This is the author's version of an article that has been published in the Computational Intelligence journal.
The final version is available at http://dx.doi.org/10.1111/coin.12169

Multimodality, Interactivity, and Crowdsourcing for Document Transcription

EMILIO GRANELL, VERÓNICA ROMERO, AND CARLOS D. MARTÍNEZ-HINAREJOS

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

Knowledge mining from documents usually use document engineering techniques that allow the user to access the information contained in documents of interest. In this framework, transcription may provide an efficient access to the contents of handwritten documents. Manual transcription is a time-consuming task that can be speed-up by using different mechanisms. A first possibility is employing state-of-the-art handwritten text recognition systems to obtain an initial draft transcription that can be manually amended. A second option is employing crowdsourcing to obtain a massive but not error-free draft transcription. In this case, when collaborators employ mobile devices, speech dictation can be used as transcription source, and speech and handwritten text recognition can be fused to provide a better draft transcription, that can be amended with even less effort. A final option is using interactive assistive frameworks, where the automatic system that provides the draft transcription and the transcriber cooperate to generate the final transcription. The novel contributions presented in this work include the study of the data fusion on a multimodal crowdsourcing framework and its integration with an interactive system. The use of the proposed solutions reduce the required transcription effort and optimises the overall performance and usability, allowing for a better transcription process.

Key words: Assistive environment, Crowdsourcing framework, Historical handwritten transcription, Multimodal combination.

1. INTRODUCTION

Knowledge mining and information extraction from documents are Artificial Intelligence fields that try to obtain associations among entities that are expressed in the documents under study, in order to find new useful patterns and correlations. These research fields typically have to employ document engineering techniques that allow us to extract relevant information from the documents of interest. This information will be posteriorly used by the usual knowledge mining and information extraction techniques.

When documents include handwritten text, the transcription of the handwritten content is one of the most relevant topics in document analysis and engineering. In contrast to Optical Character Recognition (OCR) (Impedovo et al., 1991), that nowadays can be employed to obtain an accurate transcription of printed text, transcription of handwritten text is far from being fully automatic in document engineering. This is even more relevant for documents that mostly contain handwritten text. For this case, simple image digitisation only provides, in most cases, search by image. Thus, transcription is needed to obtain an easy digital access to the contents of the documents, and it allows us to search by linguistic contents (keywords, expressions, syntactic or semantic categories, ...), which are the usual elements employed in knowledge mining and information extraction. Transcription is even more important for historical documents, since most of these documents are unique and the preservation of their contents is crucial for cultural and historical reasons. Interest in historical document transcription is reflected in the development of international projects such as IMPACT (Na-

Address correspondence to: egranell@dsic.upv.es

tional Library of the Netherlands, 2010), tranScriptorium (tranScriptorium Project, 2013), and READ (READ Project, 2016).

Usually, transcriptions are done by professionals in order to guarantee its quality. These specialists perform the transcriptions by typing the contents of the manuscripts. In the last decade, development of Handwritten Text Recognition (HTR) (Plamondon and Srihari, 2000) tools provided professional transcribers with an initial transcription that they can amend, obtaining a higher productivity in the transcription task. In the case of historical manuscripts more difficulties appear, since the vocabulary, lexical forms, expressions, and scripting style can be quite different from modern handwritten documents. For this case, specialists called paleographers are responsible for obtaining a final accurate transcription from this difficult data.

Nowadays, the rise of crowdsourcing platforms (Doan et al., 2011), where many volunteers could contribute to a given task at a very small or even null cost, made the transcription of handwritten text, and more specifically of historical texts, widespread. Apart from popular platforms such as Mechanical Turk[1] or CrowdFlower[2], several platforms (specific for historical texts) that benefit from this approach have been developed in the last years (such as AnnoTate[3], Transcribe Bentham[4], or Transkribus[5]). However, final supervision of paleographers is required most times in order to obtain accurate enough transcriptions, since difficulties presented by historical documents (degraded image quality, ancient vocabulary, or strange calligraphy) make difficult to obtain high quality transcriptions for non-expert transcribers.

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, given a speech dictation of the handwritten text contents. This modality is available in nearly all mobile devices, and would allow us to obtain a larger number of volunteers (Granell and Martínez-Hinarejos, 2017). However, voice presents an ambiguity that typed input lacks of, since Automatic Speech Recognition (ASR) (Rabiner and Juang, 1993) systems are far from perfect, although their performance has increased substantially over the recent years (Hinton et al., 2012).

In any case, since paleographers must finally supervise the transcription, the use of voice input could provide a new more accurate transcription (taking as a starting point an HTR transcription) that can reduce substantially the supervision effort of final transcribers. Moreover, it is foreseeable that the more volunteers provide their voice transcription, the more accurate the transcription would be (Granell and Martínez-Hinarejos, 2016a; Granell and Martínez-Hinarejos, 2017). Thus, since the use of mobile devices is widespread, final performance of the system could be near to that of a typed-input system.

Another possibility to increment paleographers' productivity is to provide them with a transcription tool that works in a collaborative way. This tool would have to take into account transcribers' feedback in order to offer them new hypotheses on the final transcription. In this way, correcting one error could produce a cascade of corrections, since the transcriber' feedback is used by the system to recalculate the whole hypothesis and provide a (hopefully) better transcription. In the projects IMPACT (National Library of the Netherlands, 2010) and

[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/

tranScriptorium (tranScriptorium Project, 2013), this approach was studied for transcription systems that only employ text image input. The most recent is the so called Computer Assisted Transcription of Text Images (CATTI) framework (Romero et al., 2012). This framework has demonstrated its appropriateness to increment productivity in historical text transcription by paleographers.

In this work, we study how to employ multimodal recognition (combining HTR and ASR) and crowdsourcing acquisition to provide a multimodal input to the interactive CATTI framework. We study the impact of these two new data sources in this assistive transcription tool in terms of user productivity gain. Special attention is paid to the multimodal data fusion mechanisms that work on massive data obtained by crowdsourcing, and how hypotheses offered by these processes provide better alternatives to the CATTI tool.

The paper is structured as follows: Section 2 reviews related work on multimodal recognition, interactive systems and crowdsourcing, Section 3 presents the details on the proposed crowdsourcing framework, Section 4 provides insights on the CATTI framework, Section 5 describes the experimental conditions, Section 6 shows the results, and Section 7 summarises the conclusions and future work lines.

2. RELATED WORK

This section reviews the work carried out in the three main areas of this paper: multimodality, interactivity, and crowdsourcing.

In multimodal recognition, the main idea is to use different sources of information to obtain a more accurate recognition. For instance, multimodality can be applied to audio-visual speech recognition. In the case of multimodality with speech and lips movements (Tamura et al., 2005; Hazen, 2006) the original sources are usually synchronous: on the one hand the speech signal, and on the other hand, the lips and mouth movements recorded in video images. However, in the case of multimodality with speech and gestures the process is more complicated (Miki et al., 2014), given that in this case the two signals are usually not synchronous.

The combination of speech and handwritten text (which are usually asynchronous signals) is usually used to improve the recognition of the handwritten text. For instance, it can be used for isolated words, where optical character and speech recognition are combined in order to enhance word recognition (Singh et al., 2012). Besides, it can be used for continuous recognition, where the speech dictation of the contents of text images is used to improve the transcription of these text images (Alabau et al., 2011; Granell and Martínez-Hinarejos, 2015).

Assistive technologies aim to make human users' work easier and faster. Those technologies are used in many fields of computer applications, such as the Computer Aided Design field (Machover, 1995), medical diagnosis (Doi, 2007), automatic driving (Malit, 2009), Natural Language Processing (Barrachina et al., 2009; Revuelta-Martínez et al., 2012; Silvestre-Cerdà et al., 2013), and Pattern Recognition and Image Processing (Romero et al., 2012).

Multimodality can be incorporated in those systems in order to improve the Human-Computer Interaction or to provide the system with additional sources of information. For example, recently, the user interaction in the CATTI framework (Romero et al., 2012) was improved in (Granell et al., 2016) by allowing the user to use on-line HTR as feedback. Besides, in addition to the multimodal feedback, the CATTI system can be fed with the dictation of the contents of the text line image to be transcribed; these dictations would act as an additional source of information for CATTI (Granell et al., 2017).

As previously said, crowdsourcing approaches for HTR are very useful. Speech recog-

nition is other field where crowdsourcing approaches can be applied. For instance, to acquire speech data (Parent and Eskenazi, 2011), task that can be performed easily using mobile devices (Caines et al., 2016).

