UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Confidence Measures for Automatic and Interactive Speech Recognition

Ph.D. Dissertation
by
Isaías Sánchez Cortina

~ January, 2016 ~

Advisors:
Dr. Alfons Juan i Ciscar
Dr. J. Alberto Sanchis Navarro

Resum:

El *reconeixment automàtic de la parla* (RAP) és una tasca crucial per una àmplia gamma d’aplicacions importants que no es poden dur a terme per mitjà de la transcripció manual. El RAP pot proporcionar transcripcions en escenaris de creixent impacte social com el *Cursos online oberts massius* (MOOC). Les transcripcions permeten automatitzar tasques com ara cercar, resumir, recomanar, traduir; a més a més, fa accessibles els continguts als parlants no nadius i els usuaris amb discapacitat, etc. Fins i tot, pot millorar el rendiment acadèmic de estudiants que aprenen de xerrades amb subtítols, encara que aquests subtítols no siguen perfectes. Malauradament, la tecnologia RAP actual encara està lluny de la precisió necessària.

Les transcripcions imperfectes resultants de RAP poden ser corregides manualment, però aquest l’esforç pot acabar sent superior a la transcripció manual. Per tal de resoldre aquest problema, en aquest treball es presenta un sistema nou per a *transcripció interactiva de la parla* (TIP). Aquest sistema TIP va ser reeixit en la reducció de l’esforç per quan es pot permetre una certa quantitat d’errors; així com també en en la millora dels models RAP subjacents.

Per tal d’adequar el marc proposat per a MOOCs, també es van investigar altres mètodes d’interacció intelůligents amb esforç d’’usuari limitat. A més a més, es va introduir un nou mètode que aprofita les interaccions per tal de millorar encara més les parts no supervisades (RAP amb *cerca restringida*).

La investigació en TIP duta a terme es va desplegar en una plataforma web amb la qual va ser possible produir un nombre massiu de transcripcions semi-supervisades de xerrades de repositoris ben coneguts dins del projecte de recerca europeu transLectures, videoLectures.net i poliMedia.

Finalment, el rendiment de la TIP i els sistemes de RAP es pot augmentar directament mitjançant la millora de l’estimació de la *Confiança Mesura* (MC) de les paraules transcrites. Per tant, es van desenvolupar dues contribucions: un nou model discriminatiu *logístic* (LR); i l’adaptació al locutor de la MC per casos en que és possible, per exemple amb MOOCs.

Objectius:

- Dissenyar mètodes i eines TIP per millorar les transcripcions automàtiques.
- Avaluació del mar TIP proposat per tasques realistes extretes de grans repositoris de vídeos educacionals.
- Millorar la fiabilitat del TIP i ASR mitjançant la millora de les MC.

Abstracts:

The *Automatic Speech Recognition* (ASR) is a crucial task in a broad range of important applications which could not accomplished by means of manual transcription. The ASR can provide cost-effective transcripts in scenarios of increasing social impact such as the *Massive Open Online Courses* (MOOC), for which the availability of accurate enough is crucial even if they are not flawless. The transcripts enable search-ability, summarisation, recommendation, translation; they make the contents accessible to non-native speakers and users with impairments, etc. The usefulness is such that students improve their academic performance when learning from subtitled video lectures even when transcript is not perfect. Unfortunately, the current ASR technology is still far from the necessary accuracy.

The imperfect transcripts resulting from ASR can be manually supervised and corrected, but the effort can be even higher than manual transcription. In order to alleviate this issue, a novel *Interactive Transcription of Speech* (IST) system is presented in this thesis. This IST succeeded in reducing the effort if a small quantity of errors can be allowed; and also in improving the underlying ASR models in a cost-effective way.

In other to adequate the proposed framework into real-life MOOCs, another intelligent interaction methods involving limited user effort were investigated. And also, it was introduced a new method which benefit from the user interactions to improve automatically the unsupervised parts (the *Constrained Search* (CS)).

The conducted research was deployed into a web-based IST platform with which it was possible to produce a massive number of semi-supervised lectures from two different well-known repositories, videoLectures.net and poliMedia.

Finally, the performance of the IST and ASR systems can be easily increased by improving the computation of the *Confidence Measure* (CM) of transcribed words. As so, two contributions were developed: a new particular *Logistic Regression* (LR) model; and the speaker adaption of the CM for cases in which it is possible, such with MOOCs.

Scientific Goals:

- To design IST methods and tools to tackle the problem of improving automatically generated transcripts.
- To assess the designed IST methods and tools on real-life tasks of transcription in large educational repositories of video lectures.
- To improve the reliability of the IST by improving the underlying (CM).

Resumen:

El *reconocimiento automático del habla* (RAH) es una tarea crucial en una amplia gama de aplicaciones importantes que no podrían realizarse mediante transcripción manual. El RAH puede proporcionar transcripciones en escenarios de creciente impacto social como el de los *cursos abiertos en linea masivos* (MOOC), para el que la disponibilidad de transcripciones es crucial, incluso cuando no son completamente perfectas. Las transcripciones permiten la automatización de procesos como buscar, resumir, recomendar, traducir; hacen que los contenidos sean más accesibles para hablantes no nativos y usuarios con discapacidades, etc. Incluso se ha comprobado que mejora el rendimiento de los estudiantes que aprenden de videos con subtítulos incluso cuando estos no son completamente perfectos. Desafortunadamente, la tecnología RAH actual aún está lejos de la precisión necesaria.

Las transcripciones imperfectas resultantes del RAH pueden ser supervisadas y corregidas manualmente, pero el esfuerzo puede resultar superior al de la transcripción manual. Para atajar este problema, esta tesis presenta un novedoso sistema de *transcripción interactiva del habla* (TIH) que reduce el esfuerzo de semi-supervisión siempre que sea aceptable una pequeña cantidad de errores; además de mejorar a la par los modelos RAH subyacentes.

Con objeto de transportar el marco propuesto para MOOCs, también se investigaron otros métodos de interacción inteligentes que involucran esfuerzo limitado por parte del usuario. Además, se introdujo un nuevo método que aprovecha las interacciones para mejorar aún más las partes no supervisadas (ASR con *búsqueda restringida*).

Esta investigación en TIH se desplegó en una plataforma web con el que fue posible producir un gran número de transcripciones de videos de los repositorios videoLectures.net y poliMedia.

Por último, el rendimiento de la TIH y los sistemas de RAH se puede aumentar directamente mediante la mejora de la estimación de la *medida de confianza* (MC) de las palabras transcritas. Por este motivo se desarrollaron dos contribuciones: un nuevo modelo discriminativo *logístico* (LR); y la adaptación al locutor de la MC para los casos en que es posible, como por ejemplo en MOOCs.

Objetivos:

- Diseño de métodos TIH para mejorar las transcripciones automáticas.
- Evaluar el TIH propueto con tareas de transcripción realistas extraídas de grandes repositorios de vídeos educacionales.
- Mejorar la fiabilidad del TIH mediante la mejora de las MC.

ACKNOWLEDGEMENTS

This thesis work has been possible thanks to the supervision and advice of Alfons Juan and Alberto Sanchis. And of course, thanks to the valuable collaboration with my lab colleagues.

Also, it should be brought into attention that this work has been supported by the former Spanish Ministry "Ministerio de Ciencia e Innovación" (MICINN) by means of the scholarship "FPI BES-2010-033005". And also under the iTrans2 (TIN2009-14511), Active2Trans (TIN2012-31723), the EC (FEDER/FSE) under the transLectures (FP7-ICT-2011-7-287755) and EMMA (ICT-PSP/2007-2013-621030) projects.

