Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 74 (2015) 65–75

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

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

Received 29 April 2015; received in revised form 7 September 2015; accepted 16 September 2015
Available online 26 September 2015

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 automatic speech recognition (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 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

* Corresponding author.
E-mail addresses: jvalor@dsic.upv.es (J.D. Valor Miró), jsilvestre@dsic.upv.es (J.A. Silvestre-Cerdà), jcivera@dsic.upv.es (J. Civera), turro@cc.upv.es (C. Turró), ajuan@dsic.upv.es (A. Juan).

http://dx.doi.org/10.1016/j.specom.2015.09.006
0167-6393/

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 min 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.

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.

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.

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

[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.