UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTMENT OF COMPUTER SYSTEMS AND COMPUTATION

MASTER THESIS

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Assisted Transcription of Video Lectures.

Master in Artificial Intelligence, Pattern Recognition and Digital Image.

Juan Daniel Valor Miró

Directed by:
Dr. Jorge Civera Saiz
Dr. Alfons Juan Ciscar

September 10, 2013

Thanks to everyone who have helped me
during the course of this master thesis.

Contents

Chapter 1
Introduction

1.1 Work Motivation

Online multimedia repositories are rapidly growing and becoming evermore consolidated as key knowledge assets. This is particularly true in the area of education, where large repositories of video lectures are being established. In this line, the *Universitat Politènica de València* (UPV) implemented its poliMedia platform system for the cost-effective creation and publication of quality educational video lectures [pol07]. It now has a collection of more than 9000 video lectures created by more than 1200 professors.

It is important to note that some studies about the usage of these video lectures has been performed ([LSCN13]), and all of them remark the importance of having transcriptions of the video lectures ([FII06]) for different purposes. For instance, searching on the video lecture content ([RGM08]), translate it to achieve greater scope, or facilitate the access for people with disabilities ([Wal06]).

In the present work we study the process of supervision the transcriptions for video lectures generated by an automatic speech recognition system. These supervisions have been performed by different professors, and have been duly studied to improve the time and quality of the supervision process. These supervisions were performed under the *Docencia en Red* action plan to boost the usage of digital resources at the UPV university.

The goal of this master thesis is to evaluate models, tools and the integration progress of the transLectures European project [SdAG^{+}12] in a real-life yet controlled setting. Other important goal of the present work, is to develop the best possible interaction model and user interface experience, to make easy and reduce the time of the computer-professor interaction [LMR08].

The rest of the master thesis is as follows: First, in this section we introduce some basic concepts of our work, then in Section 2 we introduce the poliMedia platform where video lectures belong, next we present our *Automatic Speech Recognition System* in Section 3, also the Web Player used for the evaluations will be presented in Section 4. The most important part of this master thesis will be explained in Sections 5 and 6, and are about the evaluation with the professors and the data and results obtained. As we will explain in detail, these evaluations were performed on three different phases, in which three different interaction models were evaluated. Finally, in Section 7 some conclusions will be draw. It is important to note that a detailed statistics of all the evaluations performed are described in the appendices A and B.

1.2 Evaluation Metrics

To carry out the performance evaluations of our system, we must use the appropriate metrics when try to measure the supervision performed by the professors. Then, based on such metrics we can perform improvements to our system in order to obtain a better supervision models. For this reason, we will use two well-spread metrics that provide a reliable measurement capability: one to measure the errors in the transcriptions, and another to measure the relative time spent on the supervisions.

The metric used to measure the errors in the transcriptions is the *Word Error Rate* or WER. In order to compute this metric we need a reference transcription (with the correct content) and the inaccurate transcription from which we need to calculate the WER. This metric (E) is computed as the number of insertions (n_i), deletions (n_d) and substitutions (n_s) between these two transcriptions divided by the number of words in the reference (n_r) as we can observe on the next equation.

$$E = \frac{n_s + n_i + n_d}{n_r} \tag{1.1}$$

On the other hand, the metric used to measure the time spent by the professors on the supervision is the *Real Time Factor* abbreviated as RTF. This measure (R) takes the time to process the input (P) and the duration of the video lecture (T), and is defined as the ratio of these two values.

$$R = \frac{P}{T} \tag{1.2}$$

Finally, it is important to note that secondary metrics as the number of times listened each segment or the seconds needed to correct one word, were used on this document but of its simplicity, these metrics not need an introduction.

Chapter 2

poliMedia

2.1 poliMedia Platform

The poliMedia platform is a recent, innovative service for the creation and distribution of multimedia educational content at the UPV [pol07]. It is designed to allow UPV professors to record and publish their own video lectures, accompanied by time-aligned slides. It serves to more than 36000 students and 2800 professors. poliMedia started on the 2008 and has been already exported to several universities. The UPV repository has 9222 lectures recorded by more than 1302 speakers, like we can see on table 2.1.

Tables 2.1: Basic statistics of the complete poliMedia repository

Number of lectures	9222
Duration (in hours)	2102
Avg. lecture length (in minutes)	13
Number of speakers	1302
Avg. Lectures per Speaker	7

The production process of the poliMedia platform has been carefully designed to achieve a high rate of production with high quality. The poliMedia studio is a 4 meter room with a white background in which all the necessary equipment for the recording is available for the professors. The studio during a recording session can be observed at Figure 2.1.

To record a Video Lecture the speakers are requested to come to the studio with their slides. The speaker performs the presentation, while the computer's screen and the speaker are recorded at the same time. Finally, a post-process is applied to the raw recordings and the final poliMedia video is generated.

All the video lectures generated on the poliMedia platform, follow a standard format, representative of the platform. This format is a join view of the professor, and the computer screen with time-aligned slides, as we can observe at Figure 2.2.

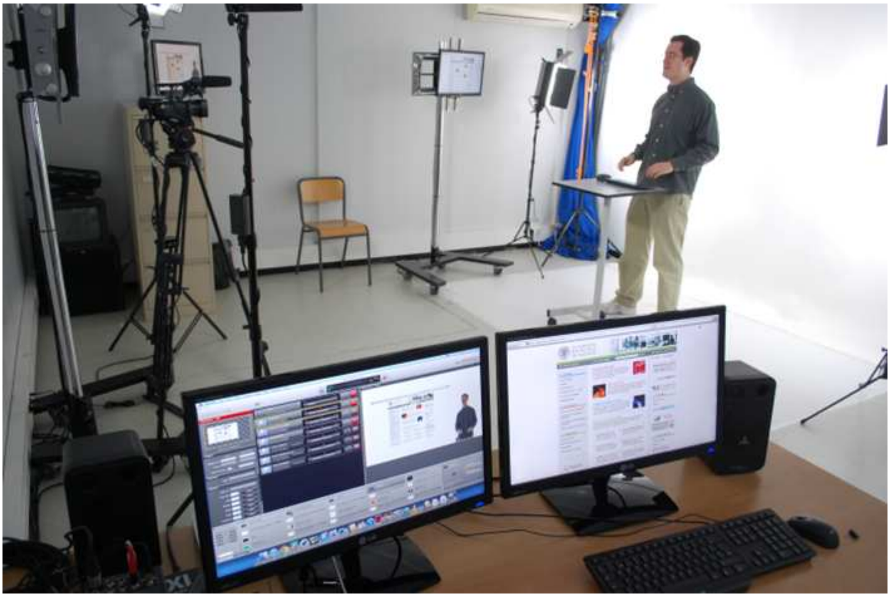


Figure 2.1: The poliMedia recording studio during a recording session

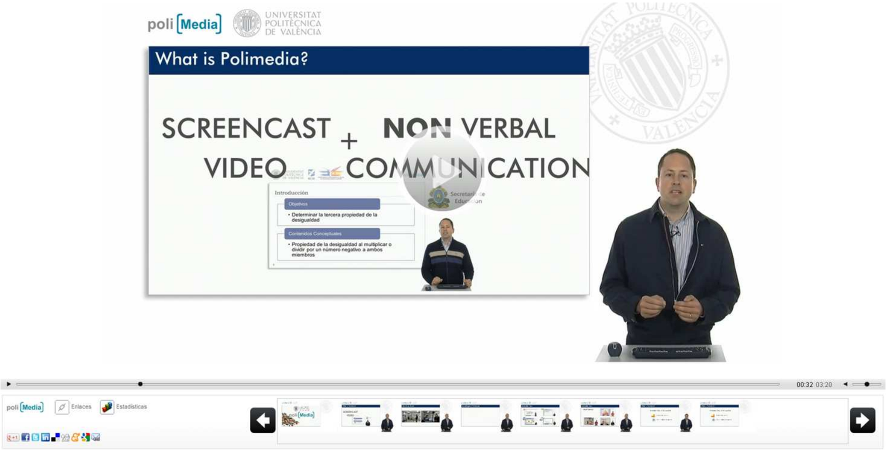


Figure 2.2: The final Video Lecture after a poliMedia recording session

Finally, it is important to note, that the poliMedia repository has been automatically transcribed using an Automatic Speech Recognition System (abbreviated as ASR) developed by the transLectures project [SdAG+12] and presented in Section 3.