CONTENTS

1 Introduction 13

1.1 Motivation and Thesis Outline 13

1.2 Document Structure 15

1.3 Scientific and Technologic Goals 16

2 Background 17

2.1 Automatic Speech Recognition 17

2.1.1 Feature Extraction: Audio Signal Analysis 18

2.1.2 Statistical approach to Automatic Speech Recognition 20

2.2 Confidence Measures 31

2.2.1 Combination of predictors 35

2.2.2 Uterance Verification as Hypothesis testing 36

2.2.3 Comparison of the different CM approaches 38

2.3 Interactive Speech Transcription 40

3 Interactive Speech Transcription 43

3.1 Introduction 43

3.2 Interactive speech transcription balancing error and supervision effort 45

3.2.1 Interactive Speech Transcription Procedure 45

3.2.2 Estimation of the Transcription Error 47

3.3 Experimentation 49

3.3.1 Experimental Setup 49

3.3.2 Results 52
3.4 Conclusions 56
3.4.1 General remarks on the behaviour of the IST method 56
3.4.2 Remarks on the large task WSJ1+0 57

4 IST prototype 59
4.1 IST prototype balancing error and user effort 59
4.2 User interface of the prototype 61
4.3 Demonstration 63
4.4 Usability 64
4.5 Conclusions 65

5 transLectures: Transcription of massive repositories of lectures 67
5.1 Introduction 67
5.2 Deployed architecture for cost-effective transcription of online repositories 68
5.2.1 the transLectures Player 71
5.2.2 ASR & SMT Systems 72
5.3 Overview of the videoLectures.net and poliMedia repositories: 73
5.4 First results on poliMedia 75
5.5 Conclusions 78

6 Intelligent interaction with limited user effort 79
6.1 Introduction 79
6.2 IILUE approach 81
6.3 Constrained search for speech recognition 83
6.4 Evaluation 85
6.4.1 Experimental setup 85
6.4.2 Results 88

6.5 Conclusions 92

7 Improved and Speaker-adapted confidence measures 95

7.1 Introduction 96

7.2 Speaker-adapted naïve Bayes classifier 98

7.2.1 Estimation of the smoothing parameters 101

7.3 Speaker-adapted logistic regression confidence estimator . . 103

7.3.1 Discriminative learning 106

7.4 Experiments 108

7.4.1 Experimental setup 108

7.4.2 Evaluation of CMs 110

7.4.3 Results 112

7.5 Interactive Speech Transcription Application 117

7.6 Conclusions 121

8 Conclusions 123

8.1 Summary 123

8.2 Publications 125

8.3 Future work 126

CHAPTER 1 INTRODUCTION

1.1 Motivation and Thesis Outline

This thesis work contributes to the field of the *Automatic Speech Recognition* (ASR). ASR is a crucial task in a broad range of important applications: hands-free/multimodal *Human Computer Interaction* (HCI); computer-aided language learning; and, of course, indexing/searching media and automatic subtitling, etc. None of these tasks might be accomplished with manual transcription since the overall process would be too slow, expensive and not embeddable into automatic systems.

The ASR is successfully approached by means of the *Pattern Recognition* (PR), which is a branch of the *Machine Learning* (ML) and *Artificial Intelligence* (AI) aiming at solving classification and regression problems. The ASR is currently a topic of intense research for which important advances have been achieved over the last decades.

Some scenarios of increasing social impact in which the ASR technology plays an essential role are the online repositories of videos, such as the *Massive Open Online Courses* (MOOC). The availability of accurate transcripts, even if not perfect, is crucial ([LABG05, Wal06, FWK07]): it enables search-ability, classification, summarisation, recommendation, translation, etc. Also, it makes the contents accessible to non-native speakers and users with impairments, etc. The usefulness is such that students can improve their academic performance when learning from subtitled video lectures even when transcript is not perfect, as it was shown in [PP12a, RTDG+13]. Unfortunately, the current ASR technology is still far from the accuracy of a professional transcriber.

The imperfect transcripts resulting from ASR can be manually supervised and corrected. However, the effort can be even higher than manual transcription. For the purpose of alleviating this issue, a novel *Interactive Transcription of Speech* (IST) system was proposed. This IST proved successful in reducing the effort whenever some small quantity of errors can be tolerated; as well as to improve existing ASR models in a cost-effective way. The produced non-perfect transcripts are sufficient to convey the meaning, so it can be used as an additional learning resource by students [MBP+06], or it can be employed for information retrieval [GVB03]. In fact, manual annotation of lectures by non-experts usually results with 10% of errors on average ([Haz06]). Details and evaluation of the proposed method can found on chap 3 and the implementation of this paradigm on chapter 4.

The introduced IST method gives rise to several advantages that make it extremely suitable for repositories of video lectures. However, some particularities inherent to these scenarios made necessary further research. For instance, repositories can count with a huge number of volunteer users who might collaborate in listening and correcting, but only for a short time and with limited effort. Being so, several intelligent interaction methods involving limited user effort were investigated in order to exploit this fact, instead of setting the desired resulting quality of the transcripts. It was also studied how to get more profit from the users interactions. Additionally, a new method was applied to further improve automatically the quality of semi-supervised transcripts: the *Constrained Search* (CS). This method makes use of the manually introduced corrections to improve automatically only the unsupervised parts. The modified IST approach and the CS method are detailed on chapter 6.

All the research conducted so far was deployed into real-life repositories of video lectures. A main product resulting from this challenging project was a web-based IST platform which deals with supervision of the automatic transcripts of a massive number of lectures from two different well-known repositories within the framework of the European project transLectures. Chapter 5 details the deployed architecture, as well as the repositories.

Finally, the conducted work showed room to easily improve the proposed IST system and its underlying ASR system by investigating the so called *Confidence Measure* (CM) of the automatically transcribed words. To do so, two different contributions were developed: a new particular *Logistic Regresion* (LR) model; and the speaker adaption of the CM whenever it is possible to identify the speakers, such as with the online video repositories. These contributions are explained and evaluated in chapter 7.

1.2 Document Structure

This thesis document is structured in eight chapters: this introductory chapter; a background on ASR and IST aspects (chap. 2); five chapters developing the main contributions of this work – Interactive Speech Transcription, IST prototype, transLectures: Transcription of massive repositories of lectures, Intelligent Interaction with limited user effort and Improved and Speaker-adapted confidence measures –. And last, the concluding chapter. The sequential reading of this document is encouraged. However, chapter 2 is optional for those readers versed on the topics treated in this work.

1.3 Scientific and Technologic Goals

The main goals of this thesis work can be summarised as follows:

1. To design IST methods and tools to tackle the problem of improving automatically generated transcripts.

2. To assess the designed IST methods and tools on real-life tasks of transcription in large educational repositories of video lectures.

3. To improve the reliability of the IST by improving the underlying *confidence measure* (CM).

The first goal is covered throughout the chapters 3 and 4. As commented before, chapter 3 proposes a novel interaction approach; while chapter 4 describes the implementation of tool following the proposed IST paradigm which greatly reduces the user effort.

The second goal corresponds to chapters 5 and 6. Chapter 5 details the proposed IST framework implemented into a web-platform with the purpose of obtaining accurate enough transcriptions for a massive number of online lectures under the EC project transLectures. Chapter 6 presents refined methods for IST more suited to this online scenarios which can count on a huge number of scarce collaborative users.

