UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Contributions to Efficient Automatic Transcription of Video Lectures

Thesis
presented by Miguel Ángel del Agua Teba
supervised by Dr. José Alberto Sanchis Navarro and Dr. Alfons Juan Císcar

June 4, 2019

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Contributions to Efficient Automatic Transcription of Video Lectures

Miguel Ángel del Agua Teba

Thesis performed under the supervision of doctors
José Alberto Sanchis Navarro and Alfons Juan Císcar
and presented at the Universitat Politècnica de València
in partial fulfilment of the requirements for the degree
Doctor en Informàtica

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
València 2019

Agradecimientos

Este trabajo representa la culminación de un proceso de aprendizaje de varios años en el que he podido desarrollar unas aptitudes académicas y personales que de otra forma no hubiera sido posible. Por este motivo, primero de todo querría agradecer a mis directores de tesis Alfons Juan y Alberto Sanchis, la oportunidad que me dieron para formar parte del grupo de investigación MLLP y así poder conocer de primera mano este apasionante campo de investigación.

Esta odisea, llena de momentos buenos y malos, tampoco hubiera llegado a buen puerto de no ser por el apoyo de las personas con las que he compartido laboratorio todos estos años. Personas brillantes en el ámbito académico, pero de un valor humano que no tiene precio y que de una manera u otra han dejado su huella. Hablo de Adrià, Adrià Agustí, Álex, Germán, Guillem, Gonçal, Ihab, Jesús, Joan Albert, Jorge, JuanDa, Nico, Pau, Rachel y Santi. Además, y pese a no haber coincidido en el día a día, esta tesis también tiene un poco de Dani Martin, Dani Ortiz, Jesús González, Joan, Luis, Paco, Radha y Ricardo.

También se lo quiero agradecer a mi familia y en particular, a quienes me han apoyado incondicionalmente en cualquier proyecto que he emprendido, a mis padres Miguel y Antonia. Y finalmente, a la persona que más ha aguantado mis noches sin dormir, los fines de semana sin salir, las innumerables horas delante del ordenador, y que pese a ello siempre me ha recibido con una sonrisa, a ti, Ana.

Abstract

During the last years, on-line multimedia repositories have become key knowledge assets thanks to the rise of Internet and especially in the area of education. Educational institutions around the world have devoted big efforts to explore different teaching methods, to improve the transmission of knowledge and to reach a wider audience. As a result, online video lecture repositories are now available and serve as complementary tools that can boost the learning experience to better assimilate new concepts. In order to guarantee the success of these repositories the transcription of each lecture plays a very important role because it constitutes the first step towards the availability of many other features. This transcription allows the searchability of learning materials, enables the translation into another languages, provides recommendation functions, gives the possibility to provide content summaries, guarantees the access to people with hearing disabilities, etc. However, the transcription of these videos is expensive in terms of time and human cost.

To this purpose, this thesis aims at providing new tools and techniques that ease the transcription of these repositories. In particular, we address the development of a complete Automatic Speech Recognition Toolkit with an special focus on the Deep Learning techniques that contribute to provide accurate transcriptions in real-world scenarios. This toolkit is tested against many other in different international competitions showing comparable transcription quality. Moreover, a new technique to improve the recognition accuracy has been proposed which makes use of Confidence Measures, and constitutes the spark that motivated the proposal of new Confidence Measures techniques that helped to further improve the transcription quality. To this end, a new speaker-adapted confidence measure approach was proposed for models based on Recurrent Neural Networks.

The contributions proposed herein have been tested in real-life scenarios in different educational repositories. In fact, the transLectures-UPV toolkit is part of a set of tools for providing video lecture transcriptions in many different Spanish and European universities and institutions.

Resum

Durant els últims anys, els repositoris multimèdia en línia s'han convertit en fonts clau de coneixement gràcies a l'expansió d'Internet, especialment en l'àrea de l'educació. Institucions educatives de tot el món han dedicat molts recursos en la recerca de nous mètodes d'ensenyament, tant per millorar l'assimilació de nous coneixements, com per poder arribar a una audiència més àmplia. Com a resultat, avui dia disposem de diferents repositoris amb classes gravades que serveixen com a eines complementàries en l'ensenyament, o fins i tot poden assentar una nova base a l'ensenyament a distància. No obstant això, han de complir amb una sèrie de requisits perquè la experiència siga totalment satisfactòria i és ací on la transcripció dels materials juga un paper fonamental. La transcripció possibilita una recerca precisa dels materials en els quals l'alumne està interessat, s'obri la porta a la traducció automàtica, a funcions de recomanació, a la generació de resums de les xerrades i el poder fer arribar el contingut a persones amb discapacitats auditives. No obstant, la generació d'aquestes transcripcions pot resultar molt costosa.

Amb això en ment, la present tesi té com a objectiu proporcionar noves eines i tècniques que faciliten la transcripció d'aquests repositoris. En particular, abordem el desenvolupament d'un conjunt d'eines de reconeixement automàtic de la parla, amb èmfasi en les tècniques d'aprenentatge profund que contribueixen a proporcionar transcripcions precises en casos d'estudi reals. A més, es presenten diferents participacions en competicions internacionals on es demostra la competitivitat del programari comparada amb altres solucions. D'altra banda, per tal de millorar els sistemes de reconeixement, es proposa una nova tècnica d'adaptació d'aquests sistemes a l'interlocutor basada en l'ús de Mesures de Confiança. A més, això va motivar el desenvolupament de tècniques per a la millora en l'estimació d'aquest tipus de mesures per mitjà de Xarxes Neuronals Recurrents.

Totes les contribucions presentades s'han provat en diferents repositoris educatius. De fet, el toolkit transLectures-UPV és part d'un conjunt d'eines que serveix per generar transcripcions de classes en diferents universitats i institucions espanyoles i europees.

Resumen

Durante los últimos años, los repositorios multimedia en línea se han convertido en fuentes clave de conocimiento gracias al auge de Internet, especialmente en el área de la educación. Instituciones educativas de todo el mundo han dedicado muchos recursos en la búsqueda de nuevos métodos de enseñanza, tanto para mejorar la asimilación de nuevos conocimientos, como para poder llegar a una audiencia más amplia. Como resultado, hoy en día disponemos de diferentes repositorios con clases grabadas que siven como herramientas complementarias en la enseñanza, o incluso pueden asentar una nueva base en la enseñanza a distancia. Sin embargo, deben cumplir con una serie de requisitos para que la experiencia sea totalmente satisfactoria y es aquí donde la transcripción de los materiales juega un papel fundamental. La transcripción posibilita una búsqueda precisa de los materiales en los que el alumno está interesado, se abre la puerta a la traducción automática, a funciones de recomendación, a la generación de resumenes de las charlas y además, el poder hacer llegar el contenido a personas con discapacidades auditivas. No obstante, la generación de estas transcripciones puede resultar muy costosa.

Con todo esto en mente, la presente tesis tiene como objetivo proporcionar nuevas herramientas y técnicas que faciliten la transcripción de estos repositorios. En particular, abordamos el desarrollo de un conjunto de herramientas de reconocimiento de automático del habla, con énfasis en las técnicas de aprendizaje profundo que contribuyen a proporcionar transcripciones precisas en casos de estudio reales. Además, se presentan diferentes participaciones en competiciones internacionales donde se demuestra la competitividad del software comparada con otras soluciones. Por otra parte, en aras de mejorar los sistemas de reconocimiento, se propone una nueva técnica de adaptación de estos sistemas al interlocutor basada en el uso Medidas de Confianza. Esto además motivó el desarrollo de técnicas para la mejora en la estimación de este tipo de medidas por medio de Redes Neuronales Recurrentes.

Todas las contribuciones presentadas se han probado en diferentes repositorios educativos. De hecho, el toolkit transLectures-UPV es parte de un conjunto de herramientas que sirve para generar transcripciones de clases en diferentes universidades e instituciones españolas y europeas.

Contents

Chapter 1

Introduction

1.1 Motivation

Nowadays Artificial Intelligence (AI) is increasingly becoming a key technology component that is changing the industry and, as such, constitutes one of the sources of progress that have a real impact on our society. During the last years, this area has experienced a very important breakthrough because of two fundamental factors: The availability of increasingly larger data repositories as a result of the popularization of cloud-based services, and greater computational power led by General Purpose GPU systems that are able to process such amount of data. Well-known Machine Learning (ML) methods based on learning data representations have specially benefited from those, and constitute the so-called Deep Learning (DL) revolution. DL comprises a set of model architectures and algorithms that have made impressive advances in different fields such as Computer Vision (CV) and Natural Language Processing (NLP). This is particularly true in the case of Automatic Speech Recognition (ASR), the research field that aims at giving computers the capability to transform an audio speech signal into text. Different authors have led the utilization of DL technologies in the context of ASR from different points of view; starting from generative and discriminative pretraining of Deep Neural Network (DNN) architectures [37, 36, 2, 5, 12, 11, 46], hybrid systems based on DNNs and Hidden Markov Models (HMMs) [4, 5, 16, 37] and their tandem counterpart [15, 13, 42] or even proposing new speaker adaptation techniques [14, 36, 19, 45, 48].

ASR research has gained a lot of interest during the last years for industry leaders, and this is mainly because of the huge amount of data in audio format that is generated nowadays and the myriad of applications this technology offers. Other than providing speech transcriptions, ASR enables analysis, classification and search functionalities on speech signals, and it also constitutes the starting point to many other NLP applications: Machine Translation (MT), Text-To-Speech synthesis (TTS), Text Summarization, Part-of-Speech tagging or Sentiment Analysis among others. It can be applied to many different scenarios such as in-car virtual assistants [20, 17, 22], health care medical documentation generation [28, 6], broadcast news, TV videos and video repositories in general [23, 21, 26, 1].

The transcription of video repositories is particularly helpful in the area of education where academic institutions have devoted big efforts to improve the way students can learn by means of new digital online media repositories. In fact, different solutions have emerged during the last years, the so-called Massive Open Online Courses (MOOCs) such as *edX* [10], *Coursera* [3] or *Udacity* [41], and video repositories such as *VideoLectures.NET* [44], *poliMedia* [30], *VideoApuntes* [43] or *TED talks* [40]. These platforms offer thousands of lecture-like videos from a wide range of topics and disciplines to reach as many users as possible. The video quality varies depending on the platform: from videos with just one speaker recorded in optimal acoustic conditions with a lapel microphone and semi-spontaneous speech, to videos recorded with a microphone array where several distant speakers interfere in a fully spontaneous manner, also known as far-field speech recognition. The availability of video transcriptions for these repositories opens new communication channels to better transfer knowledge and consequently expand its target audience, breaks down acoustic barriers for students with hearing impairments and, as mentioned before, enables lecture classification, search and analysis. However, the transcription of these videos is a costly task that cannot be easily performed by human experts as it involves a considerable expenditure in terms of time and money. At this point is where ASR systems are fundamental because they alleviate the problem by providing inexpensive yet reliable subtitles in a time efficient manner. Therefore, ASR systems constitute a powerful tool that not only provide automatic transcriptions but also expand the number of functionalities of these platforms.

Although the reliability of automatic transcriptions turns out to be crucial to guarantee a good learning experience, ASR systems are still far from producing perfect results. Different approaches can be followed to improve the system performance, such as increasing the size of the training data or adapting the systems to the speaker and domain, both in terms of acoustic or language model level. Also, confidence measures (CM) can play an important role in this regard. In its simplest aspect, CMs give an insight into the reliability of the recognized transcription by providing an score between 0 and 1 for each transcribed unit (usually at word level), which is by itself helpful and can further be used in different situations. First of all, in order to ensure good quality transcriptions a necessary step is to supervise its content by human experts. In this scenario, CMs can help in the smart selection of those speech segments that might require supervision (interactive speech transcription [34]), and therefore greatly reducing human effort. Secondly, CMs can also be used to alleviate a problem most ASR systems suffer: the mismatch between the acoustic conditions found during training and test. This mismatch is one of the main reasons behind the poor quality transcriptions that an ASR system might generate. As a result, the use of algorithms that adapt the model parameters to match the acoustic conditions to those of the testing phase have gained a lot of interest. Although, model adaptation based on the correct mapping between phoneme class labels and their corresponding sound segment should produce the best results (supervised adaptation), CM brings means to perform model parameters adaptation in an unsupervised fashion [8, 7] . Thirdly, unsupervised training of acoustic models can be also carried out using CMs by improving the data filtering and selection stage, as they can give an idea about the alignments quality between the acoustic signal and manual transcriptions. Finally, in a setup where several recognizers are run in parallel, CMs can be used to select the one that provides the best confidence score for the whole transcription.

As mentioned before, the potential benefits that ASR technologies and CMs estimation can provide on educational repositories are numerous. However, these repositories offer videos in a wide range of formats, with high variability in acoustic conditions, different speaker accents, spontaneous speech and very specific terminology. Therefore, the use of these technologies in this kind of repositories constitutes a very challenging task. On the one hand, the lack of open-source solutions and the requirement of skilled people to set them up drive away some organizations to exploit such functionalities. On the other hand, the solutions that provide video transcriptions are general purpose, and therefore, they do not take advantage of the meta data associated to each lecture (speaker, title, notes, slides, etc.) to further improve the transcription quality. In this context, this thesis aims at developing the latest ASR techniques to provide the best transcriptions to be used in real educational repositories. Moreover, another goal is to take advantage of the speaker meta data to propose new speaker adaptation techniques using Neural Networks (NNs). Finally, given that CMs have demonstrated to be very helpful in different situations when transcribing video repositories, the final goal is to propose state-of-the-art (SOTA) techniques to further improve their estimation and to propose new applications to better exploit their potential.

1.2 Framework

The work developed in this thesis has contributed to 4 different research projects: iTrans2, transLectures, EMMA and MORE. All share the continuous improvement and application of ASR technologies towards the efficient transcription of video repositories and therefore, to ease the communication channels by breaking down language barriers. Although iTrans2 overcomes the problem from a general perspective and serves as starting point, most of this thesis was developed during transLectures, EMMA and MORE which have a clear focus on educational video repositories. These last 3 projects also aspire to widen open education to any student no matter her/his mother tongue or hearing disabilities and to enrich the learning experience by facilitating new ways of learning.

The main goal of the transLectures project was to develop innovative, cost-effective tools for producing accurate transcriptions and translations of videos from different educational repositories. In this regard, the contributions of this thesis can be summarized in two main takeaways: On the one hand, the development of a new ASR toolkit to build systems capable of providing high-quality automatic transcriptions, which supported all the basic functionality to train an ASR system from scratch, including the training of Context-Dependent Deep Neural Networks (DNNs) [5, 4, 37]. On the other hand, considering ASR systems are far from producing perfect transcriptions, it was considered to develop tools to facilitate human supervision following an interactive approach. Although, the progress on CM estimation at that time was still on an early stage, it was observed that CMs could greatly reduce human supervision effort.

The project EMMA (European Multiple MOOC Aggregator) vision was to offer a unified MOOC aggregator platform capable of providing free, open and online courses from different European universities with the goal of preserving Europe's rich cultural, educational and linguistic heritage. In order to guarantee the maximum spread of these repositories, automatic translation systems play a fundamental role to break down linguistic barriers. These repositories are usually composed of Open Educational Resources (OERs) from different natures such as course materials, lecture slides or recorded lectures in audio or video format. All materials in text format are easily handled by MT systems, but in the case of video lectures, a first step in order to transcribe the audio is required. This thesis contributed to provide the best possible transcriptions for those video resources and therefore constitute a follow up work from the transLectures project, expanding the set of supported languages and also keeping the systems updated with the latest ASR techniques.

The MORE (Multilingual Open Resources for Education) project aimed to dramatically foster Open Education by providing multilingual access to OER and by enabling multilingual online communication in MOOC platforms. In this context, Spoken Language Translation (SLT), the task of translating a video from voice-to-voice plays a fundamental role. This kind of systems are built using three components: ASR, MT and TTS. It's a cascade-like task where first, the ASR system should provide high-quality transcriptions because otherwise its errors will propagate to the MT system and finally to the TTS system. Here again is fundamental to keep ASR systems updated to tackle with different accents, spontaneous speech, noise, false starts or hesitations.

The work of this thesis has contributed to achieve the goals of the different research projects by providing tools for building ASR systems and also by proposing new competitive speaker adaptation techniques that are fundamental given the heterogeneous nature of the repositories. Moreover, the progress made towards CM estimation by means of new Long Short-Term Memory (LSTM)-based models along with speaker adaptation techniques had a very positive impact given its importance in not only improving the overall transcription quality but also to carry out interactive speech transcription or to provide an smart utterance verification procedure.

1.3 Scientific and Technological Goals

The main goal of this thesis is to improve the performance of ASR systems in the context of educational video repositories. In order to guarantee its achievement, the following scientific goals can be derived:

- Improve ASR systems based on DNNs by means of unsupervised speaker adaptation (SA).
- Improve CM estimation by means of NNs and SA.

1.4 Contributions

Improve ASR systems based on DNNs by means of unsupervised SA

One of the contributions towards this goal has been the publication of the transLectures-UPV toolkit (TLK) which is presented in Paper 1. TLK has been continuously updated with the latest SOTA techniques and constitutes the basic tool used to build all the systems presented along the thesis. Its competitiveness has been particularly demonstrated in the two international competitions presented in Papers 2 and 3. It features a simple interface and provides all the functionality to build an ASR system from scratch: audio preprocessing and feature extraction (MFCCs and filter bank), training (based on HMMs or hybrid DNNs) and evaluation (following the so-called Viterbi decoding).

Regarding TLK, the main focus of this thesis has been the development of the internal DL tool for ASR. This tool was initially implemented in C++ and CUDA due to the lack of good-enough alternatives, and featured the training and evaluation of the most common NNs. At the time of writing this thesis, TLK provides support to train different NN topologies such as DNNs or Deep Convolutional NNs, common activation functions, multilingual NNs which are fundamental for languages with scarce resources, different speaker adaptation techniques, cross-entropy (CE) loss function, Maximum Mutual Information (MMI) training and support to train models using external DL libraries such as Tensorflow for recurrent-based topologies. With respect to the papers presented in this thesis, in Paper 2 TLK features hybrid-based systems with DNNs and CNNs trained following CE or MMI. Moreover, in Paper 3, Deep Bidirectional LSTMs (BLSTMs) are used for the first time using an external API, and finally in Paper 5 BLSTM-based acoustic models perform significantly better than DNNs.

Even though TLK was also updated with the latest unsupervised speaker adaptation techniques, a new one that made use of CMs is proposed in Paper 2, which constitutes another important contribution. This technique basically consists on adding an additional recognition step to the classical 2-steps (non-adapted and speaker-adapted recognitions), where the system is adapted again on a per-speaker fashion based on the output from the previous step. From that output, CMs are computed as word posterior probabilities or by means of other more advanced approaches, and finally, the acoustic model is fine-tuned over this data using as class-labels the CMs estimated. This technique provides competitive results in different setups, and also motivates the research on CM to further improve the recognition accuracy. Moreover, in Paper 5 it is demonstrated that even with high-quality ASR systems, this technique has a significant impact in the final transcription quality.

Improve CM estimation by means of NNs and SA

The contributions made towards this goal are by far the most important of this thesis. Looking at Confidence Estimation (CE) as a two-class classification problem in which class posterior probabilities are estimated combining word-level predictor features [35, 38, 33, 25], in Paper 4 a new approach is proposed where Bidirectional LSTMs models are used together with speaker adaptation techniques. The use of this kind of models is based on its modeling power of not only left-to-right context but also right-to-left. In this paper, it is shown that in a speaker independent setting, previous SOTA models such as Conditional Random Fields (CRFs) are systematically surpassed by Neural Network-based models. Other than that, it is proposed a simple yet effective speaker adaptation technique of CE models based on BLSTMs, that obtains consistent gains, obtaining the best results in the publicly available dataset LibriSpeech [27].

In Paper 5 more exhaustive experiments were carried out in 3 different datasets: LibriSpeech [27], poliMedia [30] and TED-LIUM [31]. Although CE accuracy is related to the quality of the ASR system, in this second work it is demonstrated that even for cutting-edge ASR systems (based on BLSTMs acoustic models), CE can benefit from the use of Neural Network-based models. This is shown in different tasks where RNN-based (BLSTMs and BRNNs) models outperform non-NNs models, and the combination of the two obtains the best results. Apart from that, a real adaptation setup is presented, where a general speaker-independent CE system is trained and afterwards adapted with speaker-dependent data from a different corpus. Moreover, the acoustic model is improved by means of the unsupervised speaker adaptation technique presented in Paper 2 and the best RNN-based CE system.

1.5 List of Publications

In this section, all the international publications published under the scope of this thesis where the PhD student is first author are summarized. The publications are presented in chronological order and classified according to their type (journals or international conferences) as well as whether they are listed in Journal Citation Reports[a] (JCR), in Computing and Research Education Association of Australasia[b] or GII-GRIN-SCIE Conference rating[c] (GGS).

[a] http://thomsonreuters.com/journal-citation-reports

[b] http://www.core.edu.au

[c] http://gii-grin-scie-rating.scie.es

1.5.1 Paper 1

Title	The transLectures-UPV Toolkit
Authors	M. A. Del-Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, A. Juan
Year	2014
Type	Journal
DOI	10.1007/978-3-319-13623-3
Name	Lecture Notes in Computer Science
Pages	269-278

In this paper, the transLectures-UPV toolkit (TLK) is presented, an ASR toolkit with clear focus on transcribing video lectures but general enough as to be applied in conventional ASR. TLK implements all the functionalities required to develop an ASR system from scratch. It can be applied from feature extraction (standard Mel Frequency Cepstral Coefficients or Filter Bank), acoustic model training based on Hidden-Markov-Models (HMMs) using the well-known Baum-Welch and Viterbi algorithms, training based on tied-state HMMs, DNNs training following an hybrid approach (Context Dependent HMMs-DNNs), acoustic model adaptation using the so called Maximum Likelihood Linear regression (MLLR) or its constrained version (CMLLR), and Viterbi-based decoding using an external language model. Moreover, features a simple yet effective interface as to easily perform each of the steps mentioned before.

In order to assess the toolkit performance, two case studies are presented: VideoLectures.NET, a video lecture repository mainly in English; and poliMedia, a Spanish and Catalan video repository developed at the Universitat Politècnica de València (UPV). As mentioned in Section 1.1, the generation of usable subtitles for videos that belong to this kind of repositories is not an easy task; spontaneous speech, different accents, technical terminology, etc. However, the experimental section of the paper shows that TLK is a competitive ASR toolkit which offers high quality transcriptions even in real-world educational repositories. In fact, it's shown a very good system performance in terms of Word Error Rate (WER) for video lectures in Spanish (12.8%), Catalan (20.1%) and English (22.7%).

1.5.2 Paper 2

Title	The MLLP ASR Systems for IWSLT 2015
Authors	M. A. Del-Agua, A. Martínez-Villaronga, S. Piqueras, A. Giménez, A. Sanchis, J. Civera, A. Juan
Year	2015
Type	Workshop
Name	The International Workshop on Spoken Language Translation
Pages	39-44

In this publication, a new unsupervised speaker adaptation technique is presented, which constitutes one of the contributions of this thesis related to the first goal. The systems presented were evaluated in the context of the challenging International Workshop on Spoken Language Translation (IWSLT). Its participants are encouraged to develop state-of-the-art solutions to perform the full process of SLT (ASR, MT and TTS) on real-world scenarios. In fact, the case study is composed of TED talks, which consist of videos from a set of conferences around the world carried out by the non-profit organization *Sapling Foundation*.

Two ASR systems for the IWSLT 2015 evaluation campaign were presented, which cover ASR for English and German. Most effort went into the development of the English recognition system which is based on the ROVER combination of five subsystems. Each of those subsystems was based on CD-HMM-DNNs with different input features (MFCCs and filter bank), activation functions (sigmoid and rectified linear) as well as various architectures such as CNN. These systems follow a three-step recognition approach where after the first pass recognition, CMLLR speaker adaptation applied to the input features is used (fMLLR) and finally the new unsupervised speaker adaption technique is applied. This last step is the proposed new speaker adaptation step for NNs, where the main idea is to use inexpensive yet reliable unsupervised speech data to further adapt the NNs. Thanks to the use of CMs at word level (computed as word posterior probabilities) from the second recognition output, the NNs were fine-tuned on a per-speaker basis using the CMs as new pseudo-truth class labels.

The new speaker-adaption technique provided consistent gains of about 1.2% to 3.7% relative improvements depending on the system setup. In the context of the thesis work, this new unsupervised adaptation technique opened the door of ASR system improvement through the research of new CMs estimation techniques. The final English system obtained 13.3% WER, which constitutes a very competitive performance, while the German system constituted the first large scale speech recognition system trained using TLK. The final results of the competition can be seen in Chapter 4.

