UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dept. of Computer Systems and Computation

# Efficient domain adaptation techniques for hybrid automatic speech recognition systems

Master's Thesis

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

AUTHOR: Santamaría Jordá, Jaume

Tutor: Silvestre Cerdà, Joan Albert

External cotutor: GIMENEZ PASTOR, ADRIAN

ACADEMIC YEAR: 2022/2023

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ


UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

MASTER'S THESIS

# Efficient domain adaptation techniques for hybrid automatic speech recognition

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging
Academic Course 2022/2023

Jaume Santamaría Jordà

Advisers:
Dr. Joan Albert Silvestre Cerdà
Dr. Adrià Giménez Pastor

# Abstract / Resum / Resumen

**Abstract**

Automatic Speech Recognition (ASR) is a very active natural language processing task in the area of artificial intelligence, with many primary and secondary applications, such as automatic and computer-assisted subtitling, speech translation, and voice dubbing, among others. In the last decade, this task has received a lot of attention from major technology companies and research labs because of the large performance improvements obtained by incorporating deep learning techniques. As a result, general-purpose ASR systems, trained with large amounts of data, can exhibit sufficiently good transcription quality in many, but not all, applications. Under very specific application domains, characterized by lexical (particular argots and keywords, e.g., particle physics, oncology, etc.), acoustic (e.g., far-field, reverberations, lossy audio compression, etc.) and/or linguistic (e.g., local dialects, non-native speakers, spontaneous speech, etc.) factors, general-purpose ASR systems often exhibit significant quality losses because of their lack of specialization. In this work we explore efficient domain adaptation techniques for hybrid general-purpose ASR systems under the framework of the EU-funded research project Interact-Europe (EU4Health Programme, project no. 101056995), with the aim of improving their transcription qulity under the oncology (medical) domain.

**Keywords**: Automatic Speech Recognition; Domain Adaptation; Machine Learning.

**Resum**

El reconeixement automàtic de la parla (Automatic Speech Recognition, ASR) és una tasca de processament del llenguatge natural molt activa en l'àrea de la intel·ligència artificial, amb moltes aplicacions primàries i secundàries, com el subtitulat automàtic i assistit per ordinador, la traducció de veu, i el doblatge de veu, entre d'altres. En l'última dècada, aquesta tasca ha rebut molta atenció per part de les principals empreses tecnològiques i laboratoris de recerca a causa de les grans millores de rendiment obtingudes en incorporar tècniques d'aprenentatge profund. Com a resultat, els sistemes ASR de propòsit general, entrenats amb grans quantitats de dades, poden exhibir una qualitat de transcripció suficientment acurada en moltes aplicacions, però no en totes. Sota dominis d'aplicació molt específics, caracteritzats per factors lèxics (argots i paraules clau particulars, p.e. física de partícules, oncologia, etc.), acústics (p.e. camp llunyà, reverberacions, compressió d'àudio amb pèrdua, etc.) i/o lingüístics (p.e. dialectes locals, parlants no nadius, parla espontània, etc.), els sistemes ASR d'ús general solen mostrar pèrdues significatives de qualitat a causa de la seva manca d'especialització. En aquest treball explorem tècniques eficients d'adaptació al domini per a sistemes ASR híbrids de propòsit general en anglès, en el marc del projecte de recerca finançat per la UE Interact-Europe (Programa EU4Health, projecte núm. 101056995), amb l'objectiu de millorar la seva qualitat de transcripció en el domini (mèdic) oncològic.

**Paraules clau**: Reconeixement Automàtic de la Parla; Adaptació al domini; Aprenentatge Automàtic.

**Resumen**

El reconocimiento automático del habla (Automatic Speech Recognition, ASR) es una tarea de procesamiento del lenguaje natural muy activa en el área de la inteligencia artificial, con muchas aplicaciones primarias y secundarias, como el subtitulado automático y asistido por ordenador, la traducción de voz, y el doblaje de voz, entre otros. En la última década, esta tarea ha recibido mucha atención por parte de las principales empresas tecnológicas y laboratorios de investigación a causa de las grandes mejoras de rendimiento obtenidas al incorporar técnicas de aprendizaje profundo. Como resultado, los sistemas ASR de propósito general, entrenados con grandes cantidades de datos, pueden exhibir una calidad de transcripción suficientemente buena en muchas aplicaciones, pero no en todas. Bajo dominios de aplicación muy específicos, caracterizados por factores léxicos (argots y palabras clave particulares, p.e. física de partículas, oncología, etc.), acústicos (p.e., campo lejano, reverberaciones, compresión de audio con pérdida, etc.) y/o lingüísticos (p.e. dialectos locales, hablantes no nativos, habla espontánea, etc.), los sistemas ASR de uso general suelen mostrar pérdidas significativas de calidad a causa de su carencia de especialización. En este trabajo exploramos técnicas eficientes de adaptación al dominio para sistemas ASR híbridos de propósito general en inglés, en el marco del proyecto de investigación financiado por la UE Interact-Europe (Programa EU4Health, proyecto nº 101056995), con el objetivo de mejorar su calidad de transcripción en el dominio (médico) oncológico.

**Palabras clave**: Reconocimiento Automático del Habla; Adaptación al dominio; Aprendizaje Automático.

# Contents

Chapter 1

# Introduction

This work studies and compares different techniques for efficient adaptation of hybrid Automatic Speech Recognition (ASR) systems. This chapter introduces the motivation and the main objectives of the work, as well as the context necessary to properly understand the details of the study.

## 1.1 Motivation

Nowadays, being able to understand and process people's speech is a key task in continuing to facilitate access to information, and this is where ASR comes in, as it allows computers to process information expressed in a way that is natural to humans. This opens the door to different technologies such as automatic captioning, recognition of voice commands to an electronic device, analysis of the content of a spoken speech, automatic translation or voice synthesis, among others. Thanks to the attention this field has received in recent years from large companies and research groups, it has been possible to develop general purpose ASR systems that achieve satisfactory transcription quality in many applications, such as parliamentary debates or audio books, achieving error rates similar to or lower than human error rates. Even so, there are many applications that do not reach an acceptable quality for real use. This is due to a lack of specialisation of the system, as it is not prepared to deal with the problems that characterise this type of task: specific slang and jargon, problems with audio quality or various linguistic problems, such as local dialects, spontaneous speech or non-native speakers.

This points to the need for systems adapted to specific tasks and capable of maximising their accuracy. Moreover, the techniques to achieve this should be efficient and effective, i.e. they should be able to accomplish this purpose with reasonable and affordable amounts of data and computational resources. In this work, different domain adaptation techniques are proposed and studied, motivated by the research project "Interact-Europe - Innovative collaboration for Inter-specialty cancer training

across Europe" [1], an 18-month project co-funded by the EU under the EU4Health programme as part of Europe's Beating Cancer Plan [2], being the *Machine Learning and Language Processing* [3] (MLLP) research group, integrated in the *Valencian Institute for Research in Artificial Intelligence* [4] (VRAIN) of the Universitat Politècnica de València (UPV), one of the partners in the consortium.

Being a collaborative project with different European countries involved, and focused on oncological medicine, it is a clear example of a task in which a domain-specific system is necessary, as we find all the above-mentioned features in the content that will need to be transcribed. As the role of the MLLP research group is, among others, to develop transcription and machine translation systems to facilitate access to different languages, the adaptation of its general purpose English ASR system to the project domain will be carried out in the present work.

## 1.2 Main objectives

The main objectives of this work are the following:

- To explore different techniques and solutions for adapting hybrid systems in the field of ASR.
- To generate in-domain Acoustic Models (AM) and Language Models (LM) for hybrid ASR.
- To apply the developed system for transcribing educational sessions in the context of the Interact-Europe project.
- Evaluate the performance of the hybrid ASR systems built.
- Compare the performance of these systems with a baseline, general-purpose system to clarify the role of technological improvements.

## 1.3 Document structure

The rest of the document is organised as follows:

- Chapter 2, provides the reader with the theoretical and technological knowledge necessary to understand the rest of the work.
- Chapter 3, describes the project in which this work is encapsulated, as well as the data used to adapt the ASR system and the technological tools.
- Chapter 5 details the techniques that have led to the adaptation of the models that will allow the construction of the final ASR system.

[1] `https://www.europeancancer.org/eu-projects/impact/interact-europe`
[2] `https://health.ec.europa.eu/system/files/2022-02/eu_cancer-plan_en_0.pdf`
[3] `https://mllp.upv.es/`
[4] `https://vrain.upv.es/`

- Chapter 6 reports the results of the evaluation for the differents systems built.
- Chapter 7 concludes with a brief analysis of the results, as well as ideas for future work.

Chapter 2

# Fundamentals of ASR

This chapter provides the theoretical background necessary to understand the work described throughout the document. It is structured as follows: Section 2.1 is intended to explain the basics of Automatic Speech Recognition (ASR). Section 2.2 covers some basic topics about Neural Networks that will be used in the work. Section 2.3 gives a basic overview of machine learning, as well as some of the techniques used. Once the basic elements are known, Section 2.4 details the preprocessing that must be applied to the raw data in order to train our models. Section 2.5 outlines the components that make up a hybrid ASR system, the acoustic model and the language model. The decoding of the hybrid system is then explained in Section 2.6. Section 2.7 explains the different metrics used to evaluate the above models are explained. Finally, Section 2.8 explores the state-of-the-art architectures and approaches.

## 2.1 Automatic Speech Recognition

ASR is the field of Machine Learning that studies the creation of automatic systems with the capability of obtaining the most probable sequence of words that transcribes a particular acoustic signal. This is, given a sequence of acoustic observations, $\boldsymbol{x}$, obtained from a feature extraction process over the acoustic signal, an ASR system computes the sequence of words, $\hat{w}$, that best matches it. All this can be understood from a statistical point of view, with two main approaches: first, using the posteriori probability, this can be modelled as follows:

$$\hat{w} = \underset{w \in L^*}{\operatorname{argmax}} P(w|\boldsymbol{x}) \tag{2.1}$$

where $L$ is the vocabulary of the system, and $L^*$ is the set of all the sentences that can be constructed with this vocabulary. Traditionally the Bayes Theorem is applied to the posteriori probability to decompose the model into two sub-models as seen below:

$$\hat{w} = \underset{w \in L^*}{\operatorname{argmax}} P(w|\boldsymbol{x}) = \underset{w \in L^*}{\operatorname{argmax}} \frac{P(\boldsymbol{x}|w) \cdot P(w)}{P(\boldsymbol{x})} = \underset{w \in L^*}{\operatorname{argmax}} P(\boldsymbol{x}|w) \cdot P(w) \tag{2.2}$$

where the computation of $P(\boldsymbol{x})$ can be avoided as it is a constant quantity during search. In this new equation we have two new terms, $P(\boldsymbol{x}|w)$, called the Acoustic Model (AM), which calculates the probability that the sequence of words $w$ generated the sequence of acoustic vectors $\boldsymbol{x}$, and $P(w)$, known as Language Model (LM), which estimates the probability that the sentence $w$ was part of the language $L$.

