Valor Miró, JD.; Pérez González De Martos, AM.; Civera Saiz, J.; Juan Císcar, A. (2012). Integrating a State-of-the-Art ASR System into the Opencast Matterhorn Platform. Communications in Computer and Information Science. 328:237-246. doi:10.1007/978-3-642-35292-8_25.

Integrating a state-of-the-art ASR system into the Opencast Matterhorn platform

Juan Daniel Valor, Alejandro Pérez González de Martos,
Jorge Civera, and Alfons Juan

Universitat Politècnica de València
Camino de Vera s/n, 46022 Valencia, Spain
juavami@upv.es,alpegon2@upv.es
jcivera@dsic.upv.es,ajuan@dsic.upv.es,
http://prhlt.iti.upv.es/

Abstract. In this paper we present the integration of a state-of-the-art ASR system into the Opencast Matterhorn platform, a free, open-source platform to support the management of educational audio and video content. The ASR system was trained on a novel large speech corpus, known as poliMedia, that was manually transcribed for the European project transLectures. This novel corpus contains more than 115 hours of transcribed speech that will be available for the research community. Initial results on the poliMedia corpus are also reported to compare the performance of different ASR systems based on the linear interpolation of language models. To this purpose, the in-domain poliMedia corpus was linearly interpolated with an external large-vocabulary dataset, the well-known Google N-Gram corpus. WER figures reported denote the notable improvement over the baseline performance as a result of incorporating the vast amount of data represented by the Google N-Gram corpus.

Keywords: Speech Recognition, Linear Combination, Language Modeling, Google N-Gram, Opencast Matterhorn

1 Introduction

Online educational repositories of video lectures are rapidly growing on the basis of increasingly available and standardized infrastructure. Transcription and translation of video lectures is needed to make them accessible to speakers of different languages and to people with disabilities. Automatic transcription in these domains is however a challenging task due to many factors such as unfavourable recording quality, high rate out-of-vocabulary words or multiplicity of speakers and accents. Therefore, human intervention is needed to achieve accurate transcriptions. Recently, approaches to hybrid transcription systems have been proposed based on fully manual correction of automatic transcriptions, which are not practical nor comfortable to the users who perform this time-consuming task. In this paper we present an intelligent user interactive semi-automatic speech recognition system to provide cost-efficient solutions to produce accurate

transcriptions. This speech recognition system is being developed within the framework of the European *transLectures* project [1], along the lines of other systems, such as JANUS-II [2], UPC RAMSES [3] or SPHINX-II [4]. Resulting transcriptions may be translated into other languages, as it is the case of the transLectures project, or other related project, such as SUMAT [5].

Initial results are reported on the recently created poliMedia corpus using a linear combination of language models [6–9]. This linear combination aims at alleviating the problem of out-of-vocabulary words in large-scale vocabulary tasks with a great variety of topics. The baseline automatic speech recognition (ASR) system is based on the RWTH ASR system [10, 11] and the SRILM toolkit [12], both state-of-the-art software in speech and language modeling, respectively. In this work, we present significant improvements in terms of WER over the baseline when interpolating the baseline language model with a language model trained on the well-known *Google n-gram* dataset [13]. Furthermore, details about the integration of this speech recognition system into the open-source videolecture platform Matterhorn are also provided. The integration into Matterhorn enables user-assisted corrections and therefore, it guarantees high quality transcriptions.

The rest of this paper is organised as follows. First, the novel freely available poliMedia corpus is presented in Section 2. Secondly, the Opencast Matterhorn platform is introduced in Section 3. In Section 4, the backend RWTH ASR system is described, and initial results are reported in Section 5. Finally, conclusions are drawn and future lines of research are depicted in Section 6.

