Multimodal Crowdsourcing for Transcribing Handwritten Documents

Emilio Granell, and Carlos-D. Martínez-Hinarejos

***Abstract*—Transcription of handwritten documents is an important research topic for multiple applications, such as document classification or information extraction. In the case of historical documents, their transcription allows to preserve cultural heritage because of the amount of historical data contained in those documents. The transcription process can employ state-of-the-art handwritten text recognition systems in order to obtain an initial transcription. This transcription is usually not good enough for the quality standards, but that may speed up the final transcription of the expert. In this framework, the use of collaborative transcription applications (crowdsourcing) has risen in the recent years, but these platforms are mainly limited by the use of non-mobile devices. Thus, the recruiting initiatives get reduced to a smaller set of potential volunteers. In this work, an alternative that allows the use of mobile devices is presented. The proposal consists of using speech dictation of handwritten text lines. Then, by using multimodal combination of speech and handwritten text images, a draft transcription can be obtained, presenting more quality than that obtained by only using handwritten text recognition. The speech dictation platform is implemented as a mobile device application, which allows for a wider range of population for recruiting volunteers. A real acquisition on the contents of a Spanish historical handwritten book was obtained with the platform. This data was used to perform experiments on the behaviour of the proposed framework. 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.**

***Index Terms*—Handwritten text transcription, speech recognition, multimodal combination, crowdsourcing.**

I. Introduction

TRANSCRIPTION of handwritten documents is a fundamental task for different applications that may use the contents of those documents. This is the case of information retrieval, document classification, or summarisation [15]. Transcription of the document provides an easier digital access to their contents, making possible the search by linguistic contents (keywords, expressions, categories, ...). However, when only the digitalisation of the document is provided, query by image is the most usual option, which is a less flexible and powerful option.

In the case of historical documents, transcription is even more important because of the singularity of the documents. For example, most of them are not physically accessible to avoid degradation. Moreover, their contents cover important facts on the history and culture of the context they were written in. Therefore, preserving their contents is crucial for cultural and historical reasons. The interest in this preservation by using transcription led to the development of international projects such as tranScriptorium[1] or READ[2].

Quality transcriptions are usually done by experts; in the case of historical texts, because of their special features (scripting, image quality, vocabulary, ancient language, etc.), the contribution of the expert transcribers, called paleographers, is mandatory. In the last decade, their task has benefited from the contribution of the handwritten text recognition (HTR) technology [19]. HTR provided paleographers with an initial draft transcription that can be amended to obtain the quality transcription. In general, this process is faster than producing the quality transcription from scratch and increases the productivity of the transcribers.

To a similar extent, the appearing of crowdsourcing platforms [5] has had a strong impact on the paleographers task. In these platforms, many volunteers provide a transcription of the text image at a very small (or even null) cost; the inherent difficulties of historical texts make necessary the posterior revision of the paleographer, but the workload is considerably lower than that of scratch transcription. There are several generic crowdsourcing platforms available, such as Mechanical Turk[3] or CrowdFlower[4], but for handwritten text transcription (and in particular for historical text) several platforms have been developed in the last years (such as AnnoTate[5], Transcribe Bentham[6], or Transkribus[7]).

These crowdsourcing platforms make the 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 that 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 automatic speech recognition (ASR) techniques [20], presents an ambiguity

Emilio Granell, and Carlos-D. Martínez-Hinarejos are with the Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain (e-mail: egranell@dsic.upv.es; cmartine@dsic.upv.es).
Corresponding author: Emilio Granell.

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

not present in typed input. Even the state-of-the-art techniques [11], although more accurate than a few years ago, produce a considerable amount of errors in the recognition process, which makes necessary to obtain a balance between the amount of collaborations and the quality they provide.

In any case, the need for the final supervision by a paleographer enables the possibility that, although not perfect, voice inputs combined with HTR provide an initial transcription more accurate than that given only by HTR. Thus, the employment of speech collaborators would reduce the final transcription effort.

This work 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 is based on language model interpolation [2] and Confusion Network combination [28] 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. In order to evaluate how to optimise the collaborators effort, a large acquisition was made with the client application. On this data, experiments on selecting the lines that would balance the acquisition (collaborators) and the transcription (paleographer) effort were performed.

The paper is structured as follows: Section II describes related work on multimodal recognition and crowdsourcing, Section III presents the details on the proposed crowdsourcing framework, Section IV describes the data acquisition and experimental conditions, Section V shows the results, Section VI summarises the conclusions and future work lines.

II. Related work

The idea of multimodal recognition is not new, and several previous works have cope with different approximations to multimodality. One of the most usual multimodal tasks that involve speech recognition is the audio-visual speech recognition approximation [25], [10]. In this case, speech signal and lips and mouth movements (recorded in video images) are the original sources. The two signals are usually synchronous, which makes easy to configure them in the same data stream during the recognition process; sometimes synchronicity is not perfect and some signal fitting is required (as the technique presented in [10]), but in general, for this type of multimodal recognition, asynchronicity is not a problem.

More recent works presented the case of multimodality with speech and gestures. In this case, the two signals are usually not synchronous, which makes difficult a joint process of the two sources. In [16], the asynchronicity is solved by calculating a distribution of time differences between the start of the speech and the gesture; the two modalities are separately recognised, obtaining N-best lists; finally, applying a dynamic programming algorithm on the two N-best lists along with the time differences distribution, a final hypothesis is obtained. In [13], speech and gesture keyboard movements are used to input typical e-mail sentences or web searches; each signal is recognised separately, obtaining not only the best hypothesis but a set of alternative hypothesis in the form of Confusion Networks, which are combined to obtain a better final result.