The initial multimodal crowdsourcing framework (Granell and Martínez-Hinarejos, 2016a) allowed us to use speech utterances for improving the transcription of historical manuscripts. This system was improved by adding a line selection module that allows us to optimise the collaboration effort (Granell and Martínez-Hinarejos, 2016b). Then, it was tested in a real scenario, where the speech collaborations were acquired from mobile devices (Granell and Martínez-Hinarejos, 2017).

In this work, the data fusion and lattices transformation performed by this crowdsourcing framework is studied. Moreover, we experimented with the integration of the crowdsourcing framework and the CATTI system. In this way, we studied the influence in the transcriber' effort reduction of the crowdsourcing and the interactive approaches. Therefore, the main contributions of this paper can be summarised as: the study of the data fusion and lattices transformation on a multimodal crowdsourcing framework, and the integration of crowdsourcing and interactivity, which are two different lines of research with the same objective (to reduce the human transcription effort).

3. MULTIMODAL CROWDSOURCING TRANSCRIPTION

The Handwritten Text Recognition (HTR) and Automatic Speech Recognition (ASR) problems can be formulated in a very similar way that allows their integration into a multimodal system. The unimodal formulation is: given a handwritten text image or a speech signal encoded into the feature vector sequence $x = (x_1, x_2, \ldots, x_{|x|})$, finding the most likely word sequence $\hat{w} = (w_1, w_2, \ldots, w_{|w|})$, that is:

$$\hat{w} = \arg\max_{w \in W} \mathrm{P}(w \mid x) = \arg\max_{w \in W} \frac{\mathrm{P}(x \mid w)\mathrm{P}(w)}{\mathrm{P}(x)} = \arg\max_{w \in W} \mathrm{P}(x \mid w)\mathrm{P}(w) \tag{1}$$

where W denotes the set of all permissible sentences, $\mathrm{P}(x)$ is the probability of observing x, $\mathrm{P}(w)$ is the probability of w (approximated by the language model, LM), and $\mathrm{P}(x \mid w)$ is the probability of observing x by assuming that w is the underlying word sequence for x (evaluated by the optical or acoustical model, for HTR and ASR respectively).

The main objective of this crowdsourcing framework is to reduce the transcription errors in $\hat{w}$ before giving it to a paleographer for obtaining the actual transcription. This framework is based on two ideas: using the current system output to improve the language model for the next decoding process (Alabau et al., 2011), and combining decoding outputs in order to obtain an output with lower error rate (Granell and Martínez-Hinarejos, 2015).

In this framework, HTR and ASR models are the only dependent parts, i.e., this framework is generalisable by using the corresponding lexicon, language, optical and acoustical models. Thus, this crowdsourcing framework can be used to transcribe any manuscript (even multiwriter manuscripts) in any language.

Figure 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 module.
(2) The crowdsourcing loop starts in the LM interpolation module, where the previous system output is interpolated with the original LM for obtaining an improved LM for the next ASR decoding.
(3) The reliability of the system output is evaluated, the lines are sorted by its reliability, and the collaborator is asked to read only a subset of those lines with lower reliability.

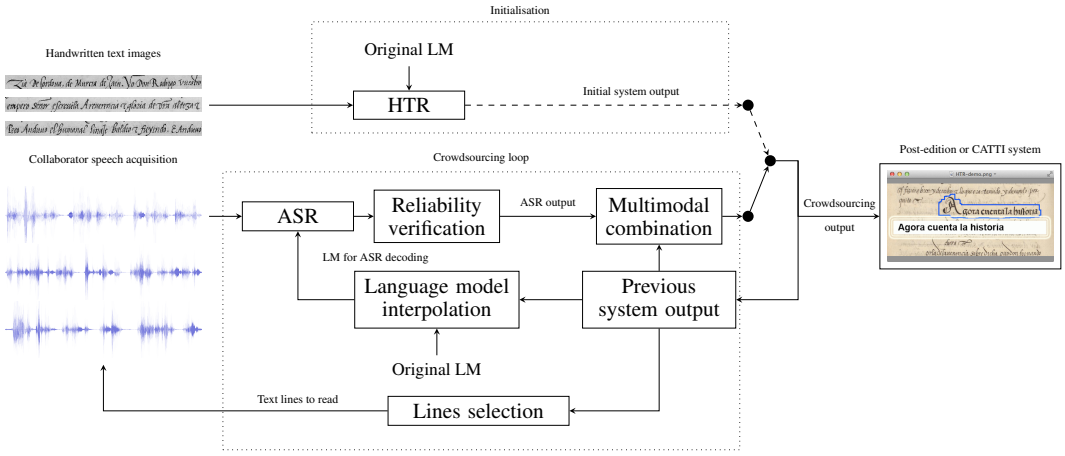
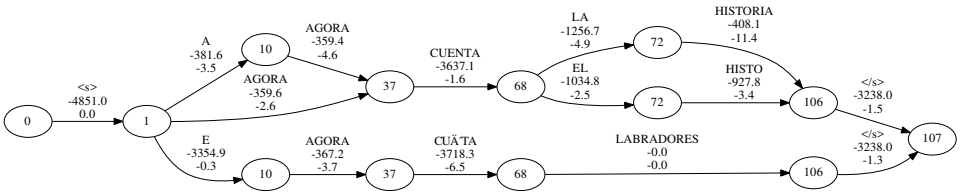
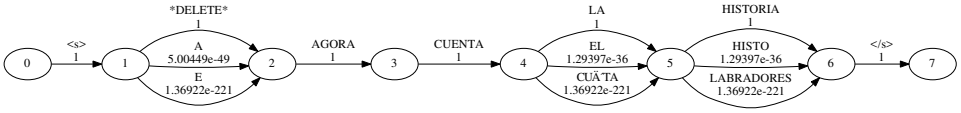


FIGURE 1. Multimodal and interactive crowdsourcing transcription framework.



(a) Word Graph.



(b) Confusion Network.

FIGURE 2. Word Graph and Confusion Network.

(4) The collaborator speech is decoded in the ASR module using the improved LM, and the reliability of the obtained ASR output is verified and filtered, i.e., only the decoding output of those utterances which reach a determined reliability threshold are accessible at the output of the reliability verification module.
(5) The system output is updated by combining the previous system output with the current and verified ASR output in the multimodal combination module.
(6) Every time a new collaborator offers to help, the crowdsourcing loop (steps from 2 to 5) is executed and the system output is improved by using the new audio samples.

The following subsections describe in detail the different modules of the framework.

3.1. Language Model Interpolation

The decoding outputs obtained from handwriting and speech recognisers can be formatted as Word Graph (WG) lattices, and as Confusion Networks (CN). In Figure 2, an example of WG and its corresponding CN is presented.

A WG is a directed, acyclic and weighted graph with an initial node q_I and a final node q_F. The nodes correspond to discrete time points for ASR and horizontal space for HTR. A

link l is defined as any edge between a starting node $s(l)$ and an ending node $e(l)$, and it represents a hypothesis word $w(l)$ with a score $f(l)$.

The language interpolation module builds a statistical language model (LM) conditioned on x that can be used to calculate the posterior probabilities of Equation (1). This adapted LM can be obtained as follows (Alabau et al., 2011):

(1) The decoding outputs are obtained from the decoding process as WG.
(2) The posterior probabilities for each WG node $\mathrm{P}(q \mid x)$ and WG link $\mathrm{P}(l \mid x)$ are computed. These probabilities are based on the forward $\alpha(q)$ and backward $\beta(q)$ probabilities of the nodes (Wessel et al., 2001).
(3) The posterior probability for a specific link l can be computed as the sum of the posterior probabilities of all hypotheses containing it. Therefore, the counts for a word sequence $x_{i-n+1}^{i} = (w_{i-n+1}, \ldots, w_i)$ can be estimated as:

$$C^*(w_{i-n+1}^{i} \mid x) = \sum_{l_1^n \in N(w_{i-n+1}^{i})} \frac{\prod_k \mathrm{P}(l_k \mid x)}{\prod_k \mathrm{P}(s(l_k) \mid x)} \tag{2}$$

where $N(w_{i-n+1}^{i})$ are all the sequences of concatenated links generating w_{i-n+1}^{i}.
(4) The word posterior probabilities associated to the current input x can be calculated after applying: a suitable discount method (for back-off estimation), a smoothing method (to avoid the out-of-vocabulary problem), and a proper normalisation:

$$\mathrm{P}^x(w) = \prod_i \frac{C^*(w_{i-n+1}^{i} \mid x)}{C^*(w_{i-n+1}^{i-1} \mid x)} \tag{3}$$

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

$$\mathrm{P}_\lambda^x(w) = \lambda \mathrm{P}^x(w) + (1-\lambda)\mathrm{P}(w) \tag{4}$$

The weight factor λ permits balancing the relative reliability in the LM interpolation between the LM estimated from the previous system output and the original LM.