Although there are modern approaches that directly model the Equation 2.1, the traditional approach offers some advantages, most notably that we can train the language model with text only, a very abundant resource and that, in the context of adaptation, makes it easier to find in-domain data to adapt, since text only can provide improvements. This results in a very robust, powerful and versatile component, which makes a decisive contribution to improving the performance of ASR systems. It is important to note that this work is based on the hybrid ASR approach, which adopts this approximation.

## 2.2 Neural Networks for ASR

ASR works with temporal sequences, whose components are highly interdependent internally. For example, in natural language production, word sequences have dependencies of order greater than 1; in phonetic production, phoneme sequences also show interdependence (it is not the same to pronounce a /k/ before an /a/ as before an /e/, for example). By using recurrent neural networks (RNNs), it is possible to exploit this contextual information thanks to their cyclic connections, since RNNs are in essence, a type of neural network with direct cycles in their neurons, which give rise to a structure of internal states or memory [48]. A RNN is a neural network that maps from an input sequence space to an output space of sequences dynamically. That is, the prediction of the output $y_t$ depends on the input $x_t$, but also on the hidden state of the system, $h_t$, which is updated over time while the sequence is processed.

One of the RNN architectures that has generated most relevance is the Long Short-Term Memory [18] (LSTM), which structure can be seen in Figure 2.1. The LSTM cell is the basic unit, responsible of connecting past information with the current one. This cell is characterised by three gates that control the flow of information: the input gate, which controls the information entering the cell, the output gate, which controls the information leaving the cell, and the forgetting gate, which is responsible for remembering or forgetting the previous information in the cell. Correctly combined, they generate an output, C, which acts in a similar way to the memory of a computer, and which allows a more or less distant context to be used. In addition to providing a scenario where the output may depend on the near or far context only, this architecture also improves the resistance to the vanishing gradient problem.

A logical evolution of LSTMs are Bidirectional-LSTMs [38] (BLSTMs), which allow dependencies with future information, that is, they can analyse data in both time directions at the same time. Basically it is two LSTM networks working at the same time, each in one time direction, with their outputs typically concatenated together. Figure 2.2 outlines the idea of a Bidirectional LSTM.

BLSTM networks have been a great improvement for ASR, but they are far from

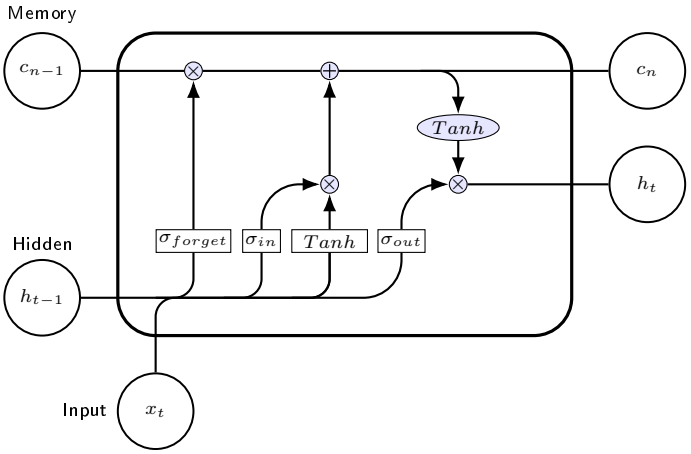


**Figure 2.1:** Internal structure of an LSTM cell. Blue circles represent element-level operations and rectangles a fully connected layer with a concrete activation function (sigmoide or tanh). Two arrows merging together represent concatenation of data, while their separation means that their content is sent to different locations.

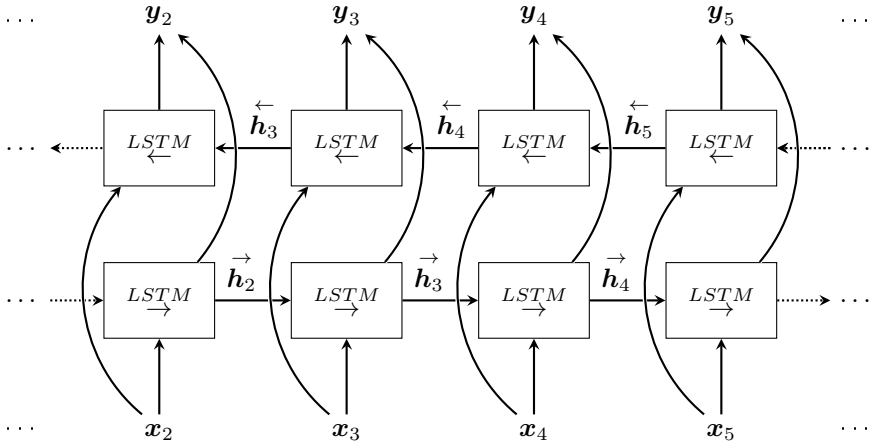


**Figure 2.2:** Representation of a Bi-directional LSTM..

perfect, since, for each time direction, they force a sequential processing of the input data, so it is not possible to exploit the graphic cards' ability to parallelise the computation. Despite this problem, language models based on LSTMs and acoustic models based on BLSTMs have proven to be very successful [28].

Due to the sequentiality of data processing, an attempt was made to analyse the input content and select only representative information for each state: the trend of attention-based models has arrived [7], providing parallel analysis of the input

without the need of sequencial data processing. This leads us to the Transformer [43] architecture. It is based on a encoder-decoder structure, whose parts are formed by attention blocks and feed-forward networks. This proposal has three key aspects that make it remarkable: first, non-sequentiality, in other words, it does not work word by word but processes the whole sequence at the same time. Secondly, the self-attention, which is nothing but attention computed on a concrete sequence to obtain its representation. Self-attention is used to relate a pair of elements in a time that is constant with respect to the relative distance of that elements. Finally, positional embeddings, whose basic idea is to use learned weights that encode the information related to a position of a specific token in the sentence. Thanks to the combination of these ideas, especially positional embeddings, the recurrence of RNNs is replaced, and a great parallelization of the computation is allowed and therefore a great increase in speed. This also provides the ability to work with an extremely long, virtually infinite context, which has proven to give superior results in this type of scenario. Thanks to its advantages and training friendliness, it has become the state-of-the-art architecture in different branches of speech technologies in recent years [33, 1], and being successfully used to construct acoustic [46] and language [21] models.

## 2.3 Regularisation techniques

Machine learning is the field of pattern recognition that is dedicated to the study and development of algorithms and techniques with the ability to generalise a type of problem and learn to solve it autonomously. To generalise in this context is to have the ability to act correctly on new, previously unseen data. Learning, on the other hand, implies that the performance, measured in a particular way, of a task is improved by using previous experience on that task [30].

The data used and the way a system is trained can have devastating effects on its performance if not chosen carefully. One of the main problems that can occur is overfitting: memorising training data. This occurs when the parameters of the trained system are so closely adjusted to the training data that they become completely senseless outside of the training set. In this way the model makes significant errors in processing data that was not in the training, and is not a useful model for its purpose.

There are many ways to combat overfitting during model training, helping the final performance of the model. But some techniques such as training the model with more (quality) data or using data augmentation may be more limited options when performing in-domain adaptation, as in-domain data is not easy to find and data augmentation can only be applied to a certain degree to spoken speech.

One technique that can be applied regardless of the data used is regularisation, a method of avoiding overfitting when training a model by adding a penalty term to the loss function. Concretely, we will talk about the L2 regularisation, which uses the L2 distance as a penalty term, as we can see below:

$$Regularisation = Loss + Penalty = Loss + \lambda \sum_{i=1}^{N} w_i^2 \tag{2.3}$$

where $\lambda$ is a scale factor that allows us to vary the impact of the regularising term. The goal is that weights have a value close to zero, but not zero, meaning that the weight of each feature should have the minimum impact on the model. Therefore, with high values of lambda, an underfitted model and a flat distribution of weights will be obtained, while with very small values of lambda, the regularising term will have no effect.

## 2.4 Data preprocessing and feature extraction

Before training ASR systems, it is necessary to preprocess the training data and to extract the feature vectors, $\boldsymbol{x}$ ,that will feed the models. In ASR systems, acoustic data, their transcriptions and, in the case of hybrid systems, monolingual text are usually available.

First, we will discuss the preprocessing of text data. When dealing with text, we usually perform a clean-up before using it in order to normalise the data. This consists of removing special characters and punctuation marks, converting numbers to text, etc. In short, leaving only the lowercase characters that make up the language to be recognised.

On the other hand, when dealing with audio data, the first step is to obtain the digital version of the original signal. This conversion is divided into two tasks: firstly, sampling must be performed, which consists of measuring the amplitude of the wave at a particular instant of time. A minimum of two samples will be necessary for each wave cycle (one for the positive part of the wave and one for the negative part). The higher the number of samples, the higher the quality, but bearing in mind that the maximum frequency we can measure is half of the sampling frequency (Nyquist frequency [15]). For ASR, 16 KHZ samples are sufficient, since the main sounds of human phonology are below the 8 KHZ spectrum. Secondly, it is necessary to quantify the information by encoding the value of each measured amplitude, typically in 16-bit integers.

Once the digital version of the waveform has been obtained, a series of time windows (frames) are then generated from which the feature vectors will be extracted. This is done because within the windows, we can assume that the signal is stationary (i.e. the statistical properties of the signal remain constant within it), contrary to what happens, normally, in spoken speech. Three parameters have to be set in these windows: the window size (in ms), the delay between a window and its adjacent window (in ms), and the window shape. Figure 2.3 represents the process of generating several windows in order to extract feature vectors. For the shape, since the rectangular window carves the audio abruptly, Hamming windows are normally used, avoiding discontinuities. This is achieved by attenuating both sides of the window

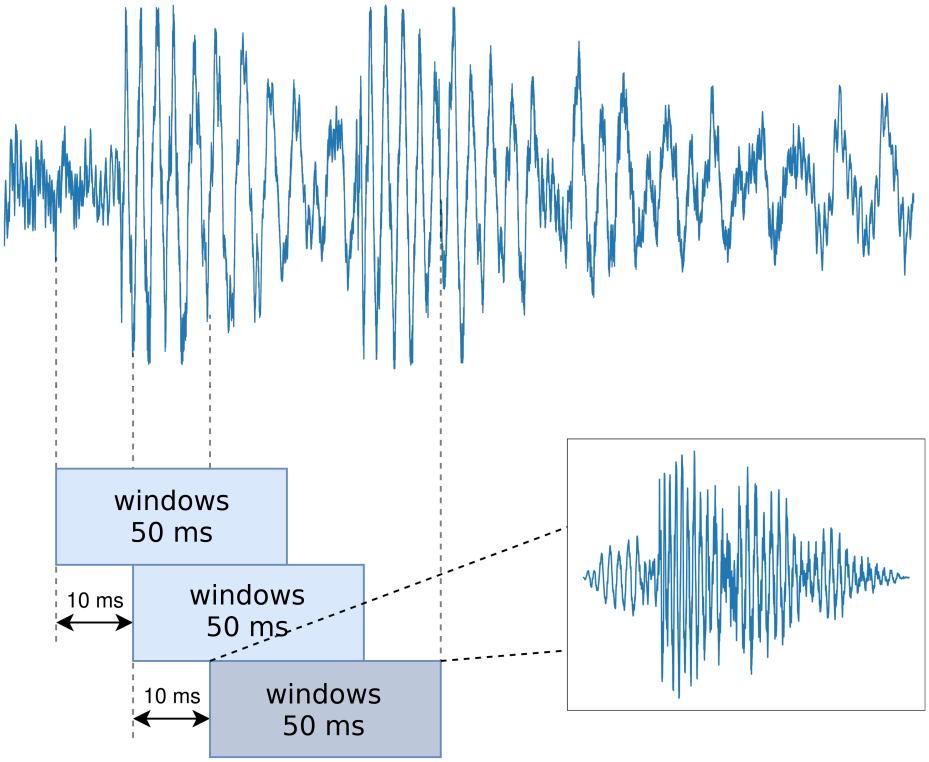


**Figure 2.3:** Windows of 10 ms, with Hamming shape and 10 ms offset.

according to the formula:

$$w(n)_{Rectangular} = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & \text{other} \end{cases} \tag{2.4}$$

$$w(n)_{Hamming} = \begin{cases} 0.54 - 0.46 \cos \dfrac{2\pi n}{L} & 0 \leq n \leq L-1 \\ 0 & \text{other} \end{cases} \tag{2.5}$$

