# TACOMA: On-line Transcription of Audiovisual Material

*Antonio Cardenal, Pablo Peso, Martín Bueno, Alberto Espiña,*[1]
*Daniel A. Rodríguez-Silva, Lilian Adkinson-Orellana, Alexandre Pellitero-Rivero*[2]

[1]Grupo de Tecnoloxías Multimedia, University of Vigo, Spain
{cardenal,ppeso,mbueno,aespina}@gts.tsc.uvigo.es
[2]GRADIANT, ETSI Telecomunicación, Campus Vigo, 36310, Spain
{darguez,ladkinson,apellitero}@gradiant.org

## Abstract

TACOMA is a project devoted to the on-line transcription of educational courses which comprises two main systems: a fully automatic and generic subtitling system intended for being applied to any audiovisual material ; and a computer-assisted transcription system intended to facilitate and speed up as much as possible the process of manual subtitling using ASR technologies.

**Index Terms**: speech recognition, computer-assisted subtitling.

## 1. Introduction

The project TACOMA [1] *Transcripción on-line de Material Audiovisual* is being developed by the GTM group of the University of Vigo and Krasis Consulting S.L. company , with the participation of Gradiant (Galician Techological Center of Telecommunications).

The main goal of the project is the automatic subtitling of the on-line courses offered by Krasis, (which is a Spanish company devoted to e-learning and e-mail direct marketing), but with the additional objective of developing and make public an on-line universal subtitling application. For these purposes the project involves the developing of two main modules:

- A universal fully-automated subtitling system.
- A computer-assisted transcriber tool.

We are aware of the fact that the objective of achieving an universal subtitling system is very ambitious and quite unrealistic nowadays. Continuous speech recognition systems still offer poor performance when there are mismatches between the statistical models and the recording conditions. To obtain good results in every reasonable condition, it is necessary to use a set of acoustic and linguistic models, covering an extensive range of conditions, which means to capture and classify a large amount of acoustic and text databases. This objective is out of the scope of the project. Here, the objective of universal subtitling may be translated in the provision of mechanisms for acoustic and linguistic model unsupervised adaptation to achieve the best results given the available models.

The computer-assisted transcriber tool is provided to be used in situations in which the automatic system offers poor performance. The idea is to give the user the possibility of taking advantage of automatic speech recognition technology in such scenario, simplifying and speeding up the otherwise expensive and slow process of manual transcription. To achieve this goal, the system uses an interactive feedback strategy, in which the information provided by the user, (basically the corrected subtitles), is employed to improve progressively the acoustic and linguistic models, eventually reaching an ideal point in which no more corrections would be needed.

It should be noted that although our ambition is to develop an universal application, some of the decisions adopted throughout this project have been dictated by the special characteristics of the on-line courses employed. These courses are mainly related with computer programming and web management, so they have a technical, very specialized vocabulary. Also, the courses were recorded while the speaker is interacting with the computer, hence the speech is almost spontaneous, with frequent repetitions, filler words and false starts.

The remainder of the paper is organized as follows. In Section 2 the overall architecture of the system is described. In Section 3 the graphic interface is presented. In Section 4, we describe in deep the recognition module. Finally, in Section 5 we present some conclusions and further work.

## 2. Description of the architecture

### 2.1. Service-Oriented architecture

The system is based on a Service-Oriented architecture [1], composed of the transcription module and the application server, both providing a transcription service as shown in figure 1. A web client was implemented to access the transcription services with a simple and powerful interface [2] to manage the generated subtitles.

At the client side, two different parts can be identified: the list of current transcriptions and a player for visualizing the media, with an integrated subtitle editor. The Web Service uses PHP as server-side script language, and includes a database implemented with MySQL. The communication between the client and the application server is performed using REST [3], providing a simple and efficient method for exchanging data. The application server is also in charge of converting the audio and video sent by the client into the appropriate formats: the transcription module needs the audio in a raw file, while the player works with .flv for videos and .mp3 for audio files.