3.2. Multimodal Combination

The multimodal combination of the ASR decoding output with the previous system output can be performed by means of Confusion Network (CN) combination methods.

A CN is a weighted directed graph, in which each hypothesis goes through all the nodes. The words and their probabilities are also stored in the edges. A subnetwork (SN) is the set of all edges between two consecutive nodes. The total probability of the words contained in a SN sum up to 1.

The bimodal CN combination method (Granell and Martínez-Hinarejos, 2015) used in the framework works as follows, starting from the system and the speech decoding outputs formatted in CN:

(1) A search for anchor subnetworks is performed in order to align the subnetworks of both CN. The algorithm searches coincidences in unigrams, bigrams and skip-bigrams in both directions (from left to right and vice versa) simultaneously, taking only as anchor subnetworks those where both searches coincide according to a gram matching value of the words in the involved subnetworks. The gram matching between words of two subnetworks (SN_A and SN_B) is assessed by using the quadratic mean of the Character Error Rate (CER) and the Phoneme Error Rate (PER) between those words.

$$\mathrm{E}(w_A, w_B) = \sqrt{\frac{\mathrm{CER}(w_A, w_B)^2 + \mathrm{PER}(w_A, w_B)^2}{2}} \tag{5}$$

where w_A and w_B represent the words of the SN_A and SN_B, respectively. CER and PER are the Levensthein distance between the words of both subnetworks, CER at character level, and PER at phoneme level (by using the phonetic transcriptions of the recognised words), and E represents the gram matching error.

(2) The new CN is built on the basis of the Bayes theorem and assuming a strong independence between both CN. The editing actions used are: combination, insertion, and deletion of subnetworks:

Combination Given two subnetworks, SN_A and SN_B, the word posterior probabilities of the combined subnetwork SN_C are obtained by applying a normalisation on the logarithmic interpolation of the smoothed word posterior probabilities of both subnetworks (SN_A and SN_B) by using a weight factor α:

$$\mathrm{P}(w \mid SN_C) = \mathrm{P}_s(w \mid SN_A)^{\alpha} \mathrm{P}_s(w \mid SN_B)^{1-\alpha} \tag{6}$$

where the smoothing of the word posterior probability $\mathrm{P}_s(w \mid SN)$ is based on Laplacian smoothing. However, since we are working with probabilities, $\mathrm{P}_s(w \mid SN)$ is calculated according to Equation (7):

$$\mathrm{P}_s(w \mid SN) = \frac{\mathrm{P}(w \mid SN) + \Theta}{1 + n\Theta} \tag{7}$$

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 SN (SN_C).

Insertion and deletion The same process is performed in both actions: the subnetwork to insert or to delete is combined with a subnetwork with an only *DELETE* arc with probability 1.0.

The weight factor α permits to balance the relative reliability in the multimodal combination between the verified ASR decoding output and the previous system output.

3.3. Reliability Verification

Given the conventional formulation of the HTR and ASR problems, the posterior probability $\mathrm{P}(w \mid x)$ is a good confidence measure for the recognition reliability. However, recognition scores are inadequate to assess the recognition confidence because most recognition systems ignore the term $\mathrm{P}(x)$, as Equation (1) shows. Nevertheless, when the recognition scores of a fairly large n-best list are re-normalised to sum up to 1, the obtained joint probability $\mathrm{P}(x, w)$ can serve as a good confidence measure, since it represents a quantitative measure of the match between x and w (Rueber, 1997; Wessel et al., 2001).

Therefore, the re-normalised 1-best joint probability is used in the reliability verification module as the confidence measure:

$$R = \frac{\max\limits_{w \in W} \mathrm{P}(x, w)}{\sum\limits_{w \in W} \mathrm{P}(x, w)} \tag{8}$$

where W denotes the set of all permissible sentences in the evaluated decoding output.

3.4. Lines Selection

Since collaborators are a scarce resource, optimisation is very important to get the maximum benefit from their efforts. One way to optimise their effort is to reduce the set of lines that they have to read; in our case, it is interesting to select the subset of lines that need more refinement. Thus, when a collaborator offers to help, the system selects the lines to read as follows:

(1) The current system output is evaluated by using the re-normalised 1-best joint probability R -Equation (8)- in order to estimate the current confidence value of each one of the lines to transcribe.
(2) The lines are ranked according to their estimated confidence value.
(3) The system selects a subset of the B (batch size) lines with the lowest confidence value.
(4) The collaborator is asked to read only the selected B lines.

4. MULTIMODAL INTERACTIVE TRANSCRIPTION

This section reviews the *Computer Assisted Transcription of Text Images* (CATTI) approach presented in (Romero et al., 2012).

4.1. Interactive Transcription Basics

As previously commented, transcription of historical documents has become an interesting research topic in the last years. However, state-of-the-art recognition systems can not suppress the need of human work when high quality transcriptions are needed. For example, Handwritten Text Recognition (HTR) systems can achieve fairly high accuracy for restricted applications with rather limited vocabulary (reading of postal addresses or bank checks) and/or form-constrained handwriting. However, in the case of historical handwritten documents, the current HTR technology typically only achieves results which do not meet the quality requirements of practical applications. The same happens when using Automatic Speech Recognition (ASR) on the dictation of the contents of a text. 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.

A way of taking advantage of recognition systems, and more specifically of a HTR system, is to combine it with the knowledge of a human transcriber, constituting the so-called "Computer Assisted Transcription of Text Images" (CATTI) framework (Romero et al., 2012). In this framework, the automatic HTR system and the human transcriber cooperate interactively to obtain the perfect transcription of the text images. At each interaction step, the system uses the information obtained from the text image and a previously validated part (prefix) of its transcription to propose an improved transcription. Then, the user finds and corrects the next system error, thereby providing a longer prefix which the system uses to suggest a new, hopefully better continuation.

Speech dictation of the handwritten text can be used as an additional or an alternative information source in the CATTI process. Taking into account both the handwritten text image and the speech signal, the system can, hopefully, propose a better transcription hypothesis in each interaction step. Using this approach, many user corrections are avoided.

In Subsection 4.2 we formalise the multimodal CATTI framework where both, text and speech sources, help each other to improve the system accuracy.

This is the author's version of an article that has been published in the Computational Intelligence journal.
The final version is available at http://dx.doi.org/10.1111/coin.12169

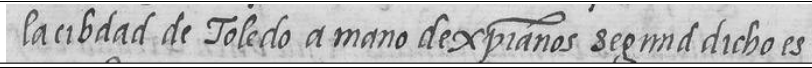
Text Image		
ITER-0	p	
ITER-1	$\widehat{s}$	la abadia de Toledo a mano de xpiãnos segun el dicho es
	m	⇑
	p	la
ITER-2	$\widehat{s}$	cibdad de Toledo a mano de xpiãnos segun el dicho es
	m	⇑
	p	la cibdad de Toledo a mano de xpiãnos
	$\widehat{s}$	sigue el dicho es
	v	**segund**
	p	la cibdad de Toledo a mano de xpiãnos segund
FINAL	$\widehat{s}$	dicho es
	v	#
	$p \equiv t$	la cibdad de Toledo a mano de xpiãnos **segund** dicho es

FIGURE 3. Example of CATTI operation using mouse actions.

4.2. CATTI Formal Framework

As previously explained, 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 more information the system has about what is written in the handwritten text line image, the better the proposed hypotheses are, and therefore, fewer user interactions are needed to obtain the perfect transcript. In this work, in addition to the handwritten text line image, we study how the CATTI system can take advantage of the speech dictation of the text that the images contain.

The process starts when the system proposes a full transcription $\widehat{s}$ of the handwritten text line image taking into account also the speech dictation, e.g., by using the multimodal combination presented in Subsection 3.2. 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 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 $\widehat{s}$) in which the first word is different from the first wrong word of the previous suffix. This way, many explicit user corrections are avoided (Romero et al., 2008). If the new suffix $\widehat{s}$ 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 enter a word v to correct the erroneous one. This action produces a new prefix p (the previously validated prefix followed by the new word). 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.

In Figure 3 we can see an example of the CATTI process. Starting with an initial recognised hypothesis $\widehat{s}$ from the input signal (a text line image, a speech utterance, or both), the user validates its longest well-recognised prefix p, making a MA m, and the system emits a new recognised hypothesis $\widehat{s}$. As the new hypothesis corrects the erroneous word, a new cycle starts. Now, the user validates the new longest error-free prefix p by making another MA m. The system provides a new suffix $\widehat{s}$ 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 p. Taking into account the new prefix, the system suggests a new hypothesis $\widehat{s}$. 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. In this example, without interaction, a user has 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: the interaction 1 only needs a mouse action, but in the interaction 2 a single mouse action does not succeed and the correct word needs to be typed.