The third main goal, which arose as necessary from the IST experimentation, is treated over chapter 7. This goal is not only useful for the IST intelligent guidance, but also for the ASR in general and many applications.

CHAPTER 2

BACKGROUND

This chapter introduces the background for the topics being treated on this work: the Automatic Speech Recognition (ASR), the Confidence Measures (CM) for ASR and the Interactive Speech Transcription (IST). Each one of these topics is further discussed on the corresponding sections to follow.

2.1 Automatic Speech Recognition

The *Automatic Speech Recognition* (ASR) is the process carried by a machine to transduce human natural speech into text[1]. In addition to the main goal of providing transcriptions, ASR has become the core technology in a broad range of novel applications. For instance, ASR enables hands-free/multimodal interaction with electronic systems (computers, automobiles, robots, telephony assistance, etc.); computer-aided language learning; subtitling, indexing, translating, etc.

The ASR has been successfully approached over the last decades as a decision/classification problem from the *Pattern Recognition* (PR) point of view. The PR approach consists in applying a mapping function from an input data vector (features) to yield a labelled class. In the particular case of the ASR, the target classes (labels) are words; the features are derived

[1] To write spoken words is also known as *to transcribe* indepently wheather the process is manual or automatic. The resulting text is the *transcription* or *transcript*.

from the audio signal; and the mapping function has always been a pipeline of complex operations and algorithms based on statistical distributions with a huge number of free parameters. The free parameters are experimentally optimised using a set of previously labelled data.

The subsections to follow describe the main steps of the ASR process: the extraction of features from the audio signal processing point of view and the statistical approach to classify from features into words.

2.1.1 Feature Extraction: Audio Signal Analysis

The features in PR are a representation of a set of observed data (the audio signal in the case of the ASR task). This representation should condense the relevant information while removing the redundant in a way that the classification algorithm performs the best. On the contrary, the use of raw data may render a problem intractable.

During several decades, the best performing features for ASR were motivated by the human auditory system and the signal processing. The most widely used features were the short-term *Mel-frequency cepstral coefficients* (MFCCs) –that even got standardised by the European Telecommunications Standards Institute–. Short term (around 30ms) MFCCs are computed by applying mel filters to the power spectrum of the signal and then the inverse discrete cosine transform. The mel filters greatly compress the information onto a discrete frequency domain in which pitches are more or less perceptually equidistant. [DM80] assessed the MFCC superior performance.

In addition to the MFCC coefficients, their first and second derivatives provide additional dynamic information about the speech signal. Or, alternatively, projecting the coefficients into a lower dimension space by means of a linear discriminant analysis, which has ben proved to perform even better. Additional features such those involving long term windows of the signal may also yield some modest improvement. While other works with improved filtering or different integral transform functions (such as the wavelet transforms) did not become so popular. More recently, a huge leap in performance has been achieved with the inclusion of features computed as the posterior probability of phoneme classifiers given the conventional features mentioned above. This is known as a *tandem approach*, and it has been successfully addressed with the *Deep/Artificial Neural Netwoks* (DNN/ANN) ([HES00, CZM04, GKKC07, VVP$^+$07]).

Obviously, the MFCCs extraction process removes unimportant information such the intensity of the speech signal, background noise. Unfortunatelly the tonal information, which is also removed, constitutes an important treat for tonal languages such as the Chinese language ([CGM$^+$97]). Another issue concerning the feature extraction is that the speaker idiosyncrasy is only partially removed. In fact, several works use the short term MFCC for speaker identification purposes. Consequently, the recognition process is severely degraded because the features are not completely speaker independent. In order to alleviate the problem several *speaker normalisation* (SN) techniques were developed. For instance, the well-known *vocal tract length normalization* (VTLN) significantly improves the performance mainly by warping the features ([PN05]).

More recent methods achieve much better performance than SN by means of the *speaker adaptation*[2] of the ASR models instead of normalising the features . For instance, the well-known method *Maximum Likelihood Linear*

[2] For some speaker-adaptation methods an equivalent feature-transformation method exists.

Regresion (MLLR) for SA is detailed in [Gan05].

2.1.2 Statistical approach to Automatic Speech Recognition

Under the framework of the pattern recognition, de-facto standard is that in which the transcription ($\hat{\vec{w}}$) is obtained as the *Maximum a Posteriori Probability* (MAP) of any possible sequence of words ($\vec{w}$) given matrix X:

$$\begin{aligned} \hat{\vec{w}} &= \underset{\forall \vec{w}}{\arg\max}\, P(\vec{w} \mid X) \quad (2.1) \\ &= \underset{\forall \vec{w}}{\arg\max}\, P(X \mid \vec{w}) \cdot P(\vec{w}) \quad (2.2) \end{aligned}$$

Where $X = \{\vec{x}_1 \ldots \vec{x}_T\}$ is build up from the concatenation of T multidimensional features ($\vec{x}_i = \{x_i^1 \ldots x_i^D\}$); which are derived directly or indirectly from the audio and they are usually ordered sequentially in time (see 2.1.1). On the other hand, $\vec{w}$ typically contains words only from a previously specified closed set *vocabulary*. Of course, the number of the all possible sequences $\vec{w}$ explodes exponentially with the size of the vocabulary and the maximum allowed length of the sequence.

Equation 2.1 renders the most probable transcription by means of a *discriminative* model. In contrast, eq. 2.2 uses a *generative* model, which has been obtained by applying the Bayes rule and then removing the prior of the features ($P(X)$) because it does not depend on $\vec{w}$.

Discriminative models usually outperform the generative ones for classification and regression tasks (which do not require to model explicitly the joint distribution), especially because of the available training methods. However, this is not the case for ASR: many pure discriminative models

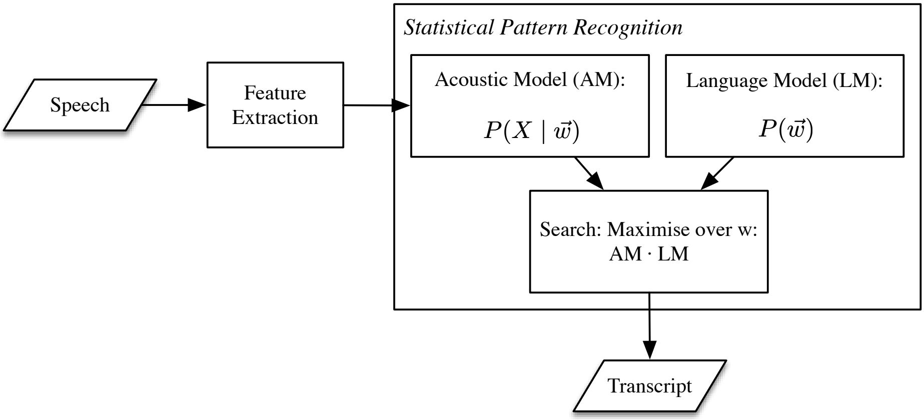


Figure 2.1: Architecture of a statistical ASR system ([Ney90]).

were tried (SVM, ANN[3], logistic regression, etc.) but they failed to properly cope with the dynamic and variable-size nature of speech signals.

Consequently, the generative formulation (eq. 2.2) has been the standard way to proceed over the last decades because it presents several advantages:

- The *Acoustic Model* (AM) (see 2.1.2), $P(X \mid \vec{w})$, can be estimated in a easier and more precisely way than its discriminative[4] counterpart.
- The *Language Model* (LM) (see 2.1.2), $P(\vec{w})$, can be estimated from text only resources independent to the speech task– which are highly available.