1.5.3 Paper 3

Title	The MLLP system for the 4th CHiME challenge
Authors	M. A. Del-Agua, A. Martínez-Villaronga, A. Giménez, A. Sanchis, J. Civera, A. Juan
Year	2016
Type	Workshop
Name	The 4th CHiME Speech Separation and Recognition Challenge
Pages	57-59

The main goal of this work is to propose new model architectures based on DL that incorporate a third unsupervised speaker adaptation step in a challenging task where the systems face audio from multiple channels. Particularly, The CHiME Speech Separation and Recognition Challenge invites participants to built ASR systems that are capable of working in challenging and real noisy conditions in a muli-channel setting. The challenge consists of 3 different tracks: 1-channel, 2-channels and 6-channels. Each of which are different depending on the number of channels available at test time. Therefore, the easiest track is the 6-channels track which offers the best setup for the application of audio enhancement techniques, while in the 1-channel or 2-channels, the systems should be able to handle audio from any of the 6 channels microphones (or apply audio enhancement in the case of 2-channels). This challenge is one of the most relevant for ASR, where the best research groups in the world participate.

Even though this challenge could be faced from audio preprocessing, language modeling or acoustic modeling point of views, our participation was focused on the acoustic modeling part. In particular, a system combination of two sub-systems based on DNNs and BLSTMs following the hybrid approach was presented. Both systems were trained on the same data, and tested in the most challenging 1-channel and 2-channel tracks. The final system combination obtained 32% and 22.7% relative improvements over the 1-channel and 2-channels baselines, which represents a good result taking into account the simplicity of the approach. This work also reflects the continuous development of the TLK toolkit, which in this case added support to train and decode using external tools, and in particular using Tensorflow for BLSTM models. A comparison of the final results can be found in Chapter 4.

1.5.4 Paper 4

Title	ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks
Authors	M. A. Del-Agua, S. Piqueras, A. Giménez, A. Sanchis, J. Civera, A. Juan
Year	2016
Type	International Conference - Core A - GGS A
DOI	10.21437/Interspeech.2016-1142
Name	InterSpeech 2016
Pages	3464-3468

In this paper, a novel approach to model CM is presented. In particular, it is proposed for the first time the use of BLSTMs in conjunction with speaker adaptation techniques, which constitutes a contribution with respect to the second goal of this thesis. The use of BLSTMs is motivated because of its ability to model long-span relations, and thanks to its bidirectional nature not only past context is taken into account but also future context. Moreover, the application of speaker adaptation techniques is proposed as a fine-tuning step starting from a general CE system where a different system is trained for each speaker using speaker-dependent input features based on speaker-dependent vocabularies.

The experiments carried out are based on the publicly available LibriSpeech [27] corpus. On the one hand, a competitive ASR system was built where the acoustic model part was trained with TLK using a subset of 100 hours, while as language model was used a pre-built 4-gram provided by the authors of the corpus. On the other hand, the speaker-independent (SI) CM system was trained using a different subset of 50 hours from the same corpus. Additionally, 20 speakers not used in the SI experiments were randomly selected in order to evaluate speaker adaptation techniques of SI CE models.

The use of BLSTM Networks along with speaker-adaptation techniques constitutes a novelty in word level CE. The results obtained in LibriSpeech show that this approach improves over previous SOTA approaches such as CRFs in CM estimation. In this work, the speaker-independent CM system based on BLSTM networks was able to produce relative reductions in terms of Classification Error Rate (CER) of 4.7%, while the speaker-adapted system was able to further reduce CER in 4.6%.

1.5.5 Paper 5

Title	Speaker-Adapted Confidence Measures for ASR Using Deep Bidirectional Recurrent Neural Networks
Authors	M. A. Del-Agua, A. Giménez, A. Sanchis, J. Civera, A. Juan
Year	2018
Type	Journal - IF 2.950 - Ranking 5/31 - Quartile 1
DOI	10.1109/TASLP.2018.2819900
Name	IEEE/ACM Transactions on Audio, Speech, and Language Processing
Pages	1198-1206

Following the same line of research from the previous paper, in this work it is presented a comprehensive study on the improvement of CM and its applications through the use of RNN-based classifiers and speaker adaptation. These classifiers and their bidirectional versions (BRNNs and BLSTMs) have demonstrated their superiority in the estimation of CMs compared to non-NN-based classifiers [25, 24, 9]. Moreover, the adaptation of CM has shown to be very effective in improving baseline system performance [33, 9, 47, 18]. In scenarios where there are scarce resources to estimate speaker-specific models, this is an important feature, since it allows to adapt a general model with limited speaker data.

In this work, new technical contributions are reported, including an improved system architecture for CE in which word embeddings and CE models are jointly trained (rather than using pre-trained Glove [29] word-vectors). Moreover, new experimental results on CE are shown using state-of-the-art ASR systems based on BLSTM acoustic models and a large speech corpus consisting of 1000 hours from the English LibriSpeech task and 800 hours from the Spanish poliMedia task [27, 39]. New speaker-adapted experiments are also carried out considering a realistic task in which CM are applied into a multi-task framework. With this in mind, RNN-based confidence classifiers trained in the LibriSpeech corpus are adapted to speakers from the TED-LIUM corpus [32]. Finally, a novel unsupervised adaptation method of the acoustic DBLSTM model based on CMs is proposed to improve the overall accuracy of the speech recognition system.

Different experiments have been carried out in order to test all the contributions. From the CE systems point of view, in both LibriSpeech and poliMedia, RNN models clearly outperform non-NN-based classifiers which is statistically significant at the 95% confidence level to a great extent. Regarding the speaker-adapted CM, it is shown that in general speaker-adapted systems improve their non-adapted counterparts and relative $2 - 8\%$ CER reductions are obtained depending on the amount of speaker data and its quality. Finally, with respect to the use of CM for acoustic model adaptation, relative reductions of 3.3%, 3.8% and 5.5% in terms of WER are achieved in LibriSpeech, poliMedia and TED-LIUM respectively compared to the non-adapted systems.

1.6 Document Structure

This document is structured in four sequential chapters that cover the topics, scientific and technological goals proposed in this thesis.

Chapter 1. Introduction: In this chapter, the motivation, goals and main contributions of the thesis are presented. Also, the research projects involved during the development of this thesis are briefly described. These projects also involve some real case studies were the tools and techniques presented are assessed. Finally, the main contributions of each selected publication are summarized.

Chapter 2. Selected Papers: In this chapter, all the resulting publications of this thesis are presented.

Chapter 3. General Discussion of the Results: This chapter presents a general discussion of the main results derived of this thesis.

Chapter 4. Conclusions and Future Work: In this last section, the conclusions and future remarks are presented.

1.7 Abbreviations and Acronyms

AI	–	Artificial Intelligence
ASR	–	Automatic Speech Recognition
BLSTM	–	Bidirectional Long Short-Term Memory
BRNN	–	Bidirectional Recurrent Neural Network
CE	–	Confidence Estimation
CER	–	Classification Error Rate
CM	–	Confidence Measure
CMLLR	–	Constrained Maximum Likelihood Linear Regression
CNN	–	Convolutional Neural Network
CRF	–	Conditional Random Field
CV	–	Computer Vision
DL	–	Deep Learning
DNN	–	Deep Neural Network
EMMA	–	European Multiple MOOC Aggregator project
fDLR	–	feature-space Discriminative Linear Regression
fMLLR	–	feature-space Maximum Likelihood Linear Regression
GGS	–	GII-GRIN-SCIE Conference Rating
GII	–	Group of Italian Professors of Computer Engineering
GRIN	–	Group of Italian Professors of Computer Science
HMM	–	Hidden Markov Model
iTrans2	–	Interactive Transcription and Translation project
IWSLT	–	International Workshop on Spoken Language Translation
JCR	–	Journal Citation Reports
LSTM	–	Long Short-Term Memory
MFCC	–	Mel Frequency Cepstral Coefficients

EMMA	–	European Multiple MOOC Aggregator project
fDLR	–	feature-space Discriminative Linear Regression
fMLLR	–	feature-space Maximum Likelihood Linear Regression
GGS	–	GII-GRIN-SCIE Conference Rating
GII	–	Group of Italian Professors of Computer Engineering
GRIN	–	Group of Italian Professors of Computer Science
HMM	–	Hidden Markov Model
iTrans2	–	Interactive Transcription and Translation project
IWSLT	–	International Workshop on Spoken Language Translation
JCR	–	Journal Citation Reports
LSTM	–	Long Short-Term Memory
MFCC	–	Mel Frequency Cepstral Coefficients
MLLP	–	Machine Learning and Language Processing Research Group
MLLR	–	Maximum Likelihood Linear Regression
MMI	–	Maximum Mutual Information
MOOC	–	Massive Open Online Course
MORE	–	Multilingual Open Resources for Education
MT	–	Machine Translation
NLP	–	Natural Language Processing
NN	–	Neural Network
OER	–	Open Educational Resource
PR	–	Pattern Recognition
ReLU	–	Rectified Linear Unit
RNN	–	Recurrent Neural Network
ROC	–	Receiver Operating Characteristic
SA	–	Speaker Adaptation
SCIE	–	Spanish Computer-Science Society
SI	–	Speaker Independent
SLT	–	Spoken Language Translation
SOTA	–	State Of The Art
TLK	–	transLectures-UPV Toolkit
transLectures	–	Transcription and Translation of video lectures
TTS	–	Test-To-Speech synthesis
UPV	–	Universitat Politècnica de València
WER	–	Word Error Rate

1.8 References

[1] Aitor Álvarez, Arantza del Pozo, and Andoni Arruti. “APyCA: Towards the Automatic Subtitling of Television Content in Spanish”. In: *International Multiconference on Computer Science and Information Technology - IMCSIT 2010, Wisla, Poland, 18-20 October 2010, Proceedings*. 2010, pp. 567–574 (cit. on p. 2).

[2] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. “Greedy layer-wise training of deep networks”. In: *Advances in neural information processing systems*. 2007, pp. 153–160 (cit. on p. 2).

[3] *Coursera: Take the World’s Best Courses, Online, For Free*. `http://www.coursera.org` (cit. on p. 2).

[4] George E Dahl, Dong Yu, Li Deng, and Alex Acero. “Large vocabulary continuous speech recognition with context-dependent DBN-HMMs”. In: *ICASSP*. IEEE. 2011, pp. 4688–4691 (cit. on pp. 2, 4).

[5] George Dahl, Dong Yu, Li Deng, and Alex Acero. “Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (receiving 2013 IEEE SPS Best Paper Award)* 20.1 (2012), pp. 30–42 (cit. on pp. 2, 4).

[6] Gary C David, Angela Cora Garcia, Anne Warfield Rawls, and Donald Chand. “Listening to what is said–transcribing what is heard: the impact of speech recognition technology (SRT) on the practice of medical transcription (MT)”. In: *Sociology of health & illness* 31.6 (2009), pp. 924–938 (cit. on p. 2).

[7] Miguel Ángel Del-Agua, Adriá Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. “Speaker-Adapted Confidence Measures for ASR Using Deep Bidirectional Recurrent Neural Networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.7 (2018), pp. 1194–1202 (cit. on p. 3).

[8] Miguel Ángel Del-Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchis, and Alfons Juan. “The transLectures-UPV toolkit”. In: *Proc. of VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*. Las Palmas de Gran Canaria (Spain), Jan. 1, 2014 (cit. on p. 3).

[9] Miguel Ángel Del-Agua, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. “ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks”. In: *Interspeech*. 2016, pp. 3464–3468 (cit. on p. 10).

[10] *edX: Access to free education for everyone*. `http://www.edx.org` (cit. on p. 2).

[11] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. “Why does unsupervised pre-training help deep learning?” In: *Journal of Machine Learning Research* 11.Feb (2010), pp. 625–660 (cit. on p. 2).

[12] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. “The difficulty of training deep architectures and the effect of unsupervised pre-training”. In: *Artificial Intelligence and Statistics*. 2009, pp. 153–160 (cit. on p. 2).

[13] Petr Fousek, Lori Lamel, and Jean-Luc Gauvain. “Transcribing broadcast data using MLP features”. In: *Ninth Annual Conference of the International Speech Communication Association*. 2008 (cit. on p. 2).

[14] Roberto Gemello, Franco Mana, Stefano Scanzio, Pietro Laface, and Renato De Mori. “Linear hidden transformations for adaptation of hybrid ANN/HMM models”. In: *Speech Communication* 49.10-11 (2007), pp. 827–835 (cit. on p. 2).

[15] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. “Tandem connectionist feature extraction for conventional HMM systems”. In: *ICASSP*. IEEE. 2000, pp. 1635–1638 (cit. on p. 2).

[16] Geoffrey Hinton, Li Deng, Dong Yu, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath George Dahl, and Brian Kingsbury. “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. In: *IEEE Signal Processing Magazine* 29.6 (Nov. 2012), pp. 82–97 (cit. on p. 2).

[17] Thomas Kuhn, Akhtar Jameel, M Stumpfle, and Afsaneh Haddadi. “Hybrid in-car speech recognition for mobile multimedia applications”. In: *Vehicular Technology Conference, 1999 IEEE 49th*. Vol. 3. IEEE. 1999, pp. 2009–2013 (cit. on p. 2).

[18] Kshitiz Kumar, Chaojun Liu, and Yifan Gong. “Normalization of ASR confidence classifier scores via confidence mapping”. In: *Interspeech*. 2014, pp. 1199–1203 (cit. on p. 10).

[19] Bo Li and Khe Chai Sim. “Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems”. In: *Eleventh Annual Conference of the International Speech Communication Association*. 2010 (cit. on p. 2).

[20] Philip Lockwood and Jérome Boudy. “Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars”. In: *Speech communication* 11.2-3 (1992), pp. 215–228 (cit. on p. 2).

[21] Hugo Meinedo, Márcio Viveiros, and João Paulo Neto. “Evaluation of a live broadcast news subtitling system for portuguese”. In: *Interspeech*. 2008, pp. 508–511 (cit. on p. 2).

[22] Rajitha Navarathna, Patrick Lucey, David Dean, Clinton Fookes, and Sridha Sridharan. “Lip detection for audio-visual speech recognition in-car environment”. In: *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE. 2010, pp. 598–601 (cit. on p. 2).

[23] João Paulo Neto, Hugo Meinedo, Márcio Viveiros, Renato Cassaca, Ciro Martins, and Diamantino Caseiro. “Broadcast news subtitling system in Portuguese”. In: *ICASSP*. IEEE. 2008, pp. 1561–1564 (cit. on p. 2).

[24] A. Ogawa and T. Hori. “ASR error detection and recognition rate estimation using deep bidirectional recurrent neural networks”. In: *ICASSP*. IEEE. 2015, pp. 4370–4374 (cit. on p. 10).

[25] Atsunori Ogawa and Takaaki Hori. “Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks”. In: *Speech Communication* 89.4 (2017), pp. 70–83 (cit. on pp. 6, 10).

[26] Alfonso Ortega, José Enrique García Laínez, Antonio Miguel, and Eduardo Lleida. “Real-time live broadcast news subtitling system for Spanish”. In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. 2009, pp. 2095–2098 (cit. on p. 2).

[27] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. “Librispeech: an ASR corpus based on public domain audio books”. In: *ICASSP*. IEEE. 2015, pp. 5206–5210 (cit. on pp. 6, 9, 10).

[28] Ronaldo Parente, Ned Kock, and John Sonsini. “An analysis of the implementation and impact of speech-recognition technology in the healthcare sector”. In: *Perspectives in Health Information Management/AHIMA, American Health Information Management Association* 1 (2004) (cit. on p. 2).

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global Vectors for Word Representation.” In: *EMNLP*. Vol. 14. 2014, pp. 1532–1543 (cit. on p. 10).

[30] *poliMedia*. `https://polimedia.upv.es/` (cit. on pp. 2, 6).

[31] Anthony Rousseau, Paul Deléglise, and Yannick Estève. “Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks.” In: *LREC*. 2014, pp. 3935–3939 (cit. on p. 6).

[32] Anthony Rousseau, Paul Deléglise, and Yannick Estève. “Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks.” In: *LREC*. 2014, pp. 3935–3939 (cit. on p. 10).

[33] Isaias Sanchez-Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, and Alfons Juan. “Speaker-adapted confidence measures for speech recognition of video lectures”. In: *Computer Speech & Language* 37 (2016), pp. 11–23 (cit. on pp. 6, 10).

[34] Isaías Sánchez-Cortina, Nicolás Serrano, Alberto Sanchis, and Alfons Juan. “A prototype for Interactive Speech Transcription Balancing Error and Supervision Effort”. In: *Proc. of the 2012 ACM IUI*. 2012, pp. 325–326 (cit. on p. 3).

[35] Alberto Sanchis, Alfons Juan, and Enrique Vidal. “A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 565–574 (cit. on p. 6).

[36] Frank Seide, Gang Li, Xie Chen, and Dong Yu. “Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription.” In: *Proc. of ASRU*. 2011, pp. 24–29. ISBN: 978-1-4673-0365-1 (cit. on p. 2).

[37] Frank Seide, Gang Li, and Dong Yu. “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks.” In: *Interspeech*. 2011, pp. 437–440 (cit. on pp. 2, 4).

[38] Matthew Stephen Seigel. “Confidence Estimation for Automatic Speech Recognition Hypotheses”. PhD thesis. Department of Engineering, University of Cambridge, 2013 (cit. on p. 6).

[39] Joan Albert Silvestre, Miguel Ángel Del-Agua, Gonzalo Vicente Garcés, Guillem Gascó, Adrián Giménez, Adrià Agustí Martínez-Villaronga, Alejandro Manuel Pérez-González, Isaías Sánchez-Cortina, Nicolás Serrano, Rachel Nadine, et al. "transLectures". In: *IberSPEECH 2012-VII Jornadas en Tecnologia del Habla and III Iberian SLTech Workshop*. IberSPEECH 2012. 2012, pp. 345–351 (cit. on p. 10).

[40] *TED: Ideas worth spreading*. `https://www.ted.com` (cit. on p. 2).

[41] *Online Courses for the Digital Economy*. `https://eu.udacity.com` (cit. on p. 2).

[42] Fabio Valente, Mathew Magimai Doss, Christian Plahl, Suman V. Ravuri, and Wen Wang. "A comparative large scale study of MLP features for Mandarin ASR". In: *Eleventh Annual Conference of the International Speech Communication Association*. 2010 (cit. on p. 2).

[43] *VideoApuntes*. `https://videoapuntes.upv.es/` (cit. on p. 2).

[44] *Videolectures.NET*. `http://www.videolectures.net/` (cit. on p. 2).

[45] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. "Adaptation of context-dependent deep neural networks for automatic speech recognition". In: *Proc. of the SLT*. 2012, pp. 366–369 (cit. on p. 2).

[46] Dong Yu, Li Deng, and George Dahl. "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition". In: *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. 2010 (cit. on p. 2).

[47] Dong Yu, Jinyu Li, and Li Deng. "Calibration of Confidence Measures in Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 19.8 (2011), pp. 2461–2473 (cit. on p. 10).

[48] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and F. Seide. "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition". In: *ICASSP*. IEEE. 2013, pp. 7893–7897 (cit. on p. 2).

Chapter 2

Selected Papers

1 The transLectures-UPV Toolkit

M. A. Del-Agua and A. Giménez and N. Serrano and J. Andrés-Ferrer and J. Civera and A. Sanchis and A. Juan

VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop. Lecture Notes. Pages 269-278

Las Palmas de Gran Canaria (Spain)

DOI 10.1007/978-3-319-13623-3

19-21 November 2014

The transLectures-UPV Toolkit

M. A. Del-Agua and A. Giménez and N. Serrano and J. Andrés-Ferrer and J. Civera and A. Sanchis and A. Juan

Abstract

Over the past few years, online multimedia educational repositories have increased in number and popularity. The main aim of the transLectures project is to develop cost-effective solutions for producing accurate transcriptions and translations for large video lecture repositories, such as VideoLectures.NET or the Universitat Politècnica de València's repository, poliMedia . In this paper, we present the transLectures-UPV toolkit (TLK), which has been specifically designed to meet the requirements of the transLectures project, but can also be used as a conventional ASR toolkit. The main features of the current release include HMM training and decoding with speaker adaptation techniques (fCMLLR). TLK has been tested on the VideoLectures.NET and poliMedia repositories, yielding very competitive results. TLK has been released under the permissive open source Apache License v2.0 and can be directly downloaded from the transLectures website.

1.1 Introduction

Online multimedia repositories are on the rise and becoming evermore consolidated as key knowledge assets. This is particularly true in the educational area where large repositories of video lectures are being established on the back of increasingly available and standardized infrastructures. A well-known example of this is VideoLectures.NET, a free and open access web portal that has so far published more than 15K educational videos. VideoLectures.NET is a major player in the diffusion of the open source Matterhorn platform currently being adopted by many institutions and organizations within the Opencast community [9]. Other examples include massive open online course (MOOCs) aggregators, such as Coursera, Udacity, EdX, Udemy, iVersity, UPV[x] and others.

The generation of subtitles for these repositories is a costly task, both in terms of time and money, which prohibits many repositories from having their videos transcribed. Most of the video lectures available on VideoLectures.NET and MOOC aggregators, for instance, are not transcribed, despite the obvious benefits of doing so, including the incorporation of search and analysis functions. In order to overcome this deficit, the transLectures project aims to develop innovative, cost-effective solutions for producing accurate transcriptions and translations for video lectures. The project has two case studies: the aforementioned VideoLectures.NET, and poliMedia, a Spanish and Catalan video lecture repository developed at the Universitat Politècnica de València (UPV).

An important area of work at transLectures is to develop solutions that can be easily transferred to other repositories beyond VideoLectures.NET and poliMedia. With this in mind, the transLectures-UPV team has developed a whole series of transferable tools, including online applications. This paper is focused on just one of these tools, the transLectures-UPV toolkit (TLK). TLK implements all the functionalities required to develop an automatic speech recognition (ASR) system. Although developed as part of the transLectures project to meet the specific requirements of video lecture transcription, it can also be used as a conventional ASR toolkit, like HTK [20], RASR [14] or KALDI [12]. In this paper, we go into more detail about this toolkit, which can be freely downloaded [18] under the permissive (for research and commercial purposes alike) Apache License v2.0.

This paper is organised as follows. Section 1.2 describes the different tools forming part of TLK that can be used either to build an ASR system or simply to transcribe input media files. A practical guide to the development of an ASR system using TLK is given in Section 1.3. Finally, the performance of TLK is assessed in Section 1.4, and some conclusions are given in Section 5.5.

1.2 Overview of the Toolkit

TLK can be divided into three major components: the library, the basic command line tools and the high-level command line tools. The library, named `libTLK`, is an ANSI C library and implements the core functionalities of TLK (feature extraction, parameter estimation, decoding, adaptation, etc.). A set of basic command line tools have been defined to use `libTLK`. Based on these basic tools, high-level command line tools have also been developed in order to carry out the main steps involved in building an HMM-based ASR system: preprocessing, training and decoding.

Building an ASR System Using TLK Tools

As illustrated in Fig. 2.1, an ASR system can be built using three high-level TLK tools: `tLtask-preprocess`, `tLtask-train` and `tLtask-recognise`.

tLtask-preprocess

This tool takes time-segmented audio signals and the corresponding transcriptions as input and performs feature extraction and phonetic annotation. It also extracts clusters from the input audio, which can be used for speaker or video adaptation, and other useful data like the original or non-punctuated text.

`tLtask-preprocess` uses the `tLextract` basic command tool to perform the Mel-Frequency Cepstral Coefficients (MFCC) feature extraction process as described in [20]. `tLextract` supports a large number of audio file formats since it uses the `libsox` library. The parameters involved in the extraction process are easy to configure: sampling frequency, duration of the extraction window, number of cepstral coefficients, etc. Furthermore, `tLextract` also allows the application of a mean variance normalization to the input samples.

The phonetic transcription is obtained using different auxiliary scripts depending on the input language. The current release supports Spanish and Catalan.

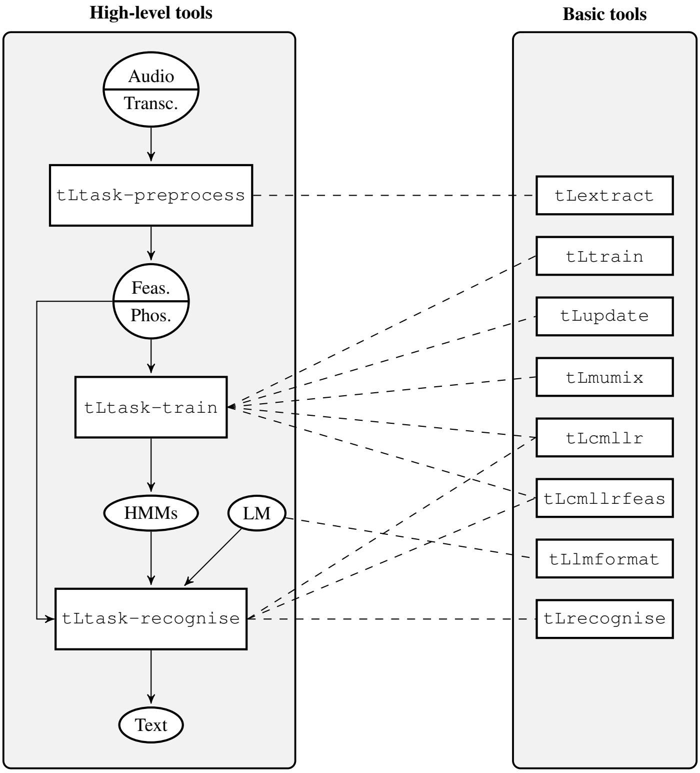

Figure 2.1: Building an ASR system using TLK tools.

tLtask-train.