### 2.2. Commumication between Web Service and transcription module

The transcription module, which will be described more in detail in Section 4, is basically formed by a server containing the recognition system and an internal database. Since this transcription module and the application server are different ma-

[1] This project has been supported by Spanish MITYC program AVANZA I+D under the project TSI-020110-2009-242

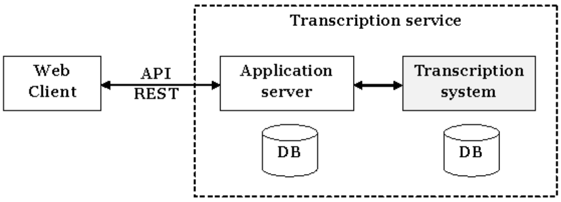


Figure 1: *System architecture overview.*

chines, the communication between them is performed using SSH protocol. This allows exchanging information efficiently in a secure way. Thus, the transcription server has available a set of scripts to manage each transcription and consult its current state.

## 3. Graphic interface

### 3.1. List of transcriptions

After logging in, the user can access to a web page where a list of his previous sessions is presented. The user may start a new session or cancel a current process. Two kind of sessions may be selected:

- Automatic, without human intervention.
- Assisted, in which the user may correct the automatic transcription provided by the system.

New transcriptions are started from a web page, in which the user may upload the audio or video file (or indicate an URL) to be transcribed. Additionally, the user can select at this point a trade-off between speed and recognition accuracy. This information will be used by the recognition module to set the number of recognition passes performed, and other minor parameters.

The interface also gives to the user the possibility of including extra information that would be useful in the recognition process. This information includes a text file, an URL pointing to a page with similar topics or a set of keywords. This information is later used by the transcription system to adapt and improve the statistical language and acoustic models.

### 3.2. Subtitles visualization and edition

After uploading the necessary information to start the transcription, the user will be redirected to a new page containing a player to visualize the audio or video, with a subtitle editor (see figure 2). This player allows the users to see the available subtitles synchronized with the audio or video in order to check the results. Every word is shown following a colour code, based on its correctness confidence, dark colour indicating higher confidence. If the word is suspected to be erroneous it will appear in red, so users can easily identify the wrong words and see if the transcription is being reliable.

If transcription module provides several alternatives for a word transcription, it will appear underlined on the subtitle editor. Making click on this word, a list will appear with the alternatives ordered from higher to lower confidence, so the user can choose the most appropriate. If the correct transcription does not appear in the list, the user has the possibility for typing it. In the case of assisted transcription, if the user considers that the subtitles received at certain stage have achieved enough quality, he can decide to change to automatic mode.

Finally, when the process finishes, the user may download the subtitles in several formats including .srt, TimedText, Subviewer 2.0, Encore, Google Video or JW FLV till 3.8. As we will explain in following sections, the corrections made by the user will be sent to the transcription server to adapt the statistical models to improve the accuracy of the following transcriptions.

## 4. Speech Recognition module

The recognition module is based on the recognition engine developed by the GTM group [4] in previous projects. This decoder is based on two stages: (1) a Viterbi algorithm which works in a synchronous way with a beam search; and (2) an $A^*$ algorithm. This recognizer was developed for large vocabulary continuous speech recognition applications.

### 4.1. System initialization

Once the audio file is received, the recognition module performs the following tasks:

- The selection of acoustic and linguistic models.
- The segmentation of the audio file using a simple voice activity detector.
- The audio file parameterization, using standard MFCC_E_D_A parameters with cepstral mean extraction.

The acoustic model selection is very simple. A phonetic recognizer is applied to the first frames of the file using each candidate model. The model which provides best acoustic score is then selected. Our experiments show that this simple mechanism provide good enough results. However, this strategy is valid if the audio file contains recordings of a single speaker, which is true for the on-line courses employed, but false in general. A speaker segmentation module will be incorporated in the future.

