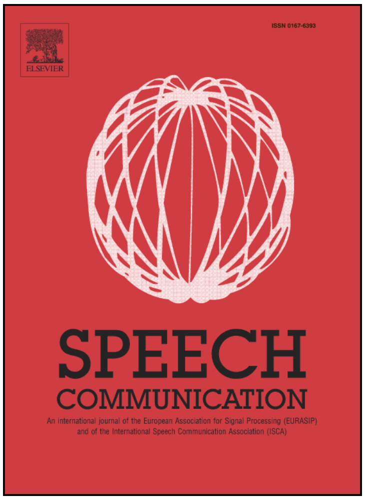

ISSN 0167-6393
ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories

Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, Alfons Juan

PII: S0167-6393(15)00101-6
DOI: http://dx.doi.org/10.1016/j.specom.2015.09.006
Reference: SPECOM 2321

To appear in: *Speech Communication*

Received Date: 29 April 2015
Revised Date: 7 September 2015
Accepted Date: 16 September 2015

Please cite this article as: Valor Miró, J.D., Silvestre-Cerdà, J.A., Civera, J., Turró, C., Juan, A., Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories, *Speech Communication* (2015), doi: http://dx.doi.org/10.1016/j.specom.2015.09.006

Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories

Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, Alfons Juan

Universitat Politècnica de València Camino de Vera s/n, 46022 Valencia, Spain

Abstract

Video lectures are widely used in education to support and complement face-to-face lectures. However, the utility of these audiovisual assets could be further improved by adding subtitles that can be exploited to incorporate added-value functionalities such as searchability, accessibility, translatability, note-taking, and discovery of content-related videos, among others. Today, automatic subtitles are prone to error, and need to be reviewed and post-edited in order to ensure that what students see on-screen are of an acceptable quality. This work investigates different user interface design strategies for this post-editing task to discover the best way to incorporate automatic transcription technologies into large educational video repositories. Our three-phase study involved lecturers from the Universitat Politècnica de València (UPV) with videos available on the poliMedia video lecture repository, which is currently over 10,000 video objects. Simply by conventional post-editing automatic transcriptions users almost reduced to half the time that would require to generate the transcription from scratch. As expected, this study revealed that the time spent by lecturers reviewing automatic transcriptions correlated directly with the accuracy of said transcriptions. However, it is also shown that the average time required to perform each individual editing operation could be precisely derived and could be applied in the definition of a user model. In addition, the second phase of this study presents a transcription review strategy based on confidence measures (CM) and compares it to the conventional post-editing strategy. Finally, a third strategy resulting from the combination of that based on CM with massive adaptation techniques for ASR achieved to improve the transcription review efficiency in comparison with the two aforementioned strategies.

Keywords: video lecture repositories, usability study, computer-assisted transcription, interface design strategies, automatic speech recognition

1. Introduction

The adoption of video lectures in higher education is a widespread phenomenon (Allen and Seaman (2010)) that is changing the landscape of formative options not only at universities, making lecturers think out of the box (Zhang et al. (2006); Ross and Bell (2007)), but also at other

Email addresses: jvalor@dsic.upv.es (Juan Daniel Valor Miró), jsilvestre@dsic.upv.es (Joan Albert Silvestre-Cerdà), jcivera@dsic.upv.es (Jorge Civera), turro@cc.upv.es (Carlos Turró), ajuan@dsic.upv.es (Alfons Juan)

institutions and private companies that understand video lectures as a possibility to train their personnel at low cost. Video lectures have been proved to be welcome by the learning community (Soong et al. (2006)). The Universitat Politènica de València (UPV) deployed in 2007 its lecture capture system for the cost-effective creation and dissemination of quality educational video (poliMedia (2007)). This collection has rapidly grown since then and currently hosts almost 20,000 mini lectures (Lyons et al. (2012)) created by over one thousand lecturers, in part incentivised by the *Docència en Xarxa* (Teaching Online) action plan to boost the use of digital resources at the UPV. poliMedia has been successfully deployed at other universities in Spain and South America. Mini-lectures with an average duration of 10 minutes are the most extended video format in Massive Open Online Courses (MOOCs), since viewers' attention rapidly drops after the first minutes being watched (Guo et al. (2014)).

From 2011 to 2014, the UPV coordinated the EU project transLectures (Silvestre et al. (2012)) to implement automatic transcription and translation systems for video lectures based on cost-effective techniques such as, *massive adaptation*[1] and *intelligent interaction*[2]. transLectures tries to give an answer to the need for transcriptions of video lectures (Dufour et al. (2005); Fujii et al. (2006)), not only for providing subtitles to non-native speakers, and the deaf and hard-of-hearing (Wald (2006)), but also to allow for lecture content searches (Repp et al. (2008)) and other advanced repository functionalities, including content summarisation to assist students in note-taking, and the discovery of related videos (Glass et al. (2007)).

In the framework of transLectures automatic subtitles in Spanish, English and Catalan have been generated for all videos in the poliMedia repository and were continuously improved during the course of the project. However, as it stands, the quality of the automatic transcriptions generated mean that lecturer intervention is required in order to guarantee the accuracy of the material ultimately made available to students (Munteanu et al. (2006)). So UPV lecturers, having filmed videos for the poliMedia repository as part of an earlier *Docència en Xarxa* call, trialled the computer-assisted transcription system *transLectures player* with editing capabilities for keyboard and mouse (Suhm et al. (2001)).

Some previous computer-assisted transcription tools are limited to batch-oriented passive user interaction strategies in which the initial transcription is manually post-edited. More precisely, Barras et al. (2001) presents the transcription tool Transcriber and some tests to measure the time needed to generate a transcription from scratch. Munteanu et al. (2008) performs an exhaustive analysis of a collaborative user post-editing system, concluding that reviewing automatic transcriptions allow to obtain useful transcriptions for educational purposes. Kolkhorst et al. (2012) proves that the usage of interactive correction methods are useful for reducing WER significantly by applying speaker adaptation techniques. However, these two latter works do not assess the impact on user effort. Papadopoulos and Pearson (2012) show a user effort reduction when transcriptions are improved with a semantic and syntactic transcription analysing tool highlighting misspelled words. Finally, Bazillon et al. (2008) tested a batch-oriented passive user interaction protocol without system participation obtaining good results in terms of user effort, similar to those obtained in the present study. However, these studies do not perform an exhaustive comparison of different user interaction methods and the relationship between quality and time devoted by the lecturer based on real-life end-user evaluations.

[1] The process whereby automatic subtitling systems can be adapted to the lecture in question using lecture-specific material such as presentation slides, related documents, or the speaker voice.

[2] The process whereby, in the subsequent post-editing stage, automatic subtitling systems direct the user to those subtitles that contain the most transcription errors.