The next step is to extract the information from each window. A Discrete Fourier Transform (DFT) is performed, which extracts the energy of the discrete-time signal in different frequency bands. This transform is defined as:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{2\pi i}{N} kn} \quad \forall k \in 0, \ldots, N-1, \tag{2.6}$$

on $i$ is the imaginary unit and $e^{-\frac{2\pi i}{N}}$ is the N-th root of the unit. Typically, the *Fast Fourier Transform (FFT)* algorithm is used to calculate the DFT. This implementation is very efficient, but only works for values of $N$ that are powers of 2. At this

point we have, for each of the N frequency bands, a complex number representing the magnitude and phase of the frequency of the original signal.

The human hearing does not perceive all frequency bands in the same way: it is much more sensitive to low frequencies than to higher ones. For this reason, this perception is modeled in order to significantly improve the quality of speech recognition. This can be achieved using the mel scale, that has the particularity of being able to separate two sounds equidistant to the human hearing at the same number of mels. The mel frequency can be computed as follows:

$$mel(f) = 1127 \cdot ln(1 + \frac{f}{700}) \tag{2.7}$$

With this is mind, it is possible to implement this idea with a Mel filter bank [24], resulting in a better resolution at low frequencies and lower resolution at higher frequencies. The number of dimensions of the resulting vector is equal to the number of filters forming the bank. Figure 2.4 exemplifies a triangular filter bank that implements this idea. At the output of this step, we have the filterbanks feature vec-

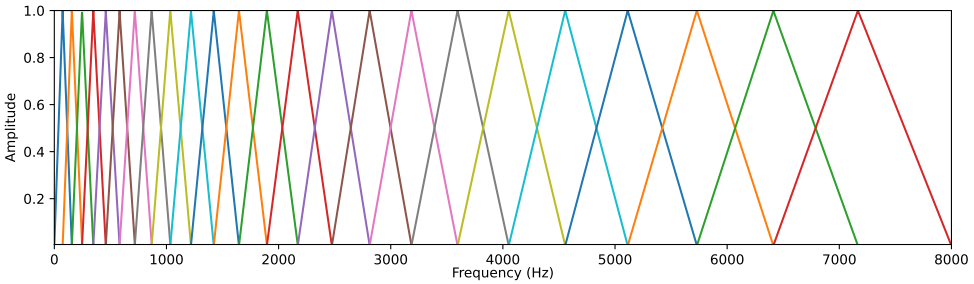


**Figure 2.4:** Mel filter bank consisting of 24 different filters. Each triangular filter is logarithmically spaced using the Mel scale.

tors that will be used to train the BLSTMs models. Traditionally an extra step is made: the *Discrete Cosine Transform (DCT)* algorithm is applied and the samples are normalised in order to minimise their differences. In this way, small differences in pronunciation and loudness become less relevant, increasing the generalisability of the model. The output of this preprocess is the MFCC (Mel Frequency Cepstral Coefficient) feature vectors, one for each frame.

## 2.5 Hybrid ASR

This section describes the basics of the hybrid approach to ASR, explaining the two components that compose it: the acoustic model and the language model.

### 2.5.1 Acoustic Model

The acoustic model, $P(\boldsymbol{x}|w)$, is the component of the hybrid ASR approach that is responsible for representing, by means of a probability, the relationship between the acoustic signal and the most basic phonetic units of human pronunciation. These are trained with acoustic data of transcribed (labelled) speech, either manually by humans or automatically with a preexisting ASR system (self-learning or pseudo-labeling).

Traditionally, this component is modelled by Hidden Markov Models, $M$, which can be formally defined as follows:

$$M = (Q, \Sigma, \pi, A, B) \tag{2.8}$$

where $Q$ is a finite set of states, $\Sigma$ is a finite set of symbols that can be emitted (also named alphabet), $\pi$ is a vector of initial probabilities, $A$ is a transition probability matrix, which holds the probability of transitioning from state $q$ into any state $q'$, and $B$ is an emission probability matrix, which holds the probabilities of emitting a symbol $x_t$ in state $q_t$ at time instant $t$.

The acoustic models of hybrid ASR systems work over an indefinite number of time windows, which is why they are based on Hidden Markov Models (HMMs). An HMM is a probabilistic model that serves to model time-dependent random processes, such as the particular case of human phonology. Specifically, HMMs are used to model phonemes and even triphonemes (phonemes with context). They are based on the two Markov assumptions: first, the probability that, at instant $t$, the model is in a state $q_t$, only depends on the state $q_{t-1}$. Moreover, the probabilities of transitioning from one state to another are stationary in time, i.e. they are independent of which instant $t$ they occur in. Second, the emission probability of an observation $x_t$ depends only on the state, $q_t$, in which the observation occurs and no other prior observation or emission. That is:

$$P(x_t|q_1, \ldots, q_t, \ldots, q_T, x_1, \ldots, x_t, \ldots, x_T) = P(x_t|q_t) \tag{2.9}$$

Figure 2.5 shows the general structure of an HMM as explained above.

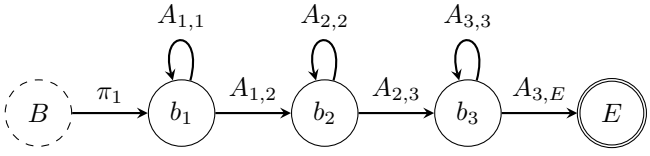


**Figure 2.5:** Representation of an HMM. We can see the three states $b_1$, $b_2$ and $b_3$ that represent the beginning, the central part and the end of the phoneme. The Initial and Final states are also present, which indicate when the recognition of the phonetic unit has started and finished.

Therefore, the probability that the HMM associated with a word $w$ generates the

acoustic sequence $\boldsymbol{x}$ is:

$$P(\boldsymbol{x}|w) = \sum_{q \in Q} \prod_{t=1}^{|q|} P_w(q_t|q_{t-1}) P_w(\boldsymbol{x}_t|q_t) \tag{2.10}$$

Where $P(q_t|q_{t-1})$ is the probability of transiting from state $q_{t-1}$ into state $q_t$, and $P(\boldsymbol{x}_t|q_t)$ is the probability of emission of vector $\boldsymbol{x}_t$ being in state $q_t$. This emission probability has traditionally been modeled with Gaussian Mixtures (GMMs), trained with the Expectation-Maximization (EM) algorithm. HMMs are actually used to model triphonemes, which are concatenated to model words. They do this by using information provided by phonetic pronunciation dictionaries (lexicons). The problem with this approach is that GMMs assume that the data follow a Gaussian distribution, which may not be true, so new approaches were proposed that replace GMMs with neural networks, since they make no assumptions about the nature of the data, which should lead to more accurate and robust models. To achieve this, the Bayes theorem is applied to the emission probabilities of Equation 2.10 as follows:

$$P(\boldsymbol{x}_t|q_t) = \frac{P(x_t)P(q_t|x_t)}{P(q_t)} \approx \frac{P(q_t|x_t)}{P(q_t)} \tag{2.11}$$

This allows us to use the discriminative power of neural networks to model $P(x_t|x_t)$ with a training set, while $P(q_t)$ is obtained by normalising the counts of each state $q_t$ observed during training. $P(x_t)$, although complex to compute, can be ignored as it will be the same value for all cases, since we are maximising $w$ according to Equation 2.2. This approaches proposed to use DNN-HMMs as an acoustic model, being in charge of modelling P(x|w), and improving considerably the results [17]. This was followed by approaches based on recurrent networks [14], especially those using BLSTMs. This is what is known as the hybrid approach, and it is the one that will be used during the present work, with an acoustic model based on BLSTM-HMMs.

### 2.5.2 Language Model

For its part, the Language Model, $P(w)$, is the component of the hybrid ASR system that represents the language structure itself. In other words, it calculates the probability that a word, $w_y$, appears given a prior history $w_1, w_2, \ldots, w_{y-1}$. The fact that language models are trained on monolingual text data (very abundant and easily accessible), allows us to train robust and powerful models. In fact, this is the main reason why it is possible to opt for a hybrid approach and still be able to compete against other state-of-the-art approaches.

There are different ways of dealing with language modelling. Traditionally, the statistical approach based on counts has been used, represented by n-gram models: contiguous sequences of $n$ words which are usually referred to by the number of words (unigram, bigram, trigram, etc.). The n-gram models allow to approximate $P(w)$ as the probability of a word, $w$, given a history $h_i = w_1 w_2 \ldots w_{i-1}$ of previous words;

this can be expressed with the conditional probability as follows:

$$P(w) \approx \prod_{i=1}^{I+1} P(w_i | w_{\max(i-n+1,\ 0)}^{i-1}) \tag{2.12}$$

where $I$ is the size of $w$ and $n$ the order of the n-gram used. $w_i^{i+j}$ represents the sequence of words $(w_i, w_{i+1}, \ldots, w_j) \in w$, and the special states $w_0$ and $w_I + 1$, that represent the tokens for Beginning of sentence (BoS) and End of Sentence (EoF). Following this idea, a third-order n-gram, or trigram, allows the contextualisation of only up to two preceding words, while a 4-gram would allow the use of the three preceding words.

Since the not so distant renaissance of neural networks, approaches based on recurrent networks [6] and transformers [45] have been used, achieving significantly better results than n-grams. In this work, transformer model has been used as a starting point for the adaptation, and different n-gram models have been used in the adaptations.

## 2.6 Hybrid Decoding

Decoding is the stage in which both acoustic and language models are combined to find the most probable word sequence for the given acoustic sample according to the decomposed version of Equation 2.2:

$$\hat{w} = \underset{w \in L^*}{\operatorname{argmax}} \sum_{q \in Q} \prod_{t=1}^{|q|} \left( P_w(q_t | q_{t-1}) \frac{P(q_t | x_t)}{P(q_t)^\beta} \right) \cdot P(w)^\alpha \tag{2.13}$$

where $\alpha$ and $\beta$ are scale factors that help to combine both models.

To do so, a directed multigraph is generated with all the words of the vocabulary, where each node is a phoneme of the word. A simplification of this structure can be seen in Figure 2.6.

It should be noted that the search space is in any case very large, since all words in the system's vocabulary are reachable from any history. Moreover, during decoding, these decisions are made at frame scale. This mechanism compromises the latency of the system and may not be suitable for real-time ASR tasks.