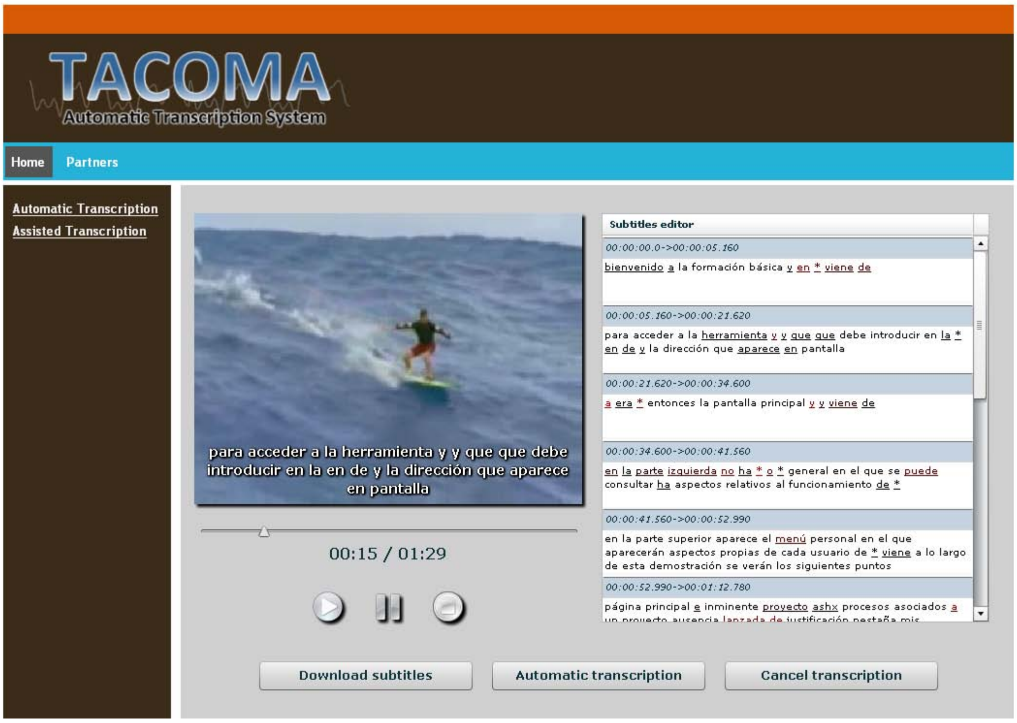
The VAD uses a simple algorithm based on energy thresholds and a small state machine. The mean length of the VAD segments may be selected by software, but our experience shows that a length of approximately one minute provides good trade-off between speed and recognition accuracy. This initial segmentation will be used throughout the remainder of the session for presentation and recognition purposes.

### 4.2. Speech recognition

Once segmented, each VAD section is recognized using a multipass strategy. In each pass an unsupervised acoustic adaptation is performed for each recognized segment. The procedure for the first pass is as follows (Figure 3:

1. Using the word-level transcription of the segment, a model-level alignment is obtained.
2. An MLLR+MAP adaptation of the HMMs is then performed, using the transcription of all segments processed until now.
3. The new models are used to recognize the next segment.

With this algorithm, the first segment is transcribed using the unadapted models, the second segment is transcribed using the models adapted with the the first segment, and so on.



para acceder a la herramienta y y que que debe introducir en la en de y la dirección que aparece en pantalla

Subtitles editor

00:00:00.0->00:00:05.160
bienvenido a la formación básica y en * viene de

00:00:05.160->00:00:21.620
para acceder a la herramienta y y que que debe introducir en la * en de y la dirección que aparece en pantalla

00:00:21.620->00:00:34.600
a era * entonces la pantalla principal y y viene de

00:00:34.600->00:00:41.560
en la parte izquierda no ha * o * general en el que se puede consultar ha aspectos relativos al funcionamiento de *

00:00:41.560->00:00:52.990
en la parte superior aparece el menú personal en el que aparecerán aspectos propias de cada usuario de * viene a lo largo de esta demostración se verán los siguientes puntos

00:00:52.990->00:01:12.780
página principal e inminente proyecto ashx procesos asociados a un proyecto ausencia lanzada de justificación pestaña mis

Figure 2: *Subtitle player with editor.*

For the second and successive passes, the procedure is similar. The main difference is that at point 2, the acoustic adaptation is performed using the alignments of pass 1 for all segments, so the first segment is recognized using acoustic models adapted with the transcription of the whole audio segment.

