 2nd, and 3rd Quartile, IQR the Interquartile Range, LIF the Lower Inner Fence, and UIF the Upper Inner Fence.

Q_1	Q_2	Q_3	IQR	LIF	UIF
54	128	183	129	-139.5	376.5

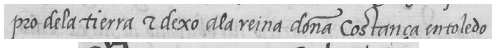
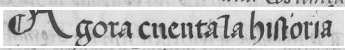
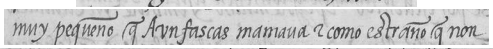
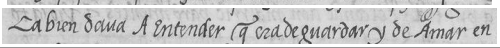
Line No.	Handwritten Text Line Image
7	
37	
45	
48	

Figure 8.12: Examples of lines that required full collaboration (7 and 48), and lines that were never refined (37 and 45). Line 7 corresponds with the 7th line of the page 515, while the lines 37, 45, and 48 correspond with the lines 12th, 20th, and 23th of page 579, respectively.

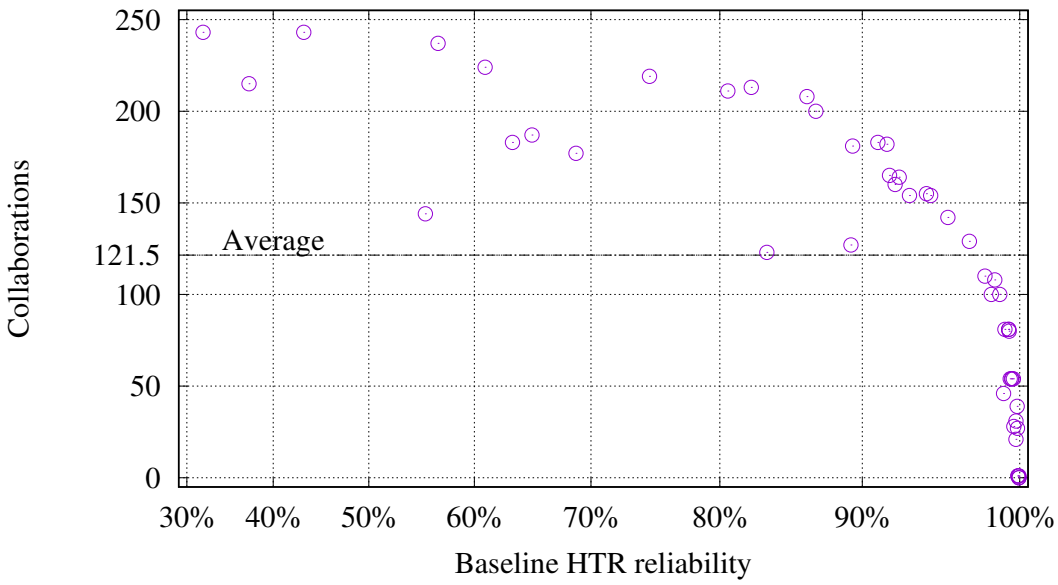


Figure 8.13: Relation between the baseline HTR reliability R and the number of collaborations for each text line in the experiments for $\tau = 0$ and $B = [5,\ldots,45]$.

R concentrated 76.9% of collaborations. Besides, all lines that needed more repetitions than the average expected number (121.5) presented a value of $R \leq 97\%$, whereas those with less repetitions than the average presented $R > 97\%$. This makes us suppose that a clear border can be established between the lines that would need more or less collaborations according to the reliability they present in the HTR recognition.

8.4 Conclusions and Future Work

This chapter presents the experimentation performed to test our proposal for a multimodal crowdsourcing framework for the transcription of historical handwritten documents. 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. This framework uses a client / server architecture, where the client application is publicly available in order to allow collaborators to decide when and where to collaborate.

The experiments showed that in this framework, the number of collaborators is more important than the order in which their speech is processed, and the speech reliability verification permits us to achieve better results. Moreover, the lines selection module analyses the transcription reliability at the output of the handwritten text lines to transcribe, and selects the set of lines with lower reliability to be presented to the collaborators. In this way, the collaboration effort is focused to the lines whose transcription needs more refinement.