2.2 poliMedia Corpus

In order to automatically transcribe all the poliMedia repository, our ASR systems needs relevant sample data of the repository. To this purpose, 114 hours of Spanish poliMedia video lectures were manually transcribed. This manually transcribed data were partitioned into training, development and test sets in order to train, tuning and evaluate our ASR systems. The statistics on these three sets are shown in Table 2.2.

Tables 2.2: Statistics of the training, development and test partitions

	Training	Development	Test
Videos	655	26	23
Speakers	73	5	5
Hours	107h	3.8h	3.4h
Sentences	39.2K	1.3K	1.1K
Words	936K	35K	31K
Vocabulary	26.9K	4.7K	4.3K

Furthermore, it is important to note that we have opted for some conventions when making these manual transcriptions, in order to adequately represent the oral disfluencies in the speech. These conventions are the follows:

1. We identify the professor on each video, as well as their gender.

2. The audio is segmented into time-aligned segments with their transcription.

3. When occurs a disfluency, it is transcribed using the following standard notation: /sound pronounced/correct word/

4. When there is a long silence a segment with the transcription [background sound] was created.

5. When occurs a short silence we add the /SF// to the transcription text.

The laborious task of manually transcribe these video lectures has an average RTF of 10, like the obtained in important publications [MPZ09] of this topic. Examples of the result transcription can be observed in Table 2.3.

Tables 2.3: Some transcribed segments extracted form the corpus

Seg.	Length	Transcription
1	3.1s	El bucle for /ich/each/ es un bucle que itera una lista.
2	1.9s	Por ejemplo, la variable entera /jota/J/.
3	4.0s	Se trata /e//, de una sentencia realmente útil.
4	5.4s	No olvideis /que que/que/ no existe en /ce/C/.

Chapter 3

Automatic Speech Recognition System

3.1 Recognition System

The Automatic Speech Recognition System used to generate the transcriptions of the poliMedia Video Lecture Repository [pol07] was developed under the transLectures European project [SdAG$^+$12]. The system is called transLectures-UPV toolkit [tUT13] abbreviated as TLK, and uses the state-of-the-art techniques of Pattern Recognition and Statistical Machine Learning [YEK$^+$02].

The process of generating the transcriptions from a video with TLK consists of five steps, as seen in Figure 3.2: audio preprocessing, extracting and segmentation; feature extraction and normalization based on the previous audio segments; an statistical speech recognition with basic models; a feature transformation with the *Constrained Maximum Likelihood Linear Regression* or CMLLR [FLBG07] technique; and a second speech recognition with speaker adapted models and the transformed features. These last two steps are part of the unsupervised massive adaptation that will be explained in Section 3.2.

The audio preprocessing consists in extracting the audio from video and convert it to a WAV format with 16KHz of frequency and a mono channel. Then, an audio segmentation process is carried out with *Hidden Markov Models* [Rab89] models trained with two classes (speech and silence) at frame level. Finally, the consecutive frames of speech type are cut off the audio signal, and passed to next step in order to extract features.

Feature extraction of audio segments is a fast process with a RTF lower than 0.1. Basically, *Mel Frequency Cepstral Coefficients* (MFCC) [RJ93] are extracted from audio and normalized by mean 0 and variance 1. Of course, the same process is applied to the training acoustic features, used to train the acoustic models needed in the next recognition step.

The speech recognition process is the most complex of our system, and uses previ-

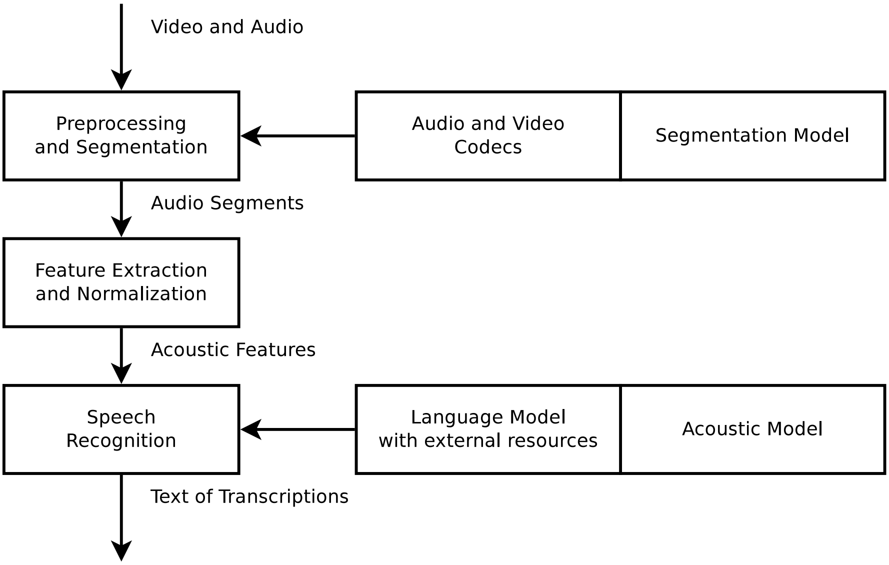


Figure 3.1: Automatic process to obtain transcriptions from a video lecture

ously trained acoustic and language models on the TLK to obtain the transcription of the audio segments passed as input. The acoustic model is a Hidden Markov Model (HMM) with 3 states for each triphoneme (group of three phonemes supposedly pronounced) which emits with a mixture of Gaussian models (for learning the variations in pronunciation). On the other hand, the language model of our baseline was trained with the poliMedia training partition, but after, we improve it using an interpolation with external resources like Google N-Grams [MSA+].

3.2 Massive Adaptation

Massive adaptation of general-purpose ASR models can be performed on the basis of video variables, such as speaker, topic, time-aligned slides (if are available) or previously supervised transcriptions. In order to make this adaptation some techniques on the state-of-the-art are used by our TLK toolkit. Basically we perform two types of adaptation: the unsupervised speaker adaptation and the supervised adaptation of speaker and topic.

The unsupervised speaker adaptation step is performed by transform the acoustic features with the CMLLR [FLBG07] technique. In order to make this linear transformation, a matrix is trained with the basic acoustic features and the result of the first speech recognition. Then, with this matrix we transform the acoustic features and generate new features that will be used on a second speech recognition process.

The final step is a second speech recognition process similar to the previous one, but using an acoustic model trained before with the transformed features. The language model remains the same of the basic recognition, because no topic adaptation is performed now. The result of this second recognition is the transcription of the input video, that we employ on all the phases of the evaluation.

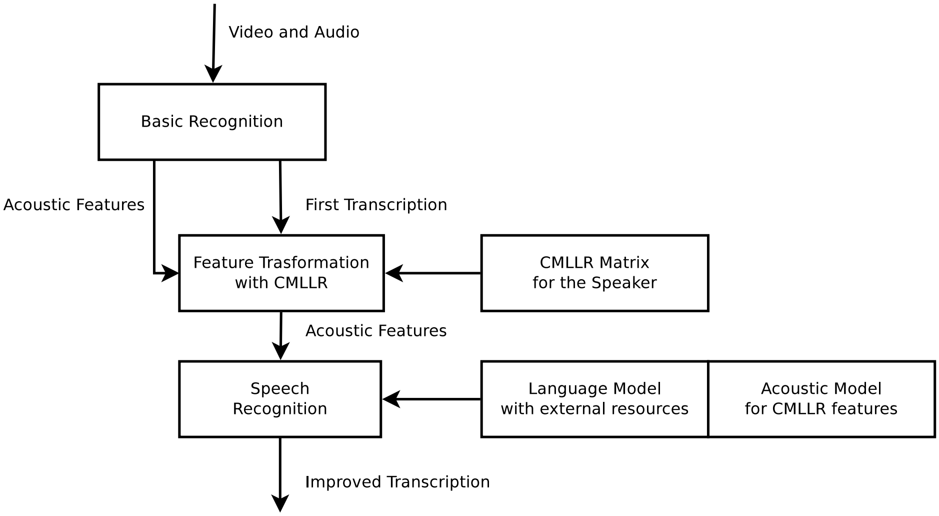

Video and Audio
Basic Recognition
Acoustic Features
First Transcription
Feature Trasformation with CMLLR
CMLLR Matrix for the Speaker
Acoustic Features
Speech Recognition
Language Model with external resources
Acoustic Model for CMLLR features
Improved Transcription

Figure 3.2: Unsupervised speaker adaptation performed to all transcriptions

On the one hand, for the supervised speaker adaptation the *Maximum Likelihood Linear Regression* or MLLR [Gal98] are performed on our acoustic models. This adaptation uses the previously supervised transcriptions of the same speaker, and the time-aligned audio of them, in order to adapt the acoustic model to the speaker.