This tool takes the output from `tLtask-preprocess` and performs the following training schema to estimate the HMMs:

1. Standard model training: monophone training, triphone training, transformation of the triphone model to a tied phoneme model, tied phoneme training.
2. Estimation of CMLLR matrices and CMLLR features.
3. CMLLR model training: CMLLR monophone training, CMLLR triphone training, CMLLR transformation of the triphone model to a tied phoneme model, CMLLR tied phoneme training.

This is the training schema for a two-step recognition system using fCMLLR features [3], and tied-state triphone HMMs. The final standard and CMLLR models are made up of Gaussian mixture distributions estimated following on an iterative training schema in which mixture components are mixed at each iteration (mixing is performed using `tLmumix`). Tied-state triphone HMMs are estimated following a phonetic decision tree approach [21]. This technique is implemented as an auxiliary Python script based on predefined linguistic rules. These rules are implemented as regular expressions in Python and can be easily defined by users. The current release includes rules for English, Spanish and Catalan.

tLtask-train uses the tLtrain basic command tool which implements Baum-Welch and Viterbi algorithms for parameter estimation [1, 19]. tLtrain has been designed to be able to properly manage large corpora by scaling in cluster environments. Specifically, tLtrain is used by tLtask-train following a Map-Reduce approach. That is, training is split into two stages: a first stage in which tLtrain is used to compute statistics, which can be split over several independent processes; and a second stage where the statistics computed in the previous stage are merged using the basic command line tool tLupdate. It is worth noting that tLupdate has support for linear interpolation of counts which might be useful in an online learning schema. Additionally, tLtrain allows samples to be packed into tar files for a better I/O latency in a cluster environment.

tLtask-train uses additional basic command tools to complete the CMLLR model training. tLcmllr is used to calculate a transformation matrix over all Gaussian mixtures of a simple HMM using the Constrained MLLR algorithm (CMLLR), while tLcmllrfeas transforms samples into fCMLLR features using a CMLLR transformation matrix.

tLtask-recognise.

This tool transcribes audio samples produced by tLtask-preprocess using HMM models estimated by tLtask-train following a two-step recognition schema:

1. Recognition using the standard tied phoneme HMMs.
2. Estimation of CMLLR matrices.
3. CMLLR transformation of input samples.
4. Recognition using the CMLLR tied phoneme HMMs.

tLtask-recognise uses the basic tool tLrecognise, which implements the well-known Viterbi algorithm, to obtain the most probable hypothesis [19]. In addition to HMMs, a language model and a pronunciation dictionary must be provided for decoding. tLrecognise allows two different language model representations. If the language model is a wordnet (without back-off), decoding is carried out over a huge finite state model built by embedding HMMs into the states of the wordnet [20]. In contrast, if the language model is in ARPA format (back-off), the decoder follows a word-conditioned tree search approach [8]. Specifically, a prefix tree with all the possible pronunciations is pre-calculated. To speed up the process, prior to decoding (tLtask-recognise or tLrecognise), the language model must be transformed into an internal format. This transformation is carried out by the basic tool tLlmformat. tLrecognise implements several well-known pruning techniques: beam search, histogram pruning, word end pruning and look-ahead. Although look-ahead is not exactly a pruning technique, its use is highly recommended when pruning techniques are applied in conjunction with a prefix tree approach [10]. tLrecognise also supports the generation of lattices following the technique described in [11]. Two formats for lattices are

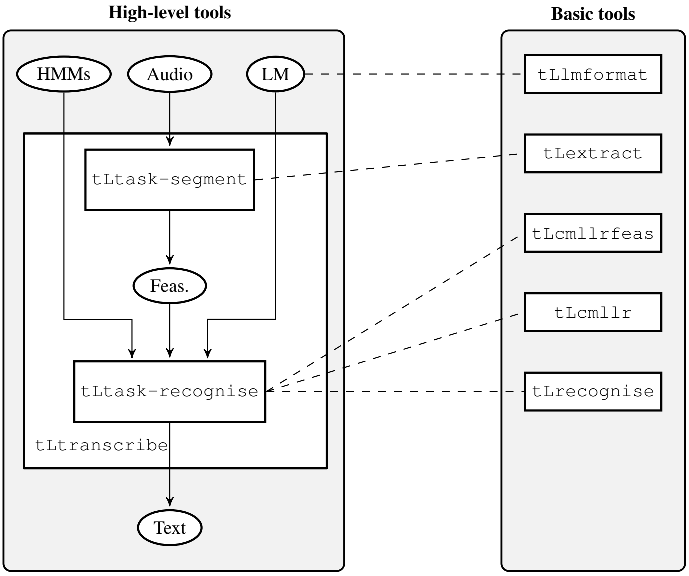

Figure 2.2: Transcribing media files with `tLtranscribe`.

supported: the TLK format and the HTK format [20]. If desired, lattices can be generated including information related to time alignment at phoneme level.

As with `tLtask-train`, `tLtask-recognise` has been designed to work well in cluster environments. Specifically, it can be configured to split recognition into parallel processes, and cache big files (like models) on host machines.

The output of `tLtask-recognise` is given in different formats: plain text, recognize output, CTM format [15], etc.

Using TLK Tools For Decoding Only

TLK includes a high-level tool named `tLtranscribe` that allows users to directly transcribe media files. This tool reads a preinstalled system, freeing the user from all the technical details. As illustrated in Fig. 2.2, `tLtranscribe` makes use of the high-level tools `tLtask-recognise` and `tLtask-segment`. The tool `tLtask-segment` uses `tLextract` to automatically perform the segmentation of the audio signal. For the purposes of testing the `tLtranscribe` tool, a system for Spanish transcription has been released under a Creative Commons Attribution 4.0 International License.

1.3 Using TLK

This section describes how an ASR system can be built using TLK following the process depicted in Fig. 2.1. A more detailed version of this tutorial is available on the transLectures website [18].

1. TLK installation and data preparation.

    - The current version of TLK runs on Linux and Mac OS X, and can be easily installed from the transLectures website.
    - Acoustic data is also available on the transLectures website and can be downloaded by executing:

      ```
      wget translectures.eu/files/tlk/tlk-tutorial-data.tgz
      tar -xzvf tlk-tutorial-data.tgz
      ```

      This will create the directory `tlk-tutorial-data`, which itself contains several directories. The `train` directory contains the data that will be used to train HMMs, while the `test` directory contains the data that will be used to asses the system. These data correspond to Spanish lectures recorded at Universitat Politècnica de València and their annotations in .trs and .dfxp format.
    - Now, running `tLtask-preprocess` the data is preprocessed obtaining the required files for training and evaluation:

      ```
      tLtask-preprocess es dfxp \\
            tlk-tutorial-data/train preprocess-train
      tLtask-preprocess es dfxp \\
            tlk-tutorial-data/test preprocess-test
      ```

      Note that the configuration options (i.e. `es` and `dfxp`) indicate the language and the file format, respectively.

2. HMM training:

    - First of all, a directory should be created to store the training files:

      ```
      mkdir training; cd training
      ```

    - Then, the two directories inside `preprocess-train` need to be linked to the training directory:

      ```
      ln -s ../preprocess-train/samples \\
            ../preprocess-train/lists .
      ```

    - Next, a template of the tool's configuration file `tLtask-train` should be generated:

      ```
      tLtask-train --write-example-config-file > \\
            config-file.ini
      ```

      This configuration file contains the default parameters needed to train standard HMMs for the Spanish language. In order to use previously preprocessed acoustic data, the `Lists` section of this configuration file has to be changed:

      ```
      [Lists]
      set_name = lists/samples
      ...
      [General]
      ...
      prefix-name = training-tutorial
      ```

- Finally, the following command runs the tool `tLtask-train` to perform the HMM training:

  ```
  tLtask-train config-file.ini --log-folder log
  ```

  The tool `tLtask-train` will execute all necessary commands to train HMMs following the training schema described in previous section Note that, although certain processes are executed in parallel depending on the computer, this process might take some time.

3. Automatic transcription:

   - As in the case of training, a directory should be created in the base directory for storing the automatic transcriptions:

     ```
     mkdir recognition; cd recognition
     ```

   - Also, some links must be created to the acoustic data and models:

     ```
     ln -s ../preprocess-test/samples ../preprocess-test/lists \
           ../preprocess-test/references ../training/models \
           ../tlk-tutorial-data/misc/mono.lex \
           ../tlk-tutorial-data/misc/mlm.gz .
     ```

   - The tool `tLtask-recognise` needs a configuration file, easily generated by running:

     ```
     tLtask-recognise --write-example-config-file > \\
           config-file.ini
     ```

     Some changes need to be made to this file in order to use previously preprocessed test data:

     ```
     [General]
     prefix-name = tutorial
     ...
     [HMM]
     prefix-name = training-tutorial
     ...
     [LM]
     language-model = mlm.gz
     lexicon = mono.lex
     ```

   - Finally, upon executing the following command, the test audio samples will be automatically transcribed following the two-step recognition schema described in previous section:

     ```
     tLtask-recognise config-file.ini --log-folder log
     ```

4. Measuring the transcription quality:

   - The sclite tool in SCTK is used to compute the Word Error Rate (WER) of the automatic transcriptions [15]:

     ```
     sclite -r references/<video_id>.stm \
            stm -h tutorial/cmllr_step2/transcription.ctm ctm
     ```

1.4 Empirical Results

TLK has been developed within the framework of the transLectures project to deal with the transLectures of video lectures. Specifically, ASR systems have been developed for three languages: English, Spanish and Catalan. The English ASR system has been developed for the transLectures of English lectures from the VideoLectures.NET repository. The Spanish and Catalan ASR systems have been developed for the poliMedia repository. For training and evaluation purposes, three databases have been developed by manually transLectures video lectures from these repositories. The main statistics of these speech databases are shown in Table 2.1.

Table 2.1: Main statistics of the English, Spanish and Catalan speech databases used in the transLectures project.

	English	Spanish	Catalan
Videos	28	704	210
Speakers	104	83	53
Hours	26.6	114.2	25.8
Sentences	7.3K	41.6K	13.7K
Running Words	192K	1M	198K
Vocabulary Size	13K	35.9K	24.4K

From each database some lectures were selected for evaluation purposes: 3.4h for Spanish and English, and 2.1h for Catalan. However, since video lectures from VideoLectures.NET are longer ($\approx$ 50min) than poliMedia lectures ($\approx$ 10min), this means just 4 videos were selected for English in absolute terms, while 23 and 16 videos were selected for Spanish and Catalan, respectively. The remaining data were used for training and development. For tasks where there was a lack of training data, as was the case for English and Catalan, the training data was increased by out-of-domain corpora.

The progress of the ASR systems developed within the transLectures project using TLK for each language is depicted in Fig. 2.3. As can be observed, the performances of the three systems have improved continuously throughout the project. In particular, very high performance levels have been achieved in Spanish (12.8% WER). Work began on the English and Catalan systems later than on the Spanish system (specifically, one year later). However, big improvements in WER have been achieved in the six-month period (20.1% in Catalan and 22.7% in English). In all languages, the performance is close or below 20% WER, which has been reported as the threshold under which ASR output becomes useful for users [7]. All these improvements can in part be explained by the fact that TLK has been under active development since the beginning of the project. This includes some features currently being tested, for example, hybrid models with deep neural networks (DNNs) [13, 2, 17], and multilingual DNNs [6]. It is worth noting that, in all cases, the language model used has about 200K words. Moreover, the percentage of out-of-vocabulary words is below 2% (1.7% for Spanish). For further details on the development of these systems, please refer to the public transLectures reports [5, 16, 4].

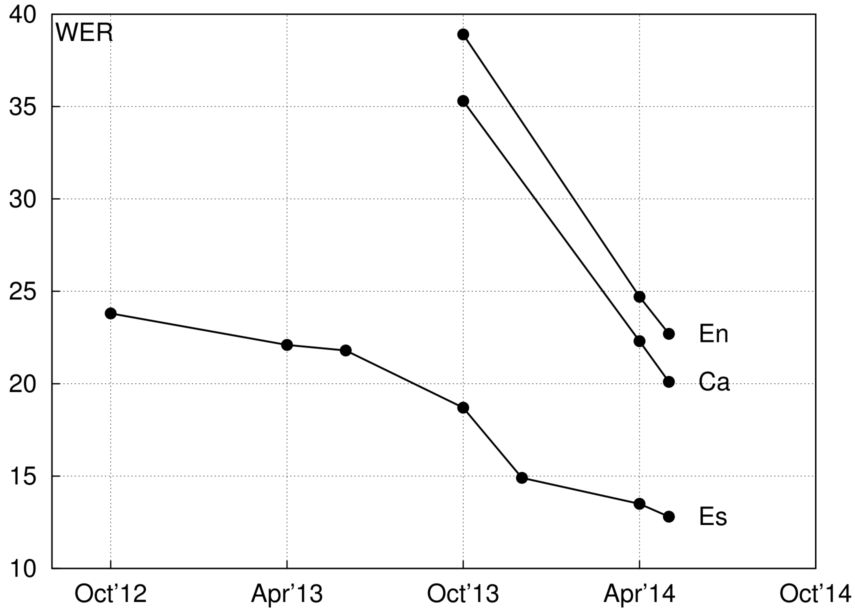

Figure 2.3: Progress measured in WER of the TLK ASR systems developed within the transLectures project for Spanish (Es), English (En) and Catalan (Ca).

1.5 Conclusions and Further Remarks

In this paper we have presented the transLectures-UPV ASR toolkit (TLK) based on HMMs. TLK implements well-known ASR features and released under the open source Apache License 2.0. The functionality of TLK has been recently extended, adding a new component that supports Deep Neural Networks (DNNs) following a hybrid decoding approach [2]. Although the current release does not include DNN training, with this still being at an experimental stage, it does include DNN support for recognition. In fact, beside the standard Gaussian HMM based Spanish system, we have also released a Spanish system based on DNNs. Both systems can be downloaded from the transLectures website [18].

As future work, we plan to improve TLK further by adding new state-of-the-art features, such as convolutional NNs or recurrent NNs. Also, we plan to carry out extensive, comparative tests with other toolkits.

Acknowledgments.

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

References

[1] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains”. In: *The Annals of Mathematical Statistics* 41.1 (1970), pp. 164–171. DOI: `http://dx.doi.org/10.2307/2239727` (cit. on p. 24).

[2] George Dahl, Dong Yu, Li Deng, and Alex Acero. “Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (receiving 2013 IEEE SPS Best Paper Award)* 20.1 (2012), pp. 30–42 (cit. on pp. 28, 29).

[3] V. Digalakis, D. Rtischev, L. Neumeyer, and Edics Sa. “Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures”. In: *IEEE Transactions on Speech and Audio Processing* 3 (1995), pp. 357–366 (cit. on p. 23).

[4] *Final report on massive adaptation (M36)*. URL: `http://www.translectures.eu//wp-content/uploads/2015/01/transLectures-D3.1.3-31Oct2014.pdf` (cit. on p. 28).

[5] *First report on massive adaptation (M12)*. URL: `https://www.translectures.eu/wp-content/uploads/2013/05/transLectures-D3.1.1-18Nov2012.pdf` (cit. on p. 28).

[6] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. “Cross-Language knowledge transfer using multilingual deep neural network with shared hidden layers”. In: *ICASSP*. IEEE. 2013, pp. 7304–7308 (cit. on p. 28).

[7] Cosmin Munteanu, Ronald Baecker, Gerald Penn, Elaine Toms, and David James. “The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives”. In: *Proc. of CHI*. 2006, pp. 493–502 (cit. on p. 28).

[8] H. Ney and S. Ortmanns. “Progress in dynamic programming search for LVCSR”. In: *Proceedings of the IEEE* 88.8 (2000), pp. 1224–1240. ISSN: 0018-9219. DOI: `10.1109/5.880081` (cit. on p. 24).

[9] *Opencast Matterhorn*. URL: `http://opencast.org/matterhorn/` (cit. on p. 21).

[10] S. Ortmanns, H. Ney, and A. Eiden. “Language-model look-ahead for large vocabulary speech recognition”. In: *Proc. of ICSLP*. Vol. 4. 1996, pp. 2095–2098. DOI: `10.1109/ICSLP.1996.607215` (cit. on p. 24).

[11] Stefan Ortmanns, Hermann Ney, and Xavier Aubert. “A word graph algorithm for large vocabulary continuous speech recognition”. In: *Computer Speech and Language* 11.1 (1997), pp. 43–72. ISSN: 0885-2308. DOI: `http://dx.doi.org/10.1006/csla.1996.0022` (cit. on p. 24).

[12] Daniel Povey et al. “The Kaldi Speech Recognition Toolkit”. In: *Proc. of ASRU*. 2011 (cit. on p. 22).

[13] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536 (cit. on p. 28).

[14] David Rybach et al. “The RWTH Aachen University Open Source Speech Recognition System”. In: *Interspeech*. 2009, pp. 2111–2114 (cit. on p. 22).

[15] *sclite - Score speech recognition system output*. URL: `http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm` (cit. on pp. 25, 27).

[16] *Second report on massive adaptation (M24)*. URL: `https://www.translectures.eu//wp-content/uploads/2014/01/transLectures-D3.1.2-15Nov2013.pdf` (cit. on p. 28).

[17] Frank Seide, Gang Li, Xie Chen, and Dong Yu. “Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription.” In: *Proc. of ASRU*. 2011, pp. 24–29. ISBN: 978-1-4673-0365-1 (cit. on p. 28).

[18] *TLK: The transLectures-UPV Toolkit*. URL: `https://www.translectures.eu/tlk/` (cit. on pp. 22, 25, 29).

[19] A.J. Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: *IEEE Transactions on Information Theory* 13.2 (1967), pp. 260–269. ISSN: 0018-9448. DOI: `10.1109/TIT.1967.1054010` (cit. on p. 24).

[20] S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 1995 (cit. on pp. 22, 24, 25).

[21] S. J. Young, J. J. Odell, and P. C. Woodland. “Tree-based state tying for high accuracy acoustic modelling”. In: *Proc. of HLT*. 1994, pp. 307–312 (cit. on p. 23).

2 The MLLP ASR Systems for IWSLT 2015

M. A. Del-Agua and A. Martínez-Villaronga and S. Piqueras and A. Giménez and A. Sanchis and J. Civera and A. Juan

Proc. of 12th International Workshop on Spoken Language Translation. Pages 39-44

Da Nang (Vietnam)

3-4 December 2015

The MLLP ASR Systems for IWSLT 2015

M. A. Del-Agua and A. Martínez-Villaronga and S. Piqueras and A. Giménez and A. Sanchis and J. Civera and A. Juan

Abstract

This paper describes the Machine Learning and Language Processing (MLLP) ASR systems for the 2015 IWSLT evaluation campaing. The English system is based on the combination of five different subsystems which consist of two types of Neural Networks architectures (Deep feed-forward and Convolutional), two types of activation functions (sigmoid and rectified linear) and two types of input features (fMLLR and FBANK). All subsystems perform an speaker adaptation step based on confidence measures the output of which is then combined with ROVER. This system achieves a Word Error Rate (WER) of 13.3% on the 2015 official IWSLT English test set.

2.1 Introduction

TED is a global set of conferences around the world carried out by the non-profit organisation *Sapling Foundation*. Its talks cover a wide range of different topics such as science, culture, economics or politics, always keeping in mind the slogan "ideas worth spreading". The speakers are given a maximum of 18 minutes to present their ideas in the most appealing way they can, typically in a storytelling format.

In order to ensure the maximum spread of these talks, turns out to be essential their transcription and translation. Big efforts have been devoted to this task, such as *The Open Translation Project* (OTP), which aims to reach out to the 4.5 billion people on the planet who do not speak English. Nevertheless, the OTP utilises crowd-based subtitling platforms, powered by volunteers to translate and caption the videos, which is still a very time-consuming task.

TED talks conform a very appropriate case study where new technologies can be applied. Particularly from the machine learning community, the International Workshop on Spoken Language Translation (IWSLT) organises a yearly challenge which aims at evaluating the core technologies in spoken language translation: automatic speech recognition (ASR), machine translation (MT) and spoken language translation (SLT). Automatically transcribing this kind of videos is still a challenging task due to the spontaneous nature of the speech; variety in acoustic conditions, the presence of disfluencies, hesitations and different accents states a great challenge even for cutting-edge technology in automatic automatic speech recognition.

This paper describes the English and German ASR systems developed in the MLLP group for the IWSLT 2015 evaluation campaign. Most effort went into the development of the English recognition system which is based on the ROVER combination of five subsystems.

Each of those subsystems was based on hybrid Deep Neural Networks Hidden Markov Models (DNN-HMM) [1] with different input features (MFCCs and filter bank), activation functions (sigmoid and rectified linear) as well as various architectures such as Deep Convolutional Neural Networks (CNN). It is worth noting that all of these systems were entirely trained using our own software; the transLectures-UPV toolkit.

The rest of this paper is organised as follows. Section 3.2 describes the ASR toolkit used for the experiments. In Section 2.3 the automatic audio segmentation technique is introduced. Section 2.4 is devoted to the English transcription system. Similarly, in Section 2.5 the German ASR system is described. Finally, conclusions are given in Section 2.6.

2.2 Translectures-UPV Toolkit

The transLectures-UPV toolkit (TLK) is composed by a set of tools that allows the development of an end-to-end speech recognition system. Its application range extends from feature extraction to HMM and DNN training and decoding. Since last state published of the toolkit [2] new state-of-the-art techniques have been added:

- DNN training and decoding hybrid based systems.
- Support to Convolutional NNs.
- Support to Multilingual NNs.
- DNN speaker adaptation techniques such as output-feature discriminant linear regression (oDLR) [11].
- DNN sequence discriminative training based on Maximum Mutual Information (MMI).

2.3 Audio Segmentation

The audio segmentation step performed by the MLLP group for English and German can be viewed as a simplified case of ASR, in which the system vocabulary is constituted by the power set of segment classes: speech and background noise.

Provided an audio stream $\vec{x}$, the segmentation problem can be stated from a statistical point of view as the search of a sequence of class labels $\hat{\vec{c}}$ so that

$$\hat{\vec{c}} = \operatorname*{argmax}_{\vec{c} \in \mathcal{C}^*} p(\vec{x} \mid \vec{c}) \, p(\vec{c}) \tag{2.1}$$

where, as in ASR, $p(\vec{x} \mid \vec{c})$ and $p(\vec{c})$ are modeled by acoustic and language models, respectively. In our case, it should be noted that each word is composed by a single phoneme.

Acoustic models were trained on MFCC feature vectors computed from acoustic samples using TLK. We used a 0.97 coefficient pre-emphasis filter and a 25 ms Hamming window that moves every 10 ms over the acoustic signal. From each 10ms frame, a feature vector of 12 MFCC coefficients is obtained using a 26 channel filter bank. Finally, the energy coefficient and the first and second time derivatives of the cepstrum coefficients are added to the feature vector.

Each segment class is represented by a single-state Hidden Markov Model (HMM) without loops, and its emission probability is modeled by a Gaussian Mixture Model (GMM). Acoustic HMM-GMM models were also trained using TLK, which implements the conventional Baum-Welch algorithm.

A 5-gram back-off language model with constant discount was trained on the sequence of class labels using the SRILM toolkit [8]. Finally, the segmentation process (search) was also carried out by the TLK toolkit.

2.4 English Transcription System

Acoustic Modeling

In this section the acoustic modeling process for the English system is described. First, the data selected for training is showed as well as the techniques used for its collection. Then, the training procedure is detailed along with all the subsystems associated.

Data Collection

This year, the IWSLT challenge allowed the use of any publicly available data for acoustic modeling, including TED talks without publication date restrictions (except those listed as disallowed). Given these requirements, roughly 400 hours of TED talks were downloaded from its official web-page [9].

The subtitles attached to a large part of the talks neither match the speaker's speech nor the timings. Therefore, a data filtering process is needed, in which those segments with a deficient or non-existent transcription must be removed. This process was performed in a similar manner to the data filtering performed for building the TEDLIUM corpus [5].

First of all, the input audio was segmented and preprocessed according to the caption timings. Secondly, a recognition step was performed using an out-of-domain acoustic model and a finite state language model. This finite state language model was built using the sequence of words from the reference with silence in-between, allowing loops (hesitations), initial state to any word transitions and from any word to final state transitions.