We propose for future studies the use of more robust modelling methods, such as Deep Neural Networks (DNN) for optical and acoustical modelling and Recurrent Neural Networks (RNN) for language modelling. Moreover, this multimodal crowdsourcing framework is open to be tested with other datasets.

Bibliography

Alabau, V., Martínez-Hinarejos, C.-D., Romero, V., and Lagarda, A. L. (2014). An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195–203. Frontiers in Handwriting Processing.

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

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, volume 1, pages 409–412.

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

Granell, E. and Martínez-Hinarejos, C.-D. (2016). 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)*, pages 411 – 417.

Kneser, R. and Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, volume 1, pages 181–184.

Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.-D., Pastor, M., Sanchis, A., and Toselli, A. (2008). iATROS: A speech and handwritting recognition system. In *Proceedings of the V Jornadas en Tecnologías del Habla (VJTH'2008)*, pages 75–78.

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

Serrano, N., Castro, F., and Juan, A. (2010). The RODRIGO Database. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2709–2712.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the 3rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 901–904.

Welch, B. L. (1947). The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35.

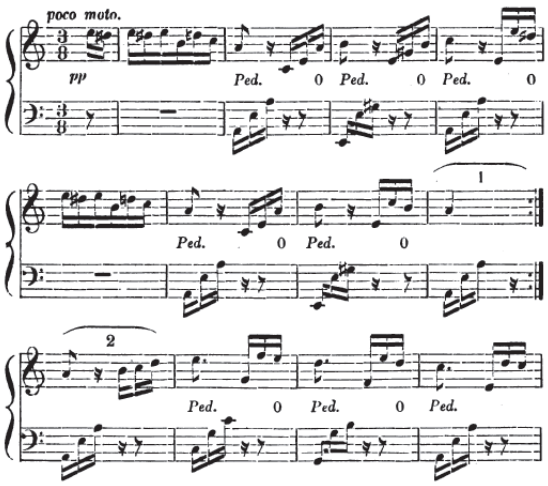
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

V

Conclusions and Future Work

"Without music, life would be a mistake."

Friedrich Nietzsche, Twilight of the Idols, 1889.



Bagatelle for solo piano. (Ludwig van Beethoven, Für Elise, 1867)

9 Conclusions and Future Work

"En qualque terra on sia pot viure maestral;
e per assò los sarrains han molt bona manera en assò que tot hom per richom que sia,
per tot assò nos lexa de mostrar a son fill alcun mester;
per so que si li fallía la riquea, que pogués viure per son mester." [1]

Ramon Llull, Doctrina pueril, 1274-1276.

Content

Conducting a doctoral thesis involves a great amount of work that culminates when a report like the one that finishes in the present chapter is completed. However, despite the large amount of work done, there will always be some additional details to be tested and improvements to be made in future work.

This chapter is structured as follows: Section 9.1 summarises the general conclusions, Section 9.2 shows the scientific work and contributions derived from this thesis, and Section 9.3 presents the future work lines.

9.1 Conclusions

In this thesis, the reduction of the required effort of the paleographer for obtaining the actual transcription of digitalised historical manuscripts have been studied in the following scenarios: multimodality, interactivity, and crowdsourcing.

In the multimodality part of the thesis (Part II), the benefits of combining additional sources of information for the transcription of historical manuscripts have been confirmed.

The proposed combination method takes advantage of the fact that different systems make different errors; thus, editing operations can correct errors. Insertion and deletion create new word sequences than enrich the resulting Confusion Network, and the combination can maximise the probability of the correct word, when both subnetworks contain the correct word even when this word has a low probability in both subnetworks. Conversely, if only one subnetwork contains the correct word and both subnetworks contain the same erroneous word, this error will be maximised at the expense of the correct word. Despite this fact, the experiments performed confirm the strengths of this Confusion Network combination technique.