In this work, we expand the preliminary results reported in (Valor Miró et al. (2014)) in order to provide an in-depth analysis of a series of more intelligent active user interaction strategies for the generation of transcriptions that are accurate enough to be useful to students while requiring the minimum effort on the part of the lecturer (Luz et al. (2008)). To this end, a three-phase evaluation process was set up to analyse alternative user interaction strategies for reviewing the automatically-generated transcription. Our first phase consisted of a conventional manual post-editing strategy. For the second we introduced the premise of intelligent interaction, before moving onto a third phase which combines the best features from phases one and two in a two-step review process.

2. System Description

The system serves two main use cases that are shown in Fig. 1. In the first use case (on the left), lecturer recordings are automatically transcribed off-line using an ASR system. While in the second use case (on the right), users interact with a web player in order to amend recognition errors found in the automatic transcriptions previously generated.

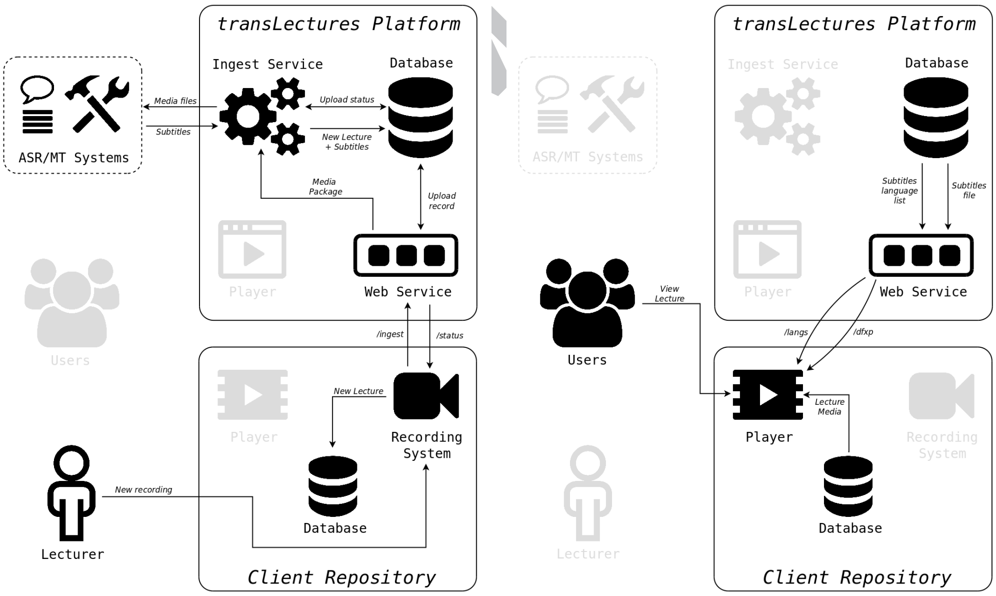


Figure 1: Main two use cases for video transcription (left side) and transcription revision by users (right side).

In the first use case, the ASR system was generated using the transLectures-UPV open source toolkit, TLK (The TransLectures-UPV team (2013)), which consists of a set of tools that allows acoustic model training and speech decoding. Besides, the SRILM toolkit (Stolcke (2002)) is used to estimated *n*-gram language models. More precisely, a Spanish ASR system based on a tied triphone HMM with Gaussian mixture models trained on the poliMedia corpus (see Table 1) was deployed. In addition, the well-known CMLLR (Gales (1998)) technique for speaker adaptation was applied. The language model was a linear mixture trained on the poliMedia transcriptions along with other external resources.

The WER results achieved with our baseline system and the CMLLR-adapted system on the evaluation sets are reported in Table 2.

Table 1: Basic statistics of the poliMedia speech corpus.

Set	Lectures	Time (h)	Phrases	Running Words	Vocabulary size
Train	655	96	41.5*k*	96.8*k*	28*k*
Development	26	3.5	1.4*k*	34*k*	4.5*k*
Test	23	3	1.1*k*	28.7*k*	4*k*

Table 2: WER (%) of the Spanish ASR system.

	Development	Test
Baseline	28.1	30.3
Baseline + CMLLR	22.2	24.6

In the second use case, the user can watch and review the transcription of a video with the transLectures web player. Corrections made by the user are sent back to the web service to update the transcription file. The transLectures player interface consists of an innovative web player with editing capabilities, complete with alternative display layout options and full keyboard support. This player was developed as part of transLectures at the UPV (Valor Miró et al. (2012)), in accordance with Nielsen's usability principles (Nielsen and Levy (1994); Nielsen (1999)); and it was iteratively improved during subsequent evaluations described in the next section.

3. User trials

Here, we describe user evaluations carried out under UPV's *Docència en Xarxa* (Online Teaching) programme. An on-going incentive-based programme to encourage university lecturers at the UPV to develop digital learning resources based on ICTs.

3.1. Methodology

A total of 27 lecturers signed up for this study, reviewing a sample of 86 video lectures organised into three phases. Most participants had degrees in different branches of engineering (17), while the rest mastered business management (6), social science (2) and biology (2).

Lecturers involved committed to reviewing the automatic transcriptions of five of their poliMedia videos. These videos were transcribed with the system described at Section 2. Lectures to be reviewed were allocated across three consecutive evaluation phases, described below.

1. Conventional post-editing: Automatic transcriptions for the first video of each lecturer are manually reviewed. Automatic transcription segments are up to 20 words long and are shown in synchrony with the video.

2. Intelligent interaction: In this phase, only a subset of probably incorrectly-recognised (low confidence) words were reviewed in the second and third videos by lecturers. These words are played within a context of one word before and one word after, being possible to expand the context to more words.
3. Two-step review: This phase organized in two consecutive rounds of evaluation for the fourth and fifth videos. The first round mimics phase two above, where the lecturer reviewed only the least confidence words. However, in this phase, least confidence words are preceded by a context of three words. Once this first round is completed, the video is then automatically re-transcribed on the basis of the lecturer's review actions preserving their corrections. In a second round, the updated transcriptions are completely reviewed as in the first phase.

Feedback from lecturers is fundamental in order to inform the design of each subsequent evaluation phase and, ultimately, of the web interface itself. The web interface being tested and evaluated by lecturers consists of the transLectures player presented at Section 2.

The transLectures player logged precise user interaction statistics, such as the duration for which the editor window is open, the number of segments (individual subtitles) edited out of the total and the display layout selected. It also logged statistics at the segment level, including the number of mouse clicks and key presses, editing time, and the number of times a segment is played. From these statistics we computed two of the main variables of this study: RTF[3] is the time spent by the lecturer reviewing transcriptions, and WER as an indicator of the minimum number of corrections required to bring the initial automatic transcriptions into line with the reviewed transcription.

However, we also assess the impact of the three aforementioned evaluation phases in terms of WER reduction per RTF unit. That is, by how many WER points the transcription error is reduced for each RTF unit spent reviewing the automatic transcription. This ratio can be understood as a review efficiency measure, i.e. error reduction per unit of time.