2 The poliMedia corpus

poliMedia [14] is a recent, innovative service for creation and distribution of multimedia educational content at the *Universitat Politècnica de València* (UPV). It is mainly designed for UPV professors to record courses on video lectures lasting 10 minutes at most. Video lectures are accompanied with time-aligned slides and recorded at specialised studios under controlled conditions to ensure maximum recording quality and homogeneity. As of today, poliMedia catalogue includes almost 8000 videos accounting for more than 1000 hours. Authors retain all intellectual property rights and thus not all videos are accessible from outside the UPV. More precisely, about 2000 videos are openly accessible.

poliMedia is one the two videolectures repositories along with Videolectures.NET[1] that are planned to be fully transcribed in the framework of the European project transLectures[2]. To this purpose, 704 videolectures in Spanish corresponding to 115 hours were manually transcribed using the tool Transcriber [15], so as to provide in-domain dataset for training, adaptation and internal evaluations in the transLectures project (see Table 1). These transcribed videolectures were selected so that authors had granted open access to their content. This fact guarantees that the poliMedia corpus can be used by the research community beyond the scope of the transLectures project.

[1] http://videolectures.net
[2] http://translectures.eu

Most of the videos in poliMedia were annotated with topic and keywords. More precisely, 94% of the videos were assigned a topic and 83% were described with keywords. However, these topics and keywords were not derived from a thesaurus, such as EuroVoc. Speakers were also identified for each transcription.

Table 1. Basic statistics on the poliMedia corpus

Videos	704
Speakers	111
Hours	115
Sentences	40K
Running words	1.1M
Vocabulary (words)	31K
Singletons (words)	13K

3 The Opencast Matterhorn platform

Matterhorn[3] is a free, open-source platform to support the management of educational audio and video content. Institutions will use Matterhorn to produce lecture recordings, manage existing video, serve designated distribution channels, and provide user interfaces to engage students with educational videos.

Matterhorn is an open source; this means that the product is fully based on open source products. The members of the Opencast Community have selected Java as programming language to create the necessary applications and a Service-Oriented Architecture (SOA) infrastructure. The overall application design is highly modularised and relies on the OSGi (dynamic module system for Java) technology. The OSGi service platform provides a standardised, component-oriented computing environment for cooperating network services.

Matterhorn is as flexible and open as possible and further extensions should not increase the overall complexity of building, maintaining and deploying the final product. To minimise the coupling of the components and third party products in the Matterhorn system, the OSGi technology provides a service-oriented architecture that enables the system to dynamically discover services for collaboration. Matterhorn uses the Apache Felix [16] implementation of the OSGI R4 Service Platform [17] to create the modular and extensible application.

The main goal in transLectures is to develop tools and models for the Matterhorn platform that can obtain accurate transcriptions by intelligent interaction with users. For that purpose, an HTML5 media player prototype has been built in order to provide a user interface to enable interactive edition and display of video transcriptions (see Figure 1). This prototype offers a main page where

[3] http://opencast.org/matterhorn

available poliMedia videolectures are listed according to some criteria. Automatic video transcriptions are obtained from the ASR system when playing a particular video.

Since automatic transcriptions are not error free, an interactive transcription editor allows intelligent user interaction to improve transcription quality. However, as users may have different preferences while watching a video, the player offers two interaction models depending on the user role: simple user and collaborative user (prosumers).

Fig. 1. HTML5 player and interactive transcription editor for collaborative users.

Simple users are allowed to interact in a very simplistic manner, just showing their liking about the transcriptions. However, collaborative users may provide richer feedback to correct transcriptions. As shown in Figure 1, collaborative users have an *edit transcription* button available on the player control bar that enables the transcription editor panel. The editor panel is situated next to the video. It basically contains the transcription text, which is shown synchronously with the video playback. Clicking on a transcription word or sentence enables the interactive content modification. User corrections are sent to the speech recognition module through a web service, so corrections are processed and new transcription hypothesis are offered back to the user. Some other user-friendly features such as keyboard shortcuts and useful editing buttons are also available.

Simple users have no edit transcription button available as they are not expected to be working on transcription editing. Instead, a *low quality transcription* button appears so they can report that the transcription quality is not good enough.