The use of Confusion Network combination is also usual for the integration of several recognisers of the same modality, in spite of their potential synchronicity, such as the systems described in [27].

With respect to the combination of speech and handwritten text (which are usually asynchronous signals), the work presented in [23] makes a combination of Optical Character Recognition (OCR) and speech recognition for enhanced ZIP codes recognition. The proposed approach performs OCR, calculates its confidence, and based on it takes the speech recognition result to make a combination and provide an alternative hypothesis. A similar approximation is that presented in [1], where speech or handwritten recognition results, in the form of word-graphs, are used to enhance the language model for recognising with the other modality. An alternative that does not use language model enhancement is proposed in [7] where Confusion Network combination (similar to that of [13]) is used for the combination of these two modalities.

Crowdsourcing approaches to the acquisition of speech data have become really popular in the last decade. In [18], 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 [4] 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. In [9], a first step on the incorporation of multimodality in crowdsourcing is shown, by presenting a framework where the acquired modality (speech) is not the one to be transcribed (handwritten text).

III. Crowdsourcing Framework

The HTR and ASR problems admit a similar formulation that makes their multimodal integration feasible. The unimodal formulation is based on taking a feature vector sequence $x = (x_1, x_2, \ldots, x_{|x|})$ (which can be derived from a handwritten text image or a speech signal) and obtaining the most likely word sequence $\hat{w}$ according to x. That is:

$$\begin{aligned} \hat{w} &= \arg\max_{w \in W} \Pr(w \mid x) \\ &= \arg\max_{w \in W} \frac{\Pr(x \mid w)\Pr(w)}{\Pr(x)} \\ &= \arg\max_{w \in W} \Pr(x \mid w)\Pr(w) \end{aligned} \tag{1}$$

where W denotes the set of all permissible sentences, $\Pr(x)$ is the probability of observing x, $\Pr(w)$, with $w = (w_1, w_2, \ldots, w_{|w|})$, is the probability of w, and $\Pr(x \mid w)$ is the probability of observing x by assuming that w is the underlying word sequence for x. $\Pr(w)$ is usually approximated by the language model (LM), whereas $\Pr(x \mid w)$ is modelled by the optical (HTR) or acoustical (ASR) models.

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 $\hat{w}$ with the lowest number of errors. This transcription will

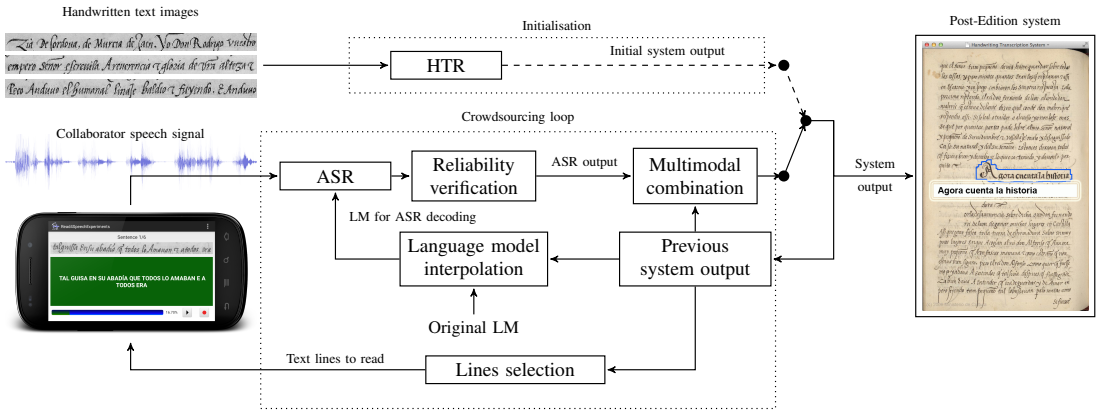


Fig. 1. Multimodal crowdsourcing transcription framework.

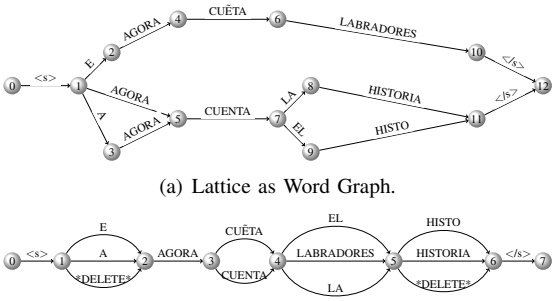
be provided to a paleographer to obtain the final quality transcription with the lowest effort.

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 [1], and combining the decoding outputs of the two modalities to obtain a final output with less errors [7].

Apart from that, the framework includes a speech reliability verification module that may exclude utterances that are considered of not enough 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 collaborators effort, a line selection module is incorporated to select lines that have a low reliability. The aim is to obtain more samples for those lines than for other lines. It is supposed that this strategy would allow to improve the global results on the whole set of lines to be transcribed.

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.
2) When a collaborator offers to help, the crowdsourcing loop starts:
   a) In the language model 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.
   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.



(a) Lattice as Word Graph.

(b) Lattice as Confusion Network.

Fig. 2. Examples of different representations of lattices, ref: *<s>AGORA CUENTA LA HISTORIA </s>*.