In addition, feedback from lecturers was collected as subjective statistics after each phase, in the form of a brief satisfaction survey based on Lewis (1995). Lecturers were asked to rate various aspects on a Likert scale from 1-10 (see Table 3). They were then asked the following three open-ended questions, allowing them to freely express their subjective impressions of using the transLectures player:

- If you were to add new features to the player, what would they be?
- If you had to work with this player on a daily basis, what would you change?
- Any additional comments.

The use of the satisfaction surveys over the three phases has proved to be a very valuable tool for collecting lecturers' subjective feedback and has led directly to the improvement and refinement of the transLectures player.

3.2. Experimental results

In this section we describe the experimental results attained over the three consecutive evaluation phases: conventional post-editing, intelligent interaction, and two-step review protocols.

[3] In our study, the Real Time Factor (RTF) is calculated as the ratio between the time spent reviewing the transcription of a video and the duration of said video. So if, for example, a video lasts twenty minutes and its review takes, by way of example only, sixty minutes, then the RTF for this video would be 3.

Table 3: Questions scored on a 1-10 Likert scale presented to lecturers after each phase.

Intuitiveness

1- I am satisfied with how easy it is to use this system.
2- It was easy to learn to use this system.
3- The help information of this system is clear.
4- The organization of information on screen is clear.

Likeability

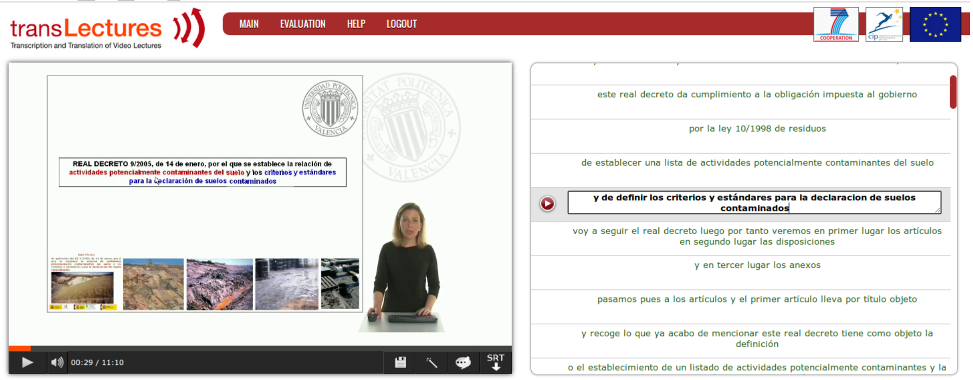
5- I feel comfortable using this system.
6- I like using the interface of this system.
7- Overall, I am satisfied with this system.

Usability

8- I can complete my work effectively using this system.
9- I can complete my work quicker than doing it from scratch.
10- This system has all the functions that I expect to have.

3.2.1. First phase: Post-editing

In the first phase, 20 UPV lecturers reviewed the automatic transcription of their first video lecture in its entirety using the transLectures player, shown in Figure 2 and described above in Section 3.1. A total of 2.6 hours in 20 video lectures were completely reviewed by the lecturers. Prior to this phase, lecturers were sent a link to a demo video explaining how to review their video transcriptions, in order to become familiar with the functionality of the transLectures player. The transLectures player plays the video and the transcription in synchrony, allowing the user to read the transcription while watching and listening to the video. When the lecturer finds a transcription error, it can be amended by clicking (or pressing *Enter*) on the incorrect segment to pause the video. With the video paused, the lecturer can easily enter their changes in the text box that opens. Lecturers save their work periodically updating both transcription and user interaction statistics.



REAL DECRETO 9/2005, de 14 de enero, por el que se establece la relación de actividades potencialmente contaminantes del suelo y los criterios y estándares para la declaración de suelos contaminados

este real decreto da cumplimiento a la obligación impuesta al gobierno

por la ley 10/1998 de residuos

de establecer una lista de actividades potencialmente contaminantes del suelo

y de definir los criterios y estándares para la declaracion de suelos contaminados

voy a seguir el real decreto luego por tanto veremos en primer lugar los artículos en segundo lugar las disposiciones

y en tercer lugar los anexos

pasamos pues a los artículos y el primer artículo lleva por título objeto

y recoge lo que ya acabo de mencionar este real decreto tiene como objeto la definición

o el establecimiento de un listado de actividades potencialmente contaminantes y la

Figure 2: transLectures web player with the side-by-side layout while the lecturer edits one of the segments.

To assess the impact of automatic transcription on the total time required to generate usable subtitles for video lectures, we first compared times to that spent performing the same task manually from scratch. We carried out the statistical two-sample Welch's t-test for RTF with the data collected in this first phase and the data collected in a previous study, in which around 100 hours of video lectures from the same repository were transcribed from scratch (Valor Miró et al. (2012)) by non-expert users (lecturers and doctoral students). We found that there was a statistically significant difference between mean RTFs (sig^4=$5.41 \cdot 10^{-10}$), with the mean RTF for subtitles generated automatically (Mean (M)=5.4, Std (S)=2.9) being notably lower than that for those generated manually from scratch (M=10.1, S=1.8). This result suggests that the automatic transcriptions (at their reported accuracy in terms of WER) allow lecturers to generate subtitles much more efficiently than manually from scratch. We should note that the background expertise of our lecturers (engineering vs. non-engineering) was not ultimately statistically significant in terms of RTF when reviewing automatic transcriptions (*sig*=0.24). In addition, we also computed the WER reduction per RTF unit (M=3.2, S=1.3) to compare the effectiveness of this interaction strategy with those proposed in the second and third phases.

As shown in Table 4, three linear regression models were evaluated to explain RTF as a function of the independent variables of our study (WER, Intuitiveness, Likeability and Usability). Model 1 revealed that WER ($beta^5$=0.285, *sig*=$4.73 \cdot 10^{-9}$) was statistically significant and accounted to a large extent for the variance observed in the data ($R^2 = 0.842$). We also considered the possibility of including the *Intercept* in this regression model, but the variance explained by the model dropped drastically.

A graphical representation of our data in terms of WER vs. RTF, and our prior knowledge of user behaviour (users essentially ignore automatic transcriptions above a certain WER threshold, preferring to transcribe from scratch) suggested that a logarithmic model might better fit our data. Consequently, the logarithmic Model 2 was proposed, resulting in a more statistically significant *beta* (*beta*=2.025, *sig*=$9.82 \cdot 10^{-12}$) and an increase in the variance explained by the model ($\Delta R^2 = 0.075$).

Table 4: Linear regression models to explain RTF using different factors.