A partial solution is to apply beam-search-like heuristics combined with Viterbi's algorithm, so only a subset of the most probable hypotheses is kept at each instant. But there is another drawback: a discarded hypothesis may actually be more accurate than one that is maintained, since during the search, only the acoustic model's opinion is taken into account, while the language model only contributes its knowledge when the decoding process transitions from a phoneme state to a word state. Traditionally this has been solved by means of two step decoding, but in [23], it is proposed the use of static look-ahead tables, a recursive algorithm that, from the search leafs (all the words in the vocabulary), assigns in each state the maximum probability that can be reached from it, according to the language model. In this way, during the search,

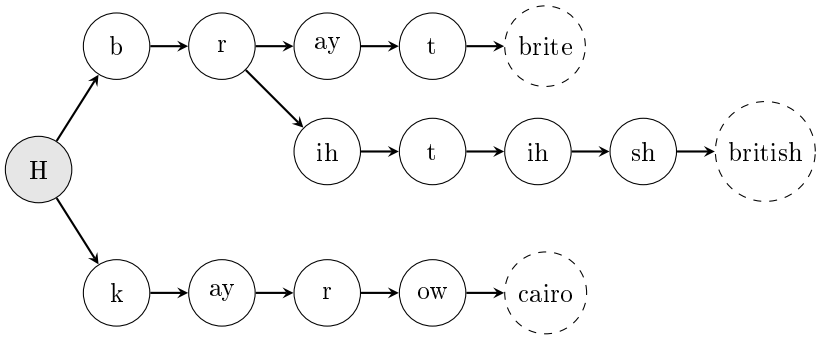


**Figure 2.6:** Example of a search graph capable of recognizing a word. For simplicity, a vocabulary of only three words is considered: "brite", "british" and "cairo". The initial state, $H$, denotes the current history. Specifically, the arrows indicate the exact instant at which the probability of a word, which involves making an explicit query to the LM; e.g. "cairo", is calculated, given the history, $P(cairo|H)$.

when there is a bifurcation, it is only necessary to subtract in the current state the current maximum probability of the LM and add the maximum achievable for each path. In this way, during the search, all hypotheses also have information provided by the language model and the number of queries made to it is drastically reduced. These improvements are implemented in the general purpose system, allowing for streaming decoding.

## 2.7 Evaluation of ASR systems

ASR systems are evaluated objectively by the Word Error Rate (WER), which is defined as the minimum distance between the transcriptions obtained by the system and the original ones, using the data from the development and test. This metric is similar to the Levenshtein distance: it allows word-scale insertions, substitutions and deletions. It is defined as:

$$\text{WER} = \frac{I + D + S}{R} \cdot 100 \tag{2.14}$$

where $I$, $D$, and $S$ are, respectively, the number of insertions, substitutions and deletions needed to make our transcription identical to the original, and $R$ is the total number of words in the reference. The WER can be understood, intuitively, as the percentage of words in the automatic transcription that must be corrected to obtain the correct reference transcription.

There are additional important metrics, as the components of a hybrid system are trained and optimised separately, and therefore require specific metrics. In the case of the acoustic model, they are trained by minimising the Frame Error Rate (FER), which is defined as the number of incorrectly classified frames divided by the total

number of frames analysed:

$$\text{FER} = \frac{F_{\text{incorrect}}}{F_{\text{total}}} \tag{2.15}$$

On the other hand, language models are optimised by minimising the perplexity (PPL) over the development set. We can define the perplexity over a sequence of words $\boldsymbol{w} = w_1...w_n$ as:

$$\text{PPL}(\boldsymbol{w}) = 2^{-\frac{1}{n}logP(\boldsymbol{w})} \tag{2.16}$$

where $-\frac{1}{n}logP(\boldsymbol{w})$ is an approximation of the cross-entropy as far as the sequence of words is reasonably long. The perplexity can be understood, in a simple way, as an estimation of the number of words that can follow our current sequence, so a smaller number implies smaller and faster searches.

## 2.8 Other approaches

Today, large companies and economically well supported research groups that can afford to train on massive amounts of data are leading the way in terms of cutting-edge systems in the world of Natural Language Processing. Several advances have been made in the field of ASR that try to fix different existing subproblems, such as latency, training with few samples or the possibility of multi-language models.

For example, with the introduction of Enformer [39], a new architecture of attention-based acoustic models is proposed, where the context size of the audio samples is reduced to decrease the computational costs of self-attention, and allowing real-time recognition with state-of-the-art results. Other approaches, like wav2vec [37, 2], introduce a contrastive loss function and improvements related to data representation, allowing to distinguish a future audio sample from a false positive, and obtaining good model adaptations with very few hours of acoustic data. There are also large language models, trained with immense amounts of hours, which allow models such as Whisper [34] to acquire large generalisation capabilities in different fields and domains, or like the recent USM [49], which goes for a multilingual model, trained with millions of hours of audio from different languages and then adapted to low-resource languages. Finally, we emphasise the approach of [26], which proposes transcription correction systems based on ideas from the field of machine translation, where we have an encoder in charge of finding errors and a decoder that corrects them.

Chapter 3

# Project description, data and tools

This chapter puts into context the project and data in which this work has been carried out, as well as the tools used to train and adapt acoustic and language the models. First, Section 3.1 describes the Interact-Europe project. In Section 3.2, the dataset used to perform the domain adaptation will be introduced. Finally, Section 3.3 gives details on the software tools used to develop ASR systems.

## 3.1 Interact-Europe

The good teamwork of different professionals from various disciplines in cancer management leads to better patient outcomes, but in many parts of the European Union our professionals may not be well versed in these matters. In fact, their education and training is mostly technically oriented and does not prepare them for this multidisciplinary care. There are some training programmes for cancer care specialties that have achieved a minimum of cooperation and mutual understanding between cancer professionals from different areas, but there is a need to take this cooperation beyond that.

The European project *Interact-Europe* [19] is committed to a patient-centred approach in order to improve the quality of cancer care by promoting multidisciplinary and multiprofessional teamwork. It aims to promote the basis of knowledge and understanding that the different professions involved have of each other, and will address the inequalities linked to cancer. For this reason, different professional organisations associated with the project will develop an inter-specialty training programme targeted at clinical oncology, surgery and radiology, including nursing services.

The project will also lay out the foundations of trainees and cancer centres recruitment to the programme, ensuring a strong understanding of their needs and their readiness for the later delivery of the programme. Elements of the training curriculum will be translated into technology-enhanced learning scenarios, to ensure

wide access to such training tools across the European Union. Finally, communication actions will ensure broad awareness of the programme to allow uptake of the project's recommendations by the cancer care community. The inclusion of representatives from the whole European cancer community in the consortium will allow for the development of a targeted and efficient multidisciplinary programme and production of recommendations to foster interdisciplinary cancer care across the European Union.

As the main idea of the project is to create an international curriculum for oncology experts in the European Union, specialists from different hospitals would agree on a number of subjects to be addressed and a series of videoconferences would be held on these topics. These videoconferences and symposiums need to be adapted to different territories and languages, and although the project's vehicular language is English, not all professionals are native speakers of this language or are fluent in it. For this reason, it is necessary to carry out automatic speech recognition, as well as its subsequent translation into different European languages. Automatic speech recognition adapted to the oncology domain and automatic translation of the transcriptions into other European languages are the responsibilities of the MLLP research group in this project.

## 3.2 e-ESO dataset

One of the project partners, the European School of Oncology [1] (ESO), has a videoconferencing platform on oncology. It has provided the consortium with a representative set of these lectures, some of them with manual transcripts and/or slides. Table 1 shows the statistics that summarise the data provided.

The shared repository had a total of 234 e-sessions (videos), with a length of approximately 176h of video, dealing with more or less technical oncology topics.

An analysis of the downloaded data showed that many of the videos did not include manual transcriptions. This can be seen in Table 3.1, which shows a detailed explanation of the available data.

**Table 3.1:** Complete distribution of data. "V" stands for videos, "S" for Slides and "T" for manual transcriptions. A sum indicates that both data are available, so "V + S" denotes videos that have slides.

| | Videos | Hours |
|---|---|---|
| V | 140 (60%) | 94.9 (54%) |
| V + S | 1 (0%) | 0.8 (1%) |
| V + T | 56 (23%) | 41.8 (23%) |
| V + T + S | 47 (17%) | 38.5 (22%) |
| Total | 234 (100%) | 176 (100%) |

Based on these data, we have generated two in-domain corpora: one of speech data for acoustic model adaptation, and the other of textual data for language model

[1] https://www.eso.net/

adaptation. On the one hand, the speech corpus comprises the whole set of the 103 videos with manual transcriptions; while the textual corpus uses the manual transcriptions extracted from the 103 videos, as well as from the available slides of 40 videos. Three possible sets will be considered for the textual dataset: transcripts, slides and transcripts + slides.

Therefore, the final dataset is made up of 73h of video for Train, and two sets with 3.5h and 3.8h of speech for Dev and Test, respectively. It is important to emphasize that not all the data used have slides, obtaining a representative sample of what could be found in a real-life scenario. In both Dev and Test sets there is only one session without slides. An analysis of the transcripted text has been carried out. Table 3.2 summarizes basic raw statistics of the speech dataset, with which the AM will be adapted: first, the number of videos in each set, and the total duration of these videos, in hours. The size of vocabulary is the number of different words. Finally, the running words (RW), are the total number of words in the reference transcription. Equivalent information for the text corpus, used to train the adapted LM, can be found in Table 3.3.

**Table 3.2:** Raw statistics of the eESO speech corpus. $|V|$ stands for vocabulary size, while RW are the Running Words.

| | # Sessions | Duration (h) | \|V\| | RW |
|---|---|---|---|---|
| Train | 93 | 73.0 | 30.0K | 2.2M |
| Dev | 5 | 3.5 | 2.8K | 27.4K |
| Test | 5 | 3.8 | 3.3K | 31.2K |

**Table 3.3:** Raw statistics of the eESO text subcorpus, with useful information for Transcriptions (T), Slides (S) and the sum of both resources (T+S). $|V|$ stands for vocabulary size, while RW are the Running Words.

| | T | | S | | T+S | |
|---|---|---|---|---|---|---|
| | \|V\| | RW | \|V\| | RW | \|V\| | RW |
| Train | 30.0K | 2.2M | 20.2K | 0.4M | 37.4K | 2.6M |
| Dev | 2.8K | 27.4K | 2.4K | 10.1K | 2.8K | 37.5K |
| Test | 3.3K | 31.2K | 2.6K | 21.1K | 3.3K | 52.3K |

It is important to emphasise that this is an adaptation to the oncologic domain and that the samples have a high variance in terms of quality and complexity: there are many different speakers in the dataset, as well as videos with more than one speaker and spontaneous speech. Besides that, most of the recordings use low-quality microphones and have complicated sound conditions (reverberation, echo, etc.), making it a potentially complex dataset.