When comparing the performance of our proposed combination technique with other hypotheses combination techniques, we observed that the Lattices Rescoring method (Stolcke et al., 1997) offered by

[1] "In any land can live the minstrel; and so the Saracens have as good habit, whatever wealth they have, to teach their sons a profession; so, in case he lacks his wealth, he can live from his work."

the SRILM toolkit (Stolcke, 2002) allows one to obtain similar results. However, the knowledge acquired by combining handwritten text and speech recognition allowed the realisation of the rest of the thesis.

In Part III, multimodality was applied on an interactive tool for transcribing historical handwritten documents (CATTI) (Romero et al., 2012). On the one hand, the multimodal hypotheses combination allows one to reduce the time and the workload of paleographers for transcribing historical books, due to the increased recognition accuracy and the quality of the alternatives contained in the multimodal lattice. On the other hand, the use of Confusion Networks combination allows one to improve the 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.

A multimodal crowdsourcing approach for the transcription of historical handwritten documents was proposed in the previous part (Part IV) of this thesis. 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.

The experiments showed that in this framework the number of collaborators is more important than the order in which their speech is processed, and the speech reliability verification permits one to achieve better results. Moreover, the lines selection module analyses the transcription reliability at the output of the handwritten text lines to transcribe, and selects the set of lines with lower reliability to be presented to the collaborators. In this way, the collaboration effort is focused to the lines whose transcription needs more refinement. 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.

9.2 Scientific Work and Contributions

During the realisation of this thesis I have participated in the following research projects:

- From 16/04/2013 until 30/06/2014 in the research project: ***Percepción**: Desarrollo de sistemas de interacción avanzada hombre-máquina*, coordinated by Factory Holding Company 25, S.L (FHC25), directed at the UPV by Dr. Carlos David Martínez Hinarejos, and supported by the *Ministerio de Industria, Energía y Turismo* under the reference TSI-020601-2012-50.
- From 01/07/14 until 31/10/2014 in the research project: ***STraDA:** Search in Transcribed Manuscripts and Document Augmentation*, directed by Dr. Joan Andreu Sánchez Peiró, and supported by the *Ministerio de Economía y Competitividad* under the reference TIN2012-37475-C02-01.
- From 01/11/2014 until 31/12/2016 in the research project: ***Smart Ways**: Desarrollo de una plataforma tecnológica orientada a la eficiencia de los recursos en el campo de las nuevas tecnologías Internet of things*, directed by Dr. Carlos David Martínez Hinarejos, and funded by *Ministerio de Economía y Competitividad* under the reference RTC-2014-1466-4.
- From 01/01/2017 in the research project: ***CoMUN-Hat**: Contexto, multimodalidad y colaboración del usuario en procesado de texto manuscrito*, directed by Dr. Carlos David Martínez Hinarejos, and supported by the *Ministerio de Economía y Competitividad* under the reference TIN2015-70924-C2-1-R.

The participation in these research projects allowed me to acquire knowledge and research experience, although it were not always directly related to this thesis. The progress of this thesis and the other scientific contributions were I have collaborated during the realisation of this thesis were disseminated through presentations and articles in national and international conferences. Moreover, I have presented parts of this thesis in several seminars and I have realised two research stays, one in *Centre de Visió per Computador (CVC) - Universitat Autònoma de Barcelona (UAB)* (Barcelona, Spain), and the second one in *Département Traitement du Signal et des Images (TSI) - École Nationale Supérieure des Télécommunications (ENST Paris - Télécom ParisTech)* (Paris, France) which is one of the top French

public institutions of higher education and research (*Grandes Écoles*) of engineering in France. These contributions are chronologically sorted.

I started presenting the idea of combining speech and handwriting recognition for improving the transcription of historical documents in the first meeting of doctoral students of the *Universitat Politècnica de València* in 2014:

- **E. Granell**, *"Multimodal Recognition: Handwriting & Speech"*, In *I Encuentro de Estudiantes de Doctorado*, Universitat Politècnica de València, Valencia (Spain), June 12, 2014.