Predictor	*beta*	*sig*
Model 1 ($\triangle R^2 = 0.842$, $R^2 = 0.842$, *sig*=$4.73 \cdot 10^{-9}$)		
WER	0.285	$4.73 \cdot 10^{-9}$
Model 2 ($\triangle R^2 = 0.075$, $R^2 = 0.917$, *sig*=$9.82 \cdot 10^{-12}$)		
$log_e(WER)$	2.025	$9.82 \cdot 10^{-12}$
Model 3 ($\triangle R^2 = 0.001$, $R^2 = 0.918$, *sig*=$1.59 \cdot 10^{-8}$)		
$log_e(WER)$	2.263	0.007
Intuitiveness	0.144	0.832
Usability	-0.302	0.665
Likeability	0.084	0.874

As expected, both Model 1 and 2 would point that WER does in fact influence lecturer review time as expressed in RTF. Finally, we decided to incorporate the subjective variables as defined in

[4] It is the probability of observing an effect given that the null hypothesis is true.

[5] It is the coefficient multiplying the predictor in the linear regression model.

the satisfaction survey in Table 3: intuitiveness (*sig*=0.832), usability (*sig*=0.665) and likeability (*sig*=0.874). However, the outcomes were ultimately not statistically significant as a means of determining RTF. This result confirms informal comments made by lecturers to the effect that transcription quality should be improved as a priority over further modifications to the user interface.

As shown in Table 5, lecturers felt (Overall Mean (OM) = 9.1) that the user interaction strategy in this phase was designed in accordance with intuitiveness (Grand Mean (GM) = 9.3), likeability ($GM = 8.8$) and usability ($GM = 8.9$) principles, with intuitiveness being the most highly rated characteristic.

Table 5: Detailed results of the satisfaction survey in the first phase.

Question	Mean
Intuitiveness	
1- I am satisfied with how easy it is to use this system.	9.4
2- It was easy to learn to use this system.	9.4
3- The help information of this system is clear.	9.2
4- The organization of information on screen is clear.	9.0
Grand Mean	*9.3*
Likeability	
5- I feel comfortable using this system.	8.7
6- I like using the interface of this system.	8.7
7- Overall, I am satisfied with this system.	9.0
Grand Mean	*8.8*
Usability	
8- I can complete my work effectively using this system.	9.0
9- I can complete my work quicker than by doing it from scratch.	8.6
10- This system has all the functions that I expect to have.	9.0
Grand Mean	*8.9*
Overall Mean	*9.1*

Comments from the three open-ended questions proved to be a valuable source of feedback for refining minor usability issues and incorporating additional new features, such as changing the font size and colour, allowing the lecturer to download the transcription file being reviewed, automatically saving the transcription file and minimising the initial loading time. All in all, results were largely positive and, as desired, lecturers were able to become familiar with the transLectures player in advance of the next two phases.

Given Model 2 that is shown in Table 4, a more detailed user model was derived in order to predict the performance of potential user interaction strategies before being tested on real users. For the sake of interpretability, variables were expressed in absolute rather than relative terms. In other words, the independent variable WER was given in terms of word-level editing operations, while the dependent variable RTF was replaced by the time taken in seconds.

As shown in Table 6, our statistically significant Model 1 ($R^2 = 0.801$, *sig*=$2.2 \cdot 10^{-16}$) correlates the time spent generating accurate subtitles with the number of correct (*beta* = 1.370, *sig*=$2.2{\cdot}10^{-16}$) and incorrect (*beta* = 4.388, *sig*=$2.2{\cdot}10^{-16}$) words in the automatic transcriptions given to our lecturers. More interesting from the point of view of the user model is the ratio between the *beta* value for independent variables (correct and incorrect words), which suggests that it takes on average three times longer to correct an incorrectly-recognised word than to confirm a correctly-recognised word.

Model 2 in Table 6 (R^2=0.808, *sig*=$2.2 \cdot 10^{-16}$) factorises the incorrect words into the three basic word edit operations: deletion (*beta*=2.059, *sig*=$3.2 \cdot 10^{-6}$), substitution (*beta* = 4.800, *sig*=$2.2 \cdot 10^{-16}$) and insertion (*beta*=5.237, *sig*=$2.2 \cdot 10^{-16}$), while the variable correct words (*beta*=1.370, *sig*=$2.2 \cdot 10^{-16}$) remains the same. The beta values can be interpreted as reflecting the relation between the time taken to perform an edit operation on an incorrect word and that taken to review a correct word, that is, essentially consisting of listening to it. As expected, simply deleting an incorrect word takes only slightly longer than reviewing a correct word. However, substitutions and insertions are more costly edit operations, requiring three to four times as long.

Table 6: Linear regression on review time provided word-level edit operations.

Predictor	*beta*	*sig*
Model 1 ($\triangle R^2$=0.801, R^2=0.801, F=2030, *sig*=$2.2 \cdot 10^{-16}$)		
Correct Words	1.370	$2.2 \cdot 10^{-16}$
Incorrect Words	4.388	$2.2 \cdot 10^{-16}$
Model 2 ($\triangle R^2$=0.007, R^2=0.808, F=1060, *sig*=$2.2 \cdot 10^{-16}$)		
Correct Words	1.370	$2.2 \cdot 10^{-16}$
Deleted Words	2.059	$3.2 \cdot 10^{-6}$
Substituted Words	4.800	$2.2 \cdot 10^{-16}$
Inserted Words	5.237	$2.2 \cdot 10^{-16}$

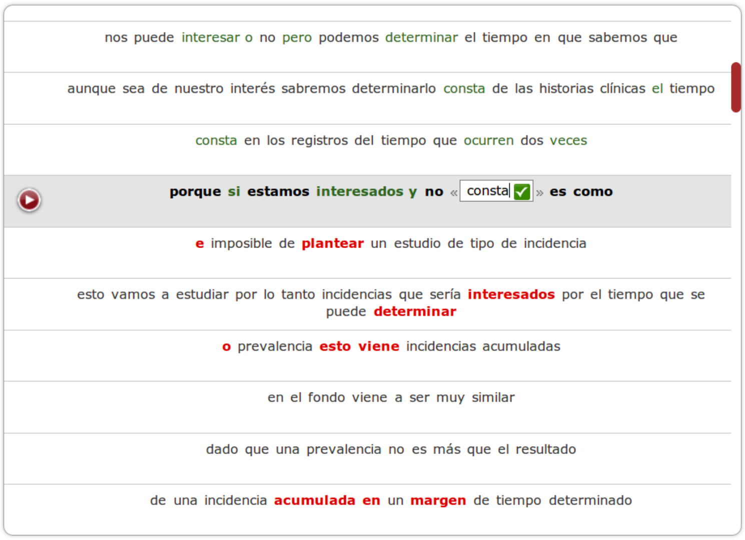
Defining this user model was a key step in exploring alternative, more time-effective user interaction strategies to post-editing for generating accurate subtitles for video lectures. These strategies are deployed in the next two phases.