[3] On the horizon, Deep NN may tackle the ASR discriminative approach. However, the state-of-the art uses Deep NN within the generative approach: tandem and hybrid approaches, commented near the end of subsections 2.1.1 and 2.1.2.

[4] During the last decade, discriminative training of generative AM models, though computianally expensier, have proved to outperform the generative learning. However, these methods should not be confused with the pure discriminative aproaches to ASR.

Also, the LM can be reliably learnt by means of the standard cheap and well-known N-grams.

- The *Search* (see 2.1.2) of the most probable transcription (or *decoding*) can be pruned early in an effective way thanks to mainly the LM, as well as other AM subunits and heuristics.

In summary, apart form the feature extraction step, the AM, LM and search has conformed the gold standard approach for ASR during decades. Fig. 2.1 depicts the basic mechanism. Each one of these ASR aspects are further explained on the following subsections.

Acoustic Modelling

As stated above, the *Acoustic Model* of the generative framework for the statistical approach to ASR provides the probability of the input signal to be generated by a hypothesised sequence of words. This model is usually estimated with supervised learning methods using a training corpus of speech segments or utterances and the associated transcriptions.

The estimation of the AM for *Large Vocabulary of Continuous Speech Recognition* (LVCSR) is rather complex. For instance, for continuous recognition, the words of the training corpora are not usually fine-grained aligned in time; for what it is necessary an iterative process starting with a coarse estimation of the alignment. Additionally, special models must be implemented to deal with the all the non-speech or useless speech portions in the audio signal; such as silences, noises, hesitations, etc.

Another issue of particular relevance with large vocabulary is that the AMs cannot be properly estimated for the vast majority of the words. This is because most words appear very few or no times on a speech corpus.

Fortunately, this problem can be addressed by augmenting the AM-words with sub-word models ([Ney90]). The most widely used sub word speech units are phonemes and tri-phonemes (phoneme and their context phonemes to model the pronunciation variability).

The sub-word unit extension is simply performed by probability concatenation, but it is necessary a *lexicon* associating each word to its (tri)phonetic transliteration. Frequently, except for languages with a direct *Grapheme to Phoneme* (G2P) correspondence such as for the Spanish language, the best results are achieved when using an automatically discovered mapping given an initial hand-made lexicon ([BN08]). Another problem to consider when using tri- or longer context phonemes as sub units is that their number increases exponentially with the context length. Thus, a large number of allophones will have no or too few observations for a reliable parameter estimation. For this reason, several states are tied together to yield more general models ([You92]). For instance, using *decision tree-based clustering* (e.g. CART) .

Also of relevance is the linguistic phenomenon of the coarticulation of words: phonemes at a word boundary may be pronounced differently depending on the predecessor and successor words. This is more prominent or not depending on the language. The solution to overcome this situation has been named *across-word* modelling ([Six03]). Nonetheless, the small improvement yielded with the across-words modelling fades off when applying new AM training techniques.

On the other hand, the modelling of the speech units is a very challenging problem by its own. Nonetheless, the *Hidden Markov Models* (HMMs) have become the standard because they are good in matching sequential patterns of high variability, like the realisation of the phonemes. More in detail, an

HMM is a *Stochastic Finite State Automaton* (SFSA) with a certain set of states (S) and probabilistic transitions between them. The states are said to be “hidden” random variables because they cannot be observed:

$$P(X \mid \vec{w}) = \sum_{\forall \vec{s} \in S^T(\vec{w})} P(X, s_1, \ldots, s_T, \mid \vec{w}) \tag{2.3}$$

The sum is over all possible state sequences of dimension $|\vec{s}| = T$ (the same as the observation sequence X). The valid state-sequences are given by a preset topology. The most common topology in ASR is that in which each recognition unit (phonemes or words) has its own reserved states that must appear in a pre-defined order, so transitioning is only allowed to self or the next and second next states. Also, the sequence must transit to special start/end states at the beginning and end of the unit been recognised. For illustration, fig. 2.1.2 depicts the state space for the recognition of 7 observation features.

Moreover, for the sake of mathematical and computational tractability of the HMMs, further simplifications must be made. The following three assumptions are standard for the ASR case (see [DHS12] for further detail):

1 *First order Markov*: a state (s_t) depends only on the previous (s_{t-1}).

2 *Stationarity*: the probabilities of state transitions are fixed over time.

3 *Output independence*: The observed set of features at a time ($\vec{x}_t$) is independent of the previous observed features ($\vec{x}_{t'}$ with $t' \neq t$). Though this assumption has a very limited validity.

The assumptions above lead to the following expression:

$$P(X \mid \vec{w}) \simeq \sum_{\forall \vec{s} \in S^T(\vec{w})} \prod_{t=1}^{T} P(\vec{x}_t \mid s_t, \vec{w}) \cdot P(s_t \mid s_{t-1}, \vec{w}) \tag{2.4}$$

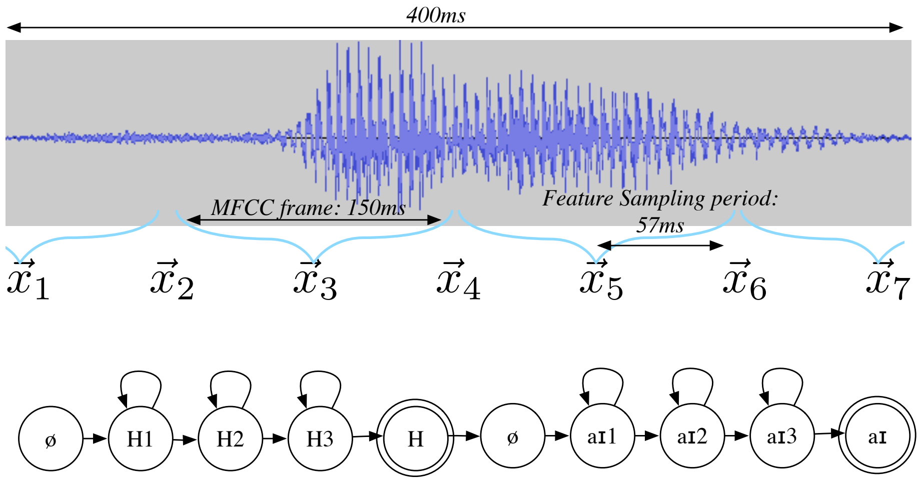
$$\approx \max_{\forall \vec{s} \in S^T(\vec{w})} \prod_{t=1}^{T} P(\vec{x}_t \mid s_t, \vec{w}) \cdot P(s_t \mid s_{t-1}, \vec{w}) \tag{2.5}$$

Where $P(\vec{x}_t \mid s_t, \vec{w})$ is known as the *emission probability*, which computes the probability to observe $\vec{x}_t$ while being in state s_t. And $P(s_t \mid s_{t-1}, \vec{w})$ is the *transition probability*. These probabilities are the result of the first and third simplifications respectively. On the other hand, eq. 2.5 presents a common solution to further reduce the computational cost by replacing the sum with the maximum. This approximation is known as *Viterbi* ([Ney90]). Both equations can be solved efficiently using the forward-backward algorithm [Bau72].