That same year, the performed work in the *Percepción* project allowed us to present three articles (related with distributed speech recognition and smart cities) in the international conference IberSPEECH 2014:

- **E. Granell** and C.-D. Martínez-Hinarejos, *"A study of the quality of automatic speech recognition in distributed system"*, In Proceedings of "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop" (IberSPEECH 2014), pp. 119-128, Las Palmas de Gran Canaria (Spain), November 19-21, 2014.
- **E. Granell**, C.-D. Martínez-Hinarejos, G. Amat, J. Fernández, Á. Arranz, Á. Ramos, J.-M. Benedí and A. Sanchis, *"Speech Recognition on the Percepción Project"*, In Proceeding of "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop" (IberSPEECH 2014), pp. 321-330, Las Palmas de Gran Canaria (Spain), November 19-21, 2014.
- C.-D. Martínez-Hinarejos, **E. Granell**, D. Rambla, A. Calia, A. Luján, G. Amat, Á. Ramos, J.-M. Benedí and A. Sanchis, *"The Percepción Smart Campus system"*, In Proceedings of "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop" (IberSPEECH 2014), pp. 359-366, Las Palmas de Gran Canaria (Spain), November 19-21, 2014.

In 2015 the main scientific contributions for the Multimodality part (Part II) were produced. Throughout that year I managed to develop a new multimodal combination technique based on Confusion Networks, that I tested mainly in two different experiments: iterative multimodal combination and combination of multiple speech and handwriting recognisers outputs. Then, in March 2015, I introduced this multimodal combination technique and the performed experiments until that moment in a seminar in the PRHLT research center:

- **E. Granell**, *"Combining Outputs from Multiple Handwritten and Speech Recognition Systems for Transcribing Historical Handwritten Documents"*, In *PRHLT seminars*, Universitat Politècnica de València, Valencia (Spain), March 9, 2015.

The criticism at the seminar served to improve the technique and I received proposals for collaboration in other fields in which this multimodal combination technique could be useful, such as the combination of *off-line* and *on-line* Handwriting Text Recognition.

Additionally, I presented this work in the second meeting of doctoral students of the *Universitat Politècnica de València* and in the summer school of the *Red Temática en Tecnologías del Habla* (RTTH):

- **E. Granell**, *"Reconocimiento multimodal: Combinando escritura manuscrita y habla"*, In *II Encuentro de Estudiantes de Doctorado*, Universitat Politècnica de València, Valencia (Spain), June 25, 2015.
- **E. Granell**, *"Multimodal Combination of Multiple Handwriting and Speech Recognition Systems"*, In *RTTH Summer School on Speech Technology: A Deep Learning Perspective*, Red Temática en Tecnologías del Habla, Universitat Politècnica de Catalunya, Barcelona (Spain), July 6-9, 2015.

The Multimodality part is supported by two articles presented in two different international conferences. Concretely, the experiments of iterative multimodal combination were presented in the "13th International Conference on Document Analysis and Recognition" (ICDAR 2015) (Core A), and the experiments of combination of multiple speech and handwriting recognisers outputs in the "16th International Conference on Computer Analysis of Images and Patterns" (CAIP 2015) (Core B):

- **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 2015), pp. 126-130, Nancy (France), August 23-26, 2015.

- **E. Granell** and C.-D. Martínez-Hinarejos, *"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, Valetta (Malta), September 2-4, 2015.

The first supporting scientific contribution for the Interactivity part (Part III) was obtained early in the year 2016. The improvement of an interactive transcription system by using Confusion Networks Combination to combine *off-line* and *on-line* Handwriting Text Recognition was presented in the "12th IAPR International Workshop on Documents Analysis Systems" (DAS 2016) (Core B):

- **E. Granell** and 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 Documents Analysis Systems (DAS 2016), pp. 269-274, Santorini (Greece), April 11-14, 2016.