3.2.2. Second phase: Intelligent Interaction

This second phase incorporates a new interaction strategy called *intelligent interaction* (Serrano et al. (2013)) in order to study if review times could be further improved. This strategy is based on the application of active learning (AL) techniques to ASR (Deng and Li (2013)). More concretely, we apply batch AL based on uncertainty sampling (Lewis and Catlett (1994)) using *confidence measures* (Wessel et al. (2001); Hakkani-Tur et al. (2002); Riccardi and Hakkani-Tur (2005)), which provide an indicator as to the probable correctness of each word appearing in the automatic transcription. In practice the lecturer may need to review (confirm) some correctly-recognised words incorrectly identified as errors (false positives), but many of the incorrectly-recognised words are spotted correctly (true positives). The idea is to focus user's review actions on incorrectly-transcribed words saving time and effort.

In this phase, lecturers are to review the subset of least confidence word according to the CAT system in increasing order of probable correctness. This subset typically constituted between 10-20% of all words transcribed using the ASR system, though lecturers could modify this range at will to as low as 5% and as high as 40%, depending on the perceived accuracy of the transcription. Each word was played in the context of one word before and one word after, in order to facilitate its comprehension and resulting correction.

Figure 3 shows a screenshot of the transcription interface in this phase. Low-confidence words are shown in red and corrected low-confidence words in green. The text box including the low-confidence word can be expanded in either direction to increase the context. For this phase, the intelligent interaction mode was activated in the transLectures player by default, though lecturers could switch back to the conventional (fully manual) post-editing strategy.



nos puede interesar o no pero podemos determinar el tiempo en que sabemos que

aunque sea de nuestro interés sabremos determinarlo consta de las historias clínicas el tiempo

consta en los registros del tiempo que ocurren dos veces

porque si estamos interesados y no « consta ✓ » **es como**

e imposible de **plantear** un estudio de tipo de incidencia

esto vamos a estudiar por lo tanto incidencias que sería **interesados** por el tiempo que se puede **determinar**

o prevalencia **esto viene** incidencias acumuladas

en el fondo viene a ser muy similar

dado que una prevalencia no es más que el resultado

de una incidencia **acumulada en** un **margen** de tiempo determinado

Figure 3: A screenshot of the transcription interface in intelligent interaction mode. Low-confidence words appear in red and reviewed low-confidence words in green. The word being edited in this example is opened for review, and the text box can be expanded to the left or right by clicking on << or >>, respectively. Clicking the green check button to the right of the text box confirms the word as correct.

Interaction statistics revealed that 12 of the 23 lecturers participating in this second phase stayed in the intelligent interaction mode for the full review of one of their poliMedia videos. In fact 2.8 hours over 18 video lectures were reviewed using that technique. In the other cases (3 hours over 22 video lectures), lecturers switched back to the conventional post-editing mode. Lecturers wanted to make sure that perfect transcriptions were obtained no matter how much time could be saved by the intelligent interaction mode. As a result, 18 videos were reviewed using intelligent interaction, while 22 videos were reviewed in the conventional post-editing mode. The RTF of the videos completely reviewed using the conventional post-editing mode (as in the first phase) was 5.2. Given the starting WER of 19.5, this time factor is comparable to results recorded in phase one.

For those lecturers that remained in the intelligent interaction mode, review time was reduced to an RTF of 2.2, though the resulting transcriptions were not error-free, unlike in phase one. That said, the residual WER of the transcriptions after being reviewed was as low as 8.0, which is not so far from that achieved by non-expert transcriptionists (Hazen (2006)). This indicates that confidence measures successfully identify approximately half of all incorrectly-recognised words. However, we should also assess the impact of the intelligent interaction strategy in terms of WER reduction per RTF unit. That is, by how many WER points the transcription is improved for each RTF unit spent reviewing the automatic transcription, compared to conventional post-editing. To do so, we carried out a statistical test between intelligent interaction (M=4.6, S=3.9) and conventional post-editing (M=3.9, S=1.3). The results indicated that there was no statistically significant difference between these two strategies in this respect (*sig*=0.486). This means that intelligent interaction is in fact just as efficient in terms of WER decrease per RTF unit as conventional post-editing.

We can see in Table 7 that lecturers showed ($OM = 7.2$) a clear preference for obtaining perfect transcriptions, irrespective of the relative time savings afforded by the intelligent interaction strategy, and insisted on an interaction mode that gave them full control over the end quality of the transcriptions. The figures collected on intuitiveness ($GM = 8.1$), likeability ($GM = 6.8$) and usability ($GM = 6.3$), dropping from the conventional post-editing phase, reflect this assessment. However, lecturers did seem to embrace confidence measures, suggesting that low confidence words denoted in red could be incorporated into the conventional post-editing strategy.

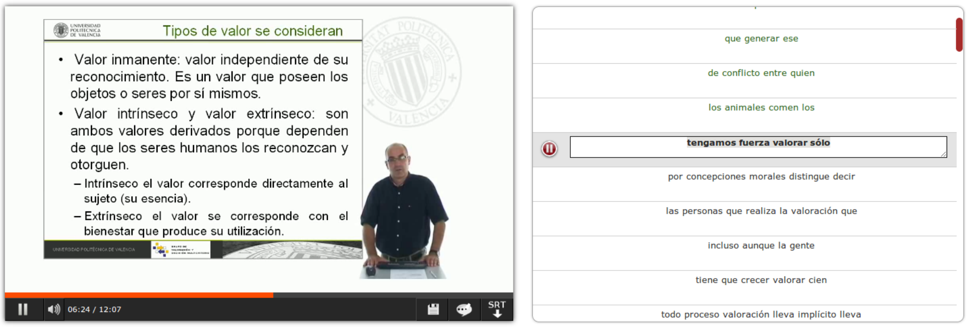
Table 7: Detailed results of the satisfaction survey for intelligent interaction.

Question	Mean
Intuitiveness	
1- I am satisfied with how easy it is to use this system.	7.8
2- It was easy to learn to use this system.	8.1
3- The help information of this system is clear.	8.1
4- The organization of information on screen is clear.	8.4
Grand Mean	*8.1*
Likeability	
5- I feel comfortable using this system.	6.5
6- I like using the interface of this system.	6.9
7- Overall, I am satisfied with this system.	6.9
Grand Mean	*6.8*
Usability	
8- I can complete my work effectively using this system.	6.7
9- I can complete my work quicker than by doing it from scratch.	6.6
10- This system has all the functions that I expect to have.	5.6
Grand Mean	*6.3*
Overall Mean	*7.2*

User satisfaction surveys statistically reflected that post-editing ($OM = 9.1$, S=1.3) was preferred over intelligent interaction ($OM = 7.2$, S=1.7) by our lecturers (sig=$4.0 \cdot 10^{-6}$). Feedback from the three open-ended questions in the satisfaction survey clearly indicated that the intelligent interaction strategy needed rethinking in order to allow the following operations: editing of words outside of the intelligent interaction text boxes, unlimited use of the text box expansion arrows (currently restricted to a given number of words before and after) in order to correct entire segments, and movement between text boxes in both directions (currently limited to moving forwards to the next only). Lecturer preferences notwithstanding, the intelligent interaction strategy based on confidence measures was proven to be an effective means of identifying incorrectly-recognised words. For this reason, we designed the third phase in such a way as to take greater advantage of the intelligent interaction strategy, while also granting lecturers full control over the final transcription quality.