This way, those segments whose recognition does not match the reference suggest that either the timings are wrongly set or the system is unable to recognise the segment due to non-speech audio. Therefore, after decoding, all of these incorrectly recognised segments were removed, which left us a total of 245 hours of clean speech distributed among 1900 talks.

Training

Regarding feature extraction, two types of acoustic features were extracted. The first type of features are Mel-frequency cepstral coefficients (MFCC), which were extracted with a Hamming window of 25 ms, shifted at 10 ms intervals. The MFCC feature consisted of 16 MFCCs and their first and second derivatives (48-dimensional feature vectors). These feature vectors were then normalised by mean and variance at speaker level. After that, a single feature-space Maximum Likelihood Linear Regression (fMLLR) transform for each training speaker was then estimated and applied to perform speaker-adaptive training (SAT).

The second type of features are log Mel filter bank (FBANK) with first and second derivatives which left 120 dimension feature vectors.

Five different acoustic models were trained in our system using TLK. All of them consisted of context-dependent Deep Neural Networks (DNNs) following an hybrid approach. To train these models, we first trained a basic context dependent triphone HMM model, after which a second-pass feature-space Maximum Likelihood Linear Regression (fMLLR) was applied. This model yielded a total of 10492 tied states, estimated following a phonetic decision tree approach [12]. It is worth noting that, in order to obtain the best transcription as to better perform fMLLR, an standard DNN was trained using the MFCCs features. The five models were build on top of these HMM acoustic model and followed a three-pass recognition approach as shown in Fig. 2.4.

From Fig.2.4, the *fMLLR CD-DNN* module can be switched among the five different acoustic models. Three of them are feed-forward DNNs and the other two are Deep Convolutional Neural Networks (CNNs). From the first set, all models took as input MFCCs feature frames with a window size of 11. Moreover, all three subsystems shared the same topology: $528 - 2048 * 7 - 10492$, i.e., an input layer with 528 neurons, 7 hidden layers with 2048 neurons and an output layer of 10492 neurons. The pre-training phase technique is also shared, which consisted of the Discriminative Pretraining [7] approach. The first system was a DNN with sigmoid activation functions, trained with the cross-entropy (CE) criterion (10 epochs) and after that, with sequence discriminative training following the MMI criterion (hereafter DNN-mmi). The second model was a DNN with rectified linear activation functions, trained following the CE criterion during 45 epochs (hereafter DNN-relu). And the third model was a DNN with sigmoid activation functions trained with the CE criterion during 45 epochs (hereafter DNN-sigm).

Two models belong to the second set of acoustic models. Both take as input FBANK features with a window size of 11 and share the same topology. It consist of one convolution layer followed by a max pooling operation, 6 feed-forward hidden layers of 2048 units each, and an output layer of 10492. The convolutional layer is composed of 128 filters with a filter size of 9 and shift of 1. Meanwhile, the max-pooling layer was configured with a pooling width and shift of 2. The difference between both models is the type of activation functions used for the feed-forward layers: sigmoid (CNN-sigm) and rectified linear (CNN-relu).

DNN Speaker Adaptation

The output from the second recognition step was used to carry out speaker adaptation of DNNs (as indicated at the lower box of Fig. 2.4). The technique used consisted of a conservative training approach, using a very small learning rate and early stopping [13].

Moreover, we made use of confidence measures at word level to exploit inexpensive yet reliable unsupervised speech data. Specifically, confidence measures are estimated from the output of the second recognition pass in order to improve the DNN adaptation step. Although there are many different ways to estimate confidence measures, here we will resort to the conventional approach by which these measures are computed as word posterior probabilities [10].

In order to take advantage of confidence measures, we decided to use them to weight the samples during the adaptation. In this approach, all samples are taken into account, but

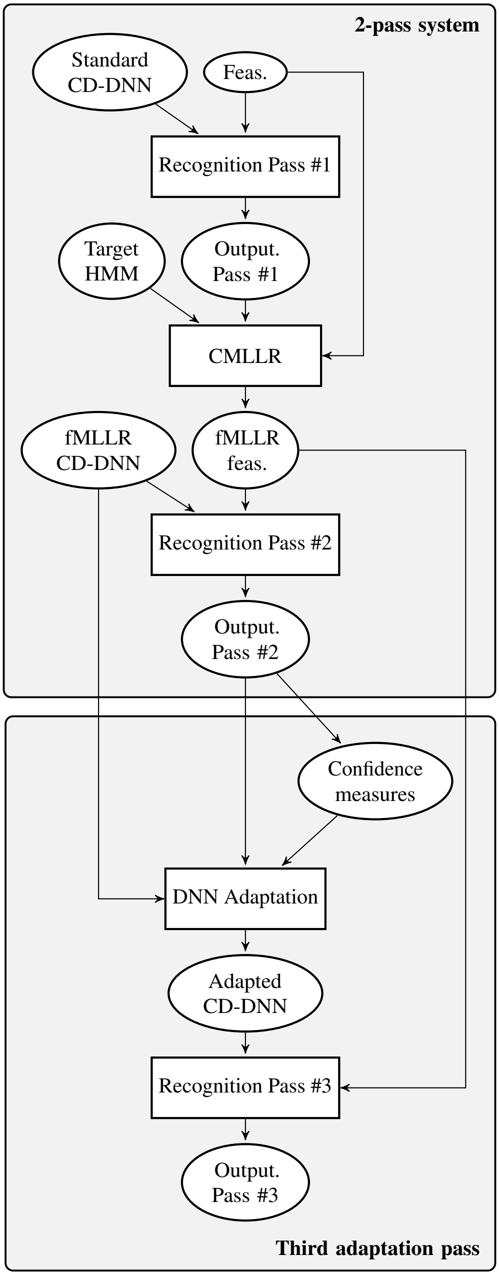

Figure 2.4: Overview of a multi-pass recognition system including DNN adaptation. Top: 2-pass recognition system using fMLLR features. Bottom: Third pass DNN adaptation.

the contribution of each sample is weighted by its corresponding confidence measure. The rationale behind this method is that only samples with high confidence measures are relevant for the adaptation process, whereas those with low confidence can be neglected. In some way, this method can be seen as a refinement of taking away those samples behind an specified threshold, avoiding the need of estimating that threshold.

Formally, adaptation with weighted samples is based on a modified cross entropy training criterion:

$$\sum_{n=1}^{N} c_n \log p(s_n \mid \mathbf{x}_n)\,, \tag{2.2}$$

where $\mathbf{x}_1^N$ is the set of frames, s_n is the senone (label) according to the output from the second pass, and $c_n \in [0,1]$ is its confidence measure. This modified criterion leads to a different way to estimate errors in the Back-Propagation algorithm. In particular, the error for the nth frame δ^n is estimated as follows

$$\delta^n = (\mathbf{y}_n - \mathbf{s}_n) \cdot c_n\,, \tag{2.3}$$

where $\mathbf{y}^n$ is the output of the last layer, and $\mathbf{s}^n$ are the target labels.

Language Modeling

We used several different text corpora to train the language models. They were preprocessed to normalise capitalisation, remove punctuation marks and transliterate numbers. We can distinguish two different types of corpora, out of domain corpora (OOD), most of them, and in domain corpora (ID), in this case only TED train set. Table 2.2 summarises the main figures of all the corpora used.

The vocabulary for the language models have been obtained by selecting the 200K most frequent words of a 1-gram LM interpolation of the OOD corpora. The words form the ID corpus are added to this selection, obtaining a final vocabulary of 209 660 words.

With this vocabulary, we trained standard Kneser-Ney smoothed n-gram models for each one of the corpora using the SRILM toolkit [8]. The order of each model is adjusted to 3 or 4 depending on the size of the corpus. The last column of Table 2.2 shows the perplexity obtained with all these models on the English development set.

All the resulting models are linearly interpolated to obtain a final powerful model adapted to the characteristics of the task, optimising the interpolation weights on the development set [3]. To reduce the size of the final model, it is pruned by removing those n-grams ($n > 1$) whose removal causes (training set) perplexity of the model to increase by less than 2×10^{-10}. This model obtained a perplexity of 126.1.

Experimental Results

In this section all the recognition experiments performed for the English transcription system are described. Recognition experiments were carried out on the IWSLT 2015 English ASR development and evaluation sets, the statistics of which are shown in Table 2.3.

Following the IWSLT evaluation requirements, tst2013 was used as development set, tst2014 as progressive evaluation set and tst2015 as evaluation.

Table 2.2: Stats of the different LM training corpora. The poliMedia [**poliM**], VideoLectures.NET and VL.NET subtitles [**VideoLectures**] corpora were generated during transLectures project.

Corpus	Sentences	Words	Perplexity
Europarl	2.2M	53M	454.3
Europarl TV	128K	1.2M	454.5
Giga 10^9	22M	557M	296.9
Google Ngrams	-	303B	1871.1
NewsCrawl	53M	1.1B	151.7
poliMedia	4K	95K	1393.1
VideoLectures.NET	5K	127K	871.4
VL.NET subtitles	85K	1.7M	371.5
Wikipedia	82M	1.5B	200.1
TED train	520K	3.7M	218.2

Table 2.3: Statistics of the English ASR development and evaluation sets.

Set	# Talks	Time
tst2013	28	4h:39m
tst2014	15	2h:22m
tst2015	12	2h:25m

The decoding was performed for all the subsystems following the scheme from Fig. 2.4. The first step was common and its output was used to perform fMLLR speaker adaptation. After that, each subsystem performed the second recognition step, the output of which was used to perform DNN speaker adaptation using confidence measures. Results from these two steps are shown in Table 2.4.

Table 2.4: Effect of DNN Speaker Adaptation on each subsystem in terms of WER. Results are shown on tst2013 data-set.

Subsystem	Non-Adapt	Adapt	R. Improvement
DNN-mmi	16.9	16.7	1.2%
DNN-sigm	17.1	16.7	2.3%
DNN-relu	18.5	17.8	3.8%
CNN-sigm	19.4	18.8	3.1%
CNN-relu	18.7	18.0	3.7%

It is worth mention that none of the above results has been subjected to a process of spelling normalisation by means of a global mapping file. As we can observe, the DNN-mmi adaptation has not performed as the rest of system's adaptations. To our knowledge this is because there is not so much room for improvement as occurs in the other systems, and also to the change in the training criterion (from MMI to CE during adaptation).

Finally, a recogniser output voting error reduction (ROVER) algorithm was applied to combine the subsystem's output and further improve the recognition results. The combination weights were estimated based on the development set, giving 2:2:1:1:1 for DNN-mmi, DNN-sigm, DNN-relu, CNN-sigm and CNN-relu. The final scoring results are shown in Table 2.5. At the time of writing this paper results on the progress test set tst2014 were not provided.

Table 2.5: The final result of the English system in terms of WER. (* means official result)

Set	ROVER
tst2013	16.2
tst2015	13.3*

2.5 German Transcription System

In this section the German ASR system is described. The first section details the data and training procedure, while the second section shows the results obtained by the system.

Training

For the acoustic modelling, we decided not to use the Euronews ASR provided corpus due to processing power constraints and its acoustic conditions being far from target conditions. Instead, we downloaded and processed the German Speechdata Corpus (GSC) [4], an open source corpus recorded and released by the LT and the Teleccoperation group from the Technical University of Darmstadt. This corpus contains 180 different speakers and 36 hours of speech, recorded under controlled conditions with many microphones in parallel. The whole corpus was used as train data. The grapheme-to-phoneme conversion was performed with the help of MaryTTS software [6].

The training procedure for German was the same as the DNN-MFCC used in the English system (Sec. 2.4). 48-dimensional MFCC acoustic vectors were extracted and normalised by speaker. A single acoustic model was estimated for German, which consists of a feed-forward DNN with a window size of 11 and 4 hidden sigmoid layers with 2048 neurons each. The output layer features 12237 senones. The network initialisation was performed with the DPT approach, and then the network was trained using the Cross-Entropy error criterion for 10 epochs.

The training and recognition follow the same three-step approach of the English system. An speaker-independent model is used in the first step. The output transcription is then used to perform unsupervised fMLLR adaptation. This second transcription is employed to perform DNN Speaker adaptation (Sec. 2.4). In the case of German, no confidence measures have been used for this third step.

The language model for our German system is made up by a standard linear interpolation of 4-gram language models. These models were estimated from different open corpus downloaded from the Internet. The corpora were normalised by lower-casing, removing punctuation marks and transliterating numbers. The corpus statistics after this process can be found in Table 2.6.

Table 2.6: Statistics of the German LM corpus.

Corpus	Sentences	Words	Perplexity
Europarl	2M	46M	515.5
News-crawl	135M	2B	352.0
Wikipedia	31M	326M	423.4

When training, the vocabulary was restricted to 200k words, selected with the same procedure described in Section 2.4. The interpolation weights were set to optimise the perplexity of the dev set. In order to improve recognition time, the interpolated model was pruned with a prune factor of 2×10^{-9}. The perplexity of the language model is 290.4.

Experimental Results

We tested our system on the tst2013 corpus, which was set as the official development corpus of the 2015 challenge. This corpus contains 9 videos from the TEDx website, with varying acoustic conditions. The results are summarised in Table 2.7. At the time of writing this work results on tst2014 set were not provided.

Table 2.7: The final results of the German system in terms of WER. (* means official result)

Set	WER
tst2013	43.6
tst2015	43.3*

Unlike the English task, we were not able to obtain state-of-the-art results for the German task. We attribute this result to the lack of relevant in-domain acoustic resources and the simplicity of the approaches employed.

2.6 Conclusions

In this paper we have described the English and German ASR systems developed for the IWSLT 2015 evaluation campaign. For the first participation of the MLLP group, the presented systems make use of the hybrid approach of HMM-DNN. Particularly, the decoding step of the English system is based on the combination of five different transcription subsystems. Each one built as a three pass recognition system and combining different types of NNs architectures, input features and activation functions. Meanwhile, the German system constitutes our first large scale speech recognition system on this language and it is based on a three pass recognition system with DNN speaker adaptation.

2.7 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 (transLectures) and ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no 621030 (EMMA), the Spanish MINECO Active2Trans (TIN2012-31723) research project, the Spanish Government with the FPU scholarship FPU13/06241 and the Generalitat Valenciana with the VALi+d scolarship ACIF/2015/082.

References

[1] George Dahl, Dong Yu, Li Deng, and Alex Acero. “Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (receiving 2013 IEEE SPS Best Paper Award)* 20.1 (2012), pp. 30–42 (cit. on p. 36).

[2] Miguel Ángel Del-Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchis, and Alfons Juan. “The transLectures-UPV toolkit”. In: *Proc. of VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*. Las Palmas de Gran Canaria (Spain), Jan. 1, 2014 (cit. on p. 36).

[3] Frederik Jelinek and Robert L. Mercer. “Interpolated estimation of Markov source parameters from sparse data”. In: *In Proceedings of the Workshop on Pattern Recognition in Practice*. Amsterdam, The Netherlands: North-Holland, May 1980, pp. 381–397 (cit. on p. 40).

[4] Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvêa, Stefan Radomski, Max Mühlhäuser, and Chris Biemann. “Open Source German Distant Speech Recognition: Corpus and Acoustic Model”. In: *Text, Speech, and Dialogue*. Springer. 2015, pp. 480–488 (cit. on p. 42).

[5] Anthony Rousseau, Paul Deléglise, and Yannick Estève. “TED-LIUM: an Automatic Speech Recognition dedicated corpus”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012. ISBN: 978-2-9517408-7-7 (cit. on p. 37).

[6] Marc Schröder and Jürgen Trouvain. "The German text-to-speech synthesis system MARY: A tool for research, development and teaching". In: *International Journal of Speech Technology* 6.4 (2003), pp. 365–377 (cit. on p. 42).

[7] Frank Seide, Gang Li, Xie Chen, and Dong Yu. "Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription." In: *Proc. of ASRU*. 2011, pp. 24–29. ISBN: 978-1-4673-0365-1 (cit. on p. 38).

[8] A. Stolcke. "SRILM – an extensible language modeling toolkit". In: *Proc. of ICSLP'02*. Denver, Colorado, USA, Sept. 2002, pp. 901–904 (cit. on pp. 37, 40).

[9] *TED: Ideas worth spreading*. `https://www.ted.com` (cit. on p. 37).

[10] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney. "Confidence measures for large vocabulary continuous speech recognition". In: *Speech and Audio Processing, IEEE Transactions on* 9.3 (2001), pp. 288–298 (cit. on p. 38).

[11] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. "Adaptation Of Context-Dependent Deep Neural Networks For Automatic Speech Recognition". In: *SLT 2012*. Dec. 2012 (cit. on p. 36).

[12] S. J. Young, J. J. Odell, and P. C. Woodland. "Tree-based state tying for high accuracy acoustic modelling". In: *Proc. of HLT*. 1994, pp. 307–312 (cit. on p. 38).

[13] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and F. Seide. "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition". In: *ICASSP*. IEEE. 2013, pp. 7893–7897 (cit. on p. 38).

3 The MLLP system for the 4th CHiME challenge

M. A. Del-Agua and A. Martínez-Villaronga and A. Giménez and A. Sanchis and J. Civera and A. Juan

Proc. of the 4th CHiME Speech Separation and Recognition Challenge (CHiME-4). Pages 57-59

San Francisco (USA)

13 September 2016

The MLLP system for the 4th CHiME challenge

M. A. Del-Agua and A. Martínez-Villaronga and A. Giménez and A. Sanchis and J. Civera and A. Juan

Abstract

The MLLP CHiME-4 system is presented in this paper. It has been built using the transLectures-UPV toolkit (TLK) developed by the MLLP research group which makes use of state-of-the-art automatic speech recognition techniques. Our best system built for the CHiME-4 challenge consists on the combination of two different sub-systems in order to deal with the variety of acoustic conditions. Each sub-system in turn, follows a hybrid approach with different acoustic models, such as Deep Neural Networks or BLSTM Networks.

3.1 Introduction

The CHiME Speech Separation and Recognition Challenge [5] encourage participants to develop innovative ASR approaches capable of dealing with challenging noisy environments that rely in speech processing, signal separation or machine learning. It is based on the Wall Street Journal corpus sentences, spoken by talkers located in real noisy environments, such as in a street junction, on the bus, or in a pedestrian area. All the audios have been recorded using a common 6-channel tablet microphone array.

In previous years, the challenge consisted of obtaining the best possible transcription from the 6 channels simultaneously, but given the successful results achieved, this year the challenge proposes two more tracks: 1-channel and 2-channels tracks. Each track only differs in the number of available channels for testing. Thus, the 6-channels track is the easiest since more favorable audio enhancement techniques can be applied. In the case of the 1-channel and 2-channels tracks, the audio enhancement techniques cannot exploit channel information at all which makes this tasks harder to deal with.

The MLLP CHiME-4 system has been developed focusing on the acoustic modeling aspect. Specifically, two different acoustic models have been trained following the hybrid approach. On the one hand, a Context-Dependent Deep Neural Network Hidden Markov Model (CD-DNN-HMM) and on the other hand, a Bidirectional Long Short Term Memory Neural Network (BLSTM). Both acoustic models will be trained on the same data and their output combined. From the proposed three tracks, this global back-end system have been tested in the 1-channel and 2-channel tracks.

The rest of this work is divided as follows. Section 3.2 describes the ASR toolkit used for the experiments. In Section 3.3 the proposed system is described and the conclusions are given in section 5.5.

3.2 The TransLectures-UPV Toolkit

The MLLP CHiME-4 system has been developed using the transLectures-UPV Toolkit (TLK) [2]. TLK comprises a set of tools for audio processing, feature extraction, HMM and DNN training and decoding. The main latest features added to the toolkit are the following:

- Multilingual and Convolutional NNs.
- Different DNN speaker adaptation techniques: output-feature discriminant linear regression (oDLR) [7] or Kullback-Leibler Divergence based [8].
- DNN sequence discriminative training based on Maximum Mutual Information (MMI).
- Online decoding.
- Gammatone feature extraction.

TLK has demonstrated to provide competitive results in challenging and well-known tasks. In [3] the TLK-based system dealt with TED video lectures, and in [4] the TLK system provided good results in the LibriSpeech [6] corpus.

3.3 Proposed System

The system proposed by the MLLP group is based on the TLK toolkit. It is composed of two transcription sub-systems that are combined following a recognizer output voting error reduction (ROVER). Each of those sub-systems are based on the HMM-NN hybrid approach. The only difference is that for the first sub-system a classical DNN is used whereas for the second sub-system a BLSTM NN is employed.

Each of those sub-systems perform a three step recognition process as can be observed in Fig. 2.5. The first and second steps are shown in the upper box. Regarding the first step, it is shared between both sub-systems, cepstral mean and variance normalization (CMVN) is applied and the decoding is performed using a standard DNN which provides the best possible transcription and a better feature-space Maximum Likelihood Linear Regression (fMLLR) transform. For the second step, each sub-system makes use of their own acoustic model (DNN or BLSTM) taking as input the transformed fMLLR features. The output of this system is used to perform a final third-pass recognition (the lower box of Fig. 2.5). During this step, an unsupervised speaker adaptation technique is applied to both, the DNN and the BLSTM. Specifically, the technique used in this work consisted of a conservative training approach using a very small learning rate and early stopping [8]. This means that a very small learning rate is estimated for a fixed number of epochs as to minimize the Word Error Rate (WER) and then this learning rate is used in evaluation. To the best of our knowledge, it is the first time that this kind of technique is applied to BLSTM NNs for acoustic modeling.

TLK allows to perform decoding efficiently with large vocabulary language models applying pruning techniques: beam search, histogram pruning, word end pruning and look-ahead. Thus, the provided 5-gram language model has been used to obtain the recognition outputs along all the steps. Once the last step is performed, the output lattices are re-scored using also the provided RNN-based language model.

BLSTM NNs have been built using TensorFlow [1]. With this purpose, a new feature has been added to TLK for decoding using TensorFlow-based graphs.

3.4 Experimental evaluation

The data used for training the acoustic models belong to the multi-condition training set defined by the CHiME-4 challenge. In our case, all data from channels 1,3,4,5 and 6 have been used to train the DNN and the BLSTM sub-systems.

Regarding feature extraction, classical Mel-frequency cepstral coefficients (MFCC) were extracted with a Hamming window of 25 ms. shifted at 10 ms. intervals. This MFCC features consisted of 16 MFCCs and their first and second derivatives (48-dimensional feature vectors). The resulting feature vectors were then normalized by mean and variance at speaker level. And after that, a single fMLLR transform for each training speaker was then estimated and applied to perform speaker-adaptive training (SAT).

In order to train the DNN and BLSTM based acoustic models, we first trained a basic context dependent triphone HMM model up to 64 component Gaussian mixtures, after which a second-pass fMLLR was applied. This model yielded a total of 9079 tied states, estimated following a phonetic decision tree approach.Both models were built on top of these HMM acoustic model. On one hand, the DNN-based acoustic model took as input the fMLLR features with a window size of 11, 5 hidden layers, sigmoid activation functions and an output layer of 9079. It was applied a discriminative pre-training stage and after that, the network was trained as to obtain the best frame accuracy on a validation set. On the other hand, the BLSTM acoustic model was trained with fMLLR input features (without windowing) with 4 hidden layers of 500 units each (both forward and backward directions) and an output layer of 9079. In this case, dropout was applied at the output of each cell with a probability of 0.1, and the Newbob strategy was also applied in order to reduce the learning rate by 0.8 each time the frame accuracy improved less than 3% relative on the validation set. Both networks were trained by minimizing the cross-entropy loss function, following the classical stochastic gradient descent algorithm. This two acoustic models were used for the 1-channel and 2-channels tracks. It is worth mentioning, that in the case of the 2-channel track, the audio enhancement beamformit was applied.

Table 2.8: WER (%) per step for the 1-channel track.