The next contribution was a book chapter where we presented a multimodal approach combining *off-line* Handwriting Recognition and Speech Recognition in the interactive transcription system, taking into account the user feed-back through touchscreen pen strokes (*on-line HTR*), traditional keyboard, and mouse operations:

- **E. Granell** and V. Romero and C.-D. Martínez-Hinarejos, *"Using Speech and Handwriting in an Interactive Approach for Transcribing Historical Documents"*, In HANDWRITING: RECOGNITION, DEVELOPMENT AND ANALYSIS, NovaPub, 2017.

Then, the scientific contributions for the Crowdsourcing part (Part IV) started with a research stay of a week (June 6-10, 2016) in the CVC research center of UAB. This research stay was partially supported by the *Ministerio de Economía y Competividad* of the Spanish Government through the researchers mobility aids of the *Multimodal Interaction in Pattern Recognition and Computer Vision Network of Excellence (R-MIPRCV)*. During this research stay I presented the developed multimodal crowdsourcing platform for the transcription of handwritten documents in a seminar:

- **E. Granell**, *"Multimodal Crowdsourcing for Transcribing Handwritten Documents"*, In *Computer Vision Center, Universitat Autònoma de Barcelona*, Bellaterra, Barcelona (Spain), June 6-10, 2016.

This multimodal crowdsourcing platform for the transcription of handwritten documents was tested through several experiments, and was presented in the "16th ACM International Symposium on Document Engineering" (DocEng 2016) (Core B):

- **E. Granell** and C.-D. Martínez-Hinarejos, *"A Multimodal Crowdsourcing Framework for Transcribing Historical Handwritten Documents"*, In Proceedings of the 16th ACM International Symposium on Document Engineering (DocEng 2016), pp. 157-163, Vienna (Austria), September 13-16, 2016.

Given that collaborators are a scarce resource, we studied how to get the maximum benefit from their effort in this crowdsourcing platform. This study and a tool that we developed for acquiring speech samples from mobile devices were presented in the international conference IberSPEECH 2016:

- **E. Granell** and C.-D. Martínez-Hinarejos, *"Collaborator Effort Optimisation in Multimodal Crowdsourcing for Transcribing Historical Manuscripts"*, Advances in Speech and Language Technologies for Iberian Languages, pp. 234-244, Springer, 2016.
- **E. Granell** and C.-D. Martínez-Hinarejos, *"Read4SpeechExperiments: A Tool for Speech Acquisition from Mobile Devices"*, In Proceedings of "IX Jornadas en Tecnologías del Habla" and "V Iberian SLTech Workshop" (IberSPEECH 2016), pp. 411 - 417, Lisboa (Portugal), November 23-25, 2016.

The next step was to study the behaviour of this crowdsourcing framework using real speech samples acquired from mobile devices. Therefore, we called for collaboration and we obtained the collaboration of a huge amount of speakers that used the *Read4SpeechExperiments* application for the acquisition of the dictation of the contents of the text lines. This data was used to perform experiments in order to analyse the behaviour of the crowdsourcing framework in a real scenario. This work was published in the international journal IEEE/ACM Transactions on Audio, Speech, and Language Processing (JCR: 1.225):

- **E. Granell** and C.-D. Martínez-Hinarejos, *"Multimodal Crowdsourcing for Transcribing Handwritten Documents"*, In IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25,

Num. 2, pp. 409-419, 2017.

In March 2017, I presented the complete crowdsourcing platform and all the performed experiments in a seminar in the PRHLT research center:

- **E. Granell**, *"Multimodal Crowdsourcing for Transcribing Historical Manuscripts."*, In *PRHLT seminars*, Universitat Politècnica de València, Valencia (Spain), March 23, 2017.