3.2.3. Third phase: Two-step Supervision

As mentioned above, the third phase was organised into two subphases or rounds and is essentially a combination of the previous two phases. In this phase, lecturers first review a subset of the least confidence words, as in the second phase. The videos are then re-transcribed (by ASR) on the basis of all previous review actions preserving those corrections made by users. These updated transcriptions are expected to be of high quality than the original transcriptions (Sanchez-Cortina et al. (2012)) reducing overall review times. In the second round of this third phase, lecturers completely review the entire re-transcription as in phase one. The fourth and fifth video of each lecturer was reviewed in this phase. Figure 4 shows a screenshot of the transLectures web player used in step one.



Tipos de valor se consideran

- Valor inmanente: valor independiente de su reconocimiento. Es un valor que poseen los objetos o seres por sí mismos.
- Valor intrínseco y valor extrínseco: son ambos valores derivados porque dependen de que los seres humanos los reconozcan y otorguen.
  - Intrínseco el valor corresponde directamente al sujeto (su esencia).
  - Extrínseco el valor se corresponde con el bienestar que produce su utilización.

que generar ese

de conflicto entre quien

los animales comen los

tengamos fuerza valorar sólo

por concepciones morales distingue decir

las personas que realiza la valoración que

incluso aunque la gente

tiene que crecer valorar cien

todo proceso valoración lleva implícito lleva

UNIVERSIDAD POLITÉCNICA DE VALENCIA
UNIVERSIDAD POLITÉCNICA DE VALENCIA
06:24 / 12:07
SRT

Figure 4: Screenshot of the transLectures web player used in step one of phase three, side-by-side layout. Each segment contains four words, of which the last word is the low-confidence word.

More concretely, the first round is devoted to review isolated segments of four words in which the last word was the low-confidence word. These segments were presented to the lecturer for review in increasing order of confidence (of the last word) until one of the following three conditions was met:

1. The total review time reached double the duration of the video itself; or
2. No corrections were entered for five consecutive segments; or
3. 20% of all words were reviewed.

The reviewed transcriptions in this phase, but also in phases one and two, were used to adapt the ASR system via a process of massive adaptation. Specifically, we adapted the acoustic models to the speaker with the MLLR technique (Gales (1998)), and the language models using a linear interpolation between the language model trained on the reviewed transcriptions and the large language model previously trained (Martínez-Villaronga et al. (2013)). Then, the automatic transcriptions were regenerated, preserving those segments already reviewed by lecturers, and using them to improve the recognition of the context words using a constrained search (Kristjansson et al. (2004); Serrano et al. (2013)). This two-step review process was successfully completed by 15 lecturers on a total of 26 video lectures with 3.7 hours of video. More precisely, a total of 1.0 and 2.7 hours were reviewed in the first and second steps, respectively.

In the first step of this phase, average review time was as low as 1.4 RTF. As reported in Table 8, WER dropped significantly from the initial 28.4 to the regenerated transcriptions 18.7. That is, almost 10 WER points over 1.4 RTF, meaning that intelligent interaction plus adaptation (M=8.6, S=5.8) achieved a higher statistically significant WER reduction per RTF unit (sig=6.9 $\cdot$

10^{-3}) than intelligent interaction alone (M=4.6, S=3.9). This suggests that intelligent interaction plus adaptation is, in fact, more effective in terms of WER decrease per RTF unit than intelligent interaction alone.

In the second step, lecturers completely reviewed the regenerated transcriptions to obtain perfect final transcriptions, as in the first phase. Average RTF for this task stood at 3.9. As expected, when comparing WER reduction per RTF unit in the first phase (M=3.2, S=1.3) and the second step of this phase three (M=5.3, S=2.0), we can observe a statistically significant learning curve (*sig*=$8.5 \cdot 10^{-5}$) in lecturers' performance. As a result, we proved that there is a learning curve involved in getting to grips with the transLectures player.

Table 8: Summary of results obtained in the two-step review phase

	WER	RTF	ΔRTF
Initial transcriptions	28.4	0.0	-
First step: Intelligent interaction	25.0	1.4	1.4
Massively adapted transcriptions	18.7	1.4	-
Second step: Complete review	0.0	5.3	3.9

In order to fairly compare the first (M=3.2, S=1.3) and third (M=6.0, S=2.0) phases in terms of WER reduction per RTF unit, we subtract the effect of the learning curve for each lecturer. To this purpose, the WER reduction per RTF unit of each lecturer in the second step of this third phase was assumed to be that of the same lecturer revising their first video. This assumption leads to a corrected WER reduction per RTF unit (M=4.7, S=2.8). Even so, we found a lower yet statistically significant difference (*sig*=0.02) in favour of the third phase explained by the application of massive adaptation. This result suggests that the two-step strategy is more efficient than the conventional post-editing strategy.

However, this statistically significant difference only holds when enough reviewed data is available for adaptation. That is, the reviewed data generated in the first step of this phase (aprox. 4 minutes per lecturer) is not sufficient to improve the ASR performance so that it reduces the user effort. In this latter scenario, the resulting WER after applying massive adaptation would be 24.0 instead of 18.7, resulting in a WER reduction per RTF unit (M=3.7, S=2.3) not statistically significant better (*sig*=0.31) than that obtained in the first phase. For this reason, as mentioned above, our experiments were carried out using video lectures reviewed in the previous phases, that accounted for up to approximately 25 minutes of audio data per lecturer. This amount of supervised data can be efficiently generated beforehand for each speaker using the conventional post-editing strategy in almost any real-life scenario, and then exploited in the application of a two-step supervision strategy in the subsequent videos of the same speaker.

In this phase, the best outcomes of both previous phases were successfully combined to obtain error-free end transcriptions at a lower RTF on the part of the lecturers, using a minimum amount of supervised data generated beforehand to perform massive adaptation.

Finally, note that the two-step supervision implied that lecturers have to put time aside on two separate occasions to review the same video. However, lecturers preferred to carry out the review process in a single step rather than in two steps (*sig*=0.06). This fact was reflected on the average score of the user satisfaction surveys (M=7.8, S=2.0), shown in Table 9. For this reason, the two-step strategy was less preferred by lecturers than the post-editing strategy.

Table 9: Detailed results from the satisfaction survey for the two-step review strategy.