System	Rec. Pass	Dev		Test	
		real	simu	real	simu
DNN	1	16.03	17.63	24.87	24.47
	2	12.66	14.52	19.80	19.92
	3	11.93	13.19	18.34	17.73
	+RNNLM	10.45	11.98	17.20	16.56
BLSTM	1	16.03	17.63	24.87	24.47
	2	15.10	17.18	23.09	23.56
	3	13.40	14.46	19.30	18.47
	+RNNLM	11.96	12.79	17.78	17.03

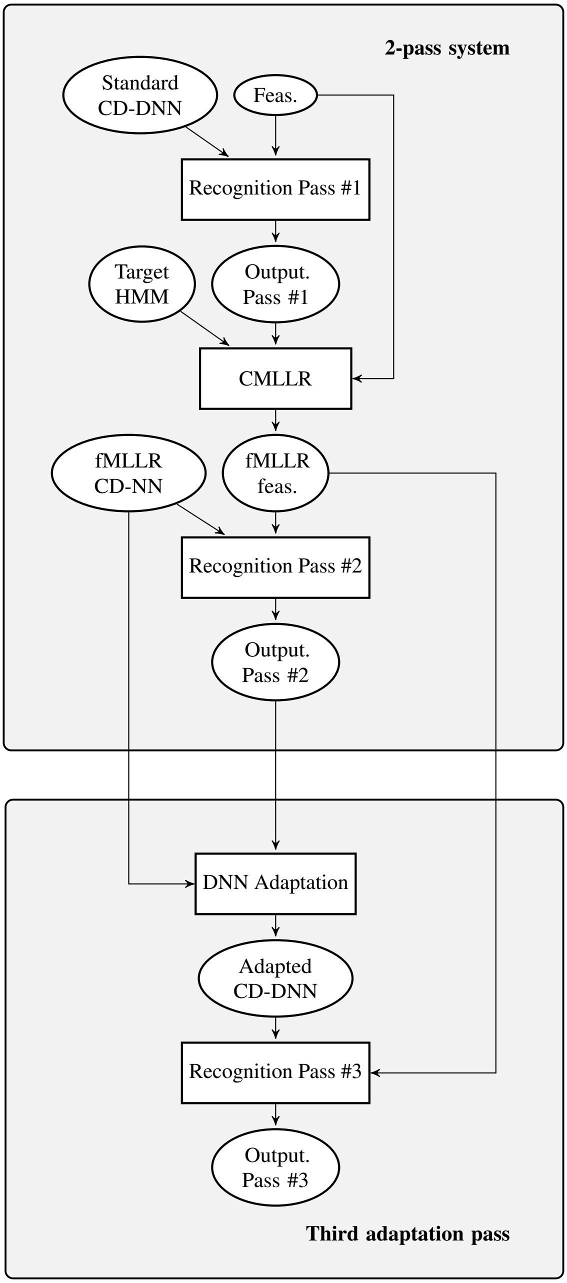

Figure 2.5: Multi-Pass recognition system with DNN adaptation. Top: 2-pass decoding using fMLLR features. Bottom: Third pass DNN adaptation.

Table 2.9: WER (%) per step for the 2-channels track.

System	Rec. Pass	Dev		Test	
		real	simu	real	simu
DNN	1	13.83	14.35	21.14	20.80
	2	10.39	11.49	16.26	15.75
	3	9.60	10.46	14.77	13.71
	+RNNLM	8.45	9.29	13.71	12.57
BLSTM	1	13.83	14.35	21.14	20.80
	2	12.81	14.22	19.09	19.64
	3	11.63	12.67	15.50	14.93
	+RNNLM	10.12	11.36	14.31	13.46

In Table 2.8 the results after each recognition step from the 1 channel track are shown, and similarly in Table 2.9 the results from the 2-channels track. As can be observed, the first recognition step is common to both sub-systems and tracks. With respect to the rest of recognition passes, very similar behaviors are observed in both tracks; the DNN performs better in all recognition steps and the BLSTM obtains a huge gain after the third step. For the first statement, we argue that the DNN is far more complex in terms of number of parameters, as we have trained a 5 hidden layer neural network of 2048 units per layer, while the BLSTM consist of 4 hidden layers of 500 units each one. Regarding the second statement, the huge WER improvement from the BLSTM at the third step comes from the fact that we are using the best transcription obtained during the previous step, i. e. the DNN, as to better perform speaker adaptation to the NN during the third step.

Once the output from both systems has been obtained, ROVER technique is applied as to combine both transcriptions. As can be seen in Table 2.10, the DNN system systematically outperforms the BLSTM-based. However, the combination of both systems yields the best result in both tracks. If we take a look to the real test set, the baseline provided by the organizers for the 1-channel track yielded 23.70% WER points whereas our system obtains 16.11%. This represents 32% relative reduction in WER for the 1-channel track. In the case of the 2-channels track, the baseline system achieved 16.58% average WER whereas our system achieves 12.82%. This represents a 22.7% relative reduction in WER for the 2-channel track. These improvements seems quite competitive, taking into account the simplicity of our system.

Table 2.11 summarizes the results obtained by the best system per environment. As shown, the most challenging has been the bus environment in all tracks for the real test set. In fact, the baseline system achieved 35.8%, while our system 21.61, which means almost 40% of relative improvement in the 1-channel track. In the case of the 2-channels track, the improvement is about 37% (from 25.37 to 16.00).

3.5 Conclusions

In this work we have described the MLLP ASR system developed for the CHiME-4 challenge built using TLK. The system is based on the combination of two sub-systems which make use of different acoustic models: DNNs and BLSTMs. The final system obtains 32% and 22.7%

Table 2.10: Average WER (%) for the tested systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	DNN	10.45	11.98	17.20	16.56
	BLSTM	11.96	12.79	17.78	17.03
	Combined	9.95	11.13	16.11	15.72
2ch	DNN	8.45	9.29	13.71	12.57
	BLSTM	10.12	11.36	14.31	13.46
	Combined	7.96	8.93	12.82	12.06

Table 2.11: WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	11.74	9.04	21.61	10.95
	CAF	11.18	14.68	18.12	19.57
	PED	7.42	9.35	13.25	15.37
	STR	9.45	11.46	11.47	16.98
2ch	BUS	8.84	7.73	16.00	8.67
	CAF	8.70	11.55	13.78	14.34
	PED	6.27	7.45	11.17	11.77
	STR	8.02	9.00	10.31	13.47

relative improvements over the 1-channel and 2-channels tracks compared to the baseline. This represents a good enough result taking into account the simplicity of our approach.

3.6 Acknowledgments

The work leading to this invention has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement no 287755 (transLectures). Also, it has received funding from the EU's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme under grant agreement no 621030 (EMMA). In addition, this work has been supported by the Spanish research project MORE TIN2015-68326-R (MINECO/FEDER) and the Spanish Government with the FPU scholarship FPU13/06241.

References

[1] Martín Abadi and et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `http://tensorflow.org/` (cit. on p. 51).

[2] Miguel Ángel Del-Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchis, and Alfons Juan. "The transLectures-UPV toolkit". In: *Proc. of VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*. Las Palmas de Gran Canaria (Spain), Jan. 1, 2014 (cit. on p. 50).

[3] Miguel Ángel Del-Agua, Adrià Martínez-Villaronga, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, J. Civera, and A. Juan. "The MLLP ASR Systems for IWSLT 2015". In: *Proc. of 12th IWSLT*. Da Nang (Vietnam), Dec. 3, 2015 (cit. on p. 50).

[4] Miguel Ángel Del-Agua, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. "ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks". In: *Interspeech*. 2016, pp. 3464–3468 (cit. on p. 50).

[5] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. "An analysis of environment, microphone and data simulation mismatches in robust speech recognition". In: *Computer Speech and Language, to appear* (2016) (cit. on p. 49).

[6] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: an ASR corpus based on public domain audio books". In: *ICASSP*. IEEE. 2015, pp. 5206–5210 (cit. on p. 50).

[7] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. "Adaptation of context-dependent deep neural networks for automatic speech recognition". In: *Proc. of the SLT*. 2012, pp. 366–369 (cit. on p. 50).

[8] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and F. Seide. "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition". In: *ICASSP*. IEEE. 2013, pp. 7893–7897 (cit. on p. 50).

4 ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks

M. A. Del-Agua and S. Piqueras and A. Giménez and A. Sanchis
and J. Civera and A. Juan

Proc. of InterSpeech. Pages 3464-3468

San Francisco (USA)

DOI 10.21437/Interspeech.2016-1142

Core A - GGS A

8-12 September 2016

ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks

M. A. Del-Agua and S. Piqueras and A. Giménez and A. Sanchis and J. Civera and A. Juan

Abstract

Confidence estimation for automatic speech recognition has been very recently improved by using Recurrent Neural Networks (RNNs), and also by speaker adaptation (on the basis of Conditional Random Fields). In this work, we explore how to obtain further improvements by combining RNNs and speaker adaptation. In particular, we explore different speaker-dependent and -independent data representations for Bidirectional Long Short Term Memory RNNs of various topologies. Empirical tests are reported on the LibriSpeech dataset showing that the best results are achieved by the proposed combination of RNNs and speaker adaptation.

Index Terms: speech recognition, speaker adaptation, confidence measures, recurrent neural networks, blstm

4.1 Introduction

Confidence estimation (CE) has been broadly investigated in automatic speech recognition (ASR) with the aim of assessing the reliability of the ASR output [3]. Over the years, an approach that has demonstrated to be very effective is to consider CE as a classical two-category (correct or incorrect) pattern recognition problem. Following this approach, CE has been gradually improved by exploring novel input features and by designing more and more accurate classifiers [3, 12, 11, 6].

Recent improvements to CE include the use of Recurrent Neural Networks (RNNs) [6] and speaker adaptation [11]. On the one hand, the use of RNNs has yielded better performance due to its ability to model context [6]. On the other hand, experimental results have shown that speaker-adapted classifiers such as naïve Bayes, logistic regression and conditional random fields outperform their non-adapted counterparts [11]. It is worth noting, however, that RNNs and speaker-adaptation have been studied separately, and thus it is still unclear whether using them in conjunction would lead to further improvements in accuracy.

In this work, we explore possible ways to use RNNs and speaker-adaptation techniques in conjunction. In particular, we propose to use the long short-term memory (LSTM) version of RNNs [4]. In this way, the vanishing gradient problem will be conveniently addressed in the case of long-span relations [1], while both history and future contexts will be modelled at the same time through its bidirectional version (BLSTMs). Furthermore, we propose to apply speaker adaptation techniques to LSTM models through the use of speaker-dependent input features based on their specific vocabulary, as well as training speaker-dependent models.

The content of the paper is organized as follows. The proposed speaker-adapted LSTM architecture is presented in Section 5.3. Empirical results on the LibriSpeech dataset are reported in Section 5.4, showing that the best results are achieved by the proposed combination of RNNs and speaker adaptation. Finally, the conclusions of this work are summarized in Section 5.5.

4.2 Speaker-Adapted LSTM Networks for Confidence Estimation

Recent work on CE [6] suggests that using temporal context by means of RNNs outperform other approximations where the sequential dependence cannot be exploited. For that reason, we propose to use LSTM networks as a further step towards context dependency. Aside from circumventing the vanishing gradient problem, LSTM networks introduce a temporal dependence over the entire segment by means of its bidirectional version. In this work, we use LSTM networks with both unidirectional and bidirectional layers, and thus we will refer to them simply as LSTMs.

What makes LSTM [4] networks different from RNNs is the use of purpose built-in memory cells which perform element-wise multiplications to control the information flow in the network. This memory cells are able to store information for a long period of time because of a gating structure that determines when the input is relevant enough to remember, when it should continue to remember or forget, and when it should yield an output. Specifically, the LSTM cells replace the activation function of a classical RNN with the following set of equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2.4}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{2.5}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2.6}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{2.7}$$

$$h_t = o_t \tanh(c_t) \tag{2.8}$$

where σ is the logistic sigmoid function and i, f, c, o, h represent five different vectors at time t from each gate: input, forget, cell memory activation, output and hidden layer, respectively. As depicted in Fig. 2.6, the LSTM Network proposed in this work follows a classical LSTM architecture. To use it in CE, input vectors at word-level are composed of two parts: a compact representation of the word identity and a set of word-level features extracted from ASR word-lattices.

Word identities have been included in the input vectors as they have been shown to be very useful in CE [12, 11, 6, 5]. To this end, we have not used a conventional one-hot encoding since this would entail a number of parameters growing linearly with the vocabulary size. Instead, we have used a more compact global word vector representation based on a “GloVe” model [9]. This is an embedding model, which tries to maintain the semantic similarities between words in their vector representation. Two very similar words will result in two very similar vectors. It is trained on the non-zero entries of a global word-word co-occurrence

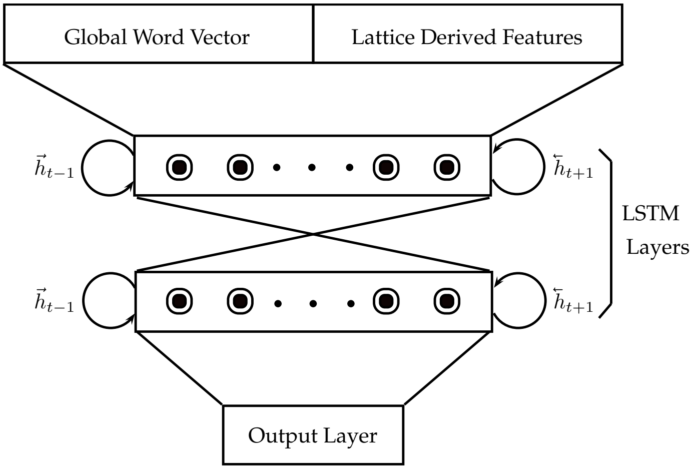

Figure 2.6: LSTM architecture for CE.

matrix which tabulates how frequently words co-occur with one another in a given corpus, in this case, the same training as the one used for CE.

Given a sequence of input vectors $\mathcal{X} = (\vec{x_1}, ..., \vec{x_T})$ which represents a sequence of T recognized words, the network produces a sequence of output vectors $\mathcal{Y} = (\vec{y_1}, ..., \vec{y_T})$ defining a probability distribution over each class c ($c = \{\text{correct}, \text{incorrect}\}$). These probabilities correspond to the network's estimate of observing each class c at time t given $\mathcal{X}$.

The LSTM network is trained to minimize the cross-entropy error of the targets using a softmax output layer with 2 output units representing the two-category class using the standard back-propagation through time algorithm (BPTT) [10]. Given a target sequence $\mathcal{Z} = (\vec{z_1}, ..., \vec{z_T})$, the network minimizes the negative log-probability of the target sequence given the input sequence:

$$-\log P(\mathcal{Z}|\mathcal{X}) = -\sum_{t=1}^{T} \log y_t^{z_t} \tag{2.9}$$

After an LSTM network has been estimated based on Eq. (2.14) using a set of N training pairs $\{\mathcal{X}, \mathcal{Z}\}_1^N$, we propose to adapt the LSTM to a new speaker by performing a few more iterations of the BPTT algorithm using a small subset of training pairs belonging to that speaker. It is worth mentioning that this adaptation also implies adapting the system to the vocabulary of the speaker, so it becomes necessary to re-estimate the global word vector model taking into account the new vocabulary of the speaker concerned. This is needed to ensure that the same word representation is used before and after adaptation.

4.3 Experiments

Experimental Setup

The proposed approach has been evaluated in the LibriSpeech ASR corpus [8]. The ASR system has been built using the transLectures-UPV toolkit [2], which is an open source set of tools for designing an ASR system from scratch. Acoustic models have been trained using the train-clean-100 LibriSpeech subset (100 hours). They consist of an hybrid HMM-DNN built on top of MFCC-CMLLR features. The DNN has been trained with a context window of 11 frames, 7 hidden layers with ReLu activation functions and 2048 units each. The number of target tied-states accounts for a total of 8132. As language model, we have used the pre-built 4-gram provided by the authors in the release of the corpus.

The official dev-other and test-other subsets of the LibriSpeech corpus have been used to adjust and evaluate CE models in a speaker-independent (SI) fashion. Also, 50h from the train-other-500 LibriSpeech subset were randomly selected for the training of the SI CE models. The main statistics of this experimental setting can be found in Table 2.12.

Table 2.12: Statistics of the speaker-independent setting.

Set	Duration (h)	Words	Vocab	WER
Train	49	475K	27K	15.6
Dev	5.3	51K	7K	21.2
Test	5.1	52K	8K	23.1

Additionally, 20 speakers not used in the SI experiments were randomly selected from the train-other-500 subset in order to evaluate speaker adaptation of the SI CE models. Specific training, development and test subsets were built for each speaker using their own speech data. Global statistics of this speaker-dependent (SD) setting are shown in Table 2.13.

Table 2.13: Statistics of the speaker-dependent setting.

Set	Duration (h)	Words	Vocab	WER
Train	5.9	54.9K	8.6K	26.2
Dev	2	19.1K	4.6K	26.3
Test	2	19.3K	4.6K	25.5

It is worth mentioning that all the speakers in LibriSpeech have almost the same amount of speech so as not to suffer from unbalanced speaker data. Therefore, in our SD partition, there is almost the same amount of data for each speaker in order to adapt, adjust parameters and evaluate.

Evaluation metrics

We have used three metrics to evaluate the performance of the CE classifiers: the area under a ROC curve (AUC), the classification error rate (CER) and the normalized cross entropy (NCE).

Let us assume that ASR output results in C correctly recognized words and I mis-recognized words. Let *False Rejection* be the number of correctly recognized words with confidence lower than a decision threshold τ ($FR(\tau)$) and, equivalently, let *True Rejection* be the number of mis-recognized words with confidence lower than τ ($TR(\tau)$). The *False Rejection Rate* (FRR(τ)) and the *True Rejection Rate* (TRR(τ)) for a decision threshold τ are computed as:

$$FRR(\tau) = \frac{FR(\tau)}{C} \qquad TRR(\tau) = \frac{TR(\tau)}{I} \tag{2.10}$$

A *Receiver Operating Characteristic* (ROC) curve represents TRR(τ) against FRR(τ) for different values of τ. The AUC provides an adequate overall estimation of the classification accuracy, being 100 a perfect classification and 50 a random classification (diagonal ROC curve).

The *Classification Error Rate* (CER) for a decision threshold τ is computed as:

$$CER(\tau) = \frac{FR(\tau) + (I - TR(\tau))}{C + I} \cdot 100 \tag{2.11}$$

A *baseline* CER can be computed by classifying all the words as correct (i.e. $\tau = 1$):

$$CER(1) = \frac{I}{C + I} \cdot 100 \tag{2.12}$$

Clearly, $\tau = 1$ is not necessarily optimal in the sense of minimizing Eq. (2.16). Therefore, it is convenient to consider the classification threshold $\tau = \tau^*$, which minimizes the CER criterion (usually that which provided the minimum CER in a *development set*):

$$\tau^* = \arg\min_{\tau} CER(\tau) \tag{2.13}$$

The *Normalized Cross Entropy* (NCE) is defined as the average log distance of the score to the real class. It attains its maximum of 1 when the system provides perfect confidence measures, that is, $0/1$ values allowing us to perfectly discriminate between correctly and incorrectly recognized words.

Results

As was mentioned in Section 5.3, a part of the input features of the LSTM Network are extracted from an ASR word-lattice. In the experiments, we used 5 word-lattice based features commonly used in CE [11]:

- SP: Word Acoustic log-score per time frame (10ms).

- D: Duration (in ms) of the word.
- NL: Length of the N-gram in which the word was decoded.
- PAvg: Word posterior probability computed as the average of frame-based posteriors [15].
- PMax: Like PAvg but using the maximum instead of the average [15].

On the other hand, a global word vector was obtained for SI and SD experiments, respectively, using the training data of each experimental setting. The optimal size of the word vectors was evaluated on the SI development set. Particularly, different vector sizes were explored, establishing the number of training epochs and window size during the global word vector model training. The best result was reached training during 30 epochs with a window size of 15 and a vector dimension of 30.

Regarding the network topology, different models were built using several types of layers and dimensions with the open source toolkit "currennt" [14]. All of them were tested on the development set and, finally, the best topology corresponded with a network with 2 hidden layers (BLSTM and LSTM) of 64 units each. This network architecture corresponds to that of Fig. 2.6.

Table 2.14 summarizes the results obtained using the SI experimental setting in terms of the different metrics presented in Section 5.4. The performance of the LSTM network is evaluated comparatively with respect to conditional random fields (CRF) and naïve Bayes (NB), which have shown to achieve very competitive results in CE [12, 13]. The experiments with CRF have been carried out using the Wapiti toolkit [7]. The best CRF models were obtained using the training algorithm *rprop-* and modelling dependencies between consecutive words.

Table 2.14: *Results on the speaker-independent test-set.*

	AUC	CER	NCE
Baseline	$--$	20.66	-
NB	84.4	16.54	-0.03
CRF	86.8	15.30	0.31
LSTM	88.3	14.58	0.35

As can be seen, LSTM models significantly achieve the best performance in terms of AUC, CER and NCE. LSTM networks stated a relative improvement of 4.7% in terms of CER with respect CRF. This statement is confirmed in Fig. 2.7, where the LSTM network outperforms consistently (for all decision thresholds τ) the rest of the classifiers. For instance, given an FRR of 20%, the LSTM classifier is the only one which can provide a TRR above 80%.

The evaluation of the speaker-adaptation technique proposed in Section 5.3 is shown in Table 2.15. This table summarizes the results obtained by different experiments using the SD experimental setting. First, the non-adapted LSTM network used in the SI experiments was evaluated in order to establish a baseline performance. Second, starting from this non-adapted LSTM network, we trained a speaker-adapted LSTM network per speaker applying a few

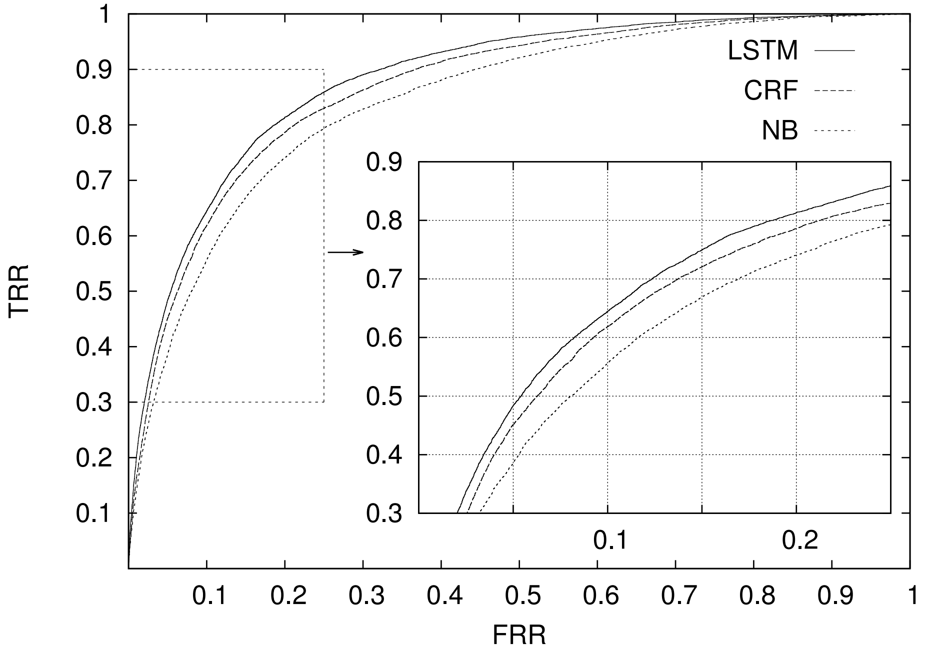

Figure 2.7: *ROC curves on the speaker-independent test-set.*

more training iterations using the BPTT algorithm with the data of each speaker. It is worth mentioning that the global word vector model was re-estimated so as to take into account the new speaker vocabulary along with the vocabulary from the SI experimental setting. Finally, a linear interpolation between both models (non-adapted and speaker-adapted) was evaluated. The optimal weights of interpolation were estimated using the development set.

Table 2.15: *Results on the speaker-dependent test-set.*

	AUC	CER	NCE
Baseline	−−	21.83	-
CRF	87.4	15.82	0.33
CRF+spkadapt	87.6	15.56	0.34
LSTM	89.3	14.48	0.38
LSTM+spkadapt	89.6	14.42	0.39
LSTM+spkadapt (interpolated)	90.0	13.81	0.41

As shown, the model interpolation results in the best model giving a relative improvement of 4.6% in CER with respect to the non-adapted model. This result is confirmed in Fig. 2.8, where the speaker-adapted model outperforms for any threshold τ their non-adapted

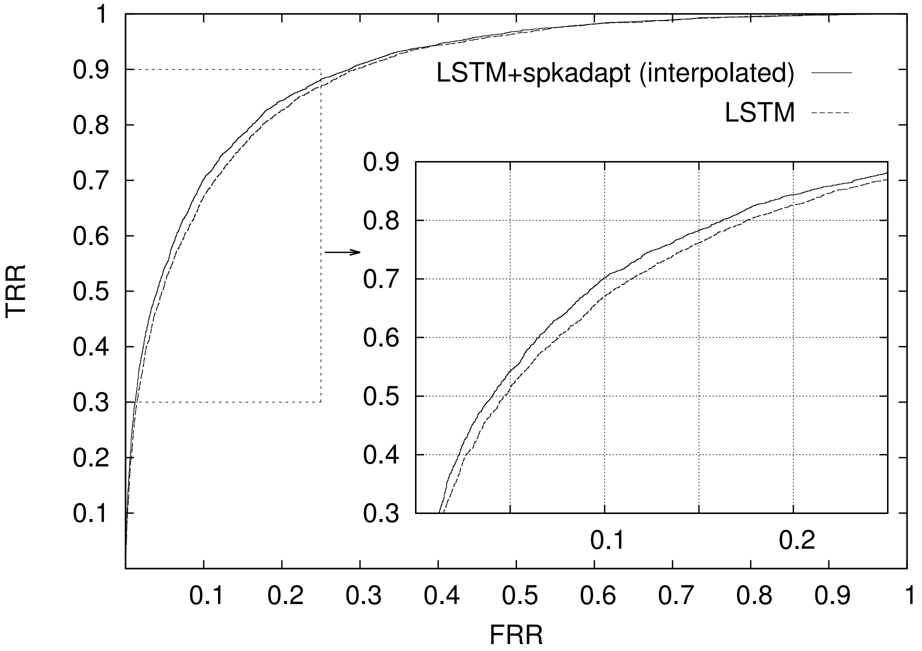

Figure 2.8: *ROC curves on the speaker-dependent test-set.*

counterpart. From our point of view, this final approximation has performed better because it has effectively prevented overfitting. This overfitting effect is usual when huge models such as LSTM networks are trained with scarce data, which is the case of adaptation to a single speaker.