The emission probabilities can be modelled by discrete probabilities [Jel76], semi-continuous probabilities [XH89] or continuous probability distributions [SL83]. In particular, the *Gaussian mixture models* (GMMs) have been the most common used continuous distributions during the last decade:

$$P(\vec{x}_t \mid s_t, \vec{w}) = \sum_{m=1}^{M_s} C_{s_t m} \cdot \left(\frac{1}{\sqrt{(2\pi)^D |\Sigma_{s_t m}|}} e^{-\frac{1}{2}(\vec{x}_t - \vec{\mu}_{s_t m})^T \Sigma_{s_t m}^{-1} (\vec{x}_t - \vec{\mu}_{s_t m})} \right) \tag{2.6}$$

Where the term inside the parenthesis is the multivariate normal density with mean $\vec{\mu}_{s_t m} \in \mathbb{R}^D$ and covariance matrix $\Sigma_{s_t m} \in \mathcal{M}_{D \times D}(\mathbb{R})$. D is the dimension of the observed feature vectors ($|\vec{x}_t| = D$); M_s is the number of predefined gaussian mixtures; and $\{C_{s_t m}\}$ are the non-negative real mixture



$$S^T(\text{‘Hi!’}) = S^T(/\text{h}/,/\text{aɪ}/) =$$

```
{(ø H1 H1 H2 H3 H ø aɪ1 aɪ2 aɪ3 aɪ),(ø H1 H2 H3 H ø aɪ1 aɪ1 aɪ2 aɪ3 aɪ),
 (ø H1 H1 H2 H3 H ø aɪ1 aɪ2 aɪ3 aɪ),(ø H1 H2 H3 H ø aɪ1 aɪ2 aɪ2 aɪ3 aɪ),
 (ø H1 H1 H2 H3 H ø aɪ1 aɪ2 aɪ3 aɪ),(ø H1 H2 H3 H ø aɪ1 aɪ2 aɪ3 aɪ3 aɪ)}
```

Figure 2.2: List of valid state sequences (bottom), corresponding to an HMMs arquitecture to regonise phonemes /h/ and /ai/ for hypothesed word "Hi!" (middle); and an input sequence of 7 set of features $X = \vec{x}_{t=1..7}$ from audio (top).

weights $\sum_m C_{s_tm} = 1$ and $C_{s_tm} \in \mathbb{R}^+$).

The most widely used method to estimate the set of parameters[5] of the GMMs ($\{C_{s_tm}, \vec{\mu}_{s_tm}, \Sigma_{s_tm}\}$ has been the *Maximum Likelihood* (ML) training criterion in combination with the *Expectation-Maximization* (EM) algo-

[5] Some state-of-the art systems assume a global pooled diagonal covariance ($\Sigma_{s_tm} \equiv \text{diag}(\Sigma_{s_t})$). These assumptions alleviate the data sparseness issue while reducing the computational cost. If so, the features must be as de-correlated as possible. Decorrelation can be achieved, for instance, applying LDA during feature extraction.

rithm ([AD77]). More recently, several works have achieve better recognition performance by means of discriminative training techniques to estimate the generative GMMs-HMMs ([Hei10, Jia10]).

On the other hand, state-of-the-art systems using Deep NN have been applied to tackle the output probabilities of HMMs. This method is known as *hybrid approach* [SLCY11, DYDA12, BSR13].

A final remark regarding the estimation of the HMMs is about the transition probabilities ($P(s_t \mid s_{t-1}, \vec{w})$). As stated above, modern ASR systems not only consider they are fixed (*stationary assumption*), but also set them to a predefined values. This is because refining their values by means of the ML-EM yields insignificant increases in recognition performance and it slows down the overall training process.

Language Modelling

The LM is the prior distribution probability for any hypothesised sequence $\vec{w}$. The LM is a key piece of the generative approach to ASR approach:

- The LM can be estimated accurately exclusively from textual data, without depending on any associated speech. This allows the use enormous quantity of data for the estimation.
- Statistically learned LMs implicitly include syntactic and semantic constraints up to some extent. The syntax and semantic constrains can be also explicit modelled within the statistical framework by concatenating probabilities for a very small improvement in performance.
- The LM is used within state-of-the-art ASR systems to prune the Viterbi search, especially if combined with look-ahead techniques.

- A vast number of works has been devoted to this topic.
- The LM can be reliably learnt with the well-known and fast N-grams.

The *N-gram* LM assumes that the probability of a word sequence can be computed as if each word depends only on the $(N-1)$ previous words. Thus, for a L-length hypothesis $\vec{w} = (w_1, \ldots, w_{n-1})$:

$$P(\vec{w}) \equiv \prod_{n=1}^{L} P(w_n \mid w_1, \ldots, w_{n-1}) \tag{2.7}$$

$$\approx \prod_{n=1}^{L} P(w_n \mid w_{\max(n-N+1\,,\,1)}, \ldots, w_{n-1}) \tag{2.8}$$

Where eq. 2.7 is the exact expression resulting from recursively applying the Bayes theorem. On the other hand, eq. 2.7 presents the N-gram approximation. For further reference, [MS99] extensively depicts the n-gram modelling.

The optimal length N of the N-grams is usually a tradeoff between a models complexity and ability to generalise. The higher N, and thus more complex, the more accurate is the modelling of the training data; but the poorer estimation of the frequencies. In order to increase the chances for more reliable estimation for high N, the size of the text corpus must be vastly increased. Moreover, the high order n-grams may are more likely to overfit the training. Thus, an independent test text may be modelled more accurately by lower order n-grams. For speech recognition the optimum is found to be within 3 and 5-grams.

Notwithstanding the many advantages of the n-gram modelling so mentioned above, the frequency estimation method of the priors in the latter equation is usually not robust for low occurrence sequences (even for low order N-grams). This problem is typically tackled using techniques such as smoothing and back-off. These techniques redistribute the mass of probability in favour of low-or-no frequency N-grams ([Kat87, MLN97]). The parameters of the smoothed[6] language model can be estimated using a cross-validation scheme like leaving-one-out ([HN94, NMW97]).

Finally, about the evaluation of the performance of the LM, the most common measure to evaluate the LM is the *perplexity*:

$$P_P = (P(\vec{w}))^{-1/L} \tag{2.9}$$

The logarithm of this measure is equivalent to the entropy of the model (for an arbitrarily large sequence $\vec{w}$ containing L words). In the particular case of the N-gram models, $P(\vec{w})$ should be computed as in eq. 2.8. In that case, the perplexity corresponds to the average number of choices to continue a word given the $N-1$ previous words. This measure is also directly used as a training criteria, yielding a closed solution which is simply the relative frequency of the sequence on the training corpus.