## 3.3 Tools

For the construction and development of the models and systems of this work, different software tools have been used, as well as multiple scripts to automate the tasks performed. Among them, we highlight:

Firstly, the TransLectures-UPV toolkit [8] (TLK) is a toolkit that provides a set of tools and libraries for the development of hybrid ASR systems. This software has been developed by the MLLP research group of the UPV, initially for the TransLectures [40] project, and has evolved over time to add the different advances in hybrid ASR technology. Among others, it has been used to preprocess the acoustic data and extract the features from the data. It has been used to perform the data alignments, as well as to create the static look-ahead tables that will be used later in the decoding, where the tool is also used.

Secondly, TensorFlow [41] is an open source library developed by the Google Brain team for deep and machine learning tasks. The internal way of working is simple, as it uses high-level directives in the form of order graphs to represent complex algorithms and architectures. Throughout the work, TensorFlow has been used to train BLSTM-HMM acoustic models, as TLK works on top of TensorFlow when working with neural networks.

As for KenML [16], it is an open-source toolkit developed by Kenneth Heafield and consists of a set of compiled tools for ASR, which have been used to train n-gram models of different order and to calculate the perplexities of the developed language models.

Finally, Fairseq [12] is a sequence modelling toolkit, developed by Facebook AI Research, that allows working with different problems related to speech technologies. In our case, it can be used for the construction of transformer language models, such as the one in the system to be adapted.

Chapter 4

# Baseline System

This chapter gives details on the MLLP's English general-purpose ASR system, built under the framework of previous research projects. It will be considered the baseline in all experiments, allowing us to assess the contribution of the different adaptation techniques in this work.

## 4.1 Baseline ASR System

The baseline system is a general-purpose hybrid English ASR system whose acoustic model has a BLSTM-HMM structure, while the language model is a transformer model. The BLSTM that constitutes the AM is composed of eight bidirectional hidden layers of 512 neurons per layer and direction, as well as a bottleneck layer with 200 neurons before the output layer. It uses a 50-frame window to work with back-propagation through time. In addition, SpecAugment, a data augmentation technique based on masking frequencies to reduce overfitting, has been used. This model has been trained with approximately 6000 hours of acoustic data 4.1. The training of the AM has been performed using cross-entropy, as well as frame-level alignments, with a total of 16K senones.

The Transformer LM uses 24 layers, which, like the embedding, has 768 neurons each, 12 attention heads are used and, finally, an FFN layer of 4096 neurons. A 4-gram model was trained, but was never used in production; despite this, it was used to build a pruned version of itself, used as static look-ahead tables during decoding, as well as in the adaptation techniques proposed in this work. The n-gram model was trained on a set of texts from different sources 4.2, with a sum of 1.2G sentences for a total of 17.9G words. The Transformer LM was trained on a subset of the previous one, with 1G words. A remarkable feature is that thanks to the work done in [22], the LM transformer can operate in streaming with very little loss of quality, which allows the ASR system to do so as well.

Taking into account the baseline ASR system, an analysis of the eESO corpus, explained in detail in Section 3.2, has been carried out in order to understand the potential margin of improvement that a domain adaptation could offer, as well as to

**Table 4.1:** Transcribed out-domain English acoustic corpora used to train out-domain English acoustic models.

| Corpus | Length (hours) |
|---|---|
| Internal: user-generated content | 1,538 |
| Internal: parliamentary | 1,008 |
| Librispeech[32] | 960 |
| Internal: TV and entertainment | 515 |
| Must-C[9] | 503 |
| TED-Lium[35] | 369 |
| How2[36] | 301 |
| Internal: educational | 218 |
| SWC[4] | 128 |
| VoxForge[44] | 120 |
| CHIME[3] | 111 |
| AMI[5] | 96 |
| Europarl-ST[20] | 91 |
| VCTK[47] | 44 |
| ELFA[10] | 38 |
| **OVERALL** | **6,040** |

**Table 4.2:** Out-domain English text corpora used to train out-domain English language models, where K=$10^3$, M=$10^6$ and G=$10^9$.

| Corpus | Sentences | Words |
|---|---|---|
| Internal: news | 816.7 M | 12.2 G |
| Wikipedia[11] | 149.9 M | 2.3 G |
| Opensubtitles[27] | 191.4 M | 1.1 G |
| Librispeech[32] | 40.7 M | 813.0 M |
| Giga[42] | 22.5 M | 616.8 M |
| Internal: mixed data | 22.5 M | 314.8 M |
| UN[50] | 11.4M | 308.3 M |
| Internal: educational | 5.0 M | 98.6 M |
| Internal: parliamentary | 3.8 M | 68.2 M |
| Europarl[25] | 1.3 M | 33.5 M |
| News Commentary[42] | 532.5 K | 11.6 M |
| **OVERALL** | **1.3 G** | **17.9 G** |

estimate the deficiencies of the model. Table 4.3 provides an overview of the results obtained from this analysis of the acoustic dataset (manual transcriptions of audio data): first, the running words (RW) and the size of vocabulary are remembered, as the remaining metrics depend on them. Then we have the *out of vocabulary* (OOV) words, that is, the number of unique words in each set that are not in the vocabulary of the baseline system. Finally, the *running OOV*, ROOVs, are the total occurrences

of OOV words.

**Table 4.3:** Statistics of the eESO acoustic dataset. $|V|$ stands for vocabulary size, RW are the Running Words, and OOV are the Out-of-vocabulary words.

| | \|V\| | OOVs | RW | ROOVs |
|---|---|---|---|---|
| Train | 30K | 8.6K (28%) | 2.15M | 33.5K (2%) |
| Dev | 2.8K | 0.2K ( 7%) | 27.4K | 0.3K (1%) |
| Test | 3.3K | 0.2K ( 6%) | 31.2K | 0.5K (2%) |

After this, the corresponding evaluation of the language model of the baseline system has been carried out: First, a calculation of the perplexities of the Transformer LM on the Dev and Test sets has been performed, where perplexities of 143 and 140, respectively, have been obtained. These values will serve as a reference point as to whether or not the language models that will be trained should be improved, where a certain correlation with the WER of each system is expected.

Subsequently, the evaluation of the system was performed on both sets, obtaining a WER of 18.1% in development and 16.1% in test. Our goal will be to improve these results by applying domain adaptation techniques to the baseline AM and LMs using the in-domain data described in Section 3.2. Our goal will be, therefore, to improve these results with the different hybrid models trained with the selected data.

Chapter 5

# Adaptation Techniques

As discussed in Section 1, adaptation to the oncology domain of the MLLP's general-purpose English ASR System is key for the success of the Interact-Europe project. For this purpose, different techniques are explored to adapt both the acoustic and language model. In addition, the possibility of adapting both models at video level has also been explored. This chapter describes the different procedures and techniques used to adapt the acoustic and language models that compose a hybrid ASR system. Section 5.1 describes the different techniques used to adapt the Acoustic Model, while the equivalent with the language model is explained in Section 5.2. Section 5.3 discusses video-level adaptation techniques for both the acoustic and language models.

## 5.1 Acoustic Model Adaptation Techniques

This section describes the different techniques and approaches used to perform the domain adaptation of the Acoustic Model, $P(q_t|x_t)$ (Eq. 2.11).

### 5.1.1 Fine-tuning

Fine-tuning is a transfer-learning technique that consists of using a previously trained model, with out-of-domain data, to continue training it but now with data from a particular domain. The idea is that the fine-tuned model loses some generalisation in exchange for being better at the new task. In practice, this consists of performing a similar training to the previous one, except that the initial weights of the model will not be random, but the weights of the baseline acoustic model. One of the particularities of this training is that, as the weights are already initialised and we only want to make small adjustments, a small learning rate has to be used, a couple of orders of magnitude smaller than in training.

### 5.1.2 L2 Regularisation

It consists on adding L2 regularisation to the loss function. As the aim of fine-tuning is not to deviate too much from the weights of the baseline acoustic model, the L2 distance is calculated on the initial weights of the pre-trained model. This is intended to accept only variations that have a high positive impact on the accuracy of the network, without needing to vary the weights too much, which already keep a good interpretation of the data and their relationships.

### 5.1.3 Linear Transformation

Motivated by the idea of adapting networks through transformations [31] and taking advantage of the existence of a bottleneck layer, it is proposed to learn a 200·200 linear transformation on the bottleneck. This training involves learning very few additional parameters. Therefore, we propose to explore the use of a linear transformation and to use a non-linear transformation, i.e. to add a non-linear ReLU transformation on the output of the linear transformation. This idea is outlined in Figure 5.1.

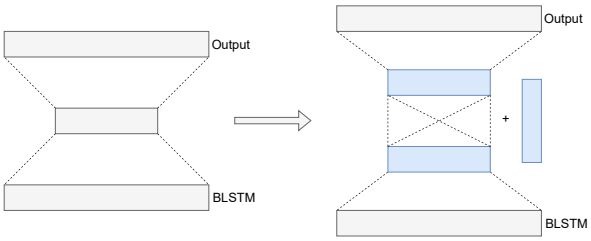


**Figure 5.1:** Before and after of the same network with a linear transformation added at the end.

### 5.1.4 Output layer training

We also studied the effect of training only the output layer. As few acoustic data were available and a fine-tuning is performed on the whole model, this could result in a model with new weights that are well adapted to the training data, but do not generalise well enough. To avoid this overfitting, we freeze all those layers already containing a good representation of the English language and only train the layer that classifies into the different senonas. In short, the network is trained with the domain-exclusive data, but only updating the weights of the output layer at each iteration.

## 5.2 Language Model Adaptation Techniques

Moreover, this section explains the techniques used to adapt the Language Model, $P(w)$ (Eq. 2.2). The language model [29] is transformer model, but there is also a

4-gram model that never went into production. This, together with the larger amount of training data available, offers a wide range of possibilities to adapt the LM. We focused on exploring n-gram adaptation techniques, that is, we worked on adapting the 4-gram of the baseline system, which was not used in production, and then we interpolated it with the TLM of the baseline system. Adaptation of the TLM has been contemplated but not implemented due to limited in-domain data.

For each of the experiments considered, the following work was done: first, a unigram model was trained in order to limit the size of the vocabulary in case it was too large. The vocabulary was small enough to not require modification, being 29.9K unique words in the case of using the manual transcriptions, 20.0K when using the slides and 37.4K unique words when using both sources combined.

Next, the 4-gram models have been trained at subset level (transcriptions, slides and trans+slides), performing a small pruning of the tokens that do not appear at least once. Finally, super-pruned versions of these models have been generated, so they can be used as static look-ahead tables during the decoding process. The trained n-gram models have been individually interpolated with the much larger and more robust n-gram model of the baseline system, resulting in three domain-adapted n-gram models. This interpolation, as well as the following ones, has been performed using an implementation of the MLLP group of the Maximum-Expectation (EM) algorithm. Finally, the new 4-gram model is interpolated with the TLM model of the baseline system.

## 5.3 Adaptation Techniques at video level

An interesting approach is to make an adaptation at video level. This approach relies on using only the data associated with a sample (video), instead of the whole domain dataset. An attractive aspect of these techniques is that, whilst they can be used as an alternative to domain adaptation, they can also be used as a complement, as we will see below.

In order to adapt the language model to the video level, it would be interesting to use automatic transcriptions, but it is true that in a real use case we would not have them, so we propose to use only the slides, which could be available beforehand even in the case of streaming scenarios. With this premise, we train small n-gram models, one for each video to be recognised, using the slides of the same video. We then interpolate them with the baseline model transformer.