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

A. 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). Figure 2 shows an example of WG and its corresponding CN.

A Word Graph (WG) is a directed, acyclic and weighted graph with an initial node q_I and a final node q_F. In ASR, the nodes correspond to discrete time points, whereas in HTR represent horizontal space. A link l is defined as any edge between a starting node $s(l)$ and an ending node $e(l)$; each link has associated a hypothesis word $w(l)$ and its likelihood $f(l)$.

The language interpolation module builds a statistical language model conditioned on a sample x as follows [1]:

1) The decoding lattices for x are formatted as WG.
2) The posterior probabilities for each WG node ($\Pr(q \mid x)$) and link ($\Pr(l \mid x)$) are computed by using the forward $\alpha(q)$ and backward $\beta(q)$ probabilities of the nodes [26].
3) The counts for a word sequence $w_{i-n+1}^{i} = (w_{i-n+1}, \ldots, w_i)$ are estimated as:

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

N-gram	$C^*(w_{i-n+1}^i \mid x)$
<s>AGORA	0.999797
AGORA	0.999797
AGORA CUENTA	0.999797
AGORA CUẼTA	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

Fig. 3. Some word sequence counts $C^*(w_{i-n+1}^i \mid x)$ estimated from the word graph in Figure 2(a).

N-gram	$\log \Pr^x(w)$	$\log \Pr(w)$	$\log \Pr_\lambda^x(w)$
...			
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 CUẼTA	–	-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
...			

Fig. 4. Example of language model interpolation from the counts in Figure 3 by using $\lambda = 0.5$ and a smoothing factor of 1^{-10}. The probabilities are in log domain.

where $N(w_{i-n+1}^i)$ are all the sequences of concatenated links that generate w_{i-n+1}^i. Figure 3 presents some of the word sequences (N-grams) and counts that could be obtained from a WG as the one presented in Figure 2(a).

4) The word posterior probabilities associated to the current input 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:

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

5) The new conditioned language model $\Pr^x(w)$ is linearly interpolated with the original language model $\Pr(w)$ by using a weight factor λ:

$$\Pr_\lambda^x(w) = \lambda \Pr^x(w) + (1-\lambda)\Pr(w) \tag{4}$$

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 4 presents an example in which a general language model is refined according to the N-grams presented in Figure 3 (by using $\lambda = 0.5$ and a smoothing factor of 1^{-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.

B. Multimodal Combination

The multimodal combination employs Confusion Networks (CN) to combine the ASR decoding output with the previous system output. A CN is a weighted directed graph, in which each hypothesis goes through all the nodes. The words and their probabilities are 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.

This framework employs the bimodal Confusion Network combination method defined in [7], [8]. Specifically, starting from the system and the speech decoding outputs in CN format, the following steps are taken:

1) Anchor subnetworks are searched in order to align the subnetworks of both Confusion Networks. The algorithm searches coincidences in unigrams, bigrams and skip-bigrams in both directions simultaneously; only those subnetworks where both searches coincide (according to a gram matching value of the words in the involved subnetworks) are taken as anchors.
The gram matching error E between the 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 are the words in SN_A and SN_B, respectively. CER and PER are the Levensthein distance between w_A and w_B (CER at character level, and PER at phoneme level according to the phonetic transcriptions of the words).
An example of the multimodal combination is presented in Figure 5. In this example, CN_A, CN_B, and CN_C are the CN for the previous system output, the ASR decoding output, and the resulting combination as the new system output, respectively. When searching for bigrams and unigrams on the most probable words 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 $SN_A^5 - SN_B^5$.

2) The new Confusion Network is composed on the basis of the Bayes theorem and assuming a strong independence between the two Confusion Networks; the composition applies the classical editing actions (combination, insertion, and deletion) on 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, using a weight factor α:

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

The smoothing of the word posterior probability $\Pr_s(w \mid SN)$ is based on Laplacian smoothing.

However, since we are working with probabilities, $\Pr_s(w|SN)$ is calculated according to Equation (7):

$$\Pr_s(w \mid SN) = \frac{\Pr(w \mid SN) + \Theta}{1 + n\Theta} \tag{7}$$

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 subnetwork (SN_C). In the example of Figure 5, SN_A^4 and SN_B^3 are combined using $\alpha = 0.5$ and $\Theta = 10^{-4}$. In the resulting SN (SN_C^4), the correct word (*LA*) becomes the most likely.

- **Insertion and deletion:** Both actions are the opposite, but they are implemented by the same process; the subnetwork to insert or to delete is combined (using the combination operation described above) with a subnetwork with an only *DELETE* arc with probability 1.0. In the example (see Figure 5), SN_B^4 and SN_A^1 are inserted and deleted, respectively.

In this case, the weight factor α balances the reliability in the multimodal combination between the ASR output and the previous system output.

Finally, a new CN is obtained as a result. As can be observed in Figure 5, in the resulting CN (CN_C) several errors have been corrected, and the correct sentence (*<s>AGORA CUENTA LA HISTORIA </s>*) has the highest probability.

C. Reliability Verification

The statistical formulation of the decoding, for both HTR and ASR problems, allows to take the posterior probability $\Pr(w \mid x)$ as a good confidence measure for the recognition reliability. However, recognition processes provide scores that are inadequate to obtain this reliability, since most recognition systems neglect the term $\Pr(x)$ (Equation (1)).