Formally, in the traditional CATTI framework (Romero et al., 2012), the system uses a given feature sequence, x_{htr}, representing a handwritten text line image and a user validated prefix p of the transcription. In this work, in addition to x_{htr}, we study how a sequence of feature vectors x_{asr} representing the speech dictation of the handwritten text line image affects the system performance. Therefore, the CATTI system should try to complete the validated prefix by searching for a most likely suffix $\widehat{s}$ taking into account both sequences of feature vectors:

$$\widehat{s} = \arg\max_{s} \mathrm{P}(s \mid x_{htr}, x_{asr}, p) \tag{9}$$

Making the naive assumption that x_{htr} does not depend on x_{asr}, and applying the Bayes' rule, we can rewrite the previous equation as:

$$\widehat{s} = \arg\max_{s} \mathrm{P}(x_{htr} \mid p, s) \cdot \mathrm{P}(x_{asr} \mid p, s) \cdot \mathrm{P}(s \mid p) \tag{10}$$

where the concatenation of p and s is w in Equation (1). As in conventional HTR and ASR, $\mathrm{P}(x_{htr} \mid p, s)$ and $\mathrm{P}(x_{asr} \mid p, s)$ can be approximated by morphological models (optical or acoustical, for HTR and ASR respectively) and $\mathrm{P}(s \mid p)$ by a language model conditioned by p. Therefore, the search must be performed over all possible suffixes of p (Romero et al., 2012).

This suffix search can be efficiently carried out in an insignificant (and not appreciable by the user) amount of time by using Word Graphs (WG) (Romero et al., 2012) or Confusion Networks (CN) (Granell et al., 2016).

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 (10). This process is repeated until a complete and correct transcription of the input text line image is obtained.

Regarding the time spent by the users to transcribe by using CATTI, 53% of *probability of improvement* (POI) (Bisani and Ney, 2004) with respect to the manual *post-edition* approach was obtained in the analysis of time of a test with real users (Romero et al., 2012).

In the following sections, the impact of crowdsourcing multimodal inputs (obtained by using the framework described in Section 3) on CATTI is analysed.

5. EXPERIMENTAL CONDITIONS

5.1. Data Sets

The *Rodrigo* corpus (Serrano et al., 2010) was the data set employed in the experiments. It was obtained from the digitisation 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 4), giving a total number of 20,356 lines. The vocabulary size is of about 11,000 words.

This corpus presents several difficulties, such as the following examples that are present in the first 5 lines of the page 515 (Figure 4):

- Text images containing abbreviations (e.g., *nrõ* in the second line) that must be pronounced as the whole word (*nuestro* ['nwes tro]).

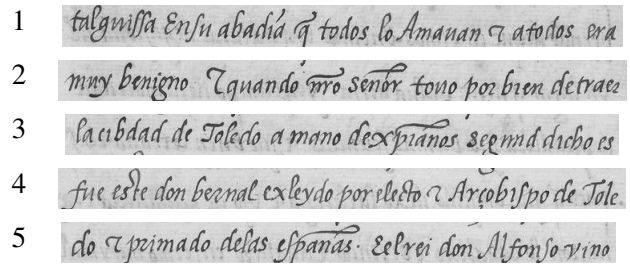


FIGURE 4. The 5 first lines of the page 515 of *Rodrigo*.

- 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).

The used corpus partition was the same that the one employed in previous works (Granell and Martínez-Hinarejos, 2016a,b). For training the optical and language models, a standard partition with a total number of 5000 lines (about 205 pages) was used. Test data for Handwritten Text Recognition (HTR) was composed of two pages that were not included in the training part (pages 515 and 579) and that were representative of the average error of the standard test set (of about 5000 lines). These two pages contain 50 lines and 514 words.

For the training of the Automatic Speech Recognition (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 a set of 4800 phonetically balanced utterances; specifically, 200 utterances read by four speakers and 25 utterances read by 160 speakers. The training data has a total length of about 4 hours. For the multimodal crowdsourcing test we obtained the collaboration of 7 different native Spanish speakers who read the 50 handwritten test lines (those of pages 515 and 579), giving a total set of 350 utterances (about 15 minutes of speech signal) acquired at 16 KHz and 16 bits. The seven collaborators (one woman and six men) were between 25 and 55 years old, they had higher education, and they were familiar with recognition of historical manuscripts. However, they had no special knowledge regarding old Spanish pronunciation.

5.2. Features

5.2.1. *HTR features.* In this work, the handwritten text features are computed in several steps from text line images following the approach presented in previous works (Pastor et al., 2004; Toselli et al., 2004). In this feature extraction process, in a first step, the background is removed and the noise is reduced by subtracting from the text line image the result of applying a median filter of size 3×3 to it (Kavallieratou and Stamatatos, 2006). After that, the foreground/background contrast is increased by applying a grey-level normalisation. Next, slant correction is performed by using the maximum variance method and a threshold of 92% (Pastor et al., 2004). Then, a size normalisation is performed and the text line

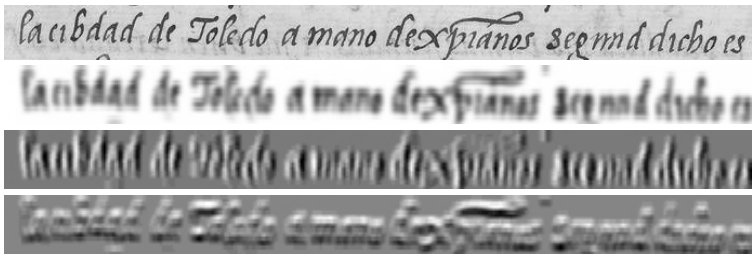


FIGURE 5. Text line sample (top) and the sequence to obtain the feature vectors, normalised grey level, horizontal derivative and vertical derivative.

image is scaled to a height of 40 pixels. Finally, the text line image is divided into a grid of squared cells, with a vertical resolution of 20 cells and 2 pixels of overlapping between 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 by a sliding window from left to right and a 60-dimensional feature vector is constructed for each frame by stacking the three features computed in its constituent cells (Toselli et al., 2004). Figure 5 presents an example of an extracted text line and the corresponding feature vector sequence obtained from it.

5.2.2. *ASR features.* Mel-Frequency Cepstral Coefficients is the most commonly used feature extraction method from audio signals in ASR (Rabiner and Juang, 1993). In this work, MFCC are extracted from the audio files following the standard ETSI-ES-201-108 for speech processing and feature extraction of the European Telecommunications Standards Institute (ETSI, 2003). In this standard, the Fourier transform is calculated every 10 ms over a window of 25 ms of a pre-emphasised signal. Next, 23 equidistant Mel scale triangular filters are applied and the filters outputs are logarithmised. Finally, to obtain the MFCC, a discrete cosine transformation is applied. We used the first 12 MFCC and log frame energy with first and second order derivatives as ASR features, which provides final feature vectors of 39 components.

5.3. Models

Optical and acoustical models were trained by using HTK (Young et al., 2006). On the one hand, symbols on the optical model are modelled by a continuous density gaussian mixture left-to-right of 106 Hidden Markov Models (HMM) with 4 states and 32 gaussians per state; on the other hand, phonemes on the acoustical model are modelled as a left-to-right gaussian mixture of 25 HMM (23 monophones, short silence, and long silence) with 3 states and 64 gaussians per state.

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 original Language Model (LM) was 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 set (about 205 pages).

5.4. Evaluation Metrics

The quality of the transcription before using the interactive system is measured by the well known Word Error Rate (WER), which is a good estimation of the user *post-edition* effort. It is defined as the minimum number of words to be substituted, deleted or inserted to convert the hypothesis into the reference, divided by the total number of reference words.

In addition, the oracle WER represents the best WER that can be obtained from the output lattices.

On the other hand, the interactivity performance is given by the Word Stroke Ratio (WSR), which can be also computed using the reference transcription. After each interactive hypothesis, the longest common prefix between the hypothesis and the reference is obtained and the first mismatching word from the hypothesis is replaced by the corresponding reference word, and a new suffix is proposed by the system. 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 the correct transcriptions using the interactive system, divided by the total number of reference words. This definition makes WER and WSR comparable. The relative difference between them gives us the effort reduction (EFR), which is an estimation of the reduction of the transcription effort that can be achieved by using the interactive system instead of the *post-editing* system. For WER and WSR, confidence intervals of 95% were calculated by using the bootstrapping method with 10,000 repetitions (Bisani and Ney, 2004).