On the other hand, for the supervised topic adaptation we improve the language model using the previously supervised transcriptions of the same topic, topic related documents automatically extracted on the web, and time-aligned slides (if are available) [MVdAAFJ13]. With this extra and topic-related data, we perform a new language model which will be interpolated with the generic model to adapt it to our particular topic.

It is important to note that the supervised adaptation of speaker and topic is used only on the third phase of our evaluation, as we explain in Section 5.4.

3.3 Intelligent Interaction

The intelligent interaction [SGC+13] is a user-computer interaction model used to supervise the transcriptions, designed to improve the commonly used manual interaction. The basis of this intelligent interaction mode are the word confidence measure [SJV12], which are a way to measure the correctness of one word recognised by our classifier. Ideally, in an automatic speech recognition system not accurate, like ours, the words with lower confidence measure shall be the incorrect.

In our experiments on the poliMedia corpus, we have concluded that these confidence measures correctly detects the 40% of all the incorrectly recognised words. Considering that only one in five words is wrong these confidence measures help us to improve the random detection, increasing the rate of accuracy form the 20% to 40%.

Returning to the intelligent interaction, comment that this model is about jump across the words with lower confidence measure, allowing to the user listen and supervise only these words (in some cases with a bit context added to improve the comprehension). In an ideal case, this strategy optimally corrects all incorrect words, avoiding to the user listen and supervise the correct parts of the transcription generated automatically.

Once we made a supervision with the intelligent interaction, we can re-transcribe the video using the supervised words to achieve even higher quality on the transcription, using a constrained recognition with the TLK. The constrained search of the ASR system [SGC$^+$13] allow to preserve such supervisions and improve the context of these, when force the ASR system to recognize them in the proper position.

3.4 System Evaluation

In order to evaluate the quality of the transcriptions generated automatically by the TLK, we perform exhaustive scientific experiments [SCPGdMJ$^+$13], evaluated in terms of WER. Our baseline system is composed with a basic language model without external resources, and only the first recognition process. Then, two improvements adding external resources to the language model and the CMLLR speaker adaptation for the acoustic, were added to our baseline system.

Tables 3.1: Evolution of the transcription quality on the poliMedia corpus

System	WER
Baseline	36.0
+External Resources	30.3
+Speaker Adaptation (CMLLR)	24.6

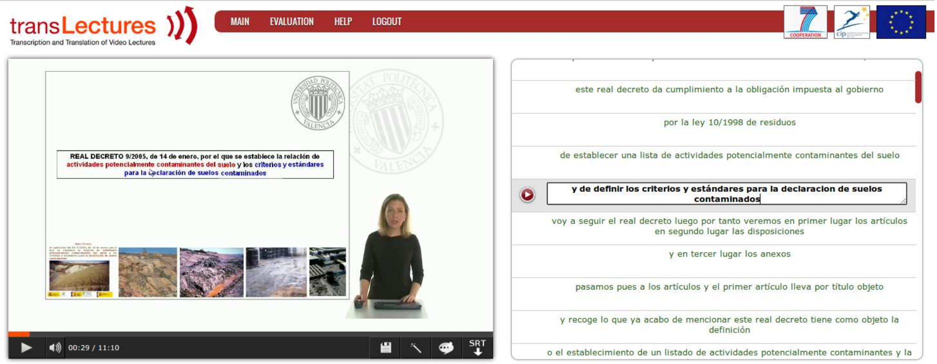
As we can observe in Table 3.1, our baseline system achieves 36.0 WER points using the poliMedia resources, that is significantly improved to 30.3 WER points when external linguistic resources are employed on the language model. The application of unsupervised adaptation techniques (CMLLR) produces a notable improvement on the transcription quality decreasing the WER to 24.6. Finally, is important to note that the RTF used to automatically transcribe one video with the TLK is about 4 in one processor, but as we use distributed computation this RTF is reduced in inverse proportion to the number of cpu's.

Chapter 4
Web Player

4.1 Player Interface

A HTML5 video player and transcription editor has been carefully designed [VMGdMCJ12] for their supervision task in order to obtain cost-effective captions of an acceptable quality in exchange for a minimum amount of user effort.

Three alternate editing layouts are available for users, to choose according to their personal preferences. The default layout is an horizontal editing interface with the video playback on the left and transcription editor on the right. The vertical layout (with the video playback on the top and transcription editor on the bottom) and the subtitle-view layout (with the transcription editor on the video) are the other two layouts available on our player. The default horizontal layout is shown at Figure 4.1.



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

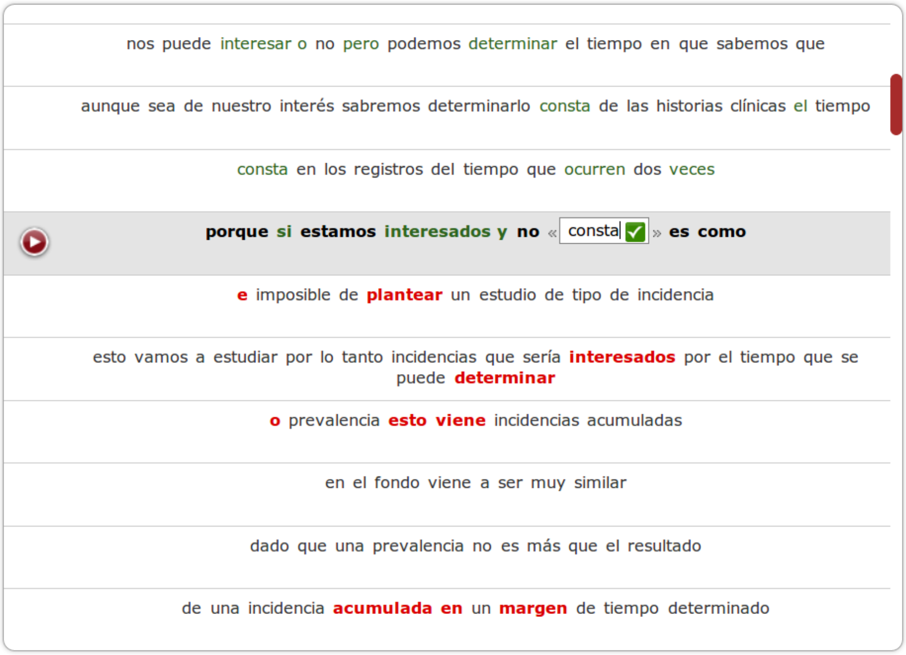
Figure 4.1: Player interface for batch interaction with horizontal layout

Additionally, a complete set of key shortcuts has been implemented to enhance expert user capabilities. Other helpful features are being continuously added to the editor in response to the user feedback.

4.2 Intelligent Interaction

Our web player has a complete implementation of the Intelligent Interaction for cost-effective supervision. Basically, as shown at Figure 4.2, this interaction mode mark in red the words with low confidence measure, and jump across that words to allow to the user perform the supervision of that words.

The rest of the player remains as explained in Section 4.1 in order to make the interface user-friendly. Also, new key shortcuts has been implemented to full support the keyboard supervision with the keyboard.



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 4.2: Player interface for the edition box with the intelligent interaction

The user can customize the time that want to spend on the supervision and the context words that need to understand correctly the asked word. With this values the system calculates the percentage of words with lower confidence measure that the user will supervise, and starts the supervision automatically.

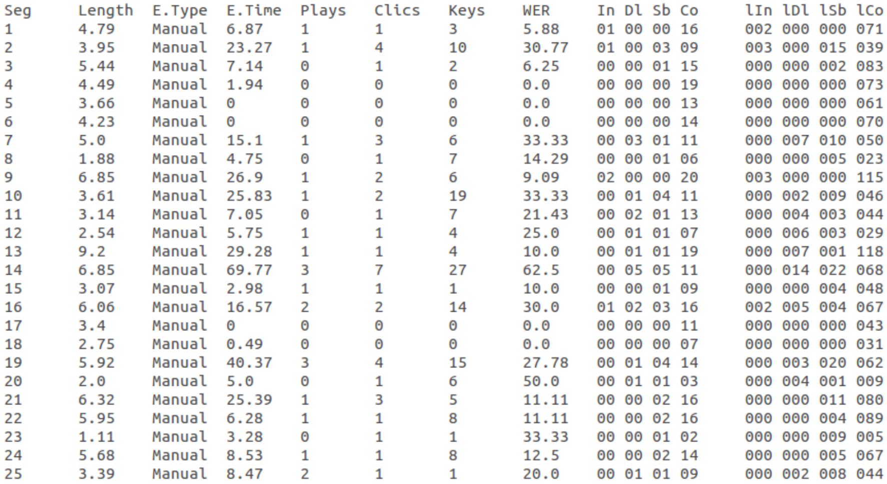
When the supervision finish, if the quality of the resulting transcription is not as accurate as the user would like, the previous parameters can be changed and the web player will automatically recalculate the words to supervise. This can be done as many times as the user want until all words become supervised.