Nevertheless, when the recognition scores of a fairly large N-best list can be re-normalised to sum up to 1, the obtained posterior probability $\Pr(w \mid x)$ can be used as a good confidence measure, since it is a measure of the match between x and w [21], [26]. An example of this confidence measure calculation is presented for a small N-best list in Figure 6. In this example the $\Pr_n(w_1 \mid x)$ is highlighted in bold.

Therefore, the reliability verification module employs the re-normalised 1-best posterior probability $\Pr_n(w_1 \mid x)$:

$$\Pr_n(w_1 \mid x) = \frac{\max\limits_{w \in W} \Pr(w \mid x)}{\sum\limits_{w \in W} \Pr(w \mid x)} \tag{8}$$

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

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 $\Pr_n(w_1 \mid x)$ 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.

D. 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 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 posterior probability - Equation (8) and example in Figure 6 -, giving an estimation of the current confidence for each line to transcribe.
2) The lines are ranked according to their estimated confidence value.
3) The system selects the subset of the B 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 experiment a higher improvement with the speech dictation. The number of lines (batch size) B is important as well, since determines the effort of a collaborator for an acquisition session.

IV. Experimental Conditions

A. Data Sets

The *Rodrigo* corpus [22] was the data set employed in the experiments. It was 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 7), giving a total number of 20,356 lines, and a vocabulary of about 11,000 words. The time required for a single paleographer to manually transcribe this manuscript was estimated in approximately 35 minutes per page on average.

This corpus presents several difficulties, such as, text images containing abbreviations (e.g., *nrõ* in line 2 of Figure 7) that must be pronounced as the whole word (*nuestro* ['nwes tro]), words written in multiple forms (e.g., *xpiãnos* -in line 3 of Figure 7- and *christianos*, or numbers as *5* and *V*) but that are pronounced in the same way ([kris 'tja nos], ['θiŋ ko]), and hyphenated words (e.g., *Toledo* in lines 4 and 5 of Figure 7, 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).

For training the optical models, a standard partition with 5000 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) 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.

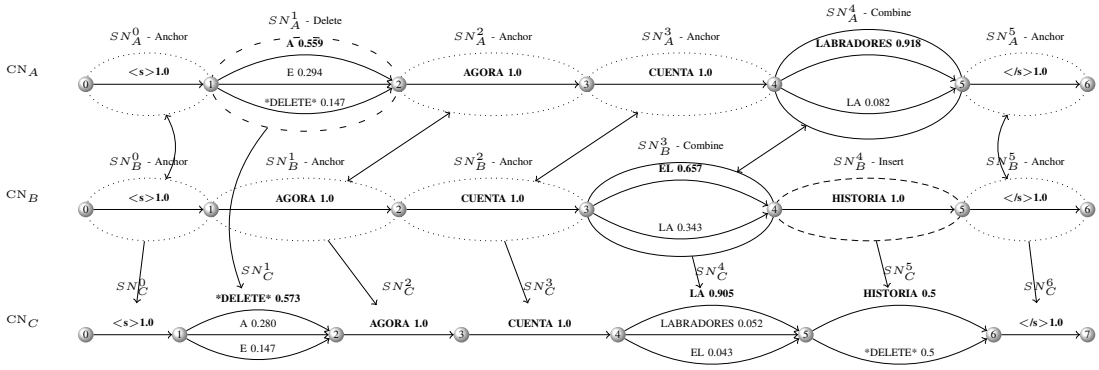


Fig. 5. Bimodal combination example, ref: *<s>AGORA CUENTA LA HISTORIA </s>*.

N-best	$\Pr(w \mid x)$	$\Pr_n(w \mid x)$
<s>Y PEQUEÑOS </s>	75.1%	**58.1%**
<s>Y NUEUE AÑOS </s>	25.8%	20.0%
<s>Y VEINTE AÑOS </s>	12.5%	9.6%
<s>Y SIETE AÑOS </s>	12.5%	9.6%
<s>Y DE DUEÑAS </s>	3.4%	2.6%

Fig. 6. Example of n-best list posterior probability re-normalisation as confidence measure.

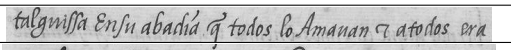
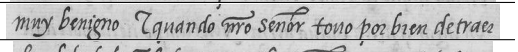
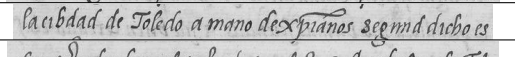
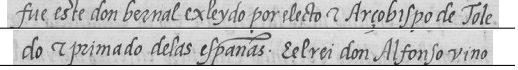
Line No.	Handwritten Text Line Image
1	
2	
3	
4	
5	

Fig. 7. The 5 first lines of the page 515 of *Rodrigo*.

For the training of the ASR acoustical models we used a partition of the Spanish phonetic corpus Albayzin [17]. 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, with a total length of about 4 hours.

B. Crowdsourcing Speech Acquisition

For the testing of the framework presented in this paper we used the application *Read4SpeechExperiments* (see Figure 8) 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, i.e. the collaborators read the text lines where and when they wanted, giving a total set of 1350 utterances (about 1 hour and 50 minutes) acquired at 16 KHz and 16 bits.