The Collaboration Effort (CE) represents the number of speech utterances used in the crowdsourcing platform to obtain a determined output (Granell and Martínez-Hinarejos, 2017). It is measured as the number of lines (batch size B) that the system asks the collaborators to read times the actual number of collaborators involved in the obtainment of a determined output.

Finally, the size of the lattices is measured by the lattice density (LD), which, in many cases, is a good indicator of the complexity of lattices and of the amount of required computation (Ström, 1997). It is defined as the total number of edges of the lattice divided by the number of reference words (Ortmanns et al., 1997).

5.5. Experimental Setup

Both 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 lattice transformation from Word Graphs to Confusion Networks were done by using the SRILM toolkit (Stolcke, 2002).

The test set presents 6.2% of out-of-vocabulary words and a perplexity of 298.4 with respect to the original LM.

In the experiments, in order to optimise the experimental results, the values of the main decoding parameters (beam, word insertion penalty, ...) were tuned.

Regarding the crowdsourcing framework adjustment, in a previous work (Granell and Martínez-Hinarejos, 2016a) we observed as in this system the speaker order and the reliability verification did not show a significant impact on the results. Besides, the highest reliability for this test set is obtained when the multimodal combination is a bit balanced to the speech output ($\alpha = 0.6$), and the LM interpolation to the original LM ($\lambda = 0.4$). Therefore, the configuration with best results was used in the next experiments, i.e. combination factor $\alpha = 0.6$, LM interpolation $\lambda = 0.4$, and speech decoding reliability threshold $R \geqslant 0.4$. These experiments were done on the speech utterances of the 6 collaborators not employed in the framework tuning.

Finally, the CATTI configuration was that employed in a previous work (Romero et al., 2009). In this work it was observed that setting a limit around 3 for the number of mouse actions allows us to obtain a significant reduction of human effort with a fairly low number of mouse actions per word. Therefore, in our experiments the number of mouse actions was limited to 3.

TABLE 1. Baseline results. EFR with respect to HTR WER.

Modality	*Post-edition*		CATTI	
	WER	Oracle WER	WSR	EFR
HTR	39.3% ± 4.1	28.0%	36.2% ± 3.6	7.9%
ASR	62.9% ± 2.2	29.5%	47.2% ± 2.3	−20.1%

TABLE 2. *Post-edition* crowdsourcing results with and without Collaboration Effort (CE) optimisation.

Collaborators	With CE optimisation		Without CE optimisation	
	WER	Oracle WER	WER	Oracle WER
1	37.4% ± 4.1	25.1%	36.8% ± 4.3	24.1%
2	34.2% ± 3.9	24.9%	32.3% ± 4.0	23.7%
3	31.5% ± 4.1	23.2%	29.2% ± 3.8	22.0%
4	30.5% ± 4.0	23.2%	27.6% ± 3.5	20.4%
5	30.5% ± 4.2	24.1%	26.7% ± 3.4	20.4%
6	30.0% ± 4.1	23.2%	26.1% ± 3.3	22.0%

6. EXPERIMENTAL RESULTS

The transformation of data in the proposed multimodal crowdsourcing framework is studied. Initially, the baseline values for both modalities were obtained for the *post-edition* and the interactive transcription approaches. Then, the effects of the optimisation of the Collaboration Effort (CE) in the crowdsourcing framework were tested on a classical *post-edition* transcription approach. Next, the data fusion and transformation performed in the crowdsourcing framework was examined. Finally, the effects of the integration of this crowdsourcing framework with a multimodal interactive transcription system were studied.

6.1. Baseline Experiments

The baseline values were obtained by using the original LM in the decoding processes of both modalities (text and speech) in the non-interactive approach (*post-edition*), and in the interactive approach (CATTI). Table 1 presents these baseline values. The *post-edition* results show the difficulty of the corpus. The HTR and ASR WER values are quite high due to the characteristics of the manuscript described in Subsection 5.1, and to the fact that the speech of the collaborators was not used in the training of the acoustical models. However, the oracle WER values for both modalities are comparable. On the other side, the obtained WSR value (36.2% ± 3.6) for the HTR modality in the interactive CATTI approach represents a relative effort reduction (EFR) of 7.9% over the baseline WER value (39.3% ± 4.1). Nevertheless, in the case of using only the speech dictation of the contents of the text lines to transcribe (ASR modality), no effort reduction can be observed.

6.2. *Post-edition* Crowdsourcing Experiments

The obtained results in the *post-edition* crowdsourcing experiments are shown in Table 2 and Figure 6. These results were presented in previous works (Granell and Martínez-Hinarejos, 2016a,b).

As can be observed, the overall *post-edition* best result is obtained when all people collaborate with full effort, i.e., giving the speech dictation of the whole set of text lines.

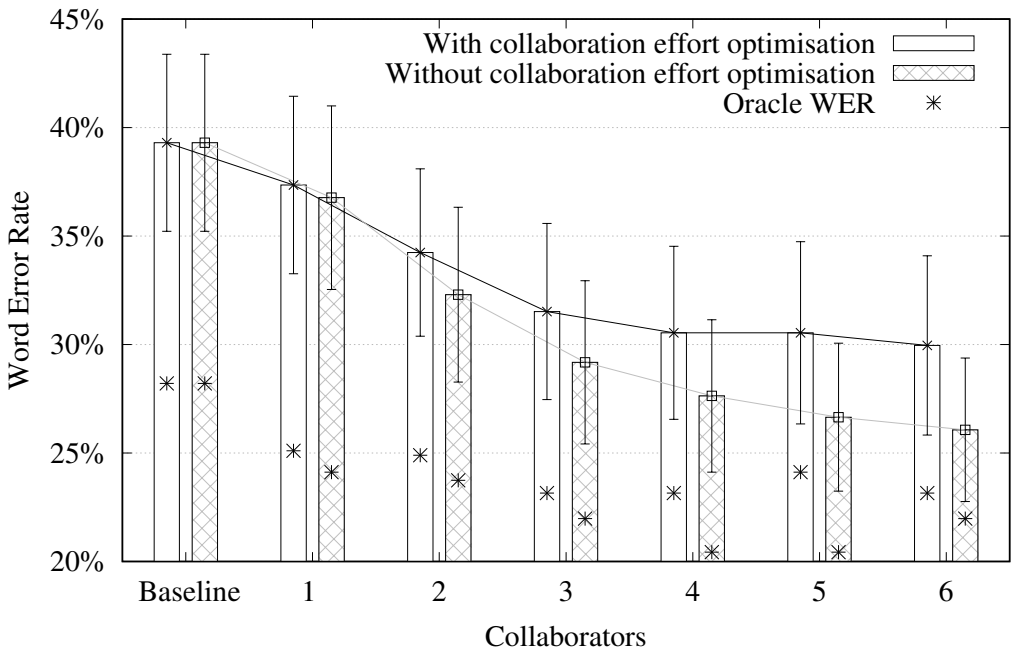


FIGURE 6. *Post-edition* experiments results.

In this case, the best result was 26.1% ± 3.3 of WER, which represents a statistically significant relative improvement of 33.6% over the baseline WER (39.3% ± 4.1). This result was obtained by using the 50 speech utterances of the 6 speakers, i.e. with a CE of 300 collaborations. It is interesting to note that the oracle WER value falls from 28.0% of the baseline to 22.0%. This improvement in the lattices hypotheses will permit to reduce the transcription effort in an interactive approach.

Similar statistically significant improvements can be achieved optimising the effort of the collaborators. In this test set, the optimal batch size B is 30 lines (Granell and Martínez-Hinarejos, 2016b). In Table 2 and Figure 6, it can be observed that optimising the collaborator effort produces 30.0% ± 4.1 of WER after the 6^{th} collaborator. Besides, with this effort optimisation, the obtained improvements are statistically significant after the collaboration of the 4^{th} collaborator. In this case, the system output presented a WER of 30.5% ± 4.0, which represents a relative improvement of 22.4% over the baseline with a CE of 120 collaborations (30 lines read by 4 collaborators). Furthermore, the difference between this value and the obtained without CE optimisation (50 lines read by the 6 speakers) is not statistically significant.

The optimisation of the collaborators effort allows us to reduce the number of global collaborations. In this case, it represents a reduction of 33.3% in the number of collaborators and 40.0% in the number of lines read by each collaborator. Moreover, given that with 120 collaborations this crowdsourcing system improved significantly the transcription of 50 lines, it can be expected that with 300 collaborations this system could improve significantly the transcription of a test set of 125 lines.