Question	Mean
Intuitiveness	
1- I am satisfied with how easy it is to use this system.	7.5
2- It was easy to learn to use this system.	8.6
3- The help information of this system is clear.	8.5
4- The organization of information on screen is clear.	8.7
Grand Mean	*8.3*
Likeability	
5- I feel comfortable using this system.	7.3
6- I like using the interface of this system.	7.4
7- Overall, I am satisfied with this system.	7.4
Grand Mean	*7.4*
Usability	
8- I can complete effectively my work using this system.	7.7
9- I can complete my work quicker than by doing it from scratch.	7.4
10- This system has all the functions that I expect to have.	7.1
Grand Mean	*7.4*
Overall Mean	*7.8*

4. Discussion and Conclusions

Provided that the review of automatic transcriptions was more efficient than generating them from scratch, alternative user interaction strategies were explored to generate subtitles from automatic transcriptions as efficiently and comfortably as possible for our lecturers (Nanjo and Kawahara (2006)). First of all, we determine that WER was the main factor involved in explaining the values of RTF. Indeed, the linear regression model derived from our data seems to generalise appropriately for transcriptions with higher WER scores than those reported here. However, it should be noted that this is a limitation of our study, since our WER figures for all video transcriptions tend to be in the range from 20 to 25.

In line with Luz et al. (2008), more sophisticated user interfaces alone, like our intelligent interaction strategy, were not proven more efficient in terms of WER decrease per RTF unit than conventional post-editing, nor were they preferred by lecturers over the simple (though more time-costly) interaction model. We find it particularly noteworthy how important it was for lecturers to be able to produce high quality (perfect) end transcriptions, prioritising this over any time-savings afforded by the more intelligent strategies (Munteanu et al. (2006); Pan et al. (2010); Favre et al. (2013)): a half of our lecturers reverted to the conventional post-editing model to complete the review of their video transcriptions.

Nevertheless, the combination of intelligent interaction with massive adaptation techniques led to statistically significant savings in user effort in comparison to intelligent interaction and to the conventional post-editing strategy when sufficient adaptation data is available. This conclusion differs from that of Luz et al. (2008) mainly because a greater amount of adaptation data has been used in our study to effectively perform the adaptation of acoustic and language models.

Our study analyses the learning curve primarily observed in the third phase as a result of lecturers having worked with the transLectures player in previous phases. WER decrease per RTF unit was statistically significantly less pronounced in the first phase than in the second step of the third phase. In this respect, Figure 5 shows the evolution of RTF as a function of WER across the three phases. It should be noted that the data points (video transcription reviews) of the second phase correspond to those lecturers that declined to use intelligent interaction and switch back to the conventional post-editing strategy. Data points of the third phase in Figure 5 are those obtained in the second step of that phase. As observed in the linear adjustment to the data points at each phase, as lecturers gain experience at reviewing transcriptions, their RTF figures improve phase-on-phase.

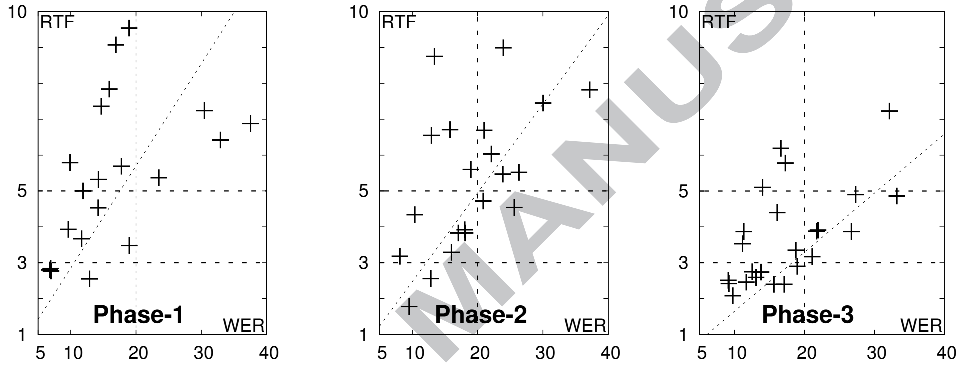


Figure 5: Evolution of RTF as a function of WER in the post-editing mode across the three phases. Data points of the second phase correspond to those lecturers that declined to use intelligent interaction and switch back to the conventional post-editing strategy. Data points of the third phase are those obtained in the second step of that phase.

Our study reveals statistically significant savings in user effort in the two-step strategy when compared to the post-editing strategy of the first phase. Intelligent interaction plus massive adaptation as a preliminary step brought significant improvements in WER to the table, that cannot solely be explained by the effect of learning curve. All in all, to our surprise, lecturers preferred the simple "one-step" post-editing strategy over the sophisticated two-step strategy.

In terms of future work, we will address some of the limitations of this study. First, alternative variants of intelligent interaction strategies which, while allowing lecturers full control over transcription quality, are better able to exploit confidence measures and visual representation (Luz et al. (2008, 2010)). Second, the most suitable interface design for transcription review could be determined on a case-by-case basis, perhaps as a function of WER. In this scenario, transcriptions with low error rates would be reviewed using an interface that focused user attention on the few words that need correcting, while a conventional post-editing interface would be loaded for transcriptions with higher error rates. However, we also believe that interface design preferences are conditioned by the user profile of our participants. As discussed, lecturers required full control over the final transcription quality, but students or casual users involved in the review process may prioritise the time devoted to review over the transcription quality. This is specially true when dealing with long video (over 30 minutes) since, as described in the second phase, the possibility of targeting only those segments that have been probably misrecognised becomes more appealing and necessary provided the limited review effort that students or casual users can devote. This latter user profile is better targeted by (Serrano et al. (2013)). A detailed study of transcription review by students or casual users of longer videos is left as future work. In addition, the improvement of the baseline ASR system incorporating HMM/DNN hybrid technology (Dahl et al. (2012)) will clearly provide higher quality transcriptions, that will further reduce user effort. Lastly, the review of translations generated from the reviewed transcriptions opens an interesting area of study that might also be taken up in future research (Casacuberta et al. (2009)).

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures) and ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no 621030 (EMMA), and the Spanish MINECO Active2Trans (TIN2012-31723) research project.

References

Allen, I. E., Seaman, J., 2010. Class differences: Online education in the United States, 2010. Babson Survey Research Group.

Barras, C., Geoffrois, E., Wu, Z., Liberman, M., 2001. Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication 33 (12), 5 – 22.

Bazillon, T., Esteve, Y., Luzzati, D., 2008. Manual vs assisted transcription of prepared and spontaneous speech. In: LREC.

Casacuberta, F., Civera, J., Cubel, E., Lagarda, A. L., Lapalme, G., Macklovitch, E., Vidal, E., Oct. 2009. Human interaction for high-quality machine translation. Communications of the ACM 52 (10), 135–138.

Dahl, G., Yu, D., Deng, L., Acero, A., January 2012. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20 (1), 30–42.

Deng, L., Li, X., May 2013. Machine learning paradigms for speech recognition: An overview. IEEE Transactions on Audio, Speech, and Language Processing 21 (5), 1060–1089.