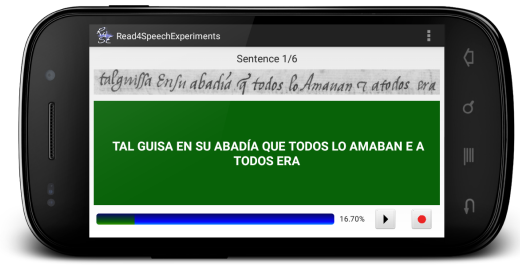


Fig. 8. Screenshot of the application *Read4SpeechExperiments*.

The set of collaborators had the following characteristics:

- They were between 25 and 60 years old.
- They were 9 women, and 18 men.
- 14 speakers were from our University, and the other came from people who knew our project by third parties.
- 24 speakers were from Spain, and the other 3 were foreigners, one of them with Arabic as mother tongue; even one of this collaborations came from abroad.

Read4SpeechExperiments is an Android free software application designed to facilitate the speech acquisition from mobile devices. The source code is available on GitLab[8], and it can be installed from the Google Play[9] and the F-Droid[10] platforms.

C. Features

1) HTR features: Handwritten text features are computed in several steps. First, a bright normalisation is performed. After that, a median filter of size 3×3 pixels is applied to the whole image. Next, slant correction is performed by using the maximum variance method and a threshold of 92%. Then, a size normalisation is performed and the final image is scaled to a height of 40 pixels. Finally, features are extracted by using the method described in [6], given vectors of 60 dimensions.

2) ASR features: Mel-Frequency Cepstral Coefficients (MFCC) are extracted from the audio files. 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, resulting in a 39 dimensional vector. Then, a Cepstral Mean Normalisation (CMN) is performed, by means of the subtraction of the cepstral mean from all the vectors. This normalisation

[8]https://gitlab.com/egranell/Read4SpeechExperiments
[9]https://play.google.com/store/apps/
[10]https://f-droid.org/repository/browse/

allows to compensate the long-term spectral effects caused by different microphones and audio channels in the final features. These features were obtained by using HTK [29].

D. Models

Optical and acoustical models were trained by using HTK [29]. On the one hand, symbols on the optical model are modelled by a continuous density gaussian mixture left-to-right of 106 HMM with 6 states and 32 gaussians per state, while 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 baseline language model was estimated as a 2-gram with Kneser-Ney back-off smoothing [12] directly from the transcriptions of the pages included on the HTR training set (about 205 pages). This model presents, with respect to the test set, a 6.2% of OOV words and a perplexity of 298.4.

E. Evaluation Metrics

The quality of the transcription is given 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. Moreover, confidence intervals of 95% were calculated by using the bootstrapping method with 10,000 repetitions [3].

The speech decoding reliability R is verified by using the re-normalised 1-best posterior probability -Equation (8)-, which is a good estimation of the decoding confidence.

Finally, 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 of collaborators involved in the obtainment of a determined output.

F. Experimental Setup

Both the HTR and the ASR systems were implemented by using the iATROS recogniser [14]. 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 [24].

V. Experimental Results

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 1350 speech utterances recorded from 27 different collaborators described in Subsections IV-A and IV-B. We started obtaining the baseline values for both modalities, 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.

TABLE I
Baseline Results.

Modality	WER
HTR	$39.3\% \pm 4.1$
ASR	$60.5\% \pm 1.3$

A. Baseline and Framework Adjustment

The baseline values were obtained by using the original language model in the decoding process of both modalities. As can be observed in Table I, the HTR and ASR WER values over the manuscript transcription reference are quite high due to the difficulty of the corpus. Moreover, 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 the collaborators can make mistakes while reading the manuscript. In order to alleviate these sources of errors, we normalised the cepstral features and the collaborators were provided with a text guide of reading along with text images, as can be observed in Figure 8.

In a previous work [9] 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 and the reliability verification did not show a significant impact on the results. Therefore, in this work the speaker ordering was defined by the order of reception of the audio utterances.

B. Preliminary Experiments

We started evaluating the performance of the multimodal crowdsourcing platform presented in this paper by using all the collaboration utterances without reliability verification. Figure 9 draws the baseline values for both modalities and the evolution of the system and ASR outputs for the whole test ASR corpus (CE $= 1350$) without reliability verification. As can be observed, the language model interpolation permits to reduce the error level in the next speech decoding process [1], and the combination with the speech decoding results allows the system output to converge to a better hypothesis with less errors to correct [7]. Besides, the ASR performance is considerably improved reducing the average WER baseline value ($60.5\% \pm 1.3$) to $33.9\% \pm 4.8$. Finally, after processing the speech of the last collaborator, the ASR and the system outputs presented $30.0\% \pm 4.1$ and $25.3\% \pm 3.9$ of WER, respectively. The $25.3\% \pm 3.9$ of WER present in this final system output represents 35.6% of relative statistically significant improvement over the HTR baseline, and an estimated time reduction for the paleographer revision of about 5 minutes per page.

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 10, 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 $44.2\% \pm 1.2$.

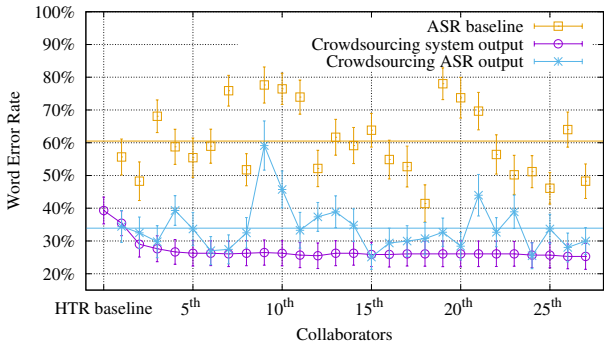


Fig. 9. Baseline values and the evolution of the system and ASR outputs for the whole test speech corpus without reliability verification nor lines selection. The horizontal lines represent the corresponding average ASR WER values.