Regarding the oracle WER values, it can be observed that the obtained values are similar in both cases, 23.2% with CE optimisation and 22.0% without CE optimisation.

TABLE 3. Average output lattice density (LD). Initial LD value 76.39.

Collaborators	With CE optimisation	Without CE optimisation
	LD	LD
1	14.18	5.40
2	13.06	3.63
3	12.26	2.07
4	12.17	1.92
5	12.13	1.90
6	12.18	1.93

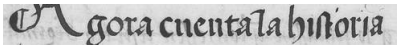


FIGURE 7. The 12th line of page 579 of the *Rodrigo* corpus.

6.3. Data Fusion and Lattices Transformation

This multimodal crowdsourcing framework allows us to merge into the output lattices the different sources of information for each text line, i.e. the HTR decoding of the text line image and the diverse collaborations with the ASR decoding of the dictation of the contents of the text line image. After each collaborative iteration, the output lattices are refined. This refinement includes the reduction of the number of incorrect hypotheses and the incorporation of correct words.

Given that this data fusion and lattice transformation produces improvements in all the graph, it not only permits to obtain a better 1-best hypothesis for a *post-edition* transcription approach, but it also permits to achieve a better performance for interactive and assistive approaches such as the CATTI system.

The baseline system output (obtained from the HTR decoding process) presented an average density of 76.39 edges per reference word. Table 3 presents the average lattice density at the crowdsourcing system output after each collaboration with and without Collaboration Effort (CE) optimisation. As it can be seen, the more collaborations, the better and smaller lattices are produced at the crowdsourcing system output. Moreover, the size of the lattices is considerably reduced after the first collaborator, and it seems to converge after the collaboration of several speakers.

As an example of the data fusion and lattice transformation, the HTR decoding of the 12th text line image of the page 579 of the *Rodrigo* corpus (Figure 7) produced the lattice presented in Figure 8. The reference for this text image is *AGORA CUENTA LA HISTORIA*. The obtained lattice presents a density of 22.25 edges per reference word, a WER value of 50%, and an oracle WER of 25% because the word *HISTORIA* is not present in this lattice, as it can be observed in Figure 9. This missing word is incorporated in the lattice after the first collaboration (Figure 10). Moreover, the number of incorrect hypotheses is reduced considerably, giving a density of 2 edges per reference word, a WER of 25% and an oracle WER of 0%. In the second collaboration (Figure 11) the erroneous word *LABRADORES* is deleted and the density is reduced to 1.75. Additional collaborations did not produce any change in this output lattice. Therefore, in this case the output lattice converged after the second collaboration.

6.4. Interactive Crowdsourcing Experiments

The interactive CATTI approach (Romero et al., 2012) applied to the proposed multimodal crowdsourcing framework permits to achieve, in most cases, an additional reduction

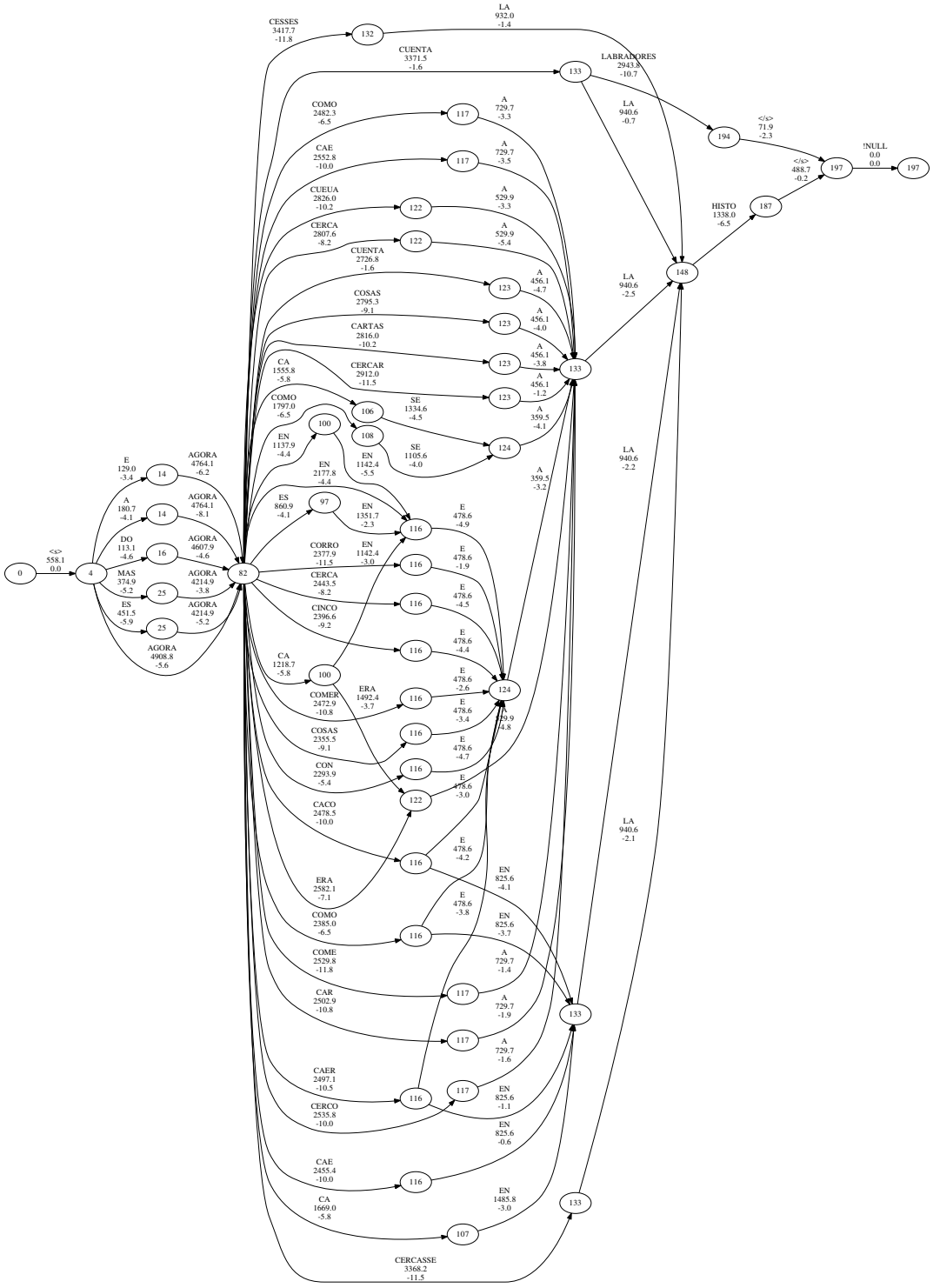


FIGURE 8. Word graph for the 12th line of page 579 at the system output before any collaboration (baseline).

This is the author's version of an article that has been published in the Computational Intelligence journal.
The final version is available at http://dx.doi.org/10.1111/coin.12169

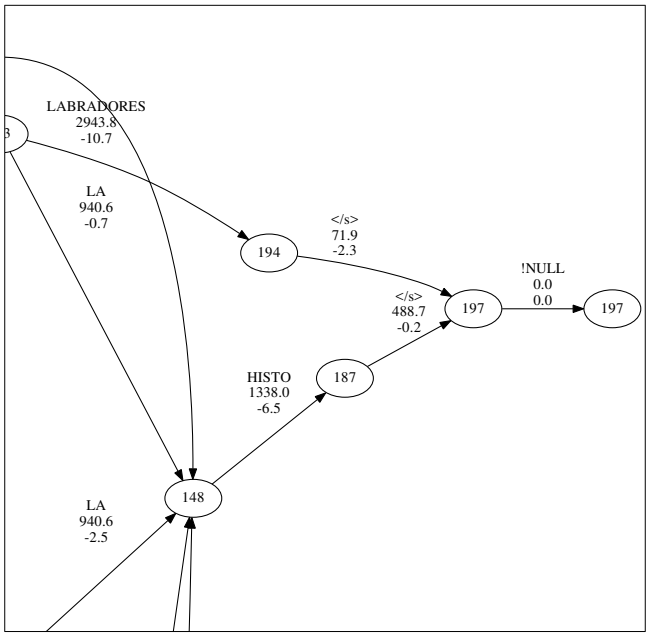


FIGURE 9. Enlargement of the final part of the word graph for the 12th line of page 579 at the system output before any collaboration. The reference word *HISTORIA* is not present in this lattice.