The current HTML5 prototype is a proof-of-concept version that works with pre-loaded transcriptions, however the version currently being developed communicates with the ASR system through a web service implemented for that purpose. Figure 2 illustrates the system architecture and the communication process.

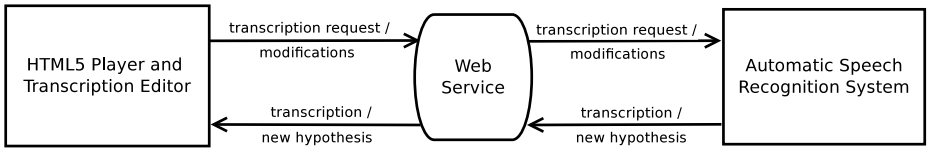

Fig. 2. HTML5 player and ASR system communication.

The next step is to integrate the developed interactive ASR system into the Matterhorn infrastructure. There are many different approaches to perform this integration. Our proposal lets an external system manage all the transcriptions, so there will not be necessary to add nor store them in any way into the current Matterhorn system[4]. In addition, two primary tasks are involved in the integration process into Matterhorn. Both of them require an interface to enable communication between Matterhorn and the ASR system. For that purpose, a RESTful Web Service has been implemented to allow media uploading, retrieve the processing status of a particular recording, request a video transcription, send transcription modifications and other functionalities.

The first task would be to define a new Matterhorn workflow operation to transfer the audio data of the new media to the ASR system through the REST service mentioned before, so as to obtain automatic transcriptions for every recording uploaded to the Matterhorn platform. This task will involve the implementation of a new Matterhorn service.

The second part is to replace or adapt the Matterhorn Engage Player to enable transcription edition, along the lines of the HTML5 player prototype indicated previously. The player must obtain and transmit every transcription-related information through the REST Web Service in a similar way as the HTML5 prototype did (see Figure 2). Here the main problem is the addition of new features to the Flash-based Matterhorn player, since it is not straightforward to implement the transcription functionalities provided by the HTML5-based player. Our solution is to use an alternative open-source Matterhorn engage player based on HTML5 called Paella Engage Player[5].

[4] http://opencast.jira.com/wiki/display/MH/MediaPackage+Overview
[5] http://unconference.opencast.org/sessions/paella-html5-matterhorn-engage-player

4 The RWTH ASR system

Our baseline ASR system is the RWTH ASR system [10, 11] along with the SRILM toolkit [12]. The RWTH ASR system includes state-of-the-art speech recognition technology for acoustic model training and decoding. It also includes speaker adaptation, speaker adaptive training, unsupervised training, a finite state automata library, and an efficient tree search decoder. SRILM toolkit is a widespread language modeling toolkit which have been applied to many different natural language processing applications.

In our case, audio data is extracted from videos and preprocessed to extract the normalized acoustic features obtaining the Mel-frequency cepstral coefficients (MFCCs) [18]. Then, triphoneme acoustic models based on a prebuilt cart tree are trained adjusting parameters such as number of states, gaussian components, etcetera on the development set. The lexicon model is obtained in the usual manner by applying a phonetic transliteration to the training vocabulary. Finally, n-gram language models are trained on the transcribed text after filtering out unwanted symbols such as punctuation marks, silence annotations and so on.

In this work, we propose to improve our baseline system by incorporating external resources to enrich the baseline language model. To this purpose, we consider the linear combination of an in-domain language model, such as that trained on the poliMedia corpus, with an external large out-domain language model computed on the Google N-Gram corpus [13]. A single parameter λ governs the linear combination between the poliMedia language model and the Google N-Gram model, being optimised in terms of perplexity on a development set.

5 Experimental Results

In order to study how the linear combination of language models affects the performance, in terms of WER, of an ASR system in the poliMedia corpus, a speaker-independent partition in training, development and test sets was defined. The statistics of this partition can be found in Table 2. Topics included in the development and test sets range from technical studies such as architecture, computer science or botany, to art studies such as law or marketing.