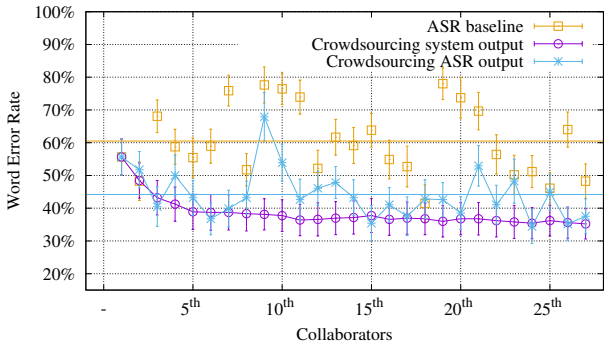


Fig. 10. ASR baseline values and the evolution of the system and ASR outputs processing only the speech, without HTR initialisation nor reliability verification nor lines selection. The horizontal lines represent the corresponding average ASR WER values.

The WER at the system output decreased to $35.2\% \pm 4.6$ from an initial value of $55.6\% \pm 5.5$. In spite of the fact that this is a remarkable improvement over the initialisation, this improvement is not statistically significant over the HTR baseline.

C. ASR Reliability Verification and Collaboration Effort

In order to analyse the behaviour of the multimodal crowdsourcing platform presented in this paper, it was tested setting different speech reliability thresholds (τ), and different amount of lines - batches (B) - to been read by the collaborators.

Figure 11 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 27$^{\text{th}}$ 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 II) over the HTR baseline ($39.3\% \pm 4.1$). On the other hand, the ASR reliability verification allows to filter the utterances that can worsen the system output, but, 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 or equal to 40%.

Figure 12 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

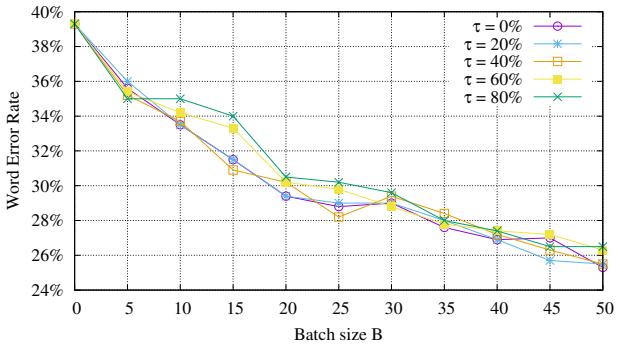


Fig. 11. Effect of the batch size B and the threshold τ on the WER of the final output.

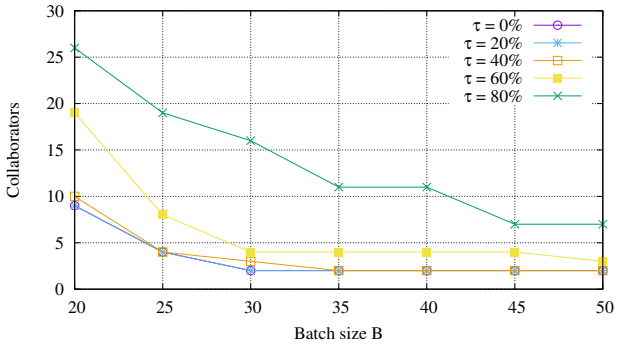


Fig. 12. Effect of the batch size B and the threshold τ on the minimum number of collaborators for improving the output significantly.

number of collaborators that allow to obtain a WER value at the system output lower than 31.2% (which represents a minimum relative improvement of 20.6%). Therefore, Figure 12 only shows results for batch sizes $B \geq 20$, where statistically significant improvements appear. The main conclusion that we can extract from Figure 12 is that high values of τ require more collaborators to refine significantly the system output, and that for τ in $0\% - 40\%$ the system presents a similar behaviour.

Table II 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 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 1350 utterances) are not statistically significant. Supposing a similar behaviour on other lines of the corpus, this means that with the whole collaboration effort (1350 utterances), 1125 lines would obtain transcription improvements.

D. Collaboration Effort per Line

We observed as some lines needed more refinement than others. Thus, we analysed the collaboration distribution over

TABLE II
COLLABORATION EFFORT (CE) EXPERIMENT RESULTS SUMMARY. IN BOLDFACE, BEST CE RESULT.