To work with the acoustic model, it is necessary to have previous transcriptions, being this technique especially useful in the case of offline speech recognition. It consists of fine-tuning the acoustic model, as explained above, but using only the acoustic data from the video itself. This may be reminiscent of MLLR adaptation techniques used in Gaussian-based models [13]. To complete the hybrid system, the language model of the system with which the transcriptions have been obtained is normally used.

These two techniques can be perfectly combined by first training an adapted language model. With it, we would obtain the necessary transcriptions and alignment

to be able to train the adapted acoustic model. Finally, both models, adapted to the same video, would be combined to obtain the final system.

Chapter 6

# Experimentation and evaluation

This chapter presents technical and experimental aspects of the adaptation and optimisation of the English ASR System presented in Section 4, as well as the results and metrics obtained in each step. Section 6.1 describes the preprocessing carried out on the data. Section 6.2 provides details on the construction and optimisation of a hybrid ASR system. Section 6.3 details the training and experiments carried out to adapt the baseline system to the oncology domain. Finally, Section 6.4 explains the experiments carried out on video-level adaptation.

## 6.1 Data preprocessing

Section 3.2 described the data used to adapt and optimise the models developed throughout this work. However, they need to be homogenised and normalised before they can be used. The steps followed to achieve this preprocessing are conceptually described in Section 2.4.

First of all, the acoustic data have been preprocessed, which require specific characteristics: to be in 16-bit little endian WAV format, single-channel and with a sample rate of 16 KHz. *Filterbank* feature vectors are then obtained, which are banks of 85 filters, without any derivates, resulting in 85-component feature vectors that will be used to retrain the BLSTM models. Normally, this is where the preprocessing of the acoustic data finishes, however, in our case an average normalisation is applied, not at the sample level, but by applying a sliding window. This allows the system to be used in streaming if necessary [22].

The next step is to preprocess the existing manual transcripts and slides. Firstly, the text from the presentations and manual transcripts was extracted to have it in plain text. Secondly, we removed all punctuation marks, and converted the text to lower case and numbers to text, among others.

## 6.2 Construction and optimization of hybrid ASR systems

Having trained acoustic and language models, we integrate them under a single system that performs the hybrid decoding. In our case, the systems will be finally constituted by executing the TLK decoder through the `tLtask-recognise` tool, which receives as arguments the acoustic and language models to be used, the corresponding static lookahead tables and the feature vectors of the samples to be recognized.

This model combination is governed by a set of hyperparameters, whose values are typically those that minimize WER over the in-domain development set. Some critical parameters are: First, the **Grammar Scale Factor (GSF)**, which scales the weight of the language model with respect to the acoustic model during decoding, and the **Prior Scale Factor (PSF)**, which weights the *a priori* probabilities, $P(q_t)$ 2.11, of the HMMs. In practice, the effect of GSF and PSF has been studied, as they have a high impact on the final system and are highly dependent on the models involved. Finally, remember that these hyperparameters are always optimised on the Development set, applying the best values obtained in Test.

## 6.3 Assessment of domain adaptation techniques

This section explores the different domain adaptation techniques outlined in Sections 5.1 and Section 5.2, which aim to train different acoustic and language models and then combine them to obtain a hybrid system adapted to the oncology domain.

### 6.3.1 Acoustic Model Adaptation

In this section we study the effects of the different AM adaptation techniques described in Section 5.1, on the performance of hybrid ASR systems that combine the resulting adapted AMs with the baseline LMs, both sourced from the baseline ASR system described in Chapter 4. As adaptation data, we used the e-ESO speech corpus described in Section 3.2. But, before starting with the training of acoustic models, it is necessary to align the data that will be used to provide the BLSTM model. This is achieved by means of the `tLtask-align` tool which, for each training sample, generates a file with the information on which senones are associated with each frame (feature vector). The baseline model explained above has been used to perform the alignment, this will allow us to obtain good results despite having little data available.

#### 6.3.1.1 From scratch

First we perform a control training, where we train a model from scratch with the same topology as the original, but only with the data available for adaptation, the 73h of our training set. The model is made up of 8 BLSTM layers, each with 1024 neurons (512 for each temporal sense), a bottleneck layer of 200 neurons and finally the output layer, with 16132 neurons, as many as we have senones. This first training

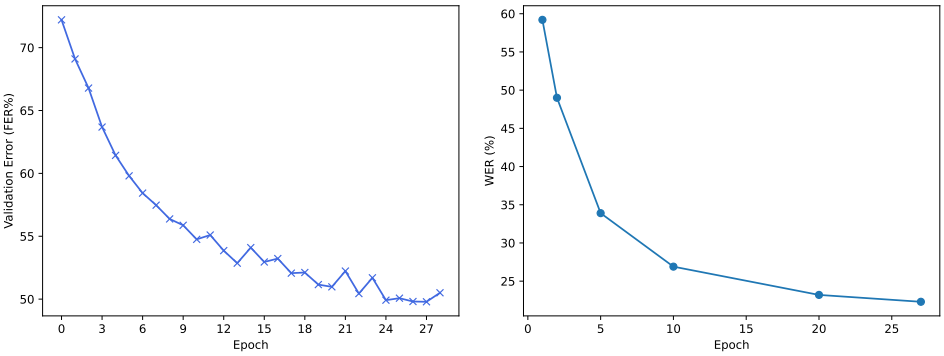


**Figure 6.1:** FER and WER computed on the development set in function of the epoch number for the system composed of the in-domain acoustic model trained from scratch and the language model.

took around 46 hours using a GeForce GTX 1080Ti which, over 29 epochs, obtained a minimum validation error (FER) of 49.8% over the development set. The best model was obtained by using values of GSF=10 and PSF=0.6, giving a WER of 22.3% in Dev set. Figure 6.1 shows the evolution of the FER and WER as a function of the epochs.

#### 6.3.1.2 Fine-tuning

Then, we performed a fine-tuning of the baseline acoustic model, so again we repeated the same network topology, but we used a small learning rate of 2.25e-5. The training was performed in parallel to the previous one, on a similar machine having the same graphics card; it was decided to pause it after 46h, having completed 36 epochs and achieving a minimum validation error of 36.7% in epoch 8, as seen in the Figure 6.3. A very oscillating pattern is observed, although it seems to converge over time. This may be due to the value of the learning rate, which could be too high. In spite of this, quite adequate values have been obtained when compared with the rest of the techniques studied. This approach achieves a minimum WER in Development of 15.4% in Development, with the hyperparameters set to 10 for GSF and 0.8 for PSF.

#### 6.3.1.3 L2 Regularisation

Following this, and as an approach to refine the results of the previous fine-tuned model, more models were trained but applying L2 regularisation over the initial weights of the AM. As explained in Section 2.3, the L2 regularisation is weighted by an $\lambda$ factor, which has a strong impact on the result. For this reason, the acoustic model's behaviour has been studied by varying this value, carrying out training for the values of $\lambda = \{0.01, 0.1, 0.5, 1.0, 10.0\}$. For the sake of brevity, the pink curve of

Figure 6.3 shows the variation of the FER for the value of $\lambda = 0.1$, the rest curves of this experiments have a similar behaviour. It can be seen that the minimum value obtained is 36.2% at epoch 5. During training, in addition to having oscillating FER values, they do not seem to converge to any result, perhaps due to the small amount of acoustic data used, which also have a series of complex acoustic conditions. Even so, the results obtained during decoding were 14.7% WER for the dev set, with a GSF of 10 and a PSF of 0.8. In Figure 6.2 we can observe the variation of the WER as a function of the epochs when using different values of lambda during regularisation.

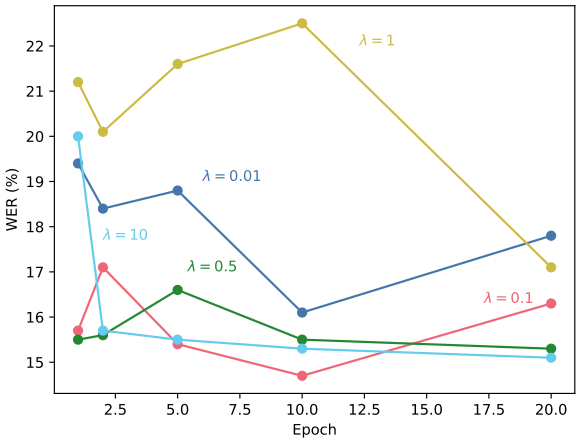


**Figure 6.2:** We found the WER as a function of epochs for each system resulting from combining the LM of the Baseline System with each of the trained AMs with different values of lambda.

#### 6.3.1.4 Linear Transformation

Thereafter, we start experimenting with the models resulting from freezing parts of the original network, starting with those that replace the bottleneck layer with a linear transformation. With this idea in mind, two acoustic models have been trained, the first one using the linear transformation described above, the second one being the same as the previous one but but adding a non-linear ReLU transformation to the output of the linear transformation. The yellow curve in Figure 6.3 show the evolution of the FER for the model with the linear transformation, while the maroon one shows the same for the model that adds the ReLU one. It can be seen that the ReLU activation function achieves a smoother convergence, but converges to worse quality values. The best results achieved were a validation error of 39.6% at epoch 32 for the first experiment, while the experiment using the ReLU obtained 50.8% at epoch 23. In both cases the hyperparameters that configured the best system were GSF=8 and PSF=0.8. Combining the linear transformation model with the LM of the baseline system yielded a WER of 18.1% in Dev, while the system using the ReLU

transformation obtained WERs of 19.4%.

#### 6.3.1.5 Output Layer

Finally, we experimented with the model trained by freezing the entire network except the output layer; throughout the section and for simplicity, we will refer to it as the Output Layer Model (OLM). The results of the training can be seen in the green curve of Figure 6.3. Although it takes several iterations, the model seems to converge, but a lot of oscillation is observed, in future experiments it would be interesting to try other learning rate values. However, this results in a minimum FER value of 37.5% in the 40th epoch. During decoding, the best model has been achieved for a value of 12 in GSF and 0.6 for PSF, with a final WER of 15.3% in Development. The evolution of the WER in terms of the epochs can be seen in the Figure 6.4, as well as for the other models trained.

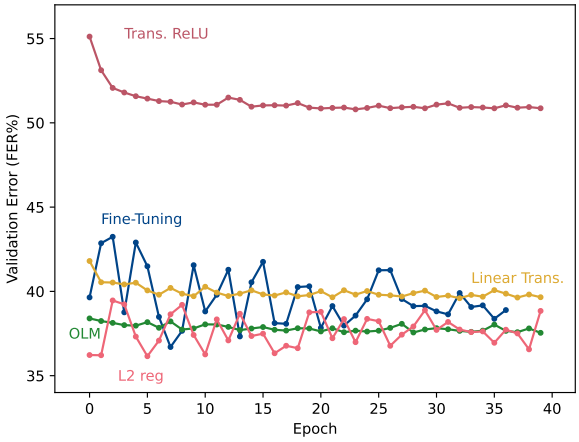


**Figure 6.3:** FER obtained per epoch in the development set when training the different models.

### 6.3.2 Summary and results on the test set

The different results obtained from the experiments carried out in this section can be found in Table 6.1, including the results in the Test set for each of the approaches studied. As can be seen, approximations based on adding a linear transformation do not seem to obtain good results, as well as training from scratch, probably because of the limited data available (remember that the acoustic model of the Baseline system has been trained with 6000h of audio).