For further analysis, Table 2.16 summarizes the performance of the interpolated model per speaker. In general, it can be stated that speaker-adapted models outperform their non-adapted counterparts in all cases in AUC, CER or both, except for speakers 4487 and 5248. For these two speakers, the adapted model achieves slightly worse CER. This could be produced by a particular vocabulary setting, quality of the adaptation data or a speaker-adapted system overfitting that could not be avoided with the interpolation.

4.4 Conclusions and Future Work

In this work, we have presented speaker-adapted confidence estimation using LSTM Networks. The use of LSTM Networks along with speaker-adaptation techniques constitutes a novelty in word confidence estimation. The results obtained over a publicly available dataset such as LibriSpeech confirm that LSTM networks improve state-of-the-art word confidence estimation models such as conditional random fields. Particularly, LSTM networks are able to produce relative reductions in CER of 4.7%. Moreover, the best speaker-adaptation technique presented is able to further reduce CER in 4.6%.

Table 2.16: *Results on the speaker-dependent test-set per speaker.*

	AUC			CER		
SPK	¬Adapt	Adapt	R. I.	¬Adapt	Adapt	R. I.
644	88.7	90.1	1.6	17.8	16.6	6.7
778	88.9	89.7	0.9	12.7	11.3	11.4
1065	88.0	88.3	0.3	14.0	13.1	6.3
1085	87.3	87.3	0.0	13.8	13.7	0.7
1544	89.9	89.8	0.0	11.9	11.2	5.5
3318	91.0	92.4	1.5	13.1	12.3	6.5
3793	92.0	92.7	0.8	12.8	11.9	7.0
3798	92.3	92.9	0.7	11.1	9.2	16.3
3992	90.8	90.8	0.1	11.3	10.9	4.1
4034	88.2	89.1	1.0	13.2	13.1	0.7
4487	87.9	88.6	0.8	13.8	14.3	-3.7
4546	86.7	87.5	0.9	12.3	11.7	4.9
5136	91.5	92.5	1.2	13.8	11.7	15.3
5248	86.9	87.2	0.3	16.1	16.2	-0.6
5993	89.4	90.1	0.7	10.4	10.4	0.0
6353	88.7	90.3	1.9	17.1	15.2	11.3
7389	91.7	92.6	1.0	12.5	11.7	6.5
7597	90.0	90.4	0.5	13.3	13.1	1.6
8042	84.8	85.3	0.6	20.2	19.7	2.5
8356	86.7	87.2	0.6	15.5	15.0	3.1

As future work, we plan to explore different word-embedding approaches. Also, we plan to study adaptation techniques for the (nearly) unsupervised case.

4.5 Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755 (transLectures) and ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no. 621030 (EMMA), the Spanish MINECO Active2Trans (TIN2012-31723) and EC FEDER Spanish MINECO MORE (TIN2015-68326-R) research projects.

References

[1] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *Neural Networks, IEEE Transactions on* 5.2 (1994), pp. 157–166 (cit. on p. 59).

[2] Miguel Ángel Del-Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchis, and Alfons Juan. “The transLectures-UPV toolkit”. In: *Proc. of VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*. Las Palmas de Gran Canaria (Spain), Jan. 1, 2014 (cit. on p. 62).

[3] H. Jiang. “Confidence Measures for Speech Recognition: A Survey”. In: *Speech Communication* 45.4 (2005), pp. 455–470 (cit. on p. 59).

[4] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on pp. 59, 60).

[5] Po-Sen Huang, Kush Kumar, Chaojun Liu, Yifan Gong, and Li Deng. “Predicting speech recognition confidence using deep learning with word identity and score features”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 7413–7417 (cit. on p. 60).

[6] Kaustubh Kalgaonkar, Chaojun Liu, Yifan Gong, and Kaisheng Yao. “Estimating confidence scores on ASR results using recurrent neural networks”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 4999–5003 (cit. on pp. 59, 60).

[7] Thomas Lavergne, Olivier Cappé, and François Yvon. “Practical Very Large Scale CRFs”. In: *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 504–513 (cit. on p. 64).

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. “Librispeech: an ASR corpus based on public domain audio books”. In: *ICASSP*. IEEE. 2015, pp. 5206–5210 (cit. on p. 62).

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global Vectors for Word Representation.” In: *EMNLP*. Vol. 14. 2014, pp. 1532–1543 (cit. on p. 60).

[10] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *Cognitive modeling* 5.3 (1988), p. 1 (cit. on p. 61).

[11] Isaias Sanchez-Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, and Alfons Juan. “Speaker-adapted confidence measures for speech recognition of video lectures”. In: *Computer Speech & Language* 37 (2016), pp. 11–23 (cit. on pp. 59, 60, 63).

[12] Alberto Sanchis, Alfons Juan, and Enrique Vidal. “A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 565–574 (cit. on pp. 59, 60, 64).

[13] Matthew Stephen Seigel. “Confidence Estimation for Automatic Speech Recognition Hypotheses”. PhD thesis. Department of Engineering, University of Cambridge, 2013 (cit. on p. 64).

[14] Felix Weninger, Johannes Bergmann, and Björn Schuller. “Introducing currennt: The munich open-source cuda recurrent neural network toolkit”. In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 547–551 (cit. on p. 64).

[15] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney. “Confidence measures for large vocabulary continuous speech recognition”. In: *Speech and Audio Processing, IEEE Transactions on* 9.3 (2001), pp. 288–298 (cit. on p. 64).

5 Speaker-Adapted Confidence Measures for ASR Using Deep Bidirectional Recurrent Neural Networks

M. A. Del-Agua and A. Giménez and A. Sanchis and J. Civera and A. Juan

IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol 26. Number 7. Pages 1198-1206

DOI 10.1109/TASLP.2018.2819900

IF 2.950 - Ranking 5/31 (Q1)

2018

Speaker-Adapted Confidence Measures for ASR Using Deep Bidirectional Recurrent Neural Networks

M. A. Del-Agua and A. Giménez and A. Sanchis and J. Civera and A. Juan

Abstract

In the last years, Deep Bidirectional Recurrent Neural Networks (DBRNN) and DBRNN with Long Short-Term Memory cells (DBLSTM) have outperformed the most accurate classifiers for confidence estimation in automatic speech recognition. At the same time, we have recently shown that speaker adaptation of confidence measures using DBLSTM yields significant improvements over non-adapted confidence measures. In accordance with these two recent contributions to the state of the art in confidence estimation, this paper presents a comprehensive study of speaker-adapted confidence measures using DBRNN and DBLSTM models. Firstly, we present new empirical evidences of the superiority of RNN-based confidence classifiers evaluated over a large speech corpus consisting of the English LibriSpeech and the Spanish poliMedia tasks. Secondly, we show new results on speaker-adapted confidence measures considering a multi-task framework in which RNN-based confidence classifiers trained with LibriSpeech are adapted to speakers of the TED-LIUM corpus. These experiments confirm that speaker-adapted confidence measures outperform their non-adapted counterparts. Lastly, we describe an unsupervised adaptation method of the acoustic DBLSTM model based on confidence measures which results in better automatic speech recognition performance.

5.1 Introduction

Confidence Estimation (CE) aims at providing *Confidence Measures* (CM) of the Automatic Speech Recognition (ASR) output at a certain level of granularity such as sub-word, word or utterance [14]. CM are represented by scores usually between 0 and 1 which reflect the reliability of any recognition output. Considering CM as probabilities of correctness, CE has been largely addressed as a two-class (correct or incorrect) pattern recognition problem [14, 35, 37, 34, 29]. To this effect, a binary classifier is trained to map input features to class posterior probabilities. Under this approach, CE has been gradually improved by exploring novel features and by designing more and more accurate classifiers [14, 35, 37, 34, 29].

Recent significant improvements to word-level CE have come from the use of Recurrent Neural Networks (RNN). Other classifiers that were considered until recently to be very effective, such as Conditional Random Fields (CRF), Logistic Regression (LR) or naïve Bayes (NB), have been clearly outperformed by RNN [29, 28, 7]. In particular, both deep bidirectional RNN (DBRNN) and DBRNN with long short-term memory units (DBLSTM) have shown their superiority when compared with non-NN-based classifiers or even with deep feedforward NN (DNN) [29, 7].

At the same time, adaptation of CM has shown to be very effective in improving baseline performance [34, 7, 47, 20]. This is a key point, from our point of view, especially for tasks with limited training data, since adaptation allows us to easily obtain accurate task-specific models from generic models trained on large, non-specific data sets. Moreover, there is an increasing number of interesting applications in which relevant information for adaptation is available, such as speaker identity in video lecture repositories. However, to the best of our knowledge, thus far there have been very few contributions in adaptation of CM. To address this, we were the first to implement speaker adaptation for CM. In [34], we evaluated speaker-dependent features into an LR model with good results. Then, in a follow-up work [7], we obtained even better preliminary results by using speaker-adapted DBLSTM.

In this paper, following our previous work [7], new technical contributions are reported, including a new architecture for CE in which word embeddings and CE models are jointly trained, and also a novel CE-based unsupervised adaptation method for acoustic BLSTMs. Furthermore, a multi-task empirical evaluation setting is applied to achieve solid empirical results which confirm our previous preliminary results for a single task.

The content of this paper is organized as follows: a brief review of recent work in CE is given in Section 5.2; the proposed speaker-adapted RNN architecture for CE is presented in Section 5.3; empirical results are reported in Section 5.4; finally, the main conclusions of this work are summarized in Section 5.5.

5.2 Recent work in confidence estimation

CE has been largely addressed following three main approaches [14]. One of them, known as *Utterance Verification* (UV), formulates CE as a statistical hypothesis testing problem [22]. The second one is based on word posterior probabilities computed over N-best lists, word lattices or confusion networks [43, 23]. The third approach considers CE as a two-class classification problem in which class posterior probabilities are estimated combining predictor features [35, 37, 34, 29]. The second approach is currently in wide use, since CM are computed in a straightforward manner from the ASR output. However, significantly better performances are generally reached using the third, classifier-based approach, mostly if word posteriors are used as input features [35, 37, 29].

In recent years, the classifier-based approach has directly benefited from the use of deep learning models outperforming the most accurate earlier classifiers such as CRF [29, 28, 7]. In a first proposal, DNN and kernel deep convex networks (K-DCN) were applied at the utterance level to discriminate between in-grammar and out-of-grammar utterances [17]. In later research, RNN have demonstrated outstanding performance in word-level CE [29, 28, 7, 19]. In particular, DBRNN and DBLSTM have confirmed their superiority over other classifiers such as CRF, DNN, DRNN and DLSTM [29, 7].

The performance of CM can be further improved by means of adaptation techniques [34, 7, 47, 20]. Significant performance gains have been reported by adapting generic CM using a small amount of transcribed adaptation data in a post-processing step called confidence calibration, based on different models such as maximum entropy, NN and deep belief networks [47]. Normalization of CM using adaptation data has also been proposed via confidence mapping to avoid decision threshold reselection when acoustic models are updated [20]. Very recently, we proposed speaker adaptation of LR models and DBLSTM for CE, showing that speaker-adapted models outperform their non-adapted counterparts [34, 7]. For instance, speaker-adapted DBLSTM produced relative reductions in *Classification Error Rate* (CER) of 4.6% when compared with non-adapted DBLSTM.

5.3 Speaker-Adapted Confidence Measures using Deep Bidirectional Recurrent Neural Networks

RNN have proven to be extremely successful in many related fields of speech processing, e.g, acoustic and language modelling, speech synthesis or spoken language understanding [42, 25, 8, 45]. RNN features recurrent connections which enable efficient modelling of temporal dependencies, outperforming other models without this capability. The most basic form of RNN was gradually improved to deal with some limitations such as the vanishing gradient problem and the use of context information in only one time direction [2, 36]. With regard to the former limitation, the LSTM architecture was proposed to overcome the vanishing gradient problem by which long-term dependencies make difficult the training of RNN [16]. Basically, LSTM differ from RNN in the use of hidden layers composed of built-in memory cells which are able to store information for long periods of time. As to the latter limitation, both past and future time directions were incorporated by extending RNN to BRNN [36]. In BRNN, hidden layers are composed of two separate forward and backward layers which are responsible for the positive and negative time directions, respectively. It is worth mentioning that BRNN with hidden layers composed of LSTM cells result in the BLSTM architecture [13]. In general, better performance can be expected from deep architectures stacking multiple BRNN or BLSTM hidden layers [49]. In this section we describe our CE model based on deep BRNN and BLSTM architectures, and the speaker adaptation process.

The architecture of the proposed CE model is depicted in Fig.2.9. For simplicity, we show an architecture based on two bidirectional recurrent hidden layers. Both the DBRNN and DBLSTM architectures are represented in this single figure, since the only difference between them is the type of recurrent cell used in the hidden layers. The input layer is composed of a set of R word-level predictor features along with a word embedding representation. Predictor features are typically computed from the speech decoding, word-lattices and from the ASR models (the features used in this work are described in Sec. 5.4).

Word embeddings are also fed into the first hidden layer, since word identities have shown to be very useful in improving CE [35, 34, 29, 17, 19, 9, 10]. To this end, we have not used a conventional one-hot encoding, as this would make the number of parameters grow linearly with the vocabulary size V. Instead, we have used a more compact representation where each word is mapped to a real word vector of a fixed dimension F [26]. In the case of NN, this word representation is learned by adding an extra layer to the NN which takes as input the one-hot representation and outputs a fixed-length vector. This means learning a projection matrix of size $V \times F$, in which the *i*th row corresponds to the embedding representation of the *i*th word in the vocabulary. In this way, words with similar behaviour can be expected to be represented by similar word embeddings. The vocabulary is typically restricted to the most frequent words. In this way, an embedding representation for unknown words is learned by labelling low-frequency words as unknown. This parameter matrix is trained jointly with the rest of the neural network parameters.

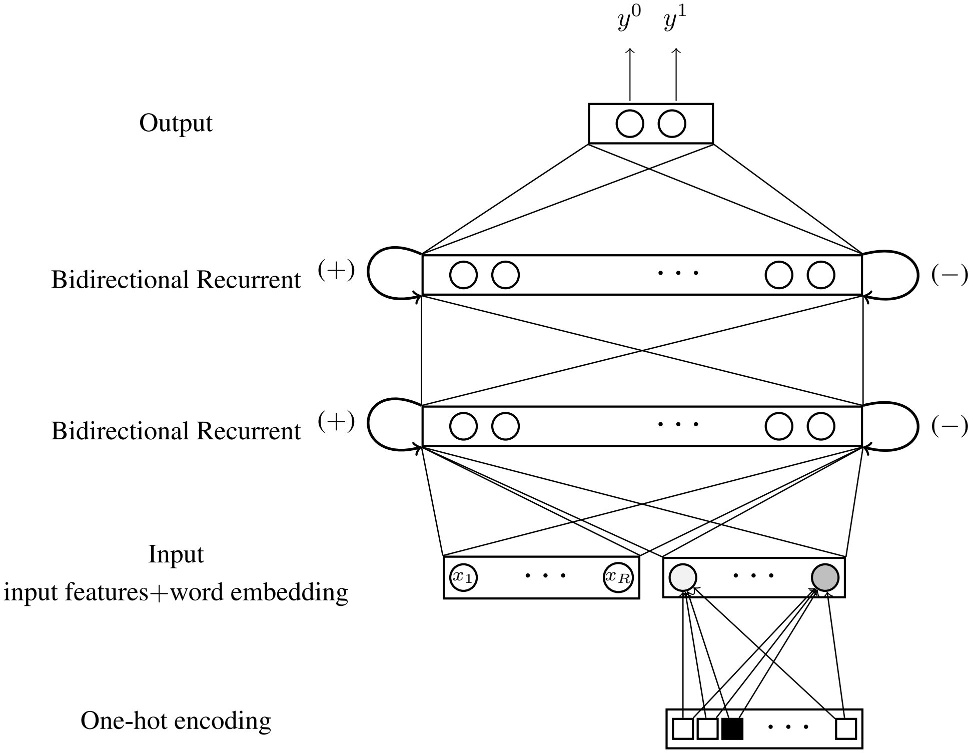

Figure 2.9: DBRNN and DBLSTM architectures of two hidden layers for CE. Positive (forward states) and negative (backward states) time directions are indicated by $(+)$ and $(-)$, respectively.

Given a sequence of N input vectors $\mathcal{X} = (\mathbf{x_1}, ..., \mathbf{x_N})$ representing N recognized words W_1^N, where each vector is composed of the R word-level predictor features along with the word embedding representation, the network produces a sequence of N output vectors $\mathcal{Y} = (\mathbf{y_1}, ..., \mathbf{y_N})$ defining a probability distribution over each class $c = \{\text{incorrect}(0), \text{correct}(1)\}$. These probabilities correspond to the network's estimation of observing each class c at word n given $\mathcal{X}$.

The network is trained to minimize the cross-entropy error of the targets using a softmax output layer with 2 output units that represent the two-category class based on the standard back-propagation through time algorithm (BPTT) [32]. Given a target sequence $\mathcal{Z} = (\mathbf{z_1}, ..., \mathbf{z_N})$, the network minimizes the negative log-probability of the target sequence given the input sequence:

$$-\log P(\mathcal{Z}|\mathcal{X}) = -\sum_{n=1}^{N} \log p(c = z_n|\mathbf{x_n}) = -\sum_{n=1}^{N} \log y_n^{z_n} \tag{2.14}$$

where $y_n^{z_n}$ is the probability estimated at word n by the output neuron that represents the target class z_n.

Once the network has been estimated based on Eq. (2.14), M new training pairs $\{\mathcal{X}, \mathcal{Z}\}_1^M$ from one speaker are used for adaptation. Adaptation is performed by following a conservative training strategy in which a very small learning rate and early stopping are used [11]. Note that this strategy has become a conventional method for regularization in deep learning because of its effectiveness and simplicity. To actually adapt models, the M given training pairs are split into an adaptation set and a validation set. Adaptation data is used to update the speaker-independent network (or part of it), whereas validation data is used to set the error of the resulting speaker-adapted network. The adaptation process finishes when the validation error stops changing significantly. Then, the final speaker-adapted network is trained from all training pairs by running an "optimal" number of epochs, as determined by the early stopping procedure.

5.4 Experiments

Experimental Setup

The experimental study was conducted over several speech tasks involving the English and Spanish languages. Accordingly, a state-of-the-art ASR system was trained for each language using the transLectures-UPV toolkit (TLK) [7, 4, 5, 6]. TLK is an open-source ASR toolkit developed at the Universitat Politècnica de València (UPV) by the MLLP research group within the framework of the EU-funded project transLectures[a]. It comprises a set of tools for audio processing, feature extraction, HMM and DNN training and decoding. Its main features include multilingual and convolutional NNs, DNN sequence discriminative training based on Maximum Mutual Information (MMI), and different DNN speaker adaptation techniques such as output-feature discriminant linear regression (oDLR) [44] or Kullback-Leibler Divergence based techniques [48]. TLK has shown to provide competitive results in challenging and well-known tasks such as TED-LIUM, LibriSpeech, IWSLT or CHiME [7, 4, 5, 6].

The English ASR system was trained using the LibriSpeech training dataset, which contains almost 1000 hours of read speech recordings from the LibriVox project's audio books [30] (statistics in Table 2.17). On the other hand, the Spanish ASR system was trained using the poliMedia speech corpus enlarged to about 800 hours for training [38]. PoliMedia is a high-quality multimedia educational repository developed by the UPV. It includes more than 15,000 Spanish video lectures lasting up to 10 minutes each, created by more that 1800 lecturers, summing up a total amount of about 3000 hours. This speech corpus was developed within the EU-funded project transLectures (statistics in Table 2.18).

The audio data was preprocessed with a Hamming window of 25 ms shifted at 10 ms intervals into 16 Mel-frequency cepstral coefficients (MFCC) plus deltas and accelerations, resulting into 48-dimensional feature vectors. Speaker-adapted features were then obtained by

[a] `https://www.translectures.eu/web/`

Table 2.17: Statistics of the LibriSpeech corpus.

Set	Duration (h)	Speakers	Words	Vocab	WER
Train	961	1210	9.4M	89K	4.7
Dev-other	5.3	33	51K	7.4K	12.5
Test-other	5.1	33	52K	7.6K	13.5

means of Cepstral Mean and Variance Normalization (CMVN) and applying a Constrained Maximum Likelihood Linear Regression transform following the simple target model approach (fMLLR) [40].

The acoustic models were based on hybrid models [3, 15, 12]. For hybrid training, forced alignments of the senone (tied-state) transcriptions to the acoustic features (MFCC and fMLLR) were obtained by training conventional context-dependent Gaussian mixture model hidden Markov models (CD-GMM-HMMs). CD-GMM-HMMs consist of three left-to-right tied-states estimated following a phonetic decision tree approach [46]. The resulting number of tied-states was 8.3K and 10K, reaching up to a total amount of 256K and 478K Gaussians for English and Spanish, respectively.

These baseline alignments were then used to train both speaker-independent and speaker-adapted CD-DNN-HMMs [15] for each language with a context window of 11 frames, 7 hidden layers with ReLU activation functions and 2048 units each. The trained speaker-adapted CD-DNN-HMMs were then used to further improve the state alignments. Using these DNN state realignments, we finally trained a speaker-adapted DBLSTM-HMM [12] for each language using the open source toolkit TensorFlow [1]. In both cases, the DBLSTM network had 5 bidirectional hidden layers with 1200 LSTM cells per layer, resulting in a total of 33.3M and 36.3M weights for English and Spanish, respectively. Relative improvements in WER of about 4.6% and 5.8% over the LibriSpeech and poliMedia test sets were achieved using DBLSTM-HMMs compared to CD-DNN-HMMs.

For the English language model (LM), we used the freely available pre-built 4-gram model released as part of the LibriSpeech corpus [30]. As for Spanish, we used the 4-gram LM built by UPV within the transLectures project [33, 41]. Both models had a vocabulary size of about 200K words, and the test set perplexities were 146 and 205, respectively.

Speech processing was carried out following a two-pass decoding setup. The speaker-independent CD-DNN-HMM ASR system was used primarily to obtain a transcription which in conjunction with a simple "target" HMM allowed for the transformation of acoustic features into speaker-adapted features. A word-lattice was then generated feeding the speaker-adapted features into the hybrid DBLSTM-HMM ASR system. Both recognition steps were carried out using a pruned version of the LMs to allow for very fast decoding. The final transcription was produced by rescoring the word-lattice with the whole LM.

Word-level predictor features

A number of $R = 20$ common word-level predictor features have been used in this work. These features have been computed from the speech decoding, word-lattices and from the

Table 2.18: Statistics of the poliMedia speech corpus.

Set	Duration (h)	Videos	Speakers	Words	Vocab	WER
Train	813	9.5K	205*	8.3M	36.6K	14.5
Dev	3.4	26	5	35K	2.6K	11.3
Test	3.2	23	5	30K	2.4K	12.5

(*) Lower estimate, since training set is not wholly speaker-annotated.

ASR models. We briefly enumerate them here:

(i) Features based on speech decoding and ASR models:

1. Decoding score: Word score produced jointly by the acoustic and language models during decoding.
2. Acoustic log-score: As in 1, but considering only the acoustic model.
3. Normalized acoustic log-score: As in 2, but normalized per time frame (10 ms).
4. Duration: Word length in ms.
5. Language model probability: N-gram language model probability for the decoded word.
6. Length of the N-gram in which the word was decoded.
7. Average number of alternative hypothesis within the decoding word boundaries.
8. Binary feature, equals 1 if the word appears in both the first and second decoding hypotheses.

(ii) Features based on word-lattices:

9-11. Forward, backward and edge posterior probabilities: The forward-backward algorithm is applied to the word-lattice to compute forward, backward and posterior probabilities for every edge in the lattice. As usual, edges in a word-lattice are associated with words occurring at specific intervals along the time axis; and probabilities are computed from acoustic and language model scores by using the (meta-)parameters set during the decoding phase. It is worth noting that edge posterior probabilities are probability sums of all paths including the given edge (normalized by the probability mass of all paths in the lattice).