Search

The search or *decoding* is the process conducted to find the most probable transcription. Putting together the considerations made above for the AM (first-order HMMs, eq. 2.4) and the LM (N-grams ,eq. 2.8), the estimation

[6] Some state of the arts systems opt to not using backoff. Thus, greatly decreasing the computation cost. This is possible because the large of the lexicons have grown so large (more than 200000 words) and they are trained with huge corpora (up to terabytes of raw text), that it can be simply considered that OOV words do no exists and a backoff model would associate an almost null probability.

of the generative model on eq. 2.2 reads as:

$$\hat{w} = \underset{\{\forall \vec{w} \in V^L \,|\forall L \in \mathbb{N}\}}{\arg\max} \prod_{n=1}^{L} P(w_n \mid w_{\max(n-N+1\,,\,1)}, \ldots, w_{n-1}) \\ \cdot \sum_{\forall \vec{s} \in S^T(\vec{w})} \prod_{t=1}^{T} P(\vec{x}_t \mid s_t, \vec{w}) \cdot P(s_t \mid s_{t-1}, \vec{w}) \tag{2.10}$$

As it can be seen from the equation above, the search of all possible sentences of any length ($\{\forall \vec{w} \in V^L \,|\forall L \in \mathbb{N}\}$) is intractable even for a small vocabularies (V). Thus, further approximations must considered, such as the Viterbi approximation (eq. 2.5):

$$\hat{w} = \underset{\{\forall \vec{w} \in V^L \,|\forall L \in \mathbb{N}\}}{\arg\max} \prod_{n=1}^{L} P(w_n \mid w_{\max(n-N+1\,,\,1)}, \ldots, w_{n-1}) \\ \cdot \max_{\forall \vec{s} \in S^T(\vec{w})} \prod_{t=1}^{T} P(\vec{x}_t \mid s_t, \vec{w}) \cdot P(s_t \mid s_{t-1}, \vec{w}) \tag{2.11}$$

The latter equation reduces considerably the complexity and efficient methods of dynamic programming can deal with computation. Still, the problem of selecting which hypothesis ($\vec{w}, \vec{s}$) should be checked, which is historically approached with classical AI strategies such as depth-first or breadth-first.

In particular, the breadth-first search allows to reduce the search space significantly when combined with pruning methods such as the *beam-search* (BS). The BS drops out all the hypotheses which their estimated likelihood differ to much from the best hypothesis found so far ([OFN97]). Of course, pruned searches do not grant the best global solution. But in practice, moderate pruning yields no significant search errors.

Another well-known methods to reduce further the computational complexity are lexical prefix tree ([NHUTO92]), look-ahead ([OVWY94]), etc. Also, more recent algorithms make use of N-best lists ([SA91]) or word lattices [ONA97] rather than looking for the best plain sequence $\vec{w}$.

Finally, it should be noted the following practical issue due to the independence assumptions made by the HMMs: the AMs are typically underestimated. Thus, the MAP decision is biased due to the multiplicative operation carried out on the probability distributions. The standard approach to alleviate this problem is to include a *word Insertion Penalty* (WIP, ρ) and a exponential *scaling factor* (γ) to the LM. Eq. 2.2 would simply turn into:

$$\hat{w} = P(\vec{w}|X) := \underset{\{\forall \vec{w} \in V^L | \forall L \in \mathbb{N}\}}{\arg\max} P(X|\vec{w}) \tag{2.12}$$

With $L = ||\vec{w}||$, the length of the hypothesised transcription. The WIP and γ parameters are included only for the decoding. As so, they could not be trained during the AM and LM learning. Nonetheless, they can be fairly estimated over a small development set by means of a coarse brute force search; or by a more refined method using recognition word lattices.

2.2 Confidence Measures

A *Confidence Measure* (CM), in the context of pattern recognition, is a measure of the reliability for any recognition decision made by an automatic system. In particular to ASR, the capability to evaluate the reliability of automatically recognised speech has become crucial to increase usefulness and intelligence of ASR systems in many practical applications. For example, *Out-of-vocabulary* (OOV) detection ([Qin13]), keyword spotting

([WMS98]), dialogue systems ([HSP02]) , system combination ([EW00a]), and for unsupervised adaptation of acoustic models and *Interactive Transcription of the Speech* (IST) [HTRT06a, SCSSJ12].

The CM has been mainly addressed as a 0-1 normalised score, computed at phoneme, word, phrase or sentence level. Nevertheless, CM at the word level has been the main focus on the literature due to its usefulness for the vast majority of applications ([WSMN01a, KK05, GZXY09, BRG07, WLW^{+}10, JY11, YP13]). Nonetheless, despite of the vast amount of works concerning CM there is still an open issue for LVCSR.

On the other hand, independently of the target level, the CM task has been approached mainly as a classification problem or by hypothesis testing. This distinction is based on the method used to learn the CM. Moreover, two prominent methods can be distinguished within the classification approach: computing a precise *posterior* ([WSMN01a]); or combining several sources of knowledge, possibly including also approximations to the posterior.

On the sections to follow the posterior, combination of predictors and *Utterance Verification* (UV) for hypothesis testing approaches are summarised; followed by a discussion on the shortcomings of each approach and finally a brief overview on the current state of the CM. For further insight on this topic, [Jia05] presents a still good introspection.

Posterior probability of a recognized word

The posterior probability of a recognised hypothesis[7] H given the input audio signal X is in fact a very good estimator. And, precisely, the posterior

[7] H is typically a sequence of phonemes, a single or several words, an arc of the recogntion lattice or the whole transcript, i.e. $H \subseteq \hat{\vec{w}}$

$P(\vec{w} \mid X)$ is the distribution in which the ASR statistical pattern recognition approach is based to obtain the most likely transcription $\hat{\vec{w}}$.

Unfortunately, the posterior of the hypothesis H of interest is not really ever computed during the recognition in the typical case of the generative approach to ASR, as remarked on sec. 2.1.2. In order to build back the posterior, $P(X)$ must be computed[8]. However, it is unfeasible to compute a prior directly from samples of features, since they cannot cover the infinite space set of features. The most obvious solution to tackle the computation of this prior is to make use of the AM and LM models:

$$P(X) = \sum_{\forall \vec{w}} P(\vec{w}) P(X \mid \vec{w}) \tag{2.13}$$

The latter sum over all potential possible recognition hypothesis ($\vec{w}$) is usually too costly because it includes all possible combinations of words, phonemes, noises and other events. Consequently, the estimation of $P(X)$ has been approached mainly with:

- *Filler-based* methods calculate $P(X)$ from a set of general filler or background models. For instance, all-phone recognition models ([YW94]), catch-all models ([KHS99]), highest word AM×LM score-based ([CR96]).
- *Lattice-based*[9] methods approximate the whole hypothesis space to be searched to the hypothesis subset present in the ASR lattice, which conform the majority of the probability mass. This approach can be efficiently computed by means of an forward-backward algorithm ([WSMN01a]).

[8] $P(X)$ is the prior distribtution of an input set of features X related to the audio signal. It is ignored during the decision-making of the generative approach because it is constant across the maximisation (eq. 2.2)

[9] A reconigtion lattice ia a compact representation of the most probable competing hypotheses generated during a recognition pass.

The lattice-based method is the most widely used since it provides superior performance compared to the more complex filler approach. The lattice-based has been addressed mainly as follows:

- *Competing hypothesis in a time-frame* ([EW00a]): the CM is the normalised ratio of the (median, maximum or mean) scores of all the paths with the exact H of interest within the same time-frame[10] and the all competing paths of any hypothesis within the same time-frame.
- *Confusion networks* (CN)[11]: the posterior is computed from a CN constructed using a clustering algorithm over the lattice ([MBS00]).

Finally, it should be noted that lattices store scores which are not properly normalised probabilities due to the simplifications that ASR carry out in practice. As so, the joint likelihood a certain path q of an hypothesis the AM (or the LM) can be approximated by an exponential scaling:

$$P(X, H, q) \simeq a^{\frac{1}{\gamma}} \cdot l \tag{2.14}$$

Where the scores a and l, computed during the recognition, are those related to the AM ($P(X|H, q)$) and LM ($P(H)$) probabilities of the arc q respectively. And γ is an scale parameter that can be optimised experimentally over a development set.

[10] Tyically there are several recognition paths with identical H bounded to the same time-frame, but for which each word or sub unit is differently aligned on time.