4.3 Statistics Collector

A very important part of our HTML5 player is the Statistics Collector, that will collect the usage statistics of the supervision without interfering with the user interaction and interface. The data are collected at three levels: global, segment and word.

At global level we extract some relevant data such the supervision time, the global WER of the video after the supervision, the layout used on the supervision, the number of function keys and clicks performed, the number of segments supervised, and other complementary data.

At segment level we have the supervision time, the number of words corrected and correct, the number of clicks and keys performed, the number of times listened the segment, the WER of the segment, the supervision type, and some complementary data. Ultimately, at word level, we obtain the supervision time, if the word has changed or not, the number of times listened, and the number of clicks and keys performed.



```
Seg     Length  E.Type  E.Time  Plays   Clics   Keys    WER     In Dl Sb Co     lIn lDl lSb lCo
1       4.79    Manual  6.87    1       1       3       5.88    01 00 00 16     002 000 000 071
2       3.95    Manual  23.27   1       4       10      30.77   01 00 03 09     003 000 015 039
3       5.44    Manual  7.14    0       1       2       6.25    00 00 01 15     000 000 002 083
4       4.49    Manual  1.94    0       0       0       0.0     00 00 00 19     000 000 000 073
5       3.66    Manual  0       0       0       0       0.0     00 00 00 13     000 000 000 061
6       4.23    Manual  0       0       0       0       0.0     00 00 00 14     000 000 000 070
7       5.0     Manual  15.1    1       3       6       33.33   00 03 01 11     000 007 010 050
8       1.88    Manual  4.75    0       1       7       14.29   00 00 01 06     000 000 005 023
9       6.85    Manual  26.9    1       2       6       9.09    02 00 00 20     003 000 000 115
10      3.61    Manual  25.83   1       2       19      33.33   00 01 04 11     000 002 009 046
11      3.14    Manual  7.05    0       1       7       21.43   00 02 01 13     000 004 003 044
12      2.54    Manual  5.75    1       1       4       25.0    00 01 01 07     000 006 003 029
13      9.2     Manual  29.28   1       1       4       10.0    00 01 01 19     000 007 001 118
14      6.85    Manual  69.77   3       7       27      62.5    00 05 05 11     000 014 022 068
15      3.07    Manual  2.98    1       1       1       10.0    00 00 01 09     000 000 004 048
16      6.06    Manual  16.57   2       2       14      30.0    01 02 03 16     002 005 004 067
17      3.4     Manual  0       0       0       0       0.0     00 00 00 11     000 000 000 043
18      2.75    Manual  0.49    0       0       0       0.0     00 00 00 07     000 000 000 031
19      5.92    Manual  40.37   3       4       15      27.78   00 01 04 14     000 003 020 062
20      2.0     Manual  5.0     0       1       6       50.0    00 01 01 03     000 004 001 009
21      6.32    Manual  25.39   1       3       5       11.11   00 00 02 16     000 000 011 080
22      5.95    Manual  6.28    1       1       8       11.11   00 00 02 16     000 000 004 089
23      1.11    Manual  3.28    0       1       1       33.33   00 00 01 02     000 000 009 005
24      5.68    Manual  8.53    1       1       8       12.5    00 00 02 14     000 000 005 067
25      3.39    Manual  8.47    2       1       1       20.0    00 01 01 09     000 002 008 044
```

Figure 4.3: Statistics file processed and extracted by our Statistics Collector

Of course with this data, we can aggregate, summarize and extract different usage statistics. We use these usage statistics in order to create user models, analyze different usage profiles, and drawing robust conclusions that will improve the whole system.

Chapter 5

Evaluation on Docencia en Red

5.1 Planning of the Evaluation

This study was carried out under the "Docencia en Red" UPV program which aims to encourage the development of learning resources to be used via new technologies. Specifically in the 2013 call, was convened a transcriptions pilot of video learning objects of the poliMedia platform (see Section 2.1).

During the program, professors must supervise and correct the transcriptions of 5 of their poliMedia videos that were generated using the state-of-the-art automatic speech recognition technology presented in Section 3.1. These 5 videos supervised by the professors were distributed in 3 distinct phases, in order to aim improvements on the interface and the interaction models used at the end of each phase.

- First phase: Professors manually supervise the first video lecture. The video automatically plays and transcriptions are displayed in a synchronized manner. When the professors detects an error, they can change the contents of the transcription that is being shown on that time.

- Second phase: We introduce a word level intelligent interaction model, where the system jumps between the words that they consider inaccurate (using word confidence measures). The professors listens and supervise only these words (with a bit of context to make easier the understanding) of the second and third video lectures.

- Third phase: Using manual supervision interface (of the first phase), we ask to professors only for the inaccurate words (like on second phase) of their fourth and fifth video lectures. This supervision is used to automatically re-transcribe better that video lectures. Then, the video lectured are resubmitted to professors for a fast manual supervision.

It is important to note that at the end of each phase, the professors filled a brief satisfaction survey in order to collect subjective information about that phase. The objective information is automatically collected by the transLectures web player, like we explain in Section 4.3.

5.2 First phase: Manual Supervision

In the first phase, 20 professors completely supervise the transcription of their first video lecture using the web player presented in Section 4. Each professor was provided with personal credentials to have access to her own private area showing their video lectures. Once the professor logs into the web player and selects a video, the web player is automatically loaded allowing to start with the supervision of the transcription. The web player shows the video lecture and its corresponding transcription in a synchronised manner, letting the user read the transcription while the video is being played. When the professor detects an error in the transcription, can modify the content of the segment the video by pressing the *Intro* key or clicking on the segment.

A preliminary round of phase one was carried out with two professors, use a draft version of the web player. It was important that the backgrounds of these two professors (computer science and architecture) differed to obtain consistent opinions of the interface usability. The two professors presented very different user interaction patterns, for instance, with the computer science professor the interaction was done primarily using the keyboard, while the architecture professor showed a clear preference for the mouse. Based on the feedback from these first two professors, we were able to improve the web player after the start of the phase one.

The remaining eighteen professors supervise the first of their video lectures using this updated version of the web player. When professor finish the supervision and saves the changes, the new correct transcription overwrites the inaccurate; and some extra information of the professor interaction is saved, as it is explained in Section 4.3.

5.3 Second phase: Intelligent Interaction

In the second phase, a the *intelligent interaction* was introduced in order to evaluate if professors could supervise their transcriptions more efficiently. This new interaction protocol based on the word-level *confidence measures* was presented in Section 3.3. The idea behind intelligent interaction is to focus professor attention on the words more likely to be wrong [SGC$^+$13]. A 10-20% of low confidence words are presented to the professor, and each word is provided with few context words, so they could be easily listened and corrected. Ideally, if perfect *confidence measure* were available, the professor would only need to supervise those words.

In this phase, professors were activated by default the intelligent interaction mode of the user interface (presented in Section 4.2), but they had the option to switch to the manual supervision mode. In our evaluation, only 12 out of 23 professors involved completely finish the supervision of at least one of their video lectures using the intelligent interaction mode. The main reason for professors to abandon this mode was that it could not guarantee to obtain perfect transcriptions, so they decided to switch to the manual supervision.

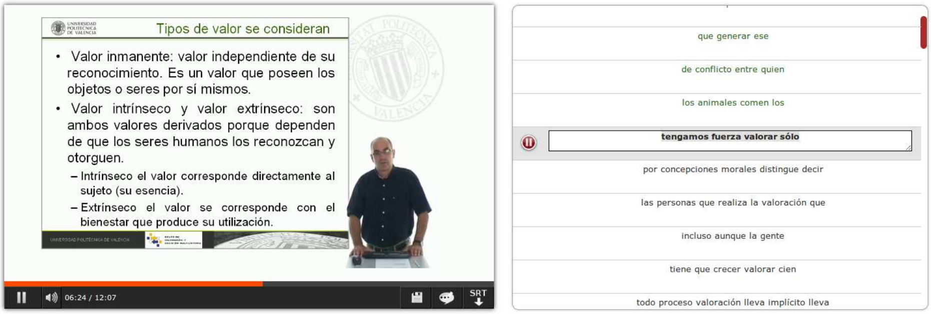
5.4 Third phase: Two-step Supervision

The third phase was divided into two rounds, essentially a combination of the previous two phases. In the first round, 15 professors partially supervise the transcription of two of their videos in a similar manner to the previous phase. Then, the massive adaptation systems presented in Section 3.2 were trained to regenerate the transcriptions with higher quality [SCSSJ12]. Finally, the second round consists in the complete supervision of the regenerated transcriptions as performed in the first phase.