12-14. Three variants of word posterior probabilities [43]: More precise word posterior probabilities can be computed by summing up the posterior probabilities of all edges containing the word in approximately the same interval time. Moreover, an appropriate scaling of acoustic model probabilities during the forward-backward algorithm is really needed to prevent (nearly) all posterior probability mass from concentrating in a few word-lattice hypotheses. In this case, given a word w which occurs at a specific point in time $t \in [s, e]$, its *accumulated* posterior probability at time t, $A(w, t)$, is computed by summing the posterior probabilities over all edges intersecting word w at time t. From this, three different variants of word posterior probabilities are computed:

12. Intersection: $P_{sec}(w, [s, e]) = \sum\limits_{t=s}^{e} A(w, t)$
13. Maximum: $P_{max}(w, [s, e]) = \max\limits_{t \in [s,e]} A(w, t)$
14. Average: $P_{avg}(w, [s, e]) = \frac{1}{e-s+1} \sum\limits_{t=s}^{e} A(w, t)$

15-17. As in 12-14, but using only acoustic scores during the forward-backward algorithm.

18-20. As in 12-14, but using only language model probabilities during the forward-backward algorithm.

Evaluation metrics

We have used three metrics to evaluate CE performance: (i) the area under a ROC curve (AUC), (ii) the classification error rate (CER), and (iii) the normalized cross entropy (NCE). We briefly explain them in this section.

Let us assume that the ASR output results in C correctly recognized words and I misrecognized words. Let *False Rejection* be the number of correctly recognized words with confidence lower than a decision threshold τ ($FR(\tau)$) and, equivalently, let *True Rejection* be the number of misrecognized words with confidence lower than τ ($TR(\tau)$). The *False Rejection Rate* (FRR(τ)) and the *True Rejection Rate* (TRR(τ)) for a decision threshold τ are computed as:

$$FRR(\tau) = \frac{FR(\tau)}{C} \qquad TRR(\tau) = \frac{TR(\tau)}{I} \tag{2.15}$$

A *Receiver Operating Characteristic* (ROC) curve represents TRR(τ) against FRR(τ) for different values of τ. The AUC provides an adequate overall estimation of the classification accuracy, 100 being a perfect classification and 50 a random classification (diagonal ROC curve).

The *Classification Error Rate* (CER) for a decision threshold τ is computed as:

$$CER(\tau) = \frac{FR(\tau) + (I - TR(\tau))}{C + I} \cdot 100 \tag{2.16}$$

A *baseline* CER can be computed by classifying all the words as correct (i.e., $\tau = 0$):

$$CER(0) = \frac{I}{C + I} \cdot 100 \tag{2.17}$$

Clearly, $\tau = 0$ is not necessarily optimal in the sense of minimizing Eq. (2.16). Therefore, it is convenient to consider the classification threshold $\tau = \tau^*$, which minimizes the CER criterion (usually that which provided the minimum CER in a *development set*):

$$\tau^* = \arg\min_{\tau} CER(\tau) \tag{2.18}$$

We have also used the *Normalized Cross Entropy* (NCE) as proposed by NIST [39]:

$$NCE = \frac{H_{max} + \sum\limits_{w \in correct} \log(cm(w)) + \sum\limits_{w \in incorrect} \log(1 - cm(w))}{H_{max}} \tag{2.19}$$

where $cm(w)$ is the CM of word w and $H_{max} = -(p \log p + (1-p)\log(1-p))$, p being the prior probability for a word to be correct. Note that the higher the NCE, the better the CM performance, with optimal classification being reached when NCE equals one. It is worth mentioning that NCE score is lower unbounded, as the logarithm of low values can occur in samples with high scores on their opposite class.

Experiments on CE

We performed experiments on CE using the LibriSpeech and poliMedia speech tasks. For each task, the training data were used to estimate DBRNN and DBLSTM models with TensorFlow following the architecture described in Sec. 5.3. The optimal numbers of hidden layers, neurons per hidden layer and word embedding size were tuned using the development set. The characteristics of the optimal topologies for each task are shown in Table 2.19.

Table 2.19: *Characteristics of the optimal DBRNN and DBLSTM topologies for the LibriSpeech and poliMedia speech tasks.*

	LibriSpeech		poliMedia	
	DBRNN	DBLSTM	DBRNN	DBLSTM
# hidden layers	3	4	2	2
# neurons per layer	512	512	64	512
Word embedding size	80	20	80	10

Table 2.20 summarizes the results obtained in terms of the different metrics presented in Section 5.4. CER(τ^*) figures in Table 2.20 correspond to the classification error attained in the test set using a threshold τ^* providing the minimum CER in validation. The performance of the RNN was comparatively evaluated with respect to word posterior probabilities (WP) and CRF [37, 43]. A linear interpolation of the RNN models was also tested aiming to further improve their individual performance (BRNN+BLSTM). The interpolation weights were tuned on the corresponding development set and fixed to 0.5 in the case of LibriSpeech and 0.3 (BRNN) and 0.7 (BLSTM) in the case of poliMedia. The experiments with CRF were carried out using the Wapiti toolkit [21]. The best CRF models were obtained using the training algorithm *rprop-* and modelling dependencies between consecutive words.

From the results in Table 2.20, it can be stated that RNN models clearly outperform CRF and WP, confirming previous results [29, 7]. Better performance is consistently achieved in all the evaluation measures using RNN models. The improvement in CER of the RNN models over CRF and WP is statistically significant at the 95% confidence level to a great extent, especially in the case of poliMedia. This better overall performance is depicted in Fig. 2.10,

Table 2.20: *AUC [%], NCE, CER [%] and* 95% *Confidence Interval (CI) of CER for the different CM on the LibriSpeech and poliMedia evaluation data. The baseline CERs (CER(0)) are* 11.99 *and* 10.90 *for LibriSpeech and poliMedia, respectively.*

Task	CM	AUC	NCE	CER(τ^*)	95%-CI CER
LibriSpeech	WP	85.3	-0.74	10.71	[10.44, 10.97]
	CRF	89.6	0.36	9.29	[9.04, 9.54]
	BRNN	91.1	0.40	8.82	[8.58, 9.07]
	BLSTM	91.0	0.38	8.85	[8.60, 9.09]
	BRNN+BLSTM	91.5	0.41	8.65	[8.41, 8.89]
poliMedia	WP	83.6	-0.57	9.67	[9.33, 10.00]
	CRF	90.0	0.40	7.69	[7.39, 7.99]
	BRNN	91.6	0.44	7.00	[6.71, 7.29]
	BLSTM	92.0	0.44	6.77	[6.48, 7.05]
	BRNN+BLSTM	92.1	0.45	6.75	[6.47, 7.04]

where the ROC curves of the RNN models clearly outperform the CRF and WP models for all decision thresholds τ.

On the other hand, the different RNN models present very similar behaviour, with no statistically significant differences visible between their performance. Even so, better figures are obtained in general using BRNN+BLSTM interpolation. This is depicted in Fig. 2.10, where small improvements can be observed when the linear interpolation of RNN models is compared to their individual performance.

Experiments on speaker-adapted CM

The evaluation of the speaker-adapted CM was conducted considering a practical scenario in which both ASR and confidence models may be used in multiple speech tasks. With this purpose, the ASR models and the BRNN+BLSTM confidence estimator trained with LibriSpeech were used to obtain the transcriptions and confidence scores of talks of eight speakers chosen from the TED-LIUM corpus [31]. The selection of speakers was made on the basis of having at least 4 talks per speaker, in order to perform a 4-fold cross-validation evaluation and also to cover a reasonable range of error between 10% and 30% of WER. The main characteristics of these talks are summarized per speaker in Table 2.21, where each speaker set is composed of exactly 4 talks.

As mentioned, speaker adaptation of CM was evaluated following the k-fold cross-validation method [27]. In this way, $k = 4$ experiments were performed per speaker, with the supervised transcriptions of 3 talks being used for adapting the LibriSpeech BRNN+BLSTM CE network, while the remaining talk was used for testing. With this strategy, each talk was used three times for adaptation and only once for testing. Moreover, the non-adapted

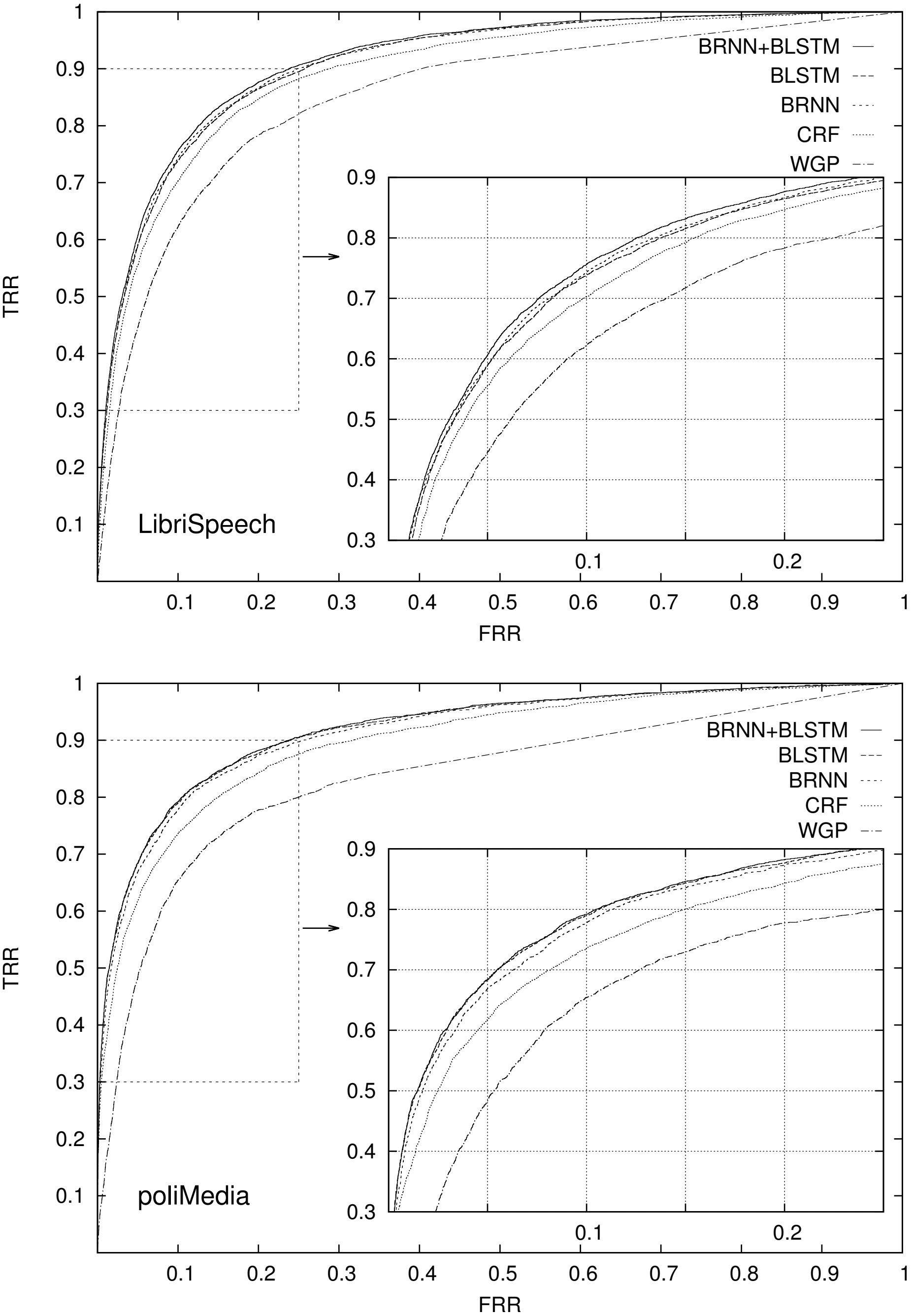

Figure 2.10: ROC curves of the different CM for the LibriSpeech (at the top) and poliMedia (at the bottom) evaluation data. TRR is the *True Rejection Rate* and FRR is the *False Rejection Rate*.

Table 2.21: *Global statistics of the* 4 *talks per speaker extracted from the TED-LIUM corpus.*

Speaker	Duration (h:mm:ss)	Running words (k)	WER [%]	CER(0) [%]
1	0:56:26	9.4	21.95	18.61
2	0:40:37	9.0	19.34	16.00
3	0:39:34	7.8	23.56	19.74
4	0:45:59	7.4	22.95	19.03
5	0:45:56	8.6	13.33	12.07
6	0:34:41	8.4	14.90	12.06
7	0:31:55	8.5	25.51	20.19
8	0:33:42	6.1	26.79	22.03
All	5:28:53	65.1	20.78	17.35

LibriSpeech BRNN+BLSTM CE network was used also to establish the baseline performance of CM without speaker adaptation.

Comparative results in terms of AUC and CER between non-adapted and adapted CM are shown in Table 2.22. It is worth noting that the non-adapted model corresponds to the BRNN+BLSTM CE network achieving the best performance on the LibriSpeech corpus in the CE experiments reported above. Also, this network was used to derive a speaker-adapted model as described in Section 5.3. CER figures were obtained using the same decision threshold τ for both non-adapted and adapted experiments. The operative τ^* for each speaker was tuned over the adaptation data.

In general, it can be stated that speaker-adapted CM outperform their non-adapted counterparts for all the speakers. Slightly better performance is achieved in terms of AUC, with the only exception of speaker number 8, for which no differences were found. The overall superiority of adapted CM is observed in Fig. 2.11, where ROC curves obtained considering all the speakers as a whole are plotted comparatively. Similarly, relative improvements in CER of about $2 - 8\%$ are produced by using adapted CM, except in the case of speaker number 7, for which CER figures were nearly identical. Overall, considering all the speakers as a whole, the improvement in CER is statistically significant at the 95% confidence level to a great extent, since the confidence intervals are $[11.94 - 12.47]$ and $[12.39 - 12.93]$ for adapted and non-adapted CM, respectively. It is worth noting that improvements notably depend on the speaker. A possible explanation of this phenomenon is that model improvement is very much dependent on the quality and amount of the speaker-dependent adaptation data.

In practice, it might not be realistic to assume that perfect transcriptions are available for at least three talks. Therefore, one could argue that results in Table 2.22 are optimistic. In order to study the CE performance in a more realistic setting, additional experiments were conducted in which the proposed adaptation approach was tested as a function of the amount of available adaptation data. We used the same 4-fold cross-validation procedure described above, though in this case it was repeated for an increasing percentage of perfect transcriptions available.

Figure 2.12 shows the CER for each speaker using increasing percentages of available adaptation data (0, 10, 25, 50, 75 and 100). Note that 0% and 100% would correspond with

Table 2.22: *AUC [%] and CER [%] for the adapted and non-adapted CM per speaker of the TED-LIUM corpus. The baseline CERs and the relative improvements (R.I.) in CER over the non-adapted CM are also shown.*

	AUC			CER(τ^*)		
Speaker	$\neg$Adapt	Adapt	CER(0)	$\neg$Adapt	Adapt	R.I. [%]
1	87.4	88.4	18.61	13.86	13.21	4.7
2	88.4	89.4	16.00	12.23	11.53	5.7
3	88.1	88.3	19.74	14.42	14.16	1.8
4	88.8	89.0	19.03	13.21	12.81	3.0
5	90.4	91.0	12.07	9.03	8.29	8.2
6	89.8	90.2	12.06	9.04	8.79	2.8
7	86.6	87.1	20.19	14.09	14.06	0.2
8	87.4	87.4	22.03	15.87	15.48	2.5
All	88.6	89.1	17.35	12.66	12.21	3.6

CER results of non-adapted and adapted models, respectively, showed in Table 2.22. As expected, Figure 2.12 confirms that the more adaptation data we use, the better CER we achieve. Although this holds visibly for speakers 1, 2, 4, 5 and 6, the results for speakers 3, 7 and 8 do not follow this pattern so clearly. It is worth mentioning that, for almost all the speakers, the CER improves already from the point where we use just 10% of the available adaptation data, and thus we can conclude that the proposed adaptation approach is really effective even when adaptation data is scarce.

Experiments on improving ASR performance

As mentioned before in Sec. 5.4, we followed a two-pass recognition strategy in which unsupervised speaker adaptation is implemented in a second step based on fMLLR transformed features. Further refinements of the second decoding hypothesis can be produced by means of an additional unsupervised adaptation step. In this extra step, the layers of the acoustic DBLSTM used in the second pass are retrained based on the senone alignments corresponding to the second decoding hypothesis. The retraining is carried out following a conservative training approach using a very small learning rate and early stopping [48]. In particular, given T acoustic vectors of fMLLR features $\mathcal{X} = (\mathbf{x_1}, ..., \mathbf{x_T})$ and the senone-level alignments from the second pass hypothesis $\mathcal{S} = (s_1, ..., s_T)$, the parameters of the acoustic DBLSTM are retrained to maximize the negative cross entropy

$$\mathcal{C}(\mathcal{X}, \mathcal{S}) = -\frac{1}{T}\sum_{t=1}^{T} \log p(s_t \mid \mathbf{x}_t) = -\frac{1}{T}\sum_{t=1}^{T} \log y_t^{s_t} \tag{2.20}$$

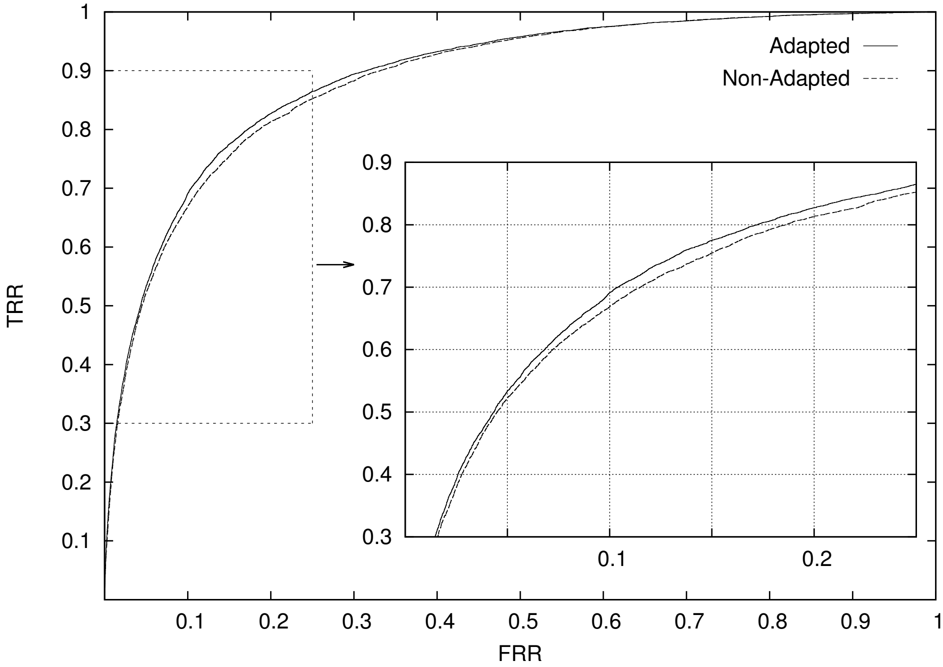

Figure 2.11: *ROC curves for the adapted and non-adapted LibriSpeech BRNN+BLSTM CE networks using the* 8 *speakers of the TED-LIUM corpus as a whole.*

where $y_t^{s_t}$ is an estimation of the probability at frame t given by the output neuron associated with the target class s_t.

Unsupervised adaptation of the acoustic DBLSTM in the additional pass can benefit from CE by adjusting the influence of the training samples as a function of CM. Formally, we propose to apply a modified cross entropy training criterion for this kind of adaptation. Following this idea, Eq. (2.20) becomes

$$\mathcal{C}(\mathcal{X}, \mathcal{S}) = -\frac{1}{T}\sum_{t=1}^{T} \log p(s_t \mid \mathbf{x}_t) \cdot cm(s_t) = -\frac{1}{T}\sum_{t=1}^{T} \log y_t^{s_t} \cdot cm(s_t) \tag{2.21}$$

where $cm(s_t)$ is the word-level CM of senone s_t.

Once the adapted acoustic DBLSTM has been retrained, a third-pass decoding is performed to produce the final hypothesis.

Table 2.23 shows the WER obtained for three different recognition settings on the LibriSpeech and poliMedia test sets and the 8 speakers of the TED-LIUM corpus. The “2-pass” setting corresponds to the baseline performance without performing the third adaptation pass. The “3-pass” setting implies performing the third pass based on Eq. (2.20). Finally, the “3-pass+CM” setting corresponds to applying Eq. (2.21) in the third pass.

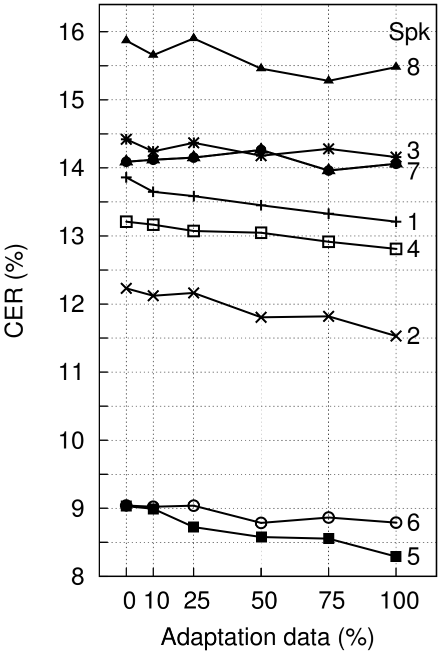

Figure 2.12: CER for different percentages of the whole adaptation data.

Table 2.23: WER [%] using different recognition settings over LibriSpeech, poliMedia test sets and the 8 speakers of the TED-LIUM corpus.

Recognition setting	LibriSpeech	poliMedia	TED-LIUM
2-pass	13.50	12.53	20.78
3-pass	13.06	12.37	20.02
3-pass+CM	13.05	12.06	19.63

As we can see, relative reductions in WER of 3.3%, 1.3% and 3.7% are obtained in LibriSpeech, poliMedia and TED-LIUM, respectively, by performing this additional adaptation pass. Moreover, in the case of poliMedia and TED-LIUM, further improvements are achieved by using the proposed third pass based on CM. These improvements reach relative reductions in WER of 2.5% and 2% in poliMedia and TED-LIUM, respectively, with respect to the "3-pass" setting. As a result, relative reductions in WER of 3.3%, 3.8% and 5.5% are achieved in LibriSpeech, poliMedia and TED-LIUM, respectively, by performing this third pass based on CM.

5.5 Conclusions and Future Work

In this paper, we have presented a comprehensive study of speaker adaptation of DBRNN and DBLSTM models for confidence estimation. The study has confirmed the superiority of RNN-based models over the CRF and WP approaches. In particular, a linear interpolation of DBRNN and DBLSTM models has obtained the best performance. Furthermore, we have shown that speaker adaptation of confidence measures is an effective approach for improving confidence estimation. This is an important practical outcome, since general-purpose confidence measures have to be applied frequently in multiple applications and adaptation becomes necessary. As a final contribution, we have proposed a novel unsupervised adaptation of the acoustic DBLSTM based on confidence measures. Relative reductions in WER in the range of $3\% - 5.5\%$ have been achieved in different speech tasks by adding an extra recognition pass of adaptation based on confidence measures into a classical two-pass ASR decoder.

As future work, we plan to apply the same approach to estimate speaker-adapted confidence measures at different levels, such as sub-word or utterance. The idea is to use a bottom-up approach (from sub-word to utterance) where class probabilities generated by lower-level RNN models are used as additional input features by RNN models at higher levels. Moreover, based on previous works [24, 18], we plan to investigate different adaptation approaches in which reestimation of specific parts of the network would be performed depending on the amount of adaptation data.

References

[1] Martín Abadi and et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `http://tensorflow.org/` (cit. on p. 78).

[2] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *Neural Networks, IEEE Transactions on* 5.2 (1994), pp. 157–166 (cit. on p. 75).

[3] Herve A. Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993. ISBN: 0792393961 (cit. on p. 78).

[4] Miguel Ángel Del-Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchis, and Alfons Juan. "The transLectures-UPV toolkit". In: *Proc. of VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*. Las Palmas de Gran Canaria (Spain), Jan. 1, 2014 (cit. on p. 77).

[5] Miguel Ángel Del-Agua, Adrià Martínez-Villaronga, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. "The MLLP system for the 4th CHiME challenge". In: *Proc. of the 4th CHiME Speech Separation and Recognition Challenge (CHiME-4)*. San Francisco (USA), Jan. 1, 2016, pp. 57–59 (cit. on p. 77).

[6] Miguel Ángel Del-Agua, Adrià Martínez-Villaronga, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, J. Civera, and A. Juan. "The MLLP ASR Systems for IWSLT 2015". In: *Proc. of 12th IWSLT*. Da Nang (Vietnam), Dec. 3, 2015 (cit. on p. 77).