[11] A CN is a linear graphical representation of the compething hypotheses in a simpler way than the lattices. All paths through a CN are constrained to pass through all nodes, resulting in a the arcs in the con- fusion network correspond to words; ϵ arcs are added for null or missing words in the hypotheses. The nodes essentially impose a segmentation of the utterance into confusion sets.

2.2.1 Combination of predictors

Several sources of information have been used to compute predictor scores for the CM estimation. These predictors have been collected within the recognition process at levels of acoustics, language model, syntax, and semantics. Some common predictors reported in the literature are:

- Pure normalised likelihood score related: acoustic score per frame.
- N-best related: count in the N-best list, N-best homogeneity score (the weighted ratio of all paths passing through the hypothesised word in N-best list), top N recognition scores, top $N-1$ difference in adjacently ranked recognition scores, etc.
- Acoustic stability: a number of alternative hypotheses are generated based on different language model weights in decoding and acoustic stability of any given word is defined as the number of times the word occurs in the list divided by the number of alternatives in the list.
- Hypothesis density: the number of alternative arcs spanning the time segment of the recognised word in word graph.
- Duration: of the HMM states, phonemes duration or words.
- Language model (LM) related: LM score, LM back-off behaviour, etc.
- Parsing related: whether or not a word is parsed by grammar in robust parsing, position of each parsed word within the semantic slot, the language model back-off mode of the whole parsed slot, etc.
- Posterior probability (explained above).
- Log-likelihood-ratio related (explained below).

Moreover, further independent features can be extracted from pure language features, such as parsing- and/or semantic-related ones. But the reported results were not compelling.

None of predictors above perform well enough by themselves. Therefore, several works attempted to combine the predictors for a better performance. Many well-known classifiers algorithms have been tried: linear ([GIY97, HSP02]), gaussian mixtures, neural networks ([WBR$^+$97, Cha97]), decision trees ([NRE97]), boosting ([MLR01]), support vector machines ([ZR01]), naïve Bayes [SJV12], maximum entropy models ([WDAO07, ESJV08]), conditional random fields ([SW11]), etc.

The combination of predictors is the most widely used approach for CM on the literature and the vast majority further simplifies the problem to a binary classification of separated words into wrong or correct independently of the degree of resemblance of them to real utterance. That is, the predictors are computed at word level, and so, they ignore some kind of possible recognition errors such as insertions.

2.2.2 Uterance Verification as Hypothesis testing

Works on utterance verification formulate the confidence measure problem as a statistical hypothesis testing problem. Mainly using the Neyman-Pearson Lemma ([LR00, LR96]) or a Bayesian approach ([JD01]).

Likelihood Ratio

A recognised hypothesis H is correct if the following likelihood ratio is lower than a threshold τ:

$$\frac{P(x|H_0)}{P(x|H)} < \tau \tag{2.15}$$

Where H_0 is the alternative hypothesis which usually represents a very complex and composite event, where the true distribution of data is unknown. Some early works modelled $P(x \mid H_0)$ using HMMs with the same structure used to recognised words (which compose H), but adapted to a general background model, or hypothesis-specific anti-model, or a set of competing models, or a combination of all the above. Parameters for these alternate recognition models have been proposed to be obtained by means of discriminative training methods (which is generally agreed they can model H_0 in a better way).

Bayesian factors

Bayes factors is a powerful statistical tool to model composite hypotheses and can be used to solve many different verification problems. Bayes factors offers a way to evaluate evidence in favour of the null hypothesis H_0 because Bayes factors is the ratio of the posterior odds of H_0 to its prior odds, regardless of the value of the prior odds. The key issues are what role the necessary prior distributions will play in utterance verification and how to use them as a flexible tool to incorporate a variety of information sources useful for UV.

One important drawback of the hypothesis testing is what data should be used to estimate these anti-models. Some heuristic methods have been probed, such as performing forced-alignment against a wrong or random transcript to generate training data for each anti-model. But an in-search data selection procedure to collect the most representative competing tokens for each HMM in the system looked more promising. Also promising was an approach proposed for parametrising the *neighborhood* of the prior models, for the Bayes factors case.

2.2.3 Comparison of the different CM approaches

– Regarding the posterior approaches to the CM, it has been widely reported that those computed from N-best[12] are significantly outperformed by the lattices or CN posterior computation approaches. Unfortunately, it is a well-known issue that posterior is not ever estimated in a precise way. This is so because only unnormalised raw posterior can be computed from the ASR decoding process, making necessary additional scale-parameter tuning and post-processing to optimise and map the scores to a suitable CM. Furthermore, lattice-based posteriors tend to be overestimations because they are computed on a subset of the hypothesis space. And, also, they are also susceptible to the independence assumptions made by the recogniser. For the CN-based case, the selected arc clustering assumptions and algorithm widely affect the estimation of the posterior.

– On the other hand, the typical methods proposed to compute the posteriors usually overpass the performance of the refined and complex likelihood ratios in UV. This is due to the intrinsic difficulty in estimating alternate hypothesis models in UV, which should be highly complex composite distributions. Possible solutions to proper alternate modelling are, for instance, the training on different data, or the use completely different modelling techniques. But they are costly to implement.

– About the combination of predictors, the predictors purely computed from CN and lattices are better than those computed from N-best lists. Furthermore, the combination of the predictors, regardless of the used method, always outperforms the best single predictor. Though the gain can be too modest if the predictors are too correlated ([KS97]). Because of that, spe-

[12] List with the most likely recognitions. This can be considered a subset or a less compact representation of the information contained in a lattice.

cial care with the design of the algorithms should be payed to not overestimate the information supplied from correlated predictors, as with the proposed algorithms in this thesis. Another positive outcome of the combination approach is that it allows combination of information at different level of granularity (phone, word, sentence); while designing and training the alternate models for competing sub units is rather complex for the UV approach. Consequently, the vast majority of works have focused on the predictor combination method.

In general terms, the CMs are not currently robust and reliable enough yet to be a solid basis for decision-making in many cases. For instance, even the simple application of effectively detecting of OOV words for LCVSR remains as an open question. Anyway, useful improvements can still be obtained with the use of the CM on in-system applications such as automatic lattice rescoring during ASR recognition (posterior-based [EW00b] or minimum risk Bayes [GKB01] and also to automatically improve existing ASR systems with active learning ([AB98, WWR00, AB98, HTRT06a, SCSSJ12]).

With regard to the precise level of the target granularity, the word-level CM target poses another issues: for example, a correctly recognised word may have a very low confidence measure because its boundary is wrong (though its identity is correct) or the context words are wrong. Nonetheless, for certain applications such as the IST presented on this work (chap. 7), the impact of this flaw is insignificant for the designed method of user interaction.

Finally, it should be noted that there are issues not related to the methods themselves that difficult the assessment of the different publications. For instance, there are very different metrics to evaluate the performance (equal error rate, confidence error rate, normalised cross entropy, etc.). Also, they are usually evaluated on very different task, instead of in a unique well-designed publicly-accepted verification tasks.

2.3 Interactive Speech Transcription

The *Interactive Speech Transcription* (IST), from a general point of view, is the process of obtaining the transcription of a speech in the core of an automatic system with the help of a user. Thereby combining human accuracy with ASR efficiency. The contribution of the user can be as much as typing the full manual transcription, or as little as a few clicks or keystrokes, or any other modal interaction. It should be recalled that, as motivated on the introductory chapter, the IST is of particular importance nowadays.