In the first round the professors supervised some segments of four words in which the last word has a low confidence measure. These segments were presented to the professor in increasing order of confidence measure, until one of this three conditions were met:

1. The supervision time reached double the duration of the video.
2. There are not corrections on five consecutive segments.
3. The 20% of words with low confidence measure were supervised.

The supervision process of this first round can be seen in Figure 5.1. Second round is a simple manual supervision of the re-transcribed video lecture like done on first phase. Ideally, the time spent on this supervision should be much lower than the needed on phase 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 POLITECNICA DE VALENCIA
UNIVERSIDAD POLITÉCNICA DE VALENCIA
06:24 / 12:07
SRT

Figure 5.1: Snapshot of the supervision on the first round of phase 3

Chapter 6

Results of the Evaluation

6.1 First phase: Manual Supervision

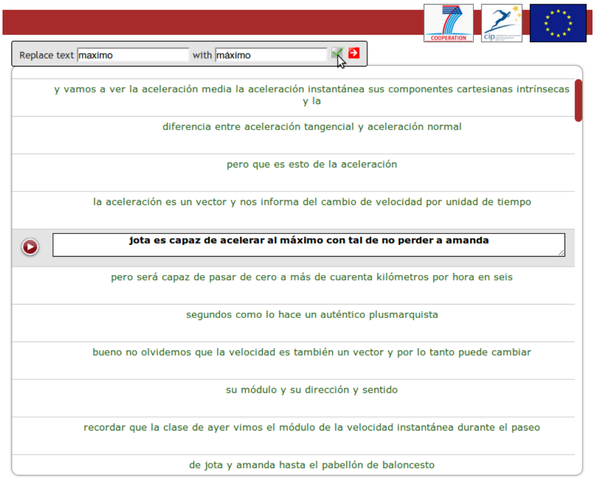
This phase involved 20 professors, who have supervise one of their poliMedia videos as explained in Section 5.2. Note that we draw these results analyzing and processing all the data extracted by the transLectures player from the supervisions.

The first two professors that perform the preliminary round, give us several important conclusions, in order to prepare the first phase. The first important conclusion is the optimum length in words of one transcription segment, which is located in the range of 10-15 words, although it is acceptable within the range 4-20. Above this range the time supervision of a segment increases exponentially, whereas below is the number of reproductions who increases exponentially. This shorter length allows the user to more easily remember what was said in the video and therefore more efficiently correct the words incorrectly recognised by our system. Secondly, a *Search & Replace* function was incorporated into the web player, at the suggestion of our computer science professor. Finally, both professors suggested that transcription segments be automatically validated as soon as the corresponding video segment has been played.

After this preliminary round, the other 18 professors also perform the supervision of their Video Lectures, and many important results derived from such supervisions. Firstly, is very important to note that the average Real Time Factor used to supervise the video lectures was around 5.6, and the average number of times listened each segment was 3. This is large improvement approximately of the 45% compared with the 10 RTF needed to make a transcription with the same quality form scratch (see Section 2.2) [MPZ09].

Also, the average Word Error Rate between the professors supervised transcription and the recognised transcription by our system is 16.9 WER points, which is consistent with the scientific results presented in Section 3.1, and effectively indicate to us that the outcome of the supervisions is a perfect transcription.

Based on the interactions at segment level, we have proposed a model for the professors, which allows to correlating the number of corrected words on the supervision with the time (T) spent on it. The variables involved on this model are the number of



Replace text maximo with máximo

y vamos a ver la aceleración media la aceleración instantánea sus componentes cartesianas intrínsecas y la

diferencia entre aceleración tangencial y aceleración normal

pero que es esto de la aceleración

la aceleración es un vector y nos informa del cambio de velocidad por unidad de tiempo

jota es capaz de acelerar al máximo con tal de no perder a amanda

pero será capaz de pasar de cero a más de cuarenta kilómetros por hora en seis

segundos como lo hace un auténtico plusmarquista

bueno no olvidemos que la velocidad es también un vector y por lo tanto puede cambiar

su módulo y su dirección y sentido

recordar que la clase de ayer vimos el módulo de la velocidad instantánea durante el paseo

de jota y amanda hasta el pabellón de baloncesto

Figure 6.1: The Search & Replace function incorporated into the web player

incorrect (w_i) and correct (w_c) words after the professor supervision on each segment, and the coefficients (a and b) estimated by a linear regression.

$$T = a * w_i + b * w_c \tag{6.1}$$

The results of the linear regression are consistent and clear: the incorrect words takes an average of 4.64 seconds to be supervised with a standard deviation of 0.16, and the correct words takes 1.16 seconds with a standard deviation of 0.05 seconds. The r-squared is 0.82 with p-value smaller than $2.2 * 10^{-16}$, so the proposed model for the interaction professor-computer is representative of the data and population. This means that correct an incorrect recognised word takes 4 times the time spent to supervise a correct word.

The final important conclusion is abut our interface and the professor-computer interaction on this phase, that can be summarized into a 9.1 out of 10 on the satisfaction surveys filled by professors. This means that this interaction mode has been well accepted by professors, and provided very good results in terms of time and quality. The proposals received from the professors reinforce this conclusion, because only small usability details and new features were reported at the end of this phase, as we can see below:

1. Allow to change the font size and color.

2. Allow to download the subtitles.

3. Automatically save the supervisions.

4. Reduce the initial loading time.

6.2 Second phase: Intelligent Interaction

The second phase has been carried out (using the intelligent interaction) by 12 professors with a total of 18 videos, who supervise one or two of their poliMedia videos as explained in Section 5.3. However, 14 professors perform the supervision of 22 of their poliMedia video lectures with the manual interaction mode used on the first phase. The main reason why these professors made this phase using the manual supervision, is because they find that could not leave their transcriptions perfectly corrected, and decided to carry out the supervision as in the first phase. On this phase we extract comparable data with the previous phase, to observe the impact of this new interaction model.

About the extracted data of the intelligent interaction, it is important to note that, the average Real Time Factor has been halved to 2.2, and the average number of times listened each segment remains similar with a value of 2.6. This drastic reduction of the RTF occurs because the professor only needs listen and supervise the words with low confidence, and not all the words of the video lecture as in the previous phase.

The main problem of this interaction model is that does not correct all the errors on the transcription. This issue is reflected in the supervision Word Error Rate which is 8.0, instead of 0.0 after the professor's supervision. In order to compute this WER, on this phase we (the transcription experts) perform a full correction of the transcriptions supervised with the intelligent interaction mode and recalculated the WER at the end, that becomes 14.5. This WER is the original transcription WER obtained by our ASR system. This means that this interaction mode effectively corrects in less than half of the previous phase time, more than the half of the inaccurate words.

About the supervision time at word level, we obtain that the average time to correct one incorrect word is 4.9 seconds, and to supervise one correct word is 3.2. If we compare these results with the extracted on the phase one, we can observe that the time to correct one incorrect word is very similar. However, the supervision time of a correct word is greater due to the interaction model assumed that the word is incorrect (because is marked by the confidence measure), and automatically enters in edition mode that it takes some time to leave. So with this interaction model, the times needed to supervise a correct word and an incorrect word tends to equalize.

The data extracted by the other 14 professors who perform the supervision of 22 video lectures as in the first phase, only confirms all that we presented in Section 6.1: The RTF remains around 5.2 and the WER of the automatic transcriptions is about 19.5, with an average number of times listened each segment about 3.1.

The main conclusion about the use of this interaction model is clear: the professors want their transcriptions to be perfect, while intelligent interaction aims at reducing the number of transcription errors in a limited amount of time. These two objectives are very different so at the end the professors declined this model. On the satisfaction surveys filled by professors this has a clear impact lowering the overall system score to 7.2 of 10. It is important to note that all the proposals done by the professors are in the line of give freedom to the intelligent interaction:

1. Allow to edit words out of the intelligent interaction guide.

2. Unlimited use of the expansion arrows to allow correct the entire segment.
3. Auto validate one correct word all the times that appears on the transcription.
4. Allow to return backward through the intelligent interaction changes.
5. Automatically remove consecutive repeated words before the supervision.