The baseline system, including acoustic, lexicon and language models, was trained only on the poliMedia corpus. System parameters were optimised in terms of WER on the development set. A significant improvement of more than 5 points of WER was observed when moving from monophoneme to triphoneme acoustic models. Triphoneme models were inferred using the conventional CART model using 800 leaves. In addition, the rest of parameters to train the acoustic model were 2^9 components per Gaussian mixture, 4 iterations per mixture and 5 states per phoneme without repetitions. The language model was an interpolated trigram model with Kneser-Ney discount. Higher order n-gram models were also assessed, but no better performance was observed.

Provided the baseline system, a set of improvements based on the language model were proposed and evaluated. The baseline language model solely trained

Table 2. Basic statistics on the poliMedia partition.

	Training	Development	Test
Videos	559	26	23
Speakers	71	5	5
Hours	99	3.8	3.4
Sentences	37K	1.3K	1.1K
Vocabulary	28K	4.7K	4.3K
Running words	931K	35K	31K
OOV (words)	-	4.6%	5.6%
Perplexity	-	222	235

on poliMedia corpus was interpolated with the Google N-Gram corpus [13]. To this purpose, we unify all Google N-Gram datasets, which are initially splitted by years, in a single, large file. Then, we train a trigram language model using Google N-Gram that was interpolated with the poliMedia language model. These two language models were interpolated to minimise perplexity on the development set. This interpolation was performed using a particular vocabulary in the case of Google N-Gram, ranging from that vocabulary matching that of poliMedia (poliMedia vocab), over the 20.000 most frequent words in the Google N-Gram corpus (20K vocab), to the 50.000 most frequent words (50K vocab). In this latter experiment, approximate values of interpolation weights are 0.65 for the poliMedia language model and 0.35 for the Google N-Gram language model.

The idea behind these experimental setups was to evaluate the effects, in terms of WER, of an increasing vocabulary coverage using external resources in the presence of a comparatively small in-domain corpus such as poliMedia. Experimental results are shown in Table 3.

Table 3. Evolution of WER above the baseline for the RWTH ASR system, as a result of interpolating the poliMedia language model with an increasingly larger vocabulary language model trained on the Google N-Gram corpus.

System	WER	OOV
baseline	39.4	5.6%
poliMedia vocab	34.6	5.6%
20K vocab	33.9	4.4%
50K vocab	33.7	3.5%

As reported in Table 3, there is a significant improvement of 5.7 points of WER over the baseline when considering a language model trained with the 50K most frequent words in the Google N-Gram corpus. As expected, the decrease in WER is directly correlated with the number of Out-Of-Vocabulary words (OOVs) in the test set, since the Google N-Gram corpus provides a better vocabulary coverage.

A similar trend is observed when comparing perplexity figures between the baseline and poliMedia vocab systems. Perplexity significantly drops from 235 to 176 just by interpolating our baseline poliMedia language model with the Google N-Gram language model that only considers the poliMedia vocabulary. Perplexity figures with 20K and 50K vocab are not comparable to the previous ones, since the size of the vocabulary is not the same. Note that by adding more vocabulary from the Google N-Gram dataset, the number of OOVs is reduced, but also more useless words are added to the final language model. This causes that the improvement in terms of WER is not so significant when going from 20K to 50K vocabulary. Further experiments with 2-gram and 4-gram language model were carried out. WER figures with 2-gram were two points below on average, while 4-gram results were similar to those obtained with 3-grams.

6 Conclusions and Future Work

In this paper we have presented the integration of a state-of-the-art ASR system into the Opencast Matterhorn platform. This system was trained on a novel large speech corpus, known as poliMedia, that was manually transcribed for the European project transLectures. This novel corpus contains more than 115 hours of transcribed speech that will be available for the research community.

Initial results on the poliMedia corpus are also provided to compare the performance of different systems based on the linear interpolation of language models. To this purpose, the in-domain poliMedia corpus was linearly interpolated with an external large-vocabulary dataset, the well-known Google N-Gram corpus. WER figures reported denote the notable improvement over the baseline performance as a result of incorporating the vast amount of data contained in the Google N-Gram corpus.