Regarding the fine-tuned model, it has obtained very good results, 2.9 absolute points in Dev with respect to the Baseline system (15.8% of relative improvement). Finally, the techniques of regularisation and training the last layer only, have proved

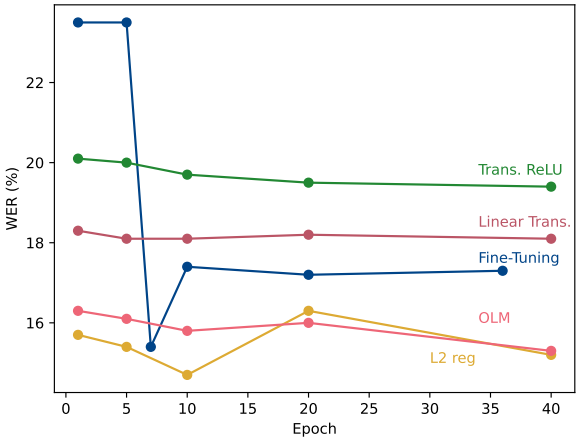


**Figure 6.4:** We found the WER as a function of the epochs for each best system resulting from applying the best acoustic model obtained with each of the fitting techniques and the LM of the Baseline System.

to be the best for this case, having a very similar performance. Although in Test the OLM has obtained better results than the L2 Regularisation, we will keep the latter, based on the results obtained in the development set. With the L2 regularisation model, relative improvements in the 6.2-19.7% range have been obtained, followed closely by the OLM, which has obtained relative improvements between 12.4-16.4%.

**Table 6.1:** Results of combining the trained acoustic model with the baseline LM.

| | eESO | |
|---|---|---|
| | Dev | Test |
| Baseline | 18.3 | 16.1 |
| From scratch | 22.3 | 22.8 |
| Fine-tuning | 15.4 | 15.5 |
| L2 Regularization | **14.7** | **15.1** |
| Linear Transformation | 18.1 | 15.7 |
| Linear Transfor. ReLU | 19.4 | 17.4 |
| OLM | 15.3 | **14.1** |

### 6.3.3 Language Model Adaptation

This section is devoted to studying how different language models trained with the Baseline System acoustic model behave. As explained in Section 2.7, when comparing language models, perplexity is a important measure of how good a system could be, which is why the perplexity of the baseline system will be used as reference in the

rest of the experiments.

We intend to perform the adaptation of the language model and two main data sources are available: manual transcriptions and slides. This allows different experiments to be carried out, observing how much the different data sets can contribute. This way, three different trainings of the N-gram model have been carried out: the first using the transcriptions, the second using the slides and the third combining both sets.

After training the 4-gram models as explained in Section 5.2, their perplexities have been calculated, to get a sense of how much they should contribute to the final system. Each of the developed models was then combined with the Baseline System Acoustic Model and the decoding was performed. In this section it has only been necessary to explore the GSF, as the PSF hyperparameter affects the acoustic model, which, being from the Baseline System, was already optimised when it was created. After conducting the experiments, the best system was the one resulting from combining the language model trained with transcripts interpolated with the TLM, with results of 16.4% in Dev and 13.7% in Test, but closely followed by the LM trained with transcripts and slides interpolated with the TLM, whose results were 16.5% in Dev and 14.0% in Test. Table 6.2 summarises the perplexities of each interpolated model trained in this section for both Dev and Test sets, as well as the resulting WER when combined with the AM of the Baseline System. Recall that the LM of the baseline system consists of a transformer model only.

**Table 6.2:** Perplexity of the different trained language models, as well as the results when combined with the AM of the Baseline System.

| | Dev | | Test | |
|---|---|---|---|---|
| | PPL | WER | PPL | WER |
| Baseline (OOD TLM) | 142 | 18.3 | 140 | 16.1 |
| $\text{ID}_{\text{transc}}$ Ng | 136 | 18.5 | 139 | 15.9 |
| + OOD TLM | 94 | **16.4** | 94 | **13.7** |
| $\text{ID}_{\text{slides}}$ Ng | 233 | 20.2 | 246 | 16.8 |
| + OOD TLM | 116 | 17.3 | 117 | 14.4 |
| $\text{ID}_{\text{all}}$ Ng | 141 | 18.1 | 143 | 15.3 |
| + OOD TLM | 95 | 16.5 | 94 | 14.0 |

It can be seen that the language models that are an interpolation are much more accurate than those that only use count models, despite the fact that the latter do not have bad performance. Analysing the results, it seems that the slide data contribute with noise rather than help, so perhaps a different pre-processing of the data should be carried out, or this data should be omitted.

### 6.3.4 Combination of adapted acoustic and language models

After carrying out these experiments, the combinations of the adapted acoustic model using L2 regularisation with the different adapted language models have been carried

out, thus obtaining different final models adapted to the domain, from which the best one will be chosen.

When performing the combination, it is necessary to re-optimise the hyperparameters; after this, the best results were obtained for a PSF of 0.8 for all the models, and a GSF of 12 for the model trained with transcriptions and its interpolation with the TLM; for the rest of the models, the best GSF value was 14. Table 6.3 shows the final results obtained for each system studied. The quality of the systems resulting from combining both adapted models has been much better than compared to the Baseline System, being the best the one resulting from combining the adapted AM with the LM trained with transcriptions only interpolated with the TLM. This system has obtained 4.8 points of absolute improvement over the Baseline System in Dev and 3.1 in Test.

**Table 6.3:** Results of combining the trained acoustic model with L2 regularisation with the different language models generated.

| | eESO | |
|---|---|---|
| | Dev | Test |
| Baseline (OOD TLM) | 18.3 | 16.1 |
| $ID_{transc}$ Ng | 15.5 | 14.9 |
| + OOD TLM | **13.5** | **13.0** |
| $ID_{slides}$ Ng | 16.1 | 15.1 |
| + OOD TLM | 14.0 | 13.4 |
| $ID_{all}$ Ng | 15.1 | 14.4 |
| + OOD TLM | 13.7 | 13.2 |

## 6.4 Video Level Adaptation

Now we have an adapted system that improves the performance of the original in a specific domain. We then explored the possibilities offered by the adaptation of this new system at the video level. In this section we will use the best system resulting from combining the best adapted models from Section 5.1 and Section 5.2. We have used the acoustic model with L2 regularisation and transformer interpolation and the 4-gram model trained with transcriptions as the language model; from now on we will refer to it as the Adapted System. As the objective is video-level adaptation, only the Development and Test sets have been used in these experiments, to conduct experimentation and to confirm or refute the hypotheses, respectively.

### 6.4.1 Language Model video level Adaptation

We will now detail the steps taken to build and evaluate the adapted language models at the video level using their slides as training data. First, we trained an n-gram

model for each video with their own slides[1] as training data. The procedure is similar to the one followed in Section 6.3.3, but with small changes: Both alignment and training have been performed at the video level, in order to obtain the initial 4-gram models. These models have been interpolated, individually, with the 4-gram model of the Adapted System, and finally, with the transformer of the baseline system, as it has proved to be the best combination. To obtain this final interpolation with the Transformer LM, we have used the EM algorithm again with the perplexities of the development videos to obtain the "average" weights to be used by the interpolation, these weights have been 0.25 for the n-gram models and 0.75 for the TLM.; for the test set these weights have been used directly, as for any future video with slides.

In order to make the fairest comparison possible with respect to the previous systems, the videos that do not have slides have been recognised using the Adapted System, as would occur in a real use case when there are no slides for adaptation. With this consideration, we combined the language models with the acoustic model of the Adapted System, and performed the recognition on each individual video, obtaining WER values of 13.5% in the Dev set and 12.9% in the Test set. If we compare this results with the ones in Table 6.3, no significant improvement is observed

### 6.4.2 Acoustic Model video level Adaptation

We continue with the experiments at the video level, now with the acoustic model. To do this, we repeated the same procedure as for the domain adaptation, but we only used the fine-tuning technique with L2 regularisation and its own transcriptions, generating a unique acoustic model for each video treated. The resulting acoustic models were combined with the Adapted System Language Model and then the decoding was performed, obtaining a WER in development of 12.1, and in test of 11.9, dropping one absolute point from the previous results.

Finally, both models trained at the video level were combined, as described in Section 5.3. Recall again, that there is one video in each set that has no slides, so it could not be adapted at the video level by this procedure; the strategy of recognising the remaining video with the Adapted System has been followed again. In this way, WERs of 12.0% in Development and 11.9% in Test were achieved, being, casually, the same as in the case of the adaptation of the acoustic model to the video level. The relative improvement over the adapted system was in the range of 7.8-10.4%, while the improvement respect the baseline system was between 32.8% and 38.6%. As there is a dependence on the results obtained from the model that is adapted at the slide level, and this does not obtain great results, it was expected that these combined systems would not outperform the others.

We conclude that the adapted system has obtained very competitive results and that, unlike the Video Level Adapted System, it can be used in a streaming setup without losing much performance.

[1] Note that there is a video in each set of Dev and Test that has no slides available, so no video-level model has been trained for these cases.

# Chapter 7
# Conclusions

In this work we have discussed possible domain adaptation techniques for hybrid ASR systems in the framework of the Interact-Europe project. To achieve this, advanced and highly specialised software tools have been used, as well as a small training set of 73h of oncological content, manually transcribed, with a total of 2.2 million words. The trained acoustic and language models have been combined with each other, forming hybrid ASR systems adapted to the domain. Furthermore, system adaptations have been made at the video level based on the previously adapted system, further improving performance.

With regard to the adaptation techniques used on the acoustic model, we can conclude that both the use of L2 regularisation on the initial weights and the training of only the last layer of the network have allowed us to deal satisfactorily with the overfitting that usually occurs with this type of training. With respect to the adapted language models, it is important to emphasise the importance of the quality of the data used, since, as has been observed, better results are not always obtained by having more task-related data, as these may introduce noise into the training.

The best adapted system built, the one which combines the trained acoustic model with fine-tuning and L2 regularisation and the language model resulting from interpolating the LM Transformer and the 4-gram model trained with the transcriptions, has obtained an absolute improvement over the Baseline System of 4.8 points in Dev and 3.1 in Test, that is, between 19.3% and 26.2% relative improvement. These are good results, and even more so if we remember that it is a system suitable for streaming environments.

As far as video adaptation is concerned, it has achieved very good relative improvements of between 32.8% and 38.6% over the baseline system, showing an improvement over the adapted system of 7.8-10.4%, which are satisfactory results.

This work has left the way clear for much future work:

- Variations of AM training based on freezing and thawing parts of the net with training in between.
- Adaptation of the LM Transformer, where techniques similar to those used in

acoustic BLSTM could be tested.

- Use more data, opening the door to being able to use unsupervised data and explore the impact it would have.
- To carry out more experiments with alternative preprocessing, as well as with other domains, in order to corroborate the results.
- Study the impact of adaptation at the video level by applying the baseline model directly.

# Bibliography


[1] Alexei Baevski and Michael Auli. *Adaptive Input Representations for Neural Language Modeling*. 2019. arXiv: 1809.10853 [cs.CL].