Despite these results, the intelligent interaction model works properly and meets its goal, so we need readjust this model, and focus it to meet the professor's objective: leave the transcription perfectly corrected. This is the line that we follow to design and perform the third phase.

6.3 Third phase: Two-step Supervision

The complete process of this two-step evaluation explained in Section 5.4, has been successfully performed by 15 professors with 26 video lectures of the poliMedia platform. First of all, it is important to note that these videos initially have 28.4 WER, so our starting point is worse than for the other two phases. This has come when we choose the video lectures for the professor's supervisions based on the confidence measure, so the videos better recognized were presented in the first two phases.

On the first round the professors supervised some segments of four words in which the last word has a low confidence measure. The average time spent on this phase is of 1.4 RTF lowering the WER to 25.0, because more strict stop conditions were imposed compared with the phase two, as we explain in Section 5.4. Then, the supervised massive adaptation systems presented in Section 3.2 were trained to regenerate the transcriptions with higher quality, automatically lowering the WER to 18.7. This means that after the first round we obtain an improvement on the transcription of the 34% with only 1.4 RTF spent by the professor.

On the second round a full manual supervision of these re-transcribed video lectures were performed by the professors, leaving the transcriptions perfect (without errors), employing an average of 3.9 RTF. This means that the whole process of supervision with this interaction model and a significantly worse starting point, has resulted in a 5.3 RTF, which is better that the RTF obtained on phase one. A brief summary of the results obtained can be observed on Table 6.1.

Tables 6.1: Summary of results obtained in Two-step Supervision

Step	WER	RTF
Recognised Transcription	28.4	-
First Round	25.0	1.4
Massive Adaptation	18.7	-
Second Round	0.0	3.9

Finally, it must be remarked that professor's comments indicate that although with this two-step supervision the professors perceive a little improvement in time and transcription quality, this improvement is not significant enough to encourage them to perform the supervision in two rounds instead of in one.

1. I spent a similar time reviewing the last two videos and the first received.
2. The transcription has been improved compared with received on first phase.

At the end, despite we reduce a bit the time needed by professors on the supervision task, the survey is not as good as in the first phase because they must do the supervision in two phases, restriction that they dislike. We obtain a mark of 7.8 of 10 for this interaction model on the satisfaction survey.

6.4 Comparison of the Interaction Models

On the previous sections we present the results of the evaluation of three interaction models, but it is important to make a comparison and define the strengths and weaknesses of each model. A brief comparison between all the interaction protocols can be found in Table 6.2, which summarizes the results presented in the previous sections.

Tables 6.2: Comparison between all the interaction models studied

Interaction Model	Initial WER	Final WER	Professor RTF
Transcribe from Scratch	24.6	0.0	10.0
Manual Supervision	16.9	0.0	5.6
Intelligent Interaction	14.5	8.0	2.2
Two-Step Supervision	28.4	0.0	5.3

The manual supervision is the most welcome by professors because is simple, effective and provide perfect transcriptions at low user effort. However, the intelligent interaction model provide very good results with the lower possible user effort, but it has the main drawback that does not allow to leave the transcription perfect. Finally, the two-step supervision allows to leave perfect transcriptions with user effort lower than needed for the manual model, but forces perform the transcription in two sessions separated in time.

Clearly, following this analysis, for the college professor user's profile the most appropriate model is the manual supervision, at least until we adapt the other two interaction models to the specific needs of this role that follow the next requirements:

- Make the interface and process as simple as possible.
- Provide perfect transcriptions after supervision.
- Minimise user time devoted to supervision.

Chapter 7
Conclusions and Future Work

7.1 General Conclusions

In this master thesis we present a full evaluation of the supervision of automatically generated transcriptions of poliMedia Video Lectures, under the *Docencia en Red* UPV action plan and the European transLectures project. First we present the poliMedia platform on which all the Video Lectures belongs. Then we present our Automatic Speech Recognition System and the Web Player used to make the supervisions. Of course, we present and focus on the evaluation with the professors. This evaluation consists of three distinct phases, on which three different interaction protocols were tested and proved with real users, that we listed below:

- Manual supervision.
- Intelligent Interaction.
- Two-Step Supervision.

As previously analysed, professors welcome manual supervisions of their transcriptions for its simplicity and the capability of providing perfect transcriptions with low effort. The intelligent interaction model provide very low user effort, but has the drawback that does not leave the transcription perfect. The two-step supervision provide perfect transcriptions with the lower user effort, but it splits the transcription process in two different sessions.

At the end, our user's profile (college professor) requires making the interface and process simple, let the transcription perfect with the minimum time possible, and make the work in one session with extensive deadlines. With this profile the best model is the manual supervision, as the evaluations confirm; but some changes on the other two interaction models can adapt them to the user's profile, as we propose in Section 7.3.

Finally, it is important to note that our system has been accepted by the professor with the three interaction protocols, as we can observe on the satisfaction surveys. All the interaction models have a mark greater than 7 of 10, pointing out that the manual exceeds the 9 out of 10. Also compared to make the transcription from scratch, our system obtains a perfect transcription in half the time, and a good quality transcription in a quarter of time. So effectively we can conclude that our system has been a hit with professors, achieving all our goals.

7.2 Contributions

As the work presented here is very extensive, it is important to highlight the contributions of this master thesis, which are listed below:

- Feature extraction (MFCC) of the Automatic Speech Recognition System.
- Testing different baselines and configurations of our ASR System.
- Developed the statistics collector of our online web player.
- Planning and execution of the user's evaluation (the three phases).
- Data analysis and knowledge extraction from the evaluations.
- Preparation, writing and submission of associated publications.

The scientific publications related to this work are listed below. The title or authors of the last publication may change slightly util we send it.

- Integrating a state-of-the-art ASR system into the Opencast Matterhorn platform. Juan Daniel Valor Miró, Alejandro Pérez González de Martos, Jorge Civera and Alfons Juan. IberSPEECH 2012, vol. CCIS 328, Springer, p. 237–246, November 2012, Madrid (Spain).
- Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. J.D. Valor Miró, R.N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera and A. Juan. EADTU Conference. Submitted. October 2013, Paris (France).
- Evaluation of innovative interaction protocols for post-editing automatic video lecture transcriptions. Juan Daniel Valor Miró, Jorge Civera and Alfons Juan. Computers & Education. Elsevier. In preparation.

7.3 Future Work

Considering the success of the evaluations, the future work will be defined by different changes and improvements in the intelligent models of interaction with the professors.

On the one hand, future lines of research on the intelligent interaction mode will focus on combining the full control allowed in manual supervision and the use of confidence measures in a way that professors find useful and usable. For instance, an interface where it was possible to switch from complete supervision mode to intelligent interaction mode, depending on the perceived quality of the present transcription segment, might be better received by professors. Also is interesting allow the supervision of words out of the guidance of the intelligent interaction protocol, to obtain the transcription perfectly corrected.

On the other hand, about the two-step interaction, it will be very interesting to merge on the first round words of all the video lectures that the professor must done with this method. This will reduce the number of working sessions to one plus the number of video lectures, instead of two multiplied by the number of video lectures. This evaluation setup is expected to reduce the *Real Time Factor* to supervise automatically generated transcriptions

Finally, another important and interesting research line will be to apply these evaluations to automatically generated translations of previously supervised video lecture transcriptions. As the translation process is more complex, probably the results will be very different. On this line it is important to note that while the correct transcription is unique, there are many possible translations which can be considered correct.

As the reader may guess, all the presented interaction models must be revised and adapted to the translation case study.

Appendix A

Detailed Usage Statistics

A.1 First phase: Manual Supervision

In Table A.1 we can see the detailed results of the first phase describing WER, RTF and the specific video lecture.