		First significant improvement			Final output
B	τ	Collaborators	CE	WER	WER
20	0%	9	180	$30.9\% \pm 4.0$	$29.4\% \pm 3.6$
	20%	9	180	$30.9\% \pm 4.2$	$29.4\% \pm 4.0$
	40%	10	200	$30.7\% \pm 4.2$	$30.2\% \pm 4.2$
25	0%	4	100	$30.5\% \pm 4.2$	$28.8\% \pm 4.0$
	20%	4	100	$30.5\% \pm 4.3$	$29.0\% \pm 4.1$
	40%	4	100	$30.9\% \pm 4.2$	$28.2\% \pm 3.8$
30	0%	2	**60**	$31.1\% \pm 3.8$	$29.0\% \pm 3.9$
	20%	4	120	$30.5\% \pm 4.3$	$29.0\% \pm 4.1$
	40%	3	90	$30.9\% \pm 3.9$	$29.4\% \pm 4.0$
35	0%	2	70	$30.2\% \pm 3.7$	$27.6\% \pm 3.5$
	20%	2	70	$30.4\% \pm 3.9$	$28.0\% \pm 3.6$
	40%	2	70	$31.1\% \pm 3.8$	$28.4\% \pm 3.8$
40	0%	2	80	$30.0\% \pm 3.8$	$26.9\% \pm 3.6$
	20%	2	80	$30.2\% \pm 3.7$	$26.9\% \pm 3.8$
	40%	2	80	$30.7\% \pm 4.0$	$27.2\% \pm 3.8$
45	0%	2	90	$29.6\% \pm 4.1$	$27.0\% \pm 4.1$
	20%	2	90	$29.8\% \pm 4.0$	$25.7\% \pm 3.7$
	40%	2	90	$30.9\% \pm 4.2$	$26.3\% \pm 3.7$
50	0%	2	100	$29.0\% \pm 3.9$	$25.3\% \pm 3.9$
	20%	2	100	$29.2\% \pm 3.8$	$25.5\% \pm 3.9$
	40%	2	100	$30.2\% \pm 4.1$	$25.5\% \pm 3.9$

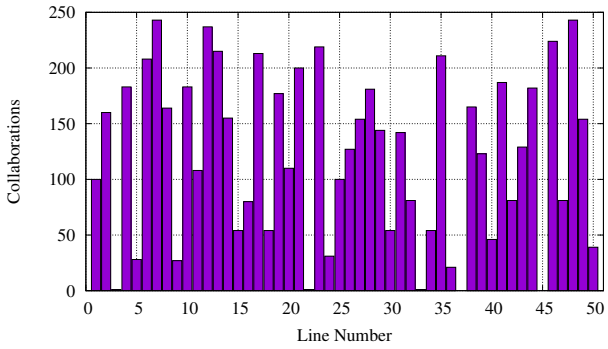


Fig. 13. Histogram representing the number of collaborations (times read) for each text line in the experiments for $\tau = 0$ and $B = [5, \ldots, 45]$.

the set of lines. Figure 13 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 III.

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

TABLE III
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

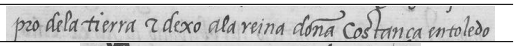
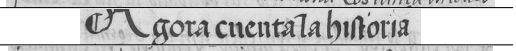
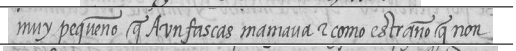
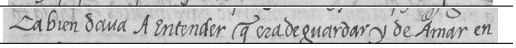
Line No.	Handwritten Text Line Image
7	
37	
45	
48	

Fig. 14. 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.

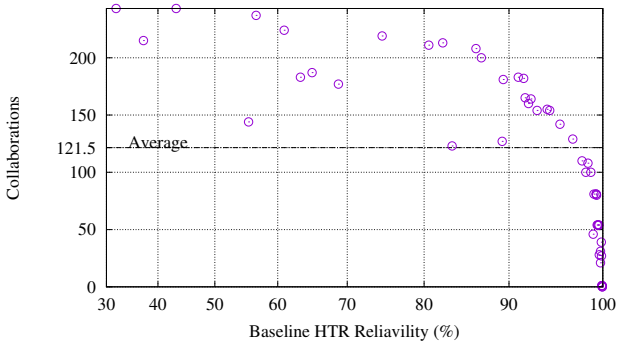


Fig. 15. Relation between the baseline HTR reliability R and the number of collaborations for each text line in the experiments for $\tau = 0$ and $B = [5, \ldots, 45]$.

in Figure 14. 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 with the collaboration effort per line. This relation is presented in Figure 15 and, as can be observed, the lines with lower R require higher amount of collaboration. Specifically, the 50% of lines with lower R concentrated 76.9% of collaborations. Besides, all lines that needed more repetitions than the average expected number (121.5) presented a value $R \leq 0.97$, whereas those with less repetitions than the average presented $R > 0.97$. 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.

VI. CONCLUSIONS AND FUTURE WORK

This paper presents a multimodal crowdsourcing framework for the transcription of historical handwritten documents. The novelties presented in this work are the client / server architecture, and the lines selection module in the server application. The client application is publicly available, and it 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 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. This two new characteristics allow to obtain more collaborations and, at the same time, to focus the collaboration effort to the lines whose transcription need more refinement.

The experiments showed as the use of speech is a good alternative for improving the transcription of historical

manuscripts, and as this modality allows people to collaborate in this task using their own mobile device. Moreover, the line selection allows to obtain similar results with a considerable collaborator effort reduction.

In view of the obtained results, we believe that there is still room for improvement. We propose for future studies the use of sentences in the handwritten text corpus instead of lines because it could make multimodality more natural for the speakers, and the use of more robust modelling methods, such as Deep Neural Networks (DNN) for optical and acoustic modelling and Recurrent Neural Networks (RNN) for language modelling. Moreover, this multimodal crowdsourcing framework is open to be tested with other datasets.

ACKNOWLEDGMENT

Work partially supported by projects READ - 674943 (European Union's H2020), SmartWays - RTC-2014-1466-4 (MINECO), CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO/FEDER), and ALMAMATER - PROMETEOII/2014/030 (Generalitat Valenciana).