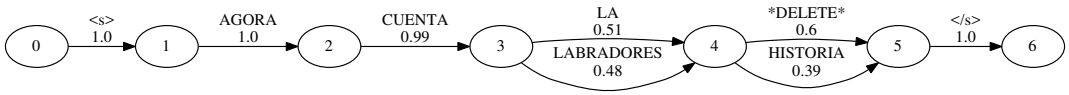


FIGURE 10. Confusion network for the 12th line of page 579 at the system output after the first collaboration.

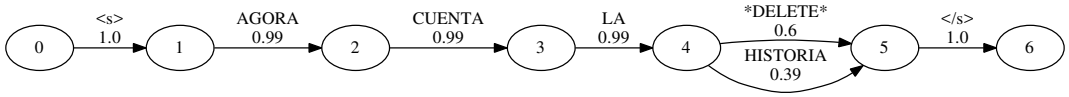


FIGURE 11. Confusion network for the 12th line of page 579 at the system output after the second collaboration.

TABLE 4. Interactive results with and without Collaboration Effort (CE) optimisation.

Collaborators	With CE optimisation		Without CE optimisation	
	WSR	EFR	WSR	EFR
1	34.2% ± 3.9	13.0%	34.0% ± 3.8	13.5%
2	32.3% ± 3.8	17.8%	32.1% ± 4.0	18.3%
3	30.4% ± 3.6	22.7%	27.2% ± 3.3	30.8%
4	30.0% ± 3.8	23.7%	27.6% ± 3.4	29.8%
5	30.4% ± 3.6	22.7%	26.5% ± 2.9	32.6%
6	29.2% ± 3.7	25.7%	25.7% ± 3.1	34.6%

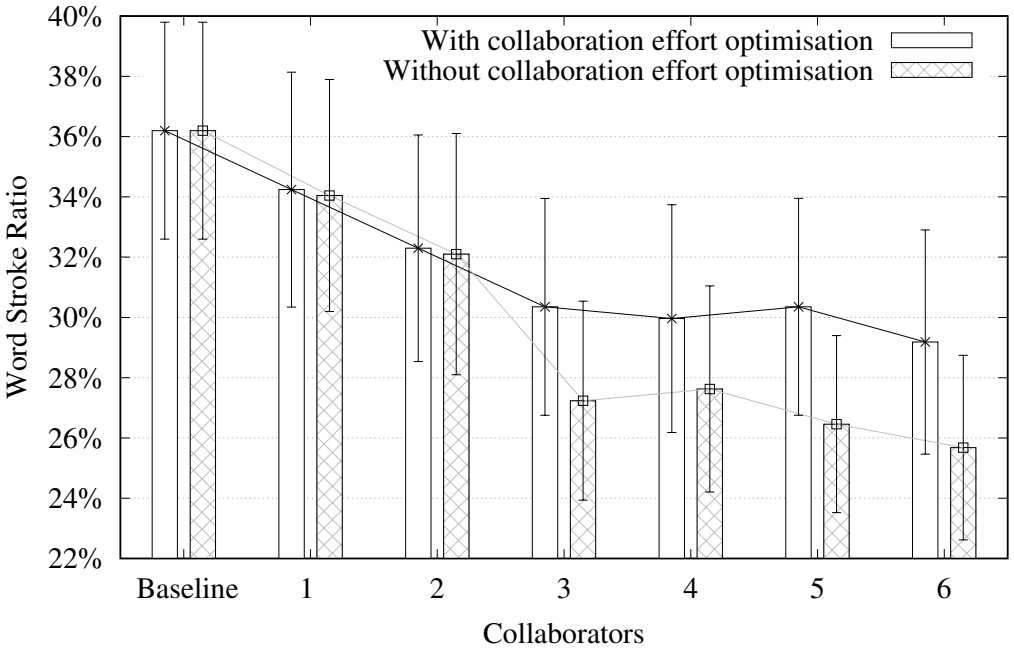


FIGURE 12. Interactive experiment results.

of the estimated human transcriber effort. The results obtained in the interactive experiments are presented in Table 4 and Figure 12. As it can be seen, the WSR obtained after the 4$^{\text{th}}$ collaborator with CE optimisation (120 collaborations) is comparable to the WSR obtained after the 6$^{\text{th}}$ collaborator without CE optimisation (300 collaborations). Both cases present an EFR higher than 20% when comparing the corresponding WSR with the *post-edition* baseline WER ($39.3\% \pm 4.1$).

Comparing each WSR of Table 4 with its corresponding *post-edition* WER presented in Table 2, the contribution of the CATTI approach to the transcriber effort reduction can be estimated. As it can be observed, the contribution of CATTI is limited by the density of the lattices provided by the crowdsourcing framework.

However, the improvements, in terms of WSR over the baseline WSR ($36.2\% \pm 3.6$), obtained by the integration of the multimodal crowdsourcing framework with the interactive system CATTI are not statistically significant with a batch size of $B = 30$ lines (value of B for the CE optimisation in the *post-edition* approach).

6.5. Interactivity and Collaboration Effort Optimisation

In this last experiment, the optimisation of the collaborator effort in the interactive approach was tested by analysing the influence of the number of collaborators and the number of lines to read in the obtained WSR.

Table 5 summarises the obtained results for the B ranges that present significant improvements with respect to baseline WSR results. As can be observed, the overall best result in terms of collaboration effort was obtained with $B = 45$. In this case, the interactive approach presented a statistically significant improvement ($27.2\% \pm 3.3$) after processing the speech of the third collaborator, i.e., with a CE of only 135 utterances. This WSR value represents a relative improvement of 24.9% over the HTR WSR baseline ($36.2\% \pm 3.6$), and

TABLE 5. Collaboration Effort (CE) Experiment Results Summary. In boldface, best CE result.

B	First significant improvement				Final output	
	Collaborators	CE	WSR	EFR	WSR	EFR
30	-	-	-	-	$29.2\% \pm 3.7$	25.7%
35	4	140	$29.0\% \pm 3.5$	26.2%	$29.2\% \pm 3.7$	25.7%
40	4	160	$27.8\% \pm 3.5$	29.3%	$27.0\% \pm 3.3$	31.3%
45	3	**135**	$27.2\% \pm 3.3$	30.8%	$26.3\% \pm 3.3$	33.1%
50	3	150	$27.2\% \pm 3.3$	30.8%	$25.7\% \pm 3.1$	34.6%

an estimated effort reduction for the paleographer revision of 30.8% over the HTR WER baseline ($39.3\% \pm 4.1$).

The contribution of CATTI to the transcriber effort reduction in the crowdsourcing framework requires a few additional contributions. Specifically, the optimal CE was 120 collaborations in the *post-edition* approach for obtaining a WER value of $30.5\% \pm 4.0$, whilst in the interactive approach, the optimal CE was 135 collaborations for obtaining a WSR of $27.2\% \pm 3.3$. This represents an increase of 12.5% of CE for an additional EFR of 10.8%. This means that with the whole collaboration effort (300 utterances), 104 lines would obtain transcription improvements (supposing a similar behaviour on other lines of the corpus), with a lower transcription effort.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have studied the integration of a multimodal crowdsourcing framework and an assistive tool for the transcription of historical handwritten documents. The experiments showed how this integration is possible given the data fusion and transformation performed by the multimodal crowdsourcing framework. The collaboration effort optimisation allows us to obtain similar results, but reducing the number of collaborators and the number of lines that they have to read. However, as it could be expected, the more collaborations, the better lattices are obtained from the crowdsourcing framework, and the better performance from the assistive transcription tool is obtained.

We propose for future studies the use of sentences in the handwritten text corpus instead of lines in order to make multimodality more natural for speakers, and the use of more robust modelling methods. An improvement that will be very useful for expanding the use of this crowdsourcing framework is the weighting of contributions according to the level of expertise on the task of the collaborators. Moreover, another line to be explored is the use of this crowdsourcing framework to obtain alternative transcriptions (different from diplomatic transcriptions), such as modernised transcriptions. Finally, this framework is open to be tested with other datasets.

ACKNOWLEDGMENTS

Work partially supported by projects READ - 674943 (European Union's H2020), Smart-Ways - RTC-2014-1466-4 (MINECO), and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO / FEDER).

REFERENCES

ALABAU, V., V. ROMERO, A.-L. LAGARDA, and C.-D. MARTÍNEZ-HINAREJOS. 2011. A Multimodal

Approach to Dictation of Handwritten Historical Documents. *In* Proc. 12th Interspeech, pp. 2245–2248.

BARRACHINA, S., O. BENDER, F. CASACUBERTA, J. CIVERA, E. CUBEL, S. KHADIVI, A. LAGARDA, H. NEY, J. TOMÁS, E. VIDAL, and OTHERS. 2009. Statistical approaches to computer-assisted translation. Computational Linguistics, **35**(1):3–28.