Tables A.1: Detailed results of the first phase supervisions

WER	RTF	Video Lecture ID
9.9	5.8	14646d99-d43c-4b4d-8e03-1ff33c550abf
14.3	5.3	173f8edb-916f-6c46-ad43-d5666037dabd
23.5	5.4	1a05718d-14e6-2346-b978-3f1780efd331
19.0	3.5	27974c54-7c48-154d-955e-eca0fb4ac6a8
15.9	7.8	2d3da0fc-54c5-a54f-9a4a-c7317cc698e7
9.6	3.9	30168dfd-1b30-764d-be5c-498987d36338
30.5	7.2	317c9ca8-301b-2a4b-b2db-410b13126a4a
32.9	6.4	3aa2e4e4-fd48-1b4b-8e39-d10de519db5e
11.7	3.7	4c3a3eae-85a0-4b41-ad21-5f6eebfc4f44
17.8	5.7	528baa97-d5b1-964b-ba0d-a17e8736ed12
6.8	2.8	5c2b1eb2-c3bf-a54d-b998-a87cfdcd510c
14.7	7.4	5ef43709-af28-0540-b8d4-e485f51baa5d
6.9	2.8	665f4176-1b1d-f24b-855b-70885ee2c315
18.9	9.5	691aee0c-8599-1b4f-bb45-cfee2704d8ff
14.2	4.5	6f109ac1-f00c-2f4d-b7c7-85d0dca076b6
12.9	2.6	7ccc89e0-120d-244f-91a1-3fbd87a5cd04
7.0	2.8	7d678e32-1faf-3c4f-a628-6c149c7cd4c6
11.9	5.0	b5cc0e61-9160-a445-b86e-80bda3513e94
37.5	6.9	cc266e1b-2047-5842-ad5d-170ab21a44a8
16.9	9.1	ebbfa8c5-4778-7548-b458-32283a887921
16.9	5.6	Using the 20 video lectures

A.2 Second phase: Intelligent Interaction

In Table A.2 we can observe the detailed results of the second phase for the professors who use the intelligent interaction. WER1 is the recognised WER of the video, and WER2 the remaining WER after the supervision by the professor with the intelligent interaction. As observed, intelligent interaction cannot guarantee perfect transcriptions, but it significantly reduces recognition errors to provide usable transcriptions that convey the meaning. Also, in Table A.3 we can observe the results for the professors who switched to the manual supervision, leaving the transcription perfect.

Tables A.2: Detailed results of the second phase with intelligent interaction

WER1	WER2	RTF	Video Lecture ID
16.3	11.5	4.3	0014de31-e740-634c-9699-96028ee4ba8a
10.9	3.1	1.0	25d72d38-8a10-084c-ac89-5c485aa32fc1
7.1	5.4	1.3	25eeb273-241f-944b-bf97-1c2905f4ba20
17.7	8.6	1.3	294d3ae5-2bd3-574e-a0ac-cd05b8e207ee
14.1	11.0	3.6	30fa1e01-0b7a-cb48-8faa-0a1a895707fe
20.2	7.4	1.1	4be3a364-2b08-cd45-aee9-52471a334ce1
25.2	11.5	2.9	4ce41182-d585-084c-b738-c4edcfd8d845
7.7	4.5	4.7	668f3d83-3f87-254e-8295-404d01767b3b
21.9	19.0	2.5	8e8fd4e9-29cd-2047-800a-0b18fe02fdd7
21.0	10.0	1.8	8f547b97-f7f1-554e-9dc9-1be76e673857
14.1	8.3	2.0	9c342ba1-51a6-1940-844f-08d30954ad35
16.4	10.4	1.3	9f485a94-c32d-5e48-995e-3744f300c711
10.1	7.2	3.7	a3ed41a4-e566-804d-96d5-5b2d0de4cb07
11.8	4.1	1.3	d558503e-a791-f440-a352-119432196012
16.8	6.8	0.9	edd4ff13-e135-b44a-8948-d89a78baee98
8.7	3.4	2.7	f1962dec-4be0-2641-a97e-4fcb3c643d43
12.3	4.2	0.9	f30af1b1-f84a-ef42-a560-c779cbbbdd2
10.9	5.3	1.8	fa5c497e-a335-6349-9f05-11c997bb418f
14.5	8.0	2.2	Using the 18 video lectures

Tables A.3: Detailed results of the second phase with manual supervision

WER	RTF	Video Lecture ID
18.0	3.9	06fa1a41-5dc7-da41-ad31-552f3ca3fd08
30.0	7.5	14c019fe-1d92-b144-a554-ae8a7718ca2c
9.5	1.8	15ae16bb-faa2-7a4a-a589-d082a2f3bb1c
21.0	6.7	219f4a5f-626b-0e4a-b004-c5c43821647b
19.0	5.6	3427dd38-7587-104e-9554-2c1940e033ac
37.2	7.8	34bdd7be-0a94-de4f-86af-466fefe264d4
10.4	4.3	353abbf4-3235-7048-8b58-658304283843
23.9	5.5	3622ddc7-a039-9d4e-9155-b2b84a383f06
26.3	5.5	382b4d71-e61e-d54d-9255-29ac5e3b723c
40.1	6.7	4362e220-633c-f947-a2eb-94d85a4514ef
8.1	3.2	4caaa4ec-fc68-704a-b0b8-88253c6221bd
15.8	6.7	6813b61b-403c-0e45-b26a-c54f9a30c7fc
23.9	9.0	6d24af05-8a7f-894d-905b-dada5ffd509c
22.1	6.0	6e7d2dc9-c8d7-1144-8db7-33c58545d059
16.0	3.3	838459ea-d46e-654e-b233-0aadd9a76b08
17.1	3.8	89e57f6e-48ff-e44c-aa02-c2fc35249640
13.4	8.8	93b26d6c-2d4c-864c-8d7e-8ae3bdf2af2f
18.1	3.8	bccd2059-fbaa-0648-b2f2-50c4de609575
20.9	4.7	dde5136b-05e4-9a4c-9512-ca5fa12ead8d
12.9	6.6	e1fa9283-d11c-7449-9d15-bddbab34e95b
12.8	2.6	e5ceaaa8-ecd2-054a-9259-c7031d80d404
25.6	4.5	f1e773cb-8dee-3e44-80fe-83f4adcaf316
19.5	5.2	Using the 22 video lectures

A.3 Third phase: Two-step Supervision

In Table A.4 we can observe the detailed results of the two-step supervision, for each video lecture supervised. W1 is the recognised WER obtained by the TLK toolkit, W2 is the WER after the first round, and W3 the WER after the re-transcription with the supervised massive adaptation. The reference is the final transcription corrected by the professor after the second round. Also, R1 is the RTF used on the first round, R2 the RTF used on the second round, and R the global RTF used on the third phase.

Tables A.4: Detailed results of the third phase with two-step supervision

W1	W2	W3	Video Lecture ID	R1	R2	R
58.1	47.7	44.9	12b81cf8-0a3d-2640-8c6c-ea83be0cc5e9	1.7	4.5	6.1
22.9	19.4	13.1	2f7d2b9e-430e-904d-99ce-be87d77f7805	0.8	2.6	3.4
47.0	36.4	19.0	3abe471f-972f-e04a-9fd0-b4a21c9dad8d	1.6	2.9	4.5
24.0	21.0	18.8	47f830c4-5a46-0d41-9037-8f71394efa46	0.7	3.4	4.0
25.0	16.3	11.2	5b168b3d-eccf-1b46-a3cf-743d77a2a3af	2.6	3.5	6.1
25.0	24.8	17.1	5b4ae1ef-ac01-9f45-85bb-f269eec2846f	0.3	2.4	2.7
35.3	28.6	21.9	734fb837-cec8-ca44-aa76-38740873bc79	1.9	3.9	5.8
40.6	30.0	15.6	797b7c07-1e25-f84e-8acf-4608a7d15778	1.6	2.4	4.0
21.8	21.3	17.3	8fb7c9dc-b70c-4b42-b26c-66791b26de69	1.2	5.8	7.0
39.9	38.2	33.2	baaac6b1-eecb-844f-ab6b-a4a676f69c02	0.8	4.9	5.7
17.6	13.8	12.5	c506ace5-df46-ea48-9431-4286595813a7	1.7	2.8	4.4
22.0	18.0	11.7	dbd9d1fc-b618-5f46-96c2-5682f2c1d53c	0.9	2.5	3.4
24.8	23.3	16.6	df22c85d-6edf-4748-9279-7a05dedd2017	0.5	6.2	6.7
28.8	21.2	16.1	e66535a1-abc6-9b48-9eac-fba3ec99bbe4	1.8	4.4	6.2
16.8	15.6	9.8	f1eeb2f7-3f7b-184e-bc1f-1a2045a0b9b8	0.9	2.1	3.0
17.7	16.0	13.8	f4a4c8f8-c885-4849-b312-e01fac526dfe	1.1	2.7	3.9
27.3	21.7	21.7	93c0367b-ce4c-224b-9864-15247a11742f	1.7	3.9	5.5
32.7	30.6	32.2	3bac5996-7696-fd4c-a638-01fb6082cb24	1.3	7.2	8.5
49.0	43.4	50.3	cd795506-8a71-3c4a-b3d8-76d2a20abf3c	1.5	4.9	6.3
28.6	24.1	26.7	514509df-4812-2546-b736-098a191b6521	2.0	3.9	5.9
33.9	24.8	14.0	8b6ac1c3-905e-764f-8fc7-752104c72e03	3.3	5.1	8.4
35.4	27.8	11.3	9095cf0a-0fcd-8749-b349-7d8ff441e16e	2.1	3.9	5.9
23.3	19.5	9.1	93edbcf1-ef9e-5644-90cd-3b6561b37df2	2.0	2.5	4.5
25.4	19.1	9.2	9f98fa1a-1ab5-9246-b9d5-5bb1f710e29d	1.9	2.4	4.3
29.9	27.5	21.1	a060bb64-7a23-1f41-b365-b0ce7465db92	0.5	3.2	3.7
27.1	24.9	27.3	ae1cc70d-2fa0-d14a-97ce-59a953a221ee	0.6	4.9	5.5
28.4	25.0	18.7	Using the 26 video lectures	1.4	3.9	5.3