The main motivation of this procedure is to provide enough feedback to the user, avoiding annoying waiting periods with no response of the system, while maintaining the best possible performance.

Our experiments show that this strategy is very effective when there is a severe acoustic mismatch. As an example, in a test with an initial 63% of WER, the algorithm provides a 48.3 % of WER in the first pass, and 43.38 % if a second pass is performed. As a comparison, a 48.0% of WER is obtained in the same experiment if the audio file is recognized with an HMMs adapted using the transcription of all segments.

### 4.3. Computer-assisted module

The workflow of the computer-assisted transcription module is very similar to the described in the previous section. The strategy of progressive model adaptation is also employed here. The main difference is that the user-corrected transcriptions are now used when available, substituting the automatic recognized word sequence. As can be expected, the adaptation is better, providing an increase of performance.

However, the availability of the correct transcription allows also the use of language model adaptation mechanisms but also produce several new problems that should be addressed.

The main problem to address is related with the phonetic transcription of the new words. The acoustic adaptation process requires the use of a model-level alignment of the audio segment. To perform this task, a forced alignment is also needed. This forced recognition may easily fail if the manual transcription is not accurate enough. The most common cause is the use of foreigner (non-Spanish) which are phonetically transcribed using Spanish rules . This problem is very common in the online courses employed in the project, which are mainly related with software programming and web management. A simple example could be the word ”windows”, which using the Spanish phonetic rules would be transcribed as ”vindovs”. Two main mechanisms are used to overcome this problem: the use of lists of word exceptions, and an automatic detector of foreign words.

The use of a list of exceptions, including several alternative pronunciations of the most common foreigner (mostly English) words used in Spanish, is straightforward, but obviously very limited. To identify foreign words we use the simple algorithm of detecting some consonant groups not used in Spanish (*th, wr, ch*, etc). If a word is detected as non-Spanish and is not in the exception file, the user is prompted to provide an approximate Spanish pronunciation. The word is then added to the exception list to be used in following sessions.

### 4.4. Language model adaptation

The second use of the user-corrected transcriptions is to improve the language modeling, for example performing topic adaptation, identifying out of vocabulary words (OOvs), etc. At the time of writing this article the mechanisms for LM adaptation and OOv identification have not been included in the prototype, and are still under development. However, the idea is to use mixture language models and topic-based language models [6]. The LM adaptation will be used in two stages of the recognition process: when the transcription session starts, using the optional information provided by the user (http links, text files or keywords) and when an user-corrected transcription is available. In both cases the tasks to perform are very similar:

- The vocabulary of the new material must be extracted

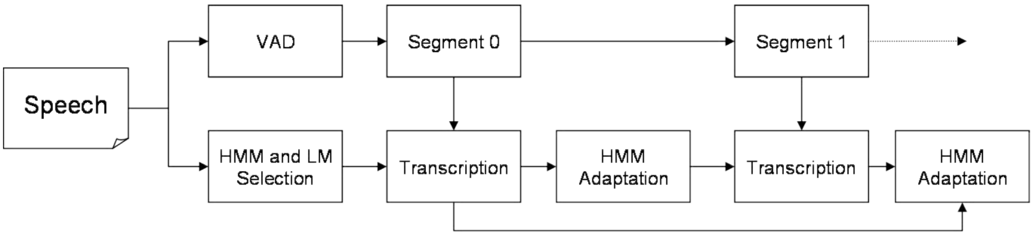


Figure 3: *Unsupervised acoustic adaptation in TACOMA.*

and if necessary, included on the LM. To perform this task, is essential that the new words appear on the text database. Otherwise the training of the new LM is not possible. We don't know any strategy to overcome this particular problem.