Dufour, C., Toms, E. G., Lewis, J., Baecker, R., 2005. User strategies for handling information tasks in webcasts. In: Proc. of ACM SIGCHI. pp. 1343–1346.

Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S., et al., 2013. Automatic human utility evaluation of ASR systems: does WER really predict performance? In: Proc. of Interspeech. pp. 3463–3467.

Fujii, A., Itou, K., Ishikawa, T., 2006. Lodem: A system for on-demand video lectures. Speech Communication 48 (5), 516 – 531.

Gales, M. J., 1998. Maximum likelihood linear transformations for hmm-based speech recognition. Computer speech & language 12 (2), 75–98.

Glass, J., Hazen, T. J., Cyphers, S., Malioutov, I., Huynh, D., Barzilay, R., 2007. Recent progress in the MIT spoken lecture processing project. In: Proc. of Interspeech 2007. Vol. 3. pp. 2553–2556.

Guo, P. J., Kim, J., Rubin, R., 2014. How video production affects student engagement: An empirical study of mooc videos. In: Proc. of Learning at Scale. pp. 41–50.

Hakkani-Tur, D., Riccardi, G., Gorin, A., 2002. Active learning for automatic speech recognition. In: Proc. of ICASSP. Vol. 4. pp. 3904–3907.

Hazen, T. J., 2006. Automatic alignment and error correction of human generated transcripts for long speech recordings. In: Proc. of Interspeech 2006.

Kolkhorst, H., Kilgour, K., Stüker, S., Waibel, A., 2012. Evaluation of interactive user corrections for lecture transcription. In: Proc. of IWSLT. pp. 217–221.

Kristjansson, T., Culotta, A., Viola, P., McCallum, A., 2004. Interactive information extraction with constrained conditional random fields. In: Proc. of AAAI. Vol. 4. pp. 412–418.

Lewis, D. D., Catlett, J., 1994. Heterogeneous uncertainty sampling for supervised learning. In: In Proc. of ICML. pp. 148–156.

Lewis, J. R., Jan. 1995. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. International Journal of Human-Computer Interaction 7 (1), 57–78.

Luz, S., Masoodian, M., Rogers, B., 2010. Supporting Collaborative Transcription of Recorded Speech with a 3D Game Interface. In: Proc. of KES: Part IV. pp. 394–401.

Luz, S., Masoodian, M., Rogers, B., Deering, C., 2008. Interface design strategies for computer-assisted speech transcription. In: Proc. of OZCHI. pp. 203–210.

Lyons, A., Reysen, S., Pierce, L., 2012. Video lecture format, student technological efficacy, and social presence in online courses. Computers in Human Behavior 28 (1), 181 – 186.

Martínez-Villaronga, A., del Agua, M., Andrés-Ferrer, J., Juan, A., 2013. Language model adaptation for video lectures transcription. In: Proc. of ICASSP. pp. 8450–8454.

Munteanu, C., Baecker, R., Penn, G., 2008. Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts. In: Proc. of ACM SIGCHI. pp. 373–382.

Munteanu, C., Baecker, R., Penn, G., Toms, E., James, D., 2006. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In: Proc. of the ACM SIGCHI. pp. 493–502.

Nanjo, H., Kawahara, T., 2006. Towards an efficient archive of spontaneous speech: Design of computer-assisted speech transcription system. The Journal of the Acoustical Society of America 120 (5), 3042–3042.

Nielsen, J., Jan. 1999. User interface directions for the web. Communications of the ACM 42 (1), 65–72.

Nielsen, J., Levy, J., 1994. Measuring usability preference vs. performance. Communications of the ACM 37 (4), 66–75.

Pan, Y., Jiang, D., Yao, L., Picheny, M., Qin, Y., 2010. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In: Proc. of the ACM SIGCHI. pp. 1725–1734.

Papadopoulos, M., Pearson, E., 2012. Improving the accessibility of the traditional lecture: an automated tool for supporting transcription. In: Proc. of BCS-HCI. British Computer Society, pp. 127–136.

poliMedia, 2007. The polimedia tool. `http://polimedia.blogs.upv.es/?lang=en`.

Repp, S., Groß, A., Meinel, C., 2008. Browsing within lecture videos based on the chain index of speech transcription. IEEE Transactions on Learning Technologies 1 (3), 145–156.

Riccardi, G., Hakkani-Tur, D., 2005. Active learning: theory and applications to automatic speech recognition. IEEE Transactions on Speech and Audio Processing 13 (4), 504–511.

Ross, T., Bell, P., 2007. "No significant difference" only on the surface. International Journal of Instructional Technology and Distance Learning 4 (7), 3–13.

Sanchez-Cortina, I., Serrano, N., Sanchis, A., Juan, A., 2012. A prototype for interactive speech transcription balancing error and supervision effort. In: Proc. of ACM IUI. pp. 325–326.

Serrano, N., Giménez, A., Civera, J., Sanchis, A., Juan, A., 2013. Interactive handwriting recognition with limited user effort. International Journal on Document Analysis and Recognition, 1–13.

Silvestre, J. A., et al., 2012. translectures. In: Proc. of IberSPEECH 2012. pp. 345–351.

Soong, S. K. A., Chan, L. K., Cheers, C., Hu, C., 2006. Impact of video recorded lectures among students. Who's learning, 789–793.

Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP. pp. 257–286.

Suhm, B., Myers, B., Waibel, A., Mar. 2001. Multimodal error correction for speech user interfaces. ACM Transactions on Computer-Human Interaction (TOCHI) 8 (1), 60–98.

The TransLectures-UPV team, 2013. The TransLectures UPV toolkit (TLK). `http://www.translectures.eu/tlk`.

Valor Miró, J., Spencer, R., Pérez González de Martos, A., Garcés Díaz-Munío, G., Turró, C., Civera, J., Juan, A., 2014. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. Open Learning: The Journal of Open, Distance and e-Learning 29 (1), 72–85.

Valor Miró, J. D., Pérez González de Martos, A., Civera, J., Juan, A., 2012. Integrating a State-of-the-Art ASR System into the Opencast Matterhorn Platform. In: Advances in Speech and Language Technologies for Iberian Languages. Vol. 328 of CCIS. Springer, pp. 237–246.

Wald, M., 2006. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. Interactive Technology and Smart Education 3 (2), 131–141.

Wessel, F., Schluter, R., Macherey, K., Ney, H., 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 9 (3), 288–298.

Zhang, D., Zhou, L., Briggs, R. O., Jr., J. F. N., 2006. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. Information and Management 43 (1), 15 – 27.

Highlights

Real-life evaluation of automatic transcriptions in a large videolecture repository.
Three different evaluation protocols are compared to minimise user supervision time.
Discovered dependencies between transcription quality and time expended in supervision.
Attained up to 70% of time reduction with supervisions compared with transcribe from scratch.
Lecturers valued simplicity over other sophisticated but more-efficient protocols.