[2] Alexei Baevski et al. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL].

[3] Jon Barker et al. "The third CHiME speech separation and recognition challenge: Dataset, task and baselines". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 2015, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837.

[4] Timo Baumann, Arne Köhn, and Felix Hennig. "The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening". In: *Language Resources and Evaluation* 53.2 (June 2019), pp. 303–329. ISSN: 1574-0218. DOI: 10.1007/s10579-017-9410-y. URL: https://doi.org/10.1007/s10579-017-9410-y.

[5] Jean Carletta. "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus". In: *Language Resources and Evaluation* 41.2 (May 2007), pp. 181–190. ISSN: 1572-8412. DOI: 10.1007/s10579-007-9040-x. URL: https://doi.org/10.1007/s10579-007-9040-x.

[6] Thomas Cherian, Akshay Badola, and Vineet Padmanabhan. *Multi-cell LSTM Based Neural Language Model*. 2018. arXiv: 1811.06477 [cs.NE].

[7] Jan Chorowski et al. "Attention-Based Models for Speech Recognition". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 577–585.

[8] M. A. del-Agua et al. "The Translectures-UPV Toolkit". In: *Advances in Speech and Language Technologies for Iberian Languages*. Ed. by Juan Luis Navarro Mesa et al. Cham: Springer International Publishing, 2014, pp. 269–278. ISBN: 978-3-319-13623-3.

[9] Mattia A. Di Gangi et al. "MuST-C: a Multilingual Speech Translation Corpus". In: *Proc. of NAACL-HLT*. 2019, pp. 2012–2017.

[10] *English as a Lingua Franca in Academic Settings (ELFA Corpus)*. Website: https://www.kielipankki.fi/corpora/elfa/?path=Corpusstructure/ELFA.imdi.

[11] *English Wikipedia)*. Website: https://dumps.wikimedia.org/backup-index.html.

[12] Facebook. https://github.com/facebookresearch/fairseq.

[13] M.J.F. Gales and P.C. Woodland. "Mean and variance adaptation within the MLLR framework". In: *Computer Speech & Language* 10.4 (1996), pp. 249–264. ISSN: 0885-2308. DOI: https://doi.org/10.1006/csla.1996.0013.

[14] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. *Speech Recognition with Deep Recurrent Neural Networks*. 2013. arXiv: 1303.5778 [cs.NE].

[15] U. Grenander, H. Cramér, and Karreman Mathematics Research Collection. *Probability and Statistics: The Harald Cramér Volume*. Wiley publications in statistics. Almqvist & Wiksell, 1959.

[16] Kenneth Heafield. https://kheafield.com/code/kenlm/.

[17] Geoffrey Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97. DOI: 10.1109/MSP.2012.2205597.

[18] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[19] Innovative collaboration for Inter-specialty cancer training across Europe Interact-Europe project. https://www.europeancancer.org/eu-projects/impact/interact-europe.

[20] Javier Iranzo-Sánchez et al. "Europarl-st: A multilingual corpus for speech translation of parliamentary debates". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 8229–8233.

[21] Kazuki Irie et al. "Language Modeling with Deep Transformers". In: *ArXiv* abs/1905.04226 (2019).

[22] Javier Jorge et al. "Live Streaming Speech Recognition Using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (Nov. 23, 2021), pp. 148–161. DOI: 10.1109/TASLP.2021.3133216. published.

[23] Javier Jorge et al. "Real-time One-pass Decoder for Speech Recognition Using LSTM Language Models". In: *Proc. of the 20th Annual Conf. of the ISCA (Interspeech 2019)*. Graz (Austria), Jan. 1, 2019, pp. 3820–3824. URL: https://www.isca-speech.org/archive/interspeech_2019/jorge19_interspeech.html. published.

[24] Hemant K. Kathania et al. "On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition". In: *2019 National Conference on Communications (NCC)*. 2019, pp. 1–5. DOI: 10.1109/NCC.2019.8732232.

[25] Philipp Koehn. "Europarl: A parallel corpus for statistical machine translation". In: *Proceedings of machine translation summit x: papers*. 2005, pp. 79–86.

[26] Yichong Leng et al. *SoftCorrect: Error Correction with Soft Detection for Automatic Speech Recognition*. 2022. arXiv: 2212.01039 [cs.CL].

[27] Pierre Lison and Jörg Tiedemann. "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. URL: https://aclanthology.org/L16-1147.

[28] Christoph Lüscher et al. "RWTH ASR Systems for LibriSpeech: Hybrid vs Attention". In: *Interspeech 2019*. ISCA, Sept. 2019. DOI: 10.21437/interspeech.2019-1780. URL: https://doi.org/10.21437%2Finterspeech.2019-1780.

[29] A. Martínez-Villaronga et al. "Language model adaptation for video lectures transcription". In: *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing ICASSP 2013*. Vancouver (Canada), Jan. 1, 2013, pp. 8450–8454. URL: http://dx.doi.org/10.1109/ICASSP.2013.6639314. published.

[30] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.

[31] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[32] V. Panayotov et al. "Librispeech: an ASR corpus based on public domain audio books". In: *Proc. of ICASSP*. 2015, pp. 5206–5210.

[33] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019.

[34] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS].

[35] Anthony Rousseau et al. "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks". In: *Proc. of LREC*. 2014, pp. 3935–3939.

[36] Ramon Sanabria et al. "How2: A Large-scale Dataset For Multimodal Language Understanding". In: *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS. 2018. URL: http://arxiv.org/abs/1811.00347.

[37] Steffen Schneider et al. *wav2vec: Unsupervised Pre-training for Speech Recognition*. 2019. arXiv: 1904.05862 [cs.CL].

[38] M. Schuster and K.K. Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093.

[39] Yangyang Shi et al. *Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition*. 2020. arXiv: 2010.10759 [cs.SD].

[40] Joan Albert Silvestre-Cerdà et al. "transLectures". In: *Proceedings of IberSPEECH 2012*. Madrid (Spain), Nov. 22, 2012, pp. 345–351. URL: http://hdl.handle.net/10251/37290%20http://www.mllp.upv.es/wp-content/uploads/2015/04/1209IberSpeech.pdf.

[41] Google Brain Team. https://www.tensorflow.org/.

[42] Jörg Tiedemann. "Parallel data, tools and interfaces in OPUS." In: *Lrec*. Vol. 2012. Citeseer. 2012, pp. 2214–2218.

[43] Ashish Vaswani et al. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[44] *VoxForge corpus*. Website: https://www.voxforge.org/.

[45] Chenguang Wang, Mu Li, and Alexander J. Smola. *Language Models with Transformers*. 2019. arXiv: 1904.09408 [cs.CL].

[46] Yongqiang Wang et al. "Transformer-Based Acoustic Modeling for Hybrid Speech Recognition". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020. DOI: 10.1109/icassp40776.2020.9054345. URL: https://doi.org/10.1109%2Ficassp40776.2020.9054345.

[47] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. 2019. DOI: https://doi.org/10.7488/ds/2645.

[48] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014. ISBN: 1447157788.

[49] Yu Zhang et al. *Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages*. 2023. arXiv: 2303.01037 [cs.CL].

[50] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. "The united nations parallel corpus v1. 0". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 3530–3534.

# List of Figures

# List of Tables

# APPENDIX

SUSTAINABLE DEVELOPMENT GOALS

Degree to which the work relates to the Sustainable Development Goals (SDGs).

| **Sustainable development goals** | **High** | **Medium** | **Low** | **Not applicable** |
|---|---|---|---|---|
| SDG 1. **No poverty.** | | | | **X** |
| SDG 2. **Zero hunger.** | | | | **X** |
| SDG 3. **Good health and well-being.** | | | **X** | |
| SDG 4. **Quality education.** | **X** | | | |
| SDG 5. **Gender equality.** | | | | **X** |
| SDG 6. **Clean water and sanitation.** | | | | **X** |
| SDG 7. **Affordable and clean energy.** | | | | **X** |
| SDG 8. **Decent work and economic growth.** | | | **X** | |
| SDG 9. **Industry, innovation and Infrastructure.** | | **X** | | |
| SDG 10. **Reduced Inequality.** | | **X** | | |
| SDG 11. **Sustainable cities and communities.** | | | | **X** |
| SDG 12. **Responsible consumption and production.** | | | | **X** |
| SDG 13. **Climate action.** | | | | **X** |
| SDG 14. **Life below water.** | | | | **X** |
| SDG 15. **Life on land.** | | | | **X** |
| SDG 16. **Peace and justice strong institutions.** | | | **X** | |
| SDG 17. **Partnerships to achieve the goal.** | | | **X** | |



ETS Enginyeria Informàtica
Camí de Vera, s/n, 46022, València
T +34 963 877 210
F +34 963 877 219
etsinf@upvnet.upv.es - www.inf.upv.es

Reflexion on the relation of the TFG/TFM with the SDGs and with the most related SDG(s).

*This work is in line with the Renewable Goals. In particular, with SDG 3 "Good health and well-being" which aims to "Ensure healthy lives and promote well-being for all at all ages" and with SDG 4 "Quality education" which aims to "Ensure inclusive, equitable and quality education". Within SDG 3, the most related targets to this work are:*

- *3.C Substantially increase health financing and the recruitment, development, training and retention of the health workforce in developing countries, especially in least developed countries and small island developing States.*
- *3.D Strengthen the capacity of all countries, in particular developing countries, for early warning, risk reduction and management of national and global health risks.*

*Within SDG 4, the most related targets to this work are:*

- *4.3 By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university.*
- *4.4 By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship.*

*This work contributes to increasing health financing and the recruitment, development and training in countries that do not have state-of-the-art knowledge and treatment in cancer medicine(target 3.C), and will enable cancer professionals to provide higher quality care to patients, thereby reducing national health risks (target 3.D).*
*This work contributes to increasing accessibility to educational resources for oncology professionals (target 4.3), aside from those whose acces to educational and formation resources requires them to learn and master a foreign language (target 4.4).*

# Agraïments

A mon pare, a ma mare i la meua germana, que sempre m'han donat el seu suport faça el que faça, comence les aventures que comence, i als quals estime molt (encara que no ho diga massa).

Gràcies a Joan Albert, que va confiar en mi el darrer any, i sense el qual tot açò no haguera sigut possible.

A Adrià, que també ha sigut un mestre durant tota aquesta etapa i al que li deguem hores d'evitar-nos patiments.

En especial a Gerard, amb qui he compartit aquest camí i que ha ajudat en gran manera a evitar que caiguera en una espiral de locura.

Als companys del grup MLLP, que són gent meravellosa i amb la que és un plaer treballar.

A totes les companyes i companys del piset, amb les que he passat un any genial, encara que un poc absent pels estudis.

A la meua CC, que sempre està ahí per a tot i que té més paciència que ningú del món mundial.

En resum a totes les amigues i amics que m'han acompanyat durant aquest llarg any, i que han sigut comprensius amb les meues absències.

Moltes gràcies a totes les mestres que han despertat en mi l'espurna de la curiositat i la necessitat de coneixement, i que d'alguna manera, m'han guiat per aquest camí tan *bonico*.

I finalment, gràcies a tu, per haver arribat fins al final!