[7] Miguel Ángel Del-Agua, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. “ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks”. In: *Interspeech*. 2016, pp. 3464–3468 (cit. on pp. 73–75, 77, 81).

[8] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong. “TTS synthesis with bidirectional LSTM based recurrent neural networks”. In: *Interspeech*. 2014 (cit. on p. 75).

[9] Sahar Ghannay, Yannick Esteve, and Nathalie Camelin. “Word embeddings combination and neural networks for robustness in ASR error detection”. In: *Signal Processing Conference (EUSIPCO)*. 2015, pp. 1671–1675 (cit. on p. 75).

[10] Sahar Ghannay, Yannick Esteve, Nathalie Camelin, and Paul Deléglise. “Acoustic Word Embeddings for ASR Error Detection.” In: *Interspeech*. 2016, pp. 1330–1334 (cit. on p. 75).

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016 (cit. on p. 77).

[12] A. Graves, N. Jaitly, and A. r. Mohamed. “Hybrid speech recognition with Deep Bidirectional LSTM”. In: *IEEE Workshop on Automatic Speech Recognition and Understanding*. 2013, pp. 273–278 (cit. on p. 78).

[13] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. “Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition”. In: *International Conference on Artificial Neural Networks: Formal Models and Their Applications*. 2005, pp. 799–804 (cit. on p. 75).

[14] H. Jiang. “Confidence Measures for Speech Recognition: A Survey”. In: *Speech Communication* 45.4 (2005), pp. 455–470 (cit. on pp. 73, 74).

[15] Geoffrey Hinton, Li Deng, Dong Yu, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath George Dahl, and Brian Kingsbury. “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. In: *IEEE Signal Processing Magazine* 29.6 (Nov. 2012), pp. 82–97 (cit. on p. 78).

[16] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 75).

[17] Po-Sen Huang, Kush Kumar, Chaojun Liu, Yifan Gong, and Li Deng. “Predicting speech recognition confidence using deep learning with word identity and score features”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 7413–7417 (cit. on pp. 74, 75).

[18] Z. Huang, J. Tang, S. Xue, and L. Dai. “Speaker adaptation of RNN-BLSTM for speech recognition based on speaker code”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 5305–5309 (cit. on p. 88).

[19] Kaustubh Kalgaonkar, Chaojun Liu, Yifan Gong, and Kaisheng Yao. “Estimating confidence scores on ASR results using recurrent neural networks”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 4999–5003 (cit. on pp. 74, 75).

[20] Kshitiz Kumar, Chaojun Liu, and Yifan Gong. “Normalization of ASR confidence classifier scores via confidence mapping”. In: *Interspeech*. 2014, pp. 1199–1203 (cit. on pp. 74, 75).

[21] Thomas Lavergne, Olivier Cappé, and François Yvon. “Practical Very Large Scale CRFs”. In: *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 504–513 (cit. on p. 81).

[22] E. Lleida and R. C. Rose. “Utterance verification in continuous speech recognition: decoding and training procedures”. In: *IEEE Transactions on Speech and Audio Processing* 8.2 (2000), pp. 126–139 (cit. on p. 74).

[23] Lidia Mangu, Eric Brill, and Andreas Stolcke. “Finding consensus in speech recognition: word error minimization and other applications of confusion networks”. In: *Computer Speech & Language* 14.4 (2000), pp. 373–400 (cit. on p. 74).

[24] Yajie Miao and Florian Metze. “On speaker adaptation of long short-term memory recurrent neural networks”. In: *Interspeech*. 2015 (cit. on p. 88).

[25] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. “Recurrent neural network based language model”. In: *Interspeech*. Vol. 2. 2010, pp. 1045–1048 (cit. on p. 75).

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 3111–3119 (cit. on p. 75).

[27] Ulisses M Braga Neto and Edward R Dougherty. *Error estimation for pattern recognition*. John Wiley & Sons, 2015 (cit. on p. 82).

[28] A. Ogawa and T. Hori. “ASR error detection and recognition rate estimation using deep bidirectional recurrent neural networks”. In: *ICASSP*. IEEE. 2015, pp. 4370–4374 (cit. on pp. 73, 74).

[29] Atsunori Ogawa and Takaaki Hori. “Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks”. In: *Speech Communication* 89.4 (2017), pp. 70–83 (cit. on pp. 73–75, 81).

[30] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. “Librispeech: an ASR corpus based on public domain audio books”. In: *ICASSP*. IEEE. 2015, pp. 5206–5210 (cit. on pp. 77, 78).

[31] Anthony Rousseau, Paul Deléglise, and Yannick Estève. “Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks.” In: *LREC*. 2014, pp. 3935–3939 (cit. on p. 82).

[32] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *Cognitive modeling* 5.3 (1988), p. 1 (cit. on p. 76).

[33] RWTH, UPVLC, XEROX, and EML. *D3.1.1: First report on massive adaptation*. Tech. rep. transLectures, 2012. URL: `http://www.translectures.eu/wp-content/uploads/2013/05/transLectures-D3.1.1-18Nov2012.pdf` (cit. on p. 78).

[34] Isaias Sanchez-Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, and Alfons Juan. "Speaker-adapted confidence measures for speech recognition of video lectures". In: *Computer Speech & Language* 37 (2016), pp. 11–23 (cit. on pp. 73–75).

[35] Alberto Sanchis, Alfons Juan, and Enrique Vidal. "A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 565–574 (cit. on pp. 73–75).

[36] M. Schuster and K.K. Paliwal. "Bidirectional Recurrent Neural Networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681 (cit. on p. 75).

[37] Matthew Stephen Seigel. "Confidence Estimation for Automatic Speech Recognition Hypotheses". PhD thesis. Department of Engineering, University of Cambridge, 2013 (cit. on pp. 73, 74, 81).

[38] Joan Albert Silvestre, Miguel Ángel Del-Agua, Gonzalo Vicente Garcés, Guillem Gascó, Adrián Giménez, Adrià Agustí Martínez-Villaronga, Alejandro Manuel Pérez-González, Isaías Sánchez-Cortina, Nicolás Serrano, Rachel Nadine, et al. "transLectures". In: *IberSPEECH 2012-VII Jornadas en Tecnologia del Habla and III Iberian SLTech Workshop*. IberSPEECH 2012. 2012, pp. 345–351 (cit. on p. 77).

[39] Manhung Siu and Herbert Gish. "Evaluation of word confidence for speech recognition systems". In: *Computer Speech & Language* 13.4 (1999), pp. 299–319 (cit. on p. 81).

[40] G. Stemmer, F. Brugnara, and D. Giuliani. "Adaptive Training Using Simple Target Models". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. 2005, pp. 997–1000 (cit. on p. 78).

[41] UPVLC, XEROX, JSI-K4A, RWTH, and EML. *D3.1.3: Final report on massive adaptation*. Tech. rep. transLectures, 2014. URL: `https://www.translectures.eu/wp-content/uploads/2015/01/transLectures-D3.1.3-31Oct2014.pdf` (cit. on p. 78).

[42] O. Vinyals, S. V. Ravuri, and D. Povey. "Revisiting Recurrent Neural Networks for robust ASR". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 4085–4088 (cit. on p. 75).

[43] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney. "Confidence measures for large vocabulary continuous speech recognition". In: *Speech and Audio Processing, IEEE Transactions on* 9.3 (2001), pp. 288–298 (cit. on pp. 74, 79, 81).

[44] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. "Adaptation of context-dependent deep neural networks for automatic speech recognition". In: *Proc. of the SLT*. 2012, pp. 366–369 (cit. on p. 77).

[45] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. "Recurrent neural networks for language understanding." In: *Interspeech*. 2013, pp. 2524–2528 (cit. on p. 75).

[46] S. J. Young, J. J. Odell, and P. C. Woodland. "Tree-based state tying for high accuracy acoustic modelling". In: *Proc. of HLT*. 1994, pp. 307–312 (cit. on p. 78).

[47] Dong Yu, Jinyu Li, and Li Deng. "Calibration of Confidence Measures in Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 19.8 (2011), pp. 2461–2473 (cit. on p. 74).

[48] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and F. Seide. “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition”. In: *ICASSP*. IEEE. 2013, pp. 7893–7897 (cit. on pp. 77, 85).

[49] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. “A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition”. In: *ICASSP*. IEEE. 2017, pp. 2462–2466 (cit. on p. 75).

Chapter 3

General discussion of the results

In this chapter the results for each of the goals established will be presented as well as some details about their impact in the research projects and beyond. The reader is invited to revisit the list of goals in Section 1.3.

Improve ASR systems based on DNNs by means of unsupervised speaker adaptation

This goal was achieved to a great extend thanks to the continuous development of the TLK toolkit. It was presented in Paper 1 and constitutes the basic tool based on which different systems have been built for all the experiments in the thesis and for real-world educational repositories. Nowadays is the core transcription tool in different platforms such as VideoLectures.NET[a], poliTrans[b] and polisubs[c]. VideoLectures.NET is a free and open access web portal that has so far published more than 20K educational videos. Politrans is a platform offered by UPV for automatic video transcription and translation available for use by any interested university or organization. Polisubs is a service used by UPV at university conference halls for real-time speech transcription of lectures. Although TLK was first released under the open source Apache License 2.0 during the transLectures project, the software was continuously improved by the addition of new SOTA techniques. These new techniques ranged from new feature extraction methods, new DL models to different speaker adaptation approaches.

TLK successfully contributed to achieve the main goal of this thesis related to the efficient transcription of video lectures. The case studies from the research projects have always constituted the best examples to test the toolkit and to obtain a good idea of its behavior in real-life scenarios. In Fig. 3.1 the different systems trained in the context of transLectures and EMMA projects are shown. Regarding the transLectures project, three systems were trained and systematically improved: one system for VideoLectures.NET (English) and two systems for poliMedia (Spanish and Catalan). With respect to EMMA, which took place

[a] http://videolectures.net

[b] https://politrans.upv.es

[c] https://polisubs.upv.es

after transLectures, five more systems were trained to transcribe courses from the *Université de Bourgogne* (French), *The Open Universiteit of the Netherlands* (Dutch), *Universidade Aberta* (Portuguese), *Università degli Studi di Napoli Federico II* (Italian) and *Tallin University* (Estonian).

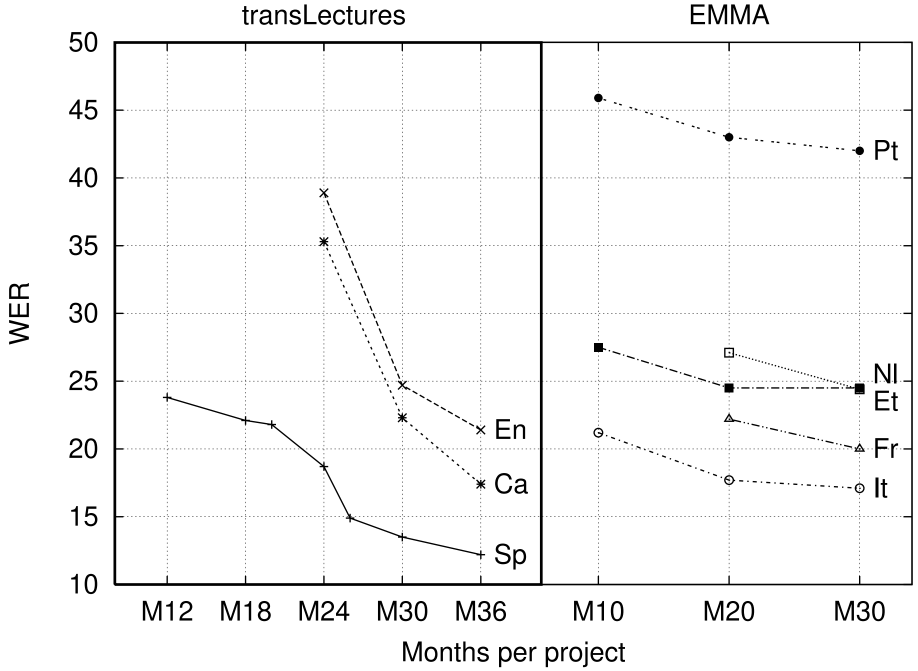

Figure 3.1: Systems developed with TLK during transLectures and EMMA.

As can be observed in Fig. 3.1, taking a look at the transLectures project during months 20 and 26, the systems experimented a big improvement which is mainly explained by the use of DNNs for acoustic modeling and adaptation. After that, this technique was gradually applied to other languages such as Catalan and English. By the end of the project, different topologies and training strategies were tested, such as CNNs [1], new regularization techniques such as dropout [5], new activation functions like Rectified Linear Unit (ReLU) [2] and speaker adaptation techniques like feature-space Discriminative Linear Regression (fDLR) [9].

In the case of EMMA, all systems started with state-of-the-art DNN-based acoustic models. Although the quality of the systems in some cases was far from perfect, it should be mentioned that in those cases the tasks were really hard. In fact, one of the main problems was the lack of in-domain data (similar acoustic conditions to the videos that constitute a course). Moreover, comparing TLK transcription quality with Youtube, TLK systematically provided transcriptions with relative improvements of 54.1% for Italian, 40.4% for Dutch and 35.6% for French. Other than that, among the new techniques applied we can highlight speaker adaptation of DNNs based on Kullback-Leibler divergence [10], system combination (i.e. DNNs and CNNs), multilingual DNNs [6] or the use of RNNs for language modeling [7].

Although MORE is not shown in the plot, during its course the existing systems were further improved thanks to the use of BLSTMs for acoustic modeling [11], the proposed unsupervised speaker adaptation technique, LSTMs for language modeling and more training data. In fact, all systems were improved, but the ones that experienced a greater leap in quality were French (from 20.0 to 16.5) and Portuguese (from 42.0 to 23.3).

From the point of view of the publications shown in previous section (presented in chronological order), it can be observed the evolution of the toolkit and the pace at which new features were added. Moreover, in Papers 2 and 3 TLK was subjected to two international competitions, the IWSLT and CHiME-4 challenges. The IWSLT, constituted the first submission of a TLK system to an ASR challenge. From a total of six research groups worldwide, the Machine Learning and Language Processing (MLLP) group was the only one to take part with a system completely based on the use of internally developed ASR technology.

As can be observed in Table. 3.1, the results of TLK were competitive enough compared to the results from other research groups. It is also worth mentioning that, as there weren't limitations on the use of acoustic training data (except for a set of videos from TED), the use of the appropriate technique for data selection and data filtering played a very important role. Therefore, the question remains whether using the same acoustic data, the gap among all results would have been greatly reduced or not. Other than that, we can consider that the English ASR performance for this task was at a very good state given the almost human performance.

Table 3.1: Results of the IWSLT 2015 evaluation campaign on English ASR.

Participant	ASR Software	Training Data (hours)	WER (test 2015)
MITLL-AFRL (USA)	KALDI + HTK	336	6.6
HLT-I2R (Singapore)	KALDI	486	8.9
KIT (Germany)	KALDI + JANUS	579	9.2
NAIST (Japan)	KALDI	439	12.0
MLLP (Spain)	**TLK**	**245**	**13.3**
IOIT (Vietnam)	KALDI	520	13.8

Regarding the CHiME-4 challenge, presented in Paper 3, up to 14 participants from around the world tried to design and train the best ASR system. There were academic research groups from Germany (Paderborn, Aachen RWTH Universities), China (University of Science and Technology among others), Japan (Tokyo Institute of Technology), USA (Georgia Institute Laboratory) or Italy (Fondazione Bruno Kessler). There were also companies such as Mitsubishi Electric, Google or Hitachi. Every participant was restricted to made use of the same set of acoustic and text data in order to train their system. Nevertheless, different approaches were proposed in order to improve the system from the acoustic or language model point of view. In addition to this, the techniques for audio preprocessing and enhancement were fundamental for tracks with more than one audio channel.

According to Table 3.2 the TLK system obtained competitive results compared to others in the 1-channel track. It made use of a rather simple system were only 2 acoustic models

Table 3.2: 1-channel results with baseline LM

Rank	Team	Dev Sim	Dev Real	Eval Sim	Eval Real
1	Heymann et al.	7.2 %	5.5 %	11.7 %	9.9 %
2	Du et al.	8.2 %	6.1 %	13.6 %	11.2 %
3	Fujita et al.	7.4 %	5.9 %	9.2 %	11.4 %
4	Alam et al.	9.3 %	6.8 %	13.7 %	12.7 %
5	Tran et al.	-	-	-	12.9 %
6	Qian and Tan	7.9 %	6.3 %	12.9 %	13.9 %
7	**Del-Agua et al.**	11.1 %	9.9 %	15.7 %	**16.1** %
8	Matassoni et al.	9.5 %	9.0 %	16.1 %	16.9 %
9	Tanaka et al.	10.9 %	9.1 %	16.5 %	17.4 %
10	Bayestehtashk and Shafran	12.1 %	9.8 %	19.1 %	18.6 %
11	Xiao et al.	14.3 %	11.4 %	21.4 %	20.9 %
12	Baseline	13.0 %	11.6 %	20.8 %	23.7 %

Table 3.3: 2-channels results with baseline LM

Rank	Team	Dev Sim	Dev Real	Eval Sim	Eval Real
1	Du et al.	4.9 %	3.6 %	7.3 %	5.4 %
2	Heymann et al.	4.5 %	3.8 %	5.4 %	6.4 %
3	Fujita et al.	5.9 %	4.2 %	7.3 %	8.6 %
4	Qian and Tan	5.7 %	4.8 %	8.7 %	9.1 %
5	Wang et al.	7.2 %	5.6 %	8.8 %	9.6 %
6	Tran et al.	-	-	-	9.8 %
7	Alam et al.	6.7 %	5.1 %	10.3 %	10.0 %
8	Xiao et al.	7.1 %	5.9 %	10.7 %	10.5 %
9	Zhang et al.	6.3 %	5.5 %	7.8 %	11.0 %
10	**Del-Agua et al.**	8.9 %	8.0 %	12.1 %	**12.8** %
11	Bayestehtashk and Shafran	8.8 %	7.3 %	13.9 %	13.8 %
12	Schrank et al.	8.5 %	6.7 %	14.5 %	14.0 %
13	Baseline	9.5 %	8.2 %	15.3 %	16.6 %

(BLSTMs and Feed-Forward based) were combined, but it was also patent that a better system fine tuning was necessary with respect the BLSTM side. Regarding the 2-channel track, better results could have been achieved by exploiting acoustic data preprocessing techniques as can be drawn from the results in Table 3.3.

In IWSLT results a new unsupervised speaker adaptation step was proposed that made use of CMs. This technique greatly contributed to improve the quality of the transcriptions without human intervention. It was applied to 5 different systems with relative gains in the range of 1.2% to 3.7% WER. Moreover, it was shown that the technique obtained good results when the CE was further improved by CM models based on NNs. In fact, it was able to consistently reduce WER even for cutting-edge ASR systems. More concretely, obtained relative reductions of 3.3%, 1.3% and 3.7% in LibriSpeech, poliMedia and TED-LIUM.

Improve CM estimation by means of NNs and speaker adaptation

This goal has been achieved thanks to the new model architecture proposed based on NNs to model CMs. Moreover, motivated by the fact that educational video repositories usually contain several videos from the same author, further improvements were achieved by means of a new speaker adaptation step which demonstrated to be successful.

In fact, it is shown that RNN based classifiers outperform all previous SOTA classifiers. In particular, both DRNNs and DBLSTMs have shown their superiority when compared with non-NN-based classifiers or even with DNNs [8, 4, 3]. This is one of the main contributions of this work, which has been described in Papers 4 and 5, where the CRF models are outperformed by up to 12% and 5% relative improvement in poliMedia and LibriSpeech tasks in terms of CER.

Regarding CM adaptation techniques, there have been different attempts in the literature such as confidence calibration or confidence measure re-normalization. In this thesis, a novel unsupervised speaker adaptation technique for RNN-based models has been proposed. This technique achieved 4.6% relative improvement in terms of CER, and it was further extended and analyzed in Paper 5 for a different task. In all experiments the speaker adaptation step consistently outperformed non-adapted systems. Moreover, in the latter work, it was also carried out an study about the amount of adaptation data required to properly adapt the system, and it was concluded that with just $3 - 5$ minutes for each speaker was enough.

References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. “Convolutional neural networks for speech recognition”. In: *IEEE/ACM Transactions on audio, speech, and language processing* 22.10 (2014), pp. 1533–1545 (cit. on p. 94).

[2] George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. “Improving deep neural networks for LVCSR using rectified linear units and dropout”. In: *ICASSP*. IEEE. 2013, pp. 8609–8613 (cit. on p. 94).

[3] Miguel Ángel Del-Agua, Adriá Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. “Speaker-Adapted Confidence Measures for ASR Using Deep Bidirectional Recurrent Neural Networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.7 (2018), pp. 1194–1202 (cit. on p. 97).

[4] Miguel Ángel Del-Agua, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. “ASR Confidence Estimation with Speaker-Adapted Recurrent Neural Networks”. In: *Interspeech*. 2016, pp. 3464–3468 (cit. on p. 97).

[5] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *CoRR* abs/1207.0580 (2012) (cit. on p. 94).

[6] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. “Cross-Language knowledge transfer using multilingual deep neural network with shared hidden layers”. In: *ICASSP*. IEEE. 2013, pp. 7304–7308 (cit. on p. 94).

[7] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. “Recurrent neural network based language model”. In: *Interspeech*. Vol. 2. 2010, pp. 1045–1048 (cit. on p. 94).

[8] Atsunori Ogawa and Takaaki Hori. “Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks”. In: *Speech Communication* 89.4 (2017), pp. 70–83 (cit. on p. 97).

[9] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. “Adaptation of context-dependent deep neural networks for automatic speech recognition”. In: *Proc. of the SLT*. 2012, pp. 366–369 (cit. on p. 94).

[10] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. “Recurrent neural networks for language understanding.” In: *Interspeech*. 2013, pp. 2524–2528 (cit. on p. 94).

[11] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. “A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition”. In: *ICASSP*. IEEE. 2017, pp. 2462–2466 (cit. on p. 95).

Chapter 4

Conclusions and Future Work

In this chapter the main conclusions of this thesis and future work are described. First of all, a new ASR tool was entirely developed from scratch within the framework of the transLectures project to facilitate the transcription of video lecture repositories. State-of-the-art techniques were implemented and particularly big efforts were devoted to add support for training and inference with DL Models. Moreover, TLK was subjected to two international ASR competitions and demonstrated to provide competitive results compared to the software used in different institutions worldwide. Finally, following the research line of ASR accuracy improvement, a novel approach for unsupervised speaker adaptation using CMs was successfully proposed.

Apart from that, a first approach based on BLSTMs to improve CE was proposed. This technique demonstrated to provide state-of-the-art results in a publicly available dataset known as LibriSpeech. Moreover, it was also proposed a novel approach for unsupervised speaker adaptation of the CM model which obtained further improvements.

A detailed analysis on the use of RNNs and LSTMs for CM estimation in different datasets was carried out. In addition, speaker adapted CMs demonstrated to be the way to follow when enough speaker data is available. Finally, an application of CMs to improve the output from an ASR system was also proposed.

In summary, the main contributions of this thesis are:

- A simple yet powerful unsupervised speaker adaptation technique of acoustic models.
- Improvements over the state-of-the-art in CE by means of RNNs and BLSTMs neural networks.
- A new approach for providing speaker-adapted CM using RNNs and BLSTMs.

Regarding future work, the technological and scientific contributions of this thesis can be further extended. In fact, TLK can be extended to support state-of-the-art NN topologies such as end-to-end systems [3], where the entire system depends only on Neural Networks. Moreover, it would be very interesting to add support for beam-forming [2] techniques as it demonstrated to be very effective for robust speech recognition. It goes without saying that the participation in new ASR challenges is fundamental to keep the software updated and conveniently compared to other toolkits within the research community.

With respect to the improvements reported in CE, it was shown that even obtaining highly competitive results in terms of recognition accuracy, these systems are still capable of improving the CMs obtained as posterior probabilities. As future work, it would be really interesting to explore different NNs topologies that have obtained competitive results in text classification tasks such as Very Deep CNNs [1]. Apart from that, it would be also interesting to explore different approaches to perform transfer learning [4], as to quickly adapt CM systems to different language domains.

References

[1] Alexis Conneau, Holger Schwenk, Loic Barrault, and Yann Lecun. “Very deep convolutional networks for text classification”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Vol. 1. 2017, pp. 1107–1116 (cit. on p. 100).

[2] Sharon Gannot, David Burshtein, and Ehud Weinstein. “Signal enhancement using beamforming and nonstationarity with applications to speech”. In: *IEEE Transactions on Signal Processing* 49.8 (2001), pp. 1614–1626 (cit. on p. 99).

[3] Alex Graves and Navdeep Jaitly. “Towards end-to-end speech recognition with recurrent neural networks”. In: *International Conference on Machine Learning*. 2014, pp. 1764–1772 (cit. on p. 99).

[4] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359 (cit. on p. 100).