BISANI, M., and H. NEY. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. *In* Proc. ICASSP, Volume 1, pp. 409–412.

CAINES, A., C. BENTZ, C. GRAHAM, T. POLZEHL, and P. BUTTERY. 2016. Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the crowdis corpora. *In* Proc. of LREC 2016. ISBN 978-2-9517408-9-1.

DOAN, A., R. RAMAKRISHNAN, and A. Y. HALEVY. 2011. Crowdsourcing systems on the world-wide web. Commun. ACM, **54**(4):86–96. ISSN 0001-0782. . http://doi.acm.org/10.1145/1924421.1924442.

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. ISSN 0895-6111. .

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).

GRANELL, E., and C. D. MARTÍNEZ-HINAREJOS. 2015. Combining Handwriting and Speech Recognition for Transcribing Historical Handwritten Documents. *In* Proc. 13th ICDAR, pp. 126–130.

GRANELL, E., and C. D. MARTÍNEZ-HINAREJOS. 2016a. A Multimodal Crowdsourcing Framework for Transcribing Historical Handwritten Documents. *In* Proc. of the 16th DocEng, pp. 157–163.

GRANELL, E., and C. D. MARTÍNEZ-HINAREJOS. 2016b. Collaborator Effort Optimisation in Multimodal Crowdsourcing for Transcribing Historical Manuscripts. *In* Proc. of IberSPEECH'2016, pp. 234–244.

GRANELL, E., and C.-D. MARTÍNEZ-HINAREJOS. 2017. Multimodal Crowdsourcing for Transcribing Handwritten Documents. IEEE/ACM Transactions on Audio, Speech, and Language Processing, **25**(2):409–419.

GRANELL, E., V. ROMERO, and C. D. MARTINEZ-HINAREJOS. 2016. An Interactive Approach with *Off-line* and *On-line* Handwritten Text Recognition Combination for Transcribing Historical Documents. *In* Proc. DAS, pp. 269–274.

GRANELL, E., V. ROMERO, and C.-D. MARTÍNEZ-HINAREJOS. 2017. Using Speech and Handwriting in an Interactive Approach for Transcribing Historical Documents. *In* Handwriting: Recognition, Development and Analysis. Nova Science, pp. 277–295.

HAZEN, TIMOTHY J. 2006. Visual model structures and synchrony constraints for audio-visual speech recognition. IEEE Trans. Audio, Speech & Language Processing, **14**(3):1082–1089. . http://dx.doi.org/10.1109/TSA.2005.857572.

HINTON, G., L. DENG, D. YU, G.E. DAHL, A. MOHAMED, N. JAITLY, A. SENIOR, V. VANHOUCKE, P. NGUYEN, T.N. SAINATH, and B. KINGSBURY. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, **29**(6):82–97. ISSN 1053-5888. .

IMPEDOVO, S., L. OTTAVIANO, and S. OCCHINEGRO. 1991. Optical character recognition - a survey. International Journal of Pattern Recognition and Artificial Intelligence, **5**(01n02):1–24.

KAVALLIERATOU, ERGINA, and EFSTATHIOS STAMATATOS. 2006. Improving the quality of degraded document images. *In* Proc. DIAL'06, IEEE, pp. 340–349.

KNESER, R., and H. NEY. 1995. Improved backing-off for m-gram language modeling. *In* Proc. ICASSP, Volume 1, pp. 181–184.

LUJÁN-MARES, M., V. TAMARIT, V. ALABAU, C. D. MARTÍNEZ-HINAREJOS, M. PASTOR, A. SANCHIS, and A. H. TOSELLI. 2008. iATROS: A speech and handwritting recognition system. *In* V Jornadas en Tecnologías del Habla, pp. 75–78.

MACHOVER, C. 1995. The CAD/CAM Handbook. McGraw-Hill. ISBN 0-07-039375-3.

MALIT, R. F. 2009. Computer assisted driving of vehicles. http://www.google.tl/patents/US7513508. US Patent 7,513,508.

MIKI, M., N. KITAOKA, C. MIYAJIMA, T. NISHINO, and K. TAKEDA. 2014. Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech. EURASIP Journal on Audio, Speech, and Music Processing, **2014**(1):2. ISSN 1687-4722. . http://dx.doi.org/10.1186/1687-4722-2014-2.

MORENO, A., D. POCH, A. BONAFONTE, E. LLEIDA, J. LLISTERRI, J. B. MARIÑO, and C. NADEU. 1993.

Albayzin speech database: design of the phonetic corpus. *In* Proc. EuroSpeech, pp. 175–178.

NATIONAL LIBRARY OF THE NETHERLANDS. 2010. IMPACT: Improving Access to Text. http://www.impact-project.eu/. Last access: November 2017.

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

PARENT, G., and M. ESKENAZI. 2011. Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. *In* Proc. of Interspeech, pp. 3037–3040.

PASTOR, M., A. H. TOSELLI, and E. VIDAL. 2004. Projection profile based algorithm for slant removal. *In* Proc. of ICIAR'04, Volume 3212 of *Lecture Notes in Computer Science*, pp. 183–190.

PLAMONDON, R., and S. N. SRIHARI. 2000. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**(1):63–84. ISSN 0162-8828. .

RABINER, L., and B.-H. JUANG. 1993. Fundamentals of Speech Recognition. Prentice Hall. ISBN 0130151572.

READ PROJECT. 2016. READ: Recognition and Enrichment of Archival Documents. https://read.transkribus.eu/. Last access: November 2017.

REVUELTA-MARTÍNEZ, A., L. RODRÍGUEZ, and I. GARCÍA-VAREA. 2012. A computer assisted speech transcription system. *In* Proc. of the 13th EACL, pp. 41–45.

ROMERO, V., A. H. TOSELLI, J. CIVERA, and E. VIDAL. 2008. Improvements in the Computer Assisted Transcription System of Handwritten Text Images. *In* Proc. 8th PRIS, pp. 103–112.

ROMERO, V., A. H. TOSELLI, and E. VIDAL. 2009. Using Mouse Feedback in Computer Assisted Transcription of Handwritten Text Images. *In* Proc. 10th ICDAR, pp. 96–100.

ROMERO, V., A. H. TOSELLI, and E. VIDAL. 2012. Multimodal Interactive Handwritten Text Transcription, Volume 80 of *in Machine Perception and Artificial Intelligence*. World Scientific Publishing. http://www.worldscientific.com/worldscibooks/10.1142/8394.

RUEBER, B. 1997. Obtaining confidence measures from sentence probabilities. *In* Proc. Eurospeech, pp. 739–742.

SERRANO, N., F. CASTRO, and A. JUAN. 2010. The RODRIGO Database. *In* Proc. 7th LREC, pp. 2709–2712. http://aclweb.org/anthology/L10-1330.

SILVESTRE-CERDÀ, J. A., A. PÉREZ, M. JIMÉNEZ, C. TURRO, A. JUAN, and J. CIVERA. 2013. A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories. *In* Proc. of SMC '13, pp. 3994–3999. .

SINGH, A., A. SANGWAN, and J. H. L. HANSEN. 2012. Improved parcel sorting by combining automatic speech and character recognition. *In* Proc. of ESPA 2012. ISBN 9781467308984. pp. 52–55. .

STOLCKE, A. 2002. SRILM-an extensible language modeling toolkit. *In* Proc. 3rd Interspeech, pp. 901–904.

STRÖM, N. 1997. Automatic continuous speech recognition with rapid speaker adaptation for human/machine interaction. Ph. D. thesis, Kungliga Tekniska Högskolan.

TAMURA, S., K. IWANO, and S. FURUI. 2005. Toward robust multimodal speech recognition. *In* Proc. of LKR 2005, pp. 163–166.

TOSELLI, A. H., A. JUAN, D. KEYSERS, J. GONZÁLEZ, I. SALVADOR, H. NEY, E. VIDAL, and F. CASACUBERTA. 2004. Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int. Journal of Pattern Recognition and Artificial Intelligence, **18**(4):519–539.

TRANSCRIPTORIUM PROJECT. 2013. tranScriptorium. http://transcriptorium.eu/. Last access: November 2017.

WESSEL, F., R. SCHLÜTER, K. MACHEREY, and H. NEY. 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. Speech and Audio Processing, **9**(3):288–298.

YOUNG, S., G. EVERMANN, M. GALES, T. HAIN, D. KERSHAW, X. LIU, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, and OTHERS. 2006. The HTK book. Cambridge University Engineering Department.