At the end of 2016 I had the opportunity to perform a research stay of three months (from September 20th to December 20th) in the TSI department of *Télécom* ParisTech. This research stay was partially supported by the *European Commission* through the *European Union programme for education, training, youth and sport Erasmus+*. The aim of this research was to improve the recognition of Out Of Vocabulary (OOV) words in the transcription of historical manuscripts. Through an extensive experimentation, we achieved to recognise a great amount of OOV words without using external resources to enrich the language model. Currently, two articles are in preparation from this experimentation.

Moreover, during the realisation of this thesis I have collaborated on some research that were not directly related with the domain of this thesis but that gave me the opportunity to continue learning and to offer other interesting contributions to the scientific community, such as the work presented in the "11th ACM International Conference on Interactive Surfaces and Spaces" (ISS 2016) (Core A):

- **E. Granell** and L. A. Leiva, *"Less Is More: Efficient Back-of-Device Tap Input Detection Using Built-in Smartphone Sensors"*, In Proceedings of the 11th ACM International Conference on Interactive Surfaces and Spaces (ISS 2016), pp. 5-11, Niagara Falls (Canada), November 6-9, 2016.

In this research, Back-of-Device (BoD) interaction using current smartphone sensors (e.g. accelerometer, microphone, or gyroscope) was studied with the aim of selecting the optimal subset of features that is a good predictor of BoD tap based input while ensuring low energy consumption.

Finally, Table 9.1 presents a summary where the publications related to this thesis are highlighted.

Table 9.1: Summary of relevant publications.

Thesis Part	Publication	Type	Ranking
Multimodality	Granell and Martínez-Hinarejos (2015a)	Conference	Core: A
	Granell and Martínez-Hinarejos (2015b)	Conference	Core: B
Interactivity	Granell et al. (2016)	Conference	Core: B
	Granell et al. (2017)	Book Chapter	
Crowdsourcing	Granell and Martínez-Hinarejos (2016a)	Conference	Core: B
	Granell and Martínez-Hinarejos (2016b)	Book Chapter	
	Granell and Martínez-Hinarejos (2016c)	Conference	
	Granell and Martínez-Hinarejos (2017)	Journal	JCR: 1.225

9.3 Future Work

We propose for future studies the use of more robust modelling methods for the natural language recognition systems, such as Deep Neural Networks (DNN) for optical, acoustical, and kinematical modelling and Recurrent Neural Networks (RNN) for language modelling. The use of context in morphological modelling is another option to be explored in the future. Using these models, it is expected that we will obtain better unimodal baseline lattices.

Regarding multimodality, the use of whole sentences instead of lines of the handwritten text corpus 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 are planning to test the use of speech not only as an

additional source of information of the handwritten text image to transcribe in the CATTI system, but as an additional modality for Computer-Human Interaction (CHI). 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 could be integrated and tested with other datasets.

Bibliography

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

Granell, E. and Martínez-Hinarejos, C.-D. (2015b). Multimodal Output Combination for Transcribing Historical Handwritten Documents. In *Proceedings of the 16th International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 246–260.

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

Granell, E. and Martínez-Hinarejos, C.-D. (2016b). Collaborator Effort Optimisation in Multimodal Crowdsourcing for Transcribing Historical Manuscripts. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 234–244. Springer.

Granell, E. and Martínez-Hinarejos, C.-D. (2016c). 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)*, pages 411 – 417.

Granell, E. and Martínez-Hinarejos, C.-D. (2017). Multimodal Crowdsourcing for Transcribing Handwritten Documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):409–419.

Granell, E., Romero, V., and Martínez-Hinarejos, C.-D. (2016). 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)*, pages 269–274.

Granell, E., Romero, V., and Martínez-Hinarejos, C.-D. (2017). Using Speech and Handwriting in an Interactive Approach for Transcribing Historical Documents. In *HANDWRITING: RECOGNITION, DEVELOPMENT AND ANALYSIS*. NovaPub.

Romero, V., Toselli, A. H., and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*, volume 80 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the 3rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 901–904.

Stolcke, A., Konig, Y., and Weintraub, M. (1997). Explicit Word Error Minimization in N-best List Rescoring. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, volume 1, pages 163–166.