- The new text is used as objective for mixture-based LM adaptation, or
- The new text is used to select a topic-based LM.

The first two mechanisms are under development and are expected to be included in the system soon. In an open-task application such this one, the third mechanism requires an unrealistic amount of data, although is been considered, does not seem very promising.

In any case, the LM adaptation mechanisms require the reconstruction of the model for each new transcription received. Unlike the acoustic model adaptation, which is a reasonably fast procedure, the LM adaptation is computationally expensive and time consuming. This is not a problem with the automatic subtitling system, in which the LM adaptation is applied only once at the beginning of the session, but it may be an issue in the computer-assisted system, in which the adaptation should be performed for each corrected transcription received.

### 4.5. Confidence measures and N-Best lists

Confidence measures are an important part of the system, specially in the computer-assisted system since are used to warn the user from potential transcription errors. We believe that this kind of mechanisms are very important for speeding up the manual transcription task.

The confidence measures employed in this project are based on word posteriori probabilities extracted using word lattices [5]. Unlike most applications we need a continuous confidence measure for employing the colour code explained in Section 3.

The confidence measures are not only computed for the main transcription, but also for the N-Best lists. This is, however, a common mechanism when using word lattices. This information is employed, as was explained in Section 3, to implement an ordered drop-down list on the client side, allowing the user to easily select the right word.

### 4.6. Conclusions and further work

In this paper we have presented a prototype of the subtitling system developed in the project TACOMA. It is composed of two modules: an automatic subtitling system and a computer-assisted transcriber. The most novel part of the project is the computer assisted transcriber, which is conceived to speed up the usually time-consuming task of manual transcription. The key points of the system are the following:

- The audio file is segmented using a voice activity detector.
- Each audio segment is automatically transcribed, using the result to perform an adaptation of the acoustic models.
- The initial transcription is presented to the user with the confidence of each word indicated by a colour code.
- The user may access an ordered list of alternative transcriptions of each word for correcting purposes.
- The corrections of the user are sent back to the recognition system for acoustic and LM adaptation, improving the transcription of the next segments.

Eventually a point should be reached in which the transcription would be accurate enough to perform an automatic subtitling of the remainder data with reasonable confidence.

The project TACOMA is still under development. In the near future we plan to incorporate some mechanisms of LM adaptation which are been tested at this time. The next improvement will be the incorporation of a speaker diarization and classification module, for dealing with multi-speaker recordings. This module is also under development in our research group [7].

## 5. References

[1] R. Perrey and M. Lycett, “Service-Oriented Architecture”, Proc. 2003 Symp. Applications and the Internet (SAINT 03), IEEE CS Press, 2003, p. 116.

[2] Accessible Rich Internet Applications (WAI-ARIA) 1.0, W3C Working Draft 16 September 2010. Online: http://www.w3.org/TR/wai-aria/, accessed on Oct 7, 2010

[3] Pautasso, Cesare; Zimmermann, Olaf; Leymann, Frank (2008-04), “RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision”, 17th International World Wide Web Conference (WWW2008) (Beijing, China)

[4] A. Cardenal-Lopez and F.J. Dieguez-Tirado and C. Garcia-Mateo, ”Fast LM look-ahead for large vocabulary continuous speech recognition using perfect hashing”. Proc. ICASSP 2002

[5] F. Wessel and R. Schlter and K. Macherey and H. Ney, “Confidence measures for large vocabulary continuous speech recognition”,In IEEE Signal Processing Letters, pages 288-298, num. 3, vol. 9, 2001.

[6] J Dieguez Tirado, C. García Mateo, A. Cardenal, ”Effective topic-tree based language model adaptation”, Proc. Interspeech 2005.

[7] Laura Docío Fernández, Paula López Otero, Carmen García Mateo. ”An adaptive Threshold Computation for Unsupervised Speaker Segmentation”. Proc. Interspeech 2005.