Appendix B

Satisfaction Survey

B.1 Content of the Satisfaction Survey

After each phase the professors have completed a brief satisfaction survey to obtain their subjective impressions about our systems. This satisfaction survey has two parts: 10 questions evaluated on the 1-10 range in order to capture the quality of the different aspects of our system, and 3 brief free questions to allow the communication of the opinion of the professors. The first 10 questions are listed below:

1. I am satisfied with how easy it is to use this system
2. I can effectively complete my work using this system
3. I can complete my work quicker than doing it from scratch
4. I feel comfortable using this system
5. It was easy to learn to use this system
6. The help information of this system is clear
7. The organization of information on screen is clear
8. I like using the interface of this system
9. This system has all the functions that I expect to have
10. Overall, I am satisfied with this system

The last 3 questions are the listed below, and can be filled with any text which wants the professor, to communicate their opinions and possible improvements to us.

1. If you were to add new features to the player, which ones would be?
2. If you had to work daily with this player, what would you like to change?
3. Additional Comments

B.2 Results of the Satisfaction Survey

In Table B.1 we can observe the detailed results of the first 10 questions of the satisfaction survey for each phase. As we can observe the best system is the used on the first phase, because it is very simple and effective. The system used in the second phase receive a lower overall score, highlighting a very low mark in question "This system has all the functions that I expect to have". This is because the professors objective is to leave perfect transcriptions and the interaction model used in this phase does not allow this. In the third phase, despite we reduce a bit the time needed by professors on the supervision task, the survey is not as good as in the first phase because they must do the supervision in two phases, restriction that they dislike.

Tables B.1: Results of the satisfaction survey on each phase

Question	Phase 1	Phase 2	Phase 3
1	9.4	7.8	7.5
2	9.4	6.7	7.7
3	9.2	6.6	7.4
4	9.0	6.5	7.3
5	9.7	8.1	8.6
6	8.7	8.1	8.5
7	9.0	8.4	8.7
8	9.0	6.9	7.4
9	8.6	5.6	7.1
10	9.0	6.9	7.4
avg	9.1	7.2	7.8

It is important to note that the novelty of the application in the first phase has also affected the results, since in the first phase the professors saw a great improvement, and show this great improvement over the other two phases compared to the first is complicated. So, probably, if we presented first the third phase, it would have received a better score. However, the relevant answers to the other 3 questions reinforce and complement these conclusions. In the first phase only small usability details and new features were reported as we can see below:

1. Limit the segments to 20 words.

2. Allow to change the font size and color.

3. Allow to download the subtitles.

4. Auto-supervise segments listened.

5. Search and Replace functionality.

6. Automatically save the supervisions.

7. Reduce the initial loading time.

In the second phase, all the reports clearly show us that the intelligent interaction model requires a complete review, since all the reports propose changes on fundamental aspects of the model.

1. Allow to edit words out of the intelligent interaction guide.

2. Unlimited use of the expansion arrows to allow correct the entire segment.

3. Auto validate one correct word all the times that appears in the transcription.

4. Allow to return backward through the intelligent interaction changes.

5. Automatically remove consecutive repeated words before the supervision.

Finally, in the third phase the suggestions indicate that although with this two-step supervision the professors perceive a little improvement on time and transcription quality, this improvement is not relevant enough to encourage them to perform the supervision in two rounds instead of in one.

1. I spent a similar time reviewing the last two videos and the first received.

2. The transcription has been improved compared with that received in the first phase.

As we can see, the use of the satisfaction survey has helped us to capture valuable information about the evaluations, in order to greatly improve all our systems.

Bibliography

[FII06] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Lodem: A system for on-demand video lectures. *Speech Communication*, 48(5):516 – 531, 2006.

[FLBG07] Marc Ferras, Cheung Chi Leung, Claude Barras, and J-L Gauvain. Constrained mllr for speaker recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–53. IEEE, 2007.

[Gal98] Mark JF Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[LMR08] Saturnino Luz, Masood Masoodian, and Bill Rogers. Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proceedings of the 9th ACM SIGCHI New Zealand Chapter's International Conference on Human-Computer Interaction: Design Centered HCI*, pages 9–16. ACM, 2008.

[LSCN13] Wendy Leadbeater, Tom Shuttleworth, John Couperthwaite, and Karl P. Nightingale. Evaluating the use and impact of lecture recording in undergraduates: Evidence for distinct approaches by different groups of students. *Computers & Education*, 61(0):185 – 192, 2013.

[MPZ09] Cosmin Munteanu, Gerald Penn, and Xiaodan Zhu. Improving automatic speech recognition for lectures through transformation-based rules learned from minimal data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 764–772. Association for Computational Linguistics, 2009.

[MSA$^+$] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. Science 14 January 2011: 331 (6014), 176-182.

[MVdAAFJ13] Adrià Martínez-Villaronga, Miguel A. del Agua, Jesús Andrés-Ferrer, and Alfons Juan. Language model adaptation for video lectures transcription. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8450–8454. IEEE, 2013.

[pol07] poliMedia. The polimedia repository, 2007.

[Rab89] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[RGM08] Stephan Repp, Andreas Groß, and Christoph Meinel. Browsing within lecture videos based on the chain index of speech transcription. *Learning Technologies, IEEE Transactions on*, 1(3):145–156, 2008.

[RJ93] L. Rabiner and B. Juang. Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs, 1993.

[SCPGdMJ$^+$13] Joan Albert Silvestre-Cerdà, Alejandro Pérez González de Martos, Manuel Jiménez, Carlos Turró, Alfons Juan, and Jorge Civera. A system architecture to support cost-effective transcription and translation of large video lecture repositories. *International Conference on Systems*, 2013.

[SCSSJ12] Isaias Sanchez-Cortina, Nicolás Serrano, Alberto Sanchis, and Alfons Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 325–326. ACM, 2012.

[SdAG$^+$12] Joan Albert Silvestre, Miguel del Agua, Gonçal Garcés, Guillem Gascó, Adrià Giménez-Pastor, Adrià Martínez, Alejandro Pérez González de Martos, Isaías Sánchez, Nicolás Serrano Martínez-Santos, Rachel Spencer, Juan Daniel Valor Miró, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchís, and Alfons Juan. translectures. In *Proceedings of IberSPEECH 2012*, 2012.

[SGC$^+$13] Nicolás Serrano, Adrià Giménez, Jorge Civera, Alberto Sanchis, and Alfons Juan. Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–13, 2013.

[SJV12] Alberto Sanchis, Alfons Juan, and Enrique Vidal. A word-based naïve bayes classifier for confidence estimation in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):565–574, 2012.

[tUT13] The transLectures UPV Team. The translectures-upv toolkit (tlk), 2013.

[VMGdMCJ12] Juan Daniel Valor Miró, Alejandro Pérez González de Martos, Jorge Civera, and Alfons Juan. Integrating a state-of-the-art asr system into the opencast matterhorn platform. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 237–246. Springer, 2012.

[Wal06] Mike Wald. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3(2):131–141, 2006.

[YEK$^+$02] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. The htk book. *Cambridge University Engineering Department*, 3:175, 2002.

List of Figures

Index of Tables