Two main different paradigms can be distinguished in IST: transcription or supervision. For the here so-called "transcription" paradigm, the user must transcribe manually almost all the speech; while for the "supervision" paradigm the user should only listen and then correct some small portions of a provided automatic transcript. The second paradigm reduces dramatically the necessary amount of time to yield a suitable transcript. Also, in case of IST embedded into online repositories of videos, it opens up the possibility of an arbitrary large number of altruistic users to contribute to the improvement of the transcriptions.

If the target IST paradigm is just facilitate the process to professional users, IST approach may consist on just a convenient interface to facilitate the manual transcription to the user. However for more refined purposes, the system should provided the user with further utilities. For instance, automatic segmentation of the speech into sentences of a length more tractable to the user ([SSHTT00]); autocompletion of text by means of automatic prediction (ex. [SSSB10]); or even performing the whole transcription with an internal ASR system and then asking the user to correct only some portions ([HTRT06a, RRCV07, SCSSJ12]).

Very few research on the supervision IST paradigm can be found on the literature (mainly the last three references just mentioned). This is probably because of the lack of proper automatic speech segmentation methods to split the audio into pieces without cropping of the utterances and suitable for a user to understand. Due to this problem, the other works in the literature focused on supervising transcriptions at a sentence level. This is, the speech was assumed to have been conveniently segmented (probably manually) into utterances lengthy enough – usually full linguistic sentences. On the contrary, on this thesis work (chap. 3), a method which deals directly at word level is presented. Supervising speech at word level implies a huge decrease of user effort, since less audio is played and listened. This segmentation issue is partially bypassed by defining the goal to reduce the user effort to yield accurate enough, although not perfect, transcripts.

CHAPTER 3

INTERACTIVE SPEECH TRANSCRIPTION

This chapter presents a novel successful interactive approach to deal with the issue of the imperfect transcripts from automatic speech recognition.

After the following introductory section, the rest of the chapter is organised as follows: sec. 3.2 details the proposed interactive approach and the method for balancing the error and user effort. Next, sec. 3.3 evaluates of the performance of the IST system in terms of the user effort over the task of transcribing the speech of two sets of the World Street Journal Speech database. Finally, section 3.4 arrises the conclusions.

3.1 Introduction

Speech transcription is a crucial task in a broad range of important applications. Speech transcriptions produced by human transcribers can provide high quality results. However, the overall process is very slow and usually expensive. One way to deal with this important drawback is to produce automatically speech transcriptions based on automatic speech recognition (ASR) technology [RSS07, SNTW^{+}11b]. However, this solution presents two main difficulties. First, the building of an ASR system implies usually an important human effort since statistical models have to be obtained. Acoustic and language models are not easily available for specific tasks and, thus, they have to be learned from manual annotated data. The second difficulty is that automatic transcriptions are still far from producing the desired quality out of some specific scenarios. Therefore, an important human effort to supervise the speech transcriptions is mandatory.

To deal with the first difficulty, active and unsupervised learning techniques have been applied to rapidly prototype ASR systems reducing significantly user effort [WN05, HTRT06b]. Upon these approaches, a little manually annotated data is used to build rapidly an ASR system. Then, this initial ASR system is used to automatically transcribe a large amount of new speech data. These new annotated data is used to improve the underlying ASR models based on active and unsupervised learning techniques. To overcome the second difficulty, an interactive paradigm has been applied to reduce significantly the supervision effort of the speech transcriptions [LMR08, RRCV07]. Following this paradigm, the system produces automatically speech transcriptions and the user is assisted by the system to amend output errors as efficiently as possible.

This chapter introduces a novel speech transcription system in which active and semi-supervised learning techniques are applied along with the interactive paradigm in a tightly coupled manner. The main goal is to significantly reduce the human effort in the transcription of a speech task by allowing a maximum tolerance error in the resulting transcriptions. The supervision is performed whenever an estimation of the transcription errors is higher than the tolerance error. Word-level confidence measures computed over the recognition word-graph are used to suggest which words should be supervised [WSMN01b]. During the transcription process, supervised and high-confidence parts of the transcriptions are used to improve incrementally the underlying statistical ASR models. This will likely improve the subsequent recognitions lowering the necessary number of supervisions.

This scenario has been already successfully applied to handwriting transcription [SSJ10] and here is adopted for speech transcription.

3.2 Interactive speech transcription balancing error and supervision effort

Interactive speech transcription systems places a human operator at the centre of the transcription process and embeds an ASR system within an interactive editing environment. The ASR system and the human transcriber tightly cooperate to generate the final transcription, thereby combining human accuracy with ASR efficiency. Different approaches have been proposed to reduce user effort for the purpose of obtaining completely accurate transcripts of speech [LMR08, RRCV07]. Here, however, the focus here is on reaching a balance between the user effort on supervision and the final residual transcription error for what the procedure on the following section is proposed.

3.2.1 Interactive Speech Transcription Procedure

At the beginning, the speech task to transcribe should be split into several smaller audio blocks, each one containing several utterances (samples). Moreover, the user must decide the acceptable minimum quality on the final transcriptions by establishing a maximum tolerance error (W^*). Then, automatic transcription, interaction and learning must take place sequentially from the first to the last block as follows:

0. **Initialization**:

a) The task should be split into smaller audio blocks.
(*Preferably, each one containing several utterances/samples*).

b) The user must establish the acceptable minimum quality on the final transcriptions by means of a tolerance error (W^*).

The semi-supervision takes then place iteratively, block after block:

1. **Recognition**: A block is transcribed using the last trained models.
Unless provided, the block should be segmented into smaller utterances.

2. **Interaction**: For each new automatically transcribed utterance:

a) $\hat{W}^-$ is estimated including the whole utterance under study.
(*With $\hat{W}^-$ the predicted error over the yet unsupervised parts*).

b) While $\hat{W}^- > W^*$:

i. The user is asked to supervise the next lowest-confidence word in the utterance.

ii. $\hat{W}^-$ is updated accordingly.

3. **Learning**:

a) New samples are extracted from the supervised and high-confidence parts, and added to the training set.

b) A new ASR system is trained using the available training data.

It should be noted that the method needs an initial ASR system at the beginning. Initial ASR models may be obtained from external resources similar to the task. However, this is not usually feasible mainly for acoustic models. There are more chances of building a suitable external language model since there exists a high amount of text data available. In any case, an initial ASR system can be built using a small part of the speech fully manually transcribed.

ASR models will be re-estimated each time after a new recognised audio block is semi-supervised. Models are expected to progressively adapt and perform better, because the training corpus is iteratively augmented with new samples obtained from the supervised and high-confidence parts of the last semi-supervised block. High-confidence parts are those words with a confidence measure greater or equal than an estimated confidence threshold (C_τ). The word-level confidence measure is based on the word posterior probabilities computed over the recognition word graph [WSMN01b].

3.2.2 Estimation of the Transcription Error

The transcription error can be measured in terms of the well-known Word Error Rate (WER). WER accounts for the average number of elementary editing operations needed to transform a faulty text into the correct reference text. The WER (W) of the transcription from the first transcribed utterance to the current about to be supervised can be defined as:

$$W = \frac{E}{N} \tag{3.1}$$

where N is the total number of reference words from the first utterance of the task up to the one currently being semi-supervised. And E is the sum of the edit cost of each utterance from first to current one, as well. The edit cost of a utterance is calculated here as the Levenshtein distance with unitary costs compared to the corresponding reference.

Assuming user corrections are flawless, then WER is only be due to the errors on the unsupervised parts:

$$W