REFERENCES

[1] V. Alabau, V. Romero, A. L. Lagarda, and C. D. Martínez-Hinarejos. A Multimodal Approach to Dictation of Handwritten Historical Documents. In *Proc. 12th Interspeech*, pages 2245–2248, 2011.
[2] J. R. Bellegarda. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93 – 108, 2004.
[3] M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. ICASSP*, volume 1, pages 409–412, 2004.
[4] A. Caines, C. Bentz, C. Graham, T. Polzehl, and P. Buttery. Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the crowdis corpora. In *Proceedings of LREC 2016*, 2016.
[5] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, Apr. 2011.
[6] P. Dreuw, S. Jonas, and H. Ney. White-space models for offline Arabic handwriting recognition. In *Proc. 19th ICPR*, pages 1–4, 2008.
[7] E. Granell and C. D. Martínez-Hinarejos. Combining Handwriting and Speech Recognition for Transcribing Historical Handwritten Documents. In *Proc. 13th ICDAR*, pages 126–130, 2015.
[8] E. Granell and C. D. Martínez-Hinarejos. Multimodal Output Combination for Transcribing Historical Handwritten Documents. In *Proc. 16th CAIP*, pages 246–260, 2015.
[9] E. Granell and C. D. Martínez-Hinarejos. A Multimodal Crowdsourcing Framework for Transcribing Historical Handwritten Documents. In *Proc. of the 16th DocEng*, pages 157–163, 2016.
[10] T. J. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Trans. Audio, Speech & Language Processing*, 14(3):1082–1089, 2006.
[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, Nov 2012.
[12] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184, 1995.
[13] P. O. Kristensson and K. Vertanen. Asynchronous multimodal text entry using speech and gesture keyboards. In *INTERSPEECH*, pages 581–584. ISCA, 2011.
[14] M. Luján-Mares, V. Tamarit, V. Alabau, C. D. Martínez-Hinarejos, M. Pastor, A. Sanchis, and A. H. Toselli. iATROS: A speech and handwritting recognition system. In *V Jornadas en Tecnologías del Habla*, pages 75–78, 2008.
[15] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.
[16] M. Miki, N. Kitaoka, C. Miyajima, T. Nishino, and K. Takeda. 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, 2014.
[17] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu. Albayzin speech database: design of the phonetic corpus. In *Proc. EuroSpeech*, pages 175–178, 1993.
[18] G. Parent and M. Eskenazi. Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *INTERSPEECH*, pages 3037–3040. ISCA, 2011.
[19] R. Plamondon and S. N. Srihari. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84, January 2000.
[20] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
[21] B. Rueber. Obtaining confidence measures from sentence probabilities. In *Proc. Eurospeech*, pages 739–742, 1997.
[22] N. Serrano, F. Castro, and A. Juan. The RODRIGO Database. In *Proc. 7th LREC*, pages 2709–2712, 2010.
[23] A. Singh, A. Sangwan, and J. Hansen. Improved parcel sorting by combining automatic speech and character recognition. In *2012 IEEE International Conference on Emerging Signal Processing Applications, ESPA 2012 - Proceedings*, pages 52–55, 3 2012.
[24] A. Stolcke. SRILM-an extensible language modeling toolkit. In *Proc. 3rd Interspeech*, pages 901–904, 2002.
[25] S. Tamura, K. Iwano, and S. Furui. Toward robust multimodal speech recognition. In *Symposium on Large Scale Knowledge Resources (LKR2005)*, pages 163–166, 2005.
[26] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, 9(3):288–298, 2001.
[27] S. Xie and Y. Liu. Using n-best lists and confusion networks for meeting summarization. *Trans. Audio, Speech and Lang. Proc.*, 19(5):1160–1169, July 2011.
[28] J. Xue and Y. Zhao. Improved confusion network algorithm and shortest path search from word lattice. In *Proc. of Int. Conf. in Acoustics, Speech and Signal Processing*, volume 1, pages 853–856, 2005.
[29] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al. The HTK book. *Cambridge university engineering department*, 2006.

Emilio Granell Emilio Granell obtained his BSc degree in Telecommunications Engineering with the speciality in Sound and Image in 2006 from Universitat Politècnica de València (UPV), then he worked in a telecommunications consulting company in France from 2006 to 2007. In 2011 he obtained his MSc degree in Artificial Intelligence, Pattern Recognition, and Digital Image also from the UPV. Currently, he is working on his thesis for obtaining a Ph.D. degree.

Mr. Emilio Granell pertains to the Pattern Recognition and Human Language Technology (PRHLT) research center, where he develops his research on the topics of speech recognition, dialogue systems, and interactive and multimodal systems. From 2010 he has participated in several research projects related with artificial intelligence, speech and handwritten text recognition, and smart cities.

Carlos-D. Martínez-Hinarejos Carlos-David Martínez-Hinarejos obtained his BSc in Computer Science in 1998, his Ph.D. degree in Pattern Recognition and Artificial Intelligence in 2003, and his BSc in Biotechnology in 2012, all from Universitat Politècnica de València (UPV). He joined to the UPV staff (Computation and Computer Systems Department, DSIC) in 2000.

Dr. Martínez-Hinarejos pertains to the Pattern Recognition and Human Language Technology (PRHLT) research center, where he develops his research on the topics of speech recognition, dialogue systems, multimodal systems, text classification, and bioinformatics. He has participated in many European and Spanish projects, and is an active member of the Spanish Network for Speech Technologies (RTTH).