Regarding the backend ASR system, various aspects need to be considered for future research. A simple manner to improve our initial results is to perform an intelligent data selection from the Google N-Gram corpus based on a chronological criteria such as the year of publication, or inspired on a simple, yet effective, method such that presented in [19]. In this latter case, only infrequent n-grams in poliMedia will be enriched with counts computed in large external resources such as the Google N-Gram corpus. Obviously, the extension of the vocabulary size to 100K words or greater may provide little reductions in WER values, but not significants compared to the computational cost required to run such an experiment.

In any case, ASR accuracy is still far from producing fully automatic high-quality transcriptions, and human intervention is still needed in order to improve transcriptions quality. However, user feedback can be exploited to minimise user effort in future interactions with the system [20]. New features need to be developed and integrated into the Matterhorn platform to achieve an effective user interaction. The resulting prototype will not only be evaluated under controlled laboratory conditions, but also in real-life conditions in the framework of the transLectures project.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 287755. Also supported by the Spanish Government (MIPRCV "Consolider Ingenio 2010" and iTrans2 TIN2009-14511) and the Generalitat Valenciana (Prometeo/2009/014).

References

1. UPVLC, XEROX, JSI-K4A, RWTH, EML, and DDS. transLectures: Transcription and Translation of Video Lectures. In *Proc. of EAMT*, page 204, 2012.
2. P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie, and A. Waibel. JANUS-II: towards spontaneous Spanish speech recognition. 4:2285–2288, 1996.
3. A. Nogueiras, J. A. R. Fonollosa, A. Bonafonte, and J. B. Mariño. RAMSES: El sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC. In *VIII Jornadas de I+D en Telecomunicaciones*, pages 399–408, 1998.
4. Xuedong Huang, Fileno Alleva, Hsiao wuen Hon, Mei yuh Hwang, and Ronald Rosenfeld. The SPHINX-II Speech Recognition System: An Overview. *Computer, Speech and Language*, 7:137–148, 1992.
5. Speech and Language Technology Group. Sumat: An online service for subtitling by machine translation. http://www.sumat-project.eu, May 2012.
6. Simo Broman and Mikko Kurimo. Methods for combining language models in speech recognition. In *Proc. of Interspeech*, pages 1317–1320, 2005.
7. X. Liu, M. Gales, J. Hieronymous, and P. Woodland. Use of contexts in language model interpolation and adaptation. volume Proc. of Interspeech, 2009.
8. X. Liu, M. Gales, J. Hieronymous, and P. Woodland. Language model combination and adaptation using weighted finite state transducers. 2010.
9. Joshua T. Goodman. Putting it all together: Language model combination. In *Proc. of ICASSP*, pages 1647–1650, 2000.
10. J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney. The rwth 2007 tc-star evaluation system for european english and spanish. In *Proc. of Interspeech*, pages 2145–2148, 2007.
11. D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The rwth aachen university open source speech recognition system. In *Proc. of Interspeech*, pages 2111–2114, 2009.
12. A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of ICSLP*, 2002.
13. J. B. Michel et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
14. C. Turro, A. Cañero, and J. Busquets. Video learning objects creation with polimedia. Multimedia (ISM), 2010 IEEE International Symposium on pp.371-376, 13-15, Dec 2010.
15. C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication special issue on Speech Annotation and Corpus Tools*, 33(1–2), 2000.
16. Apache. Apache felix. http://felix.apache.org/site/index.html, May 2012.
17. Osgi alliance. osgi r4 service platform. http://www.osgi.org/Main/HomePage, May 2012.

18. M. Sahidullah and G. Saha. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. 54(4):543–565, 2012.
19. Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. Does more data always yield better translations? In *Proc. of EACL*, pages 152–161, 2012.
20. I. Sánchez-Cortina, N. Serrano, A. Sanchis, and A. Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proc. of IUI*, pages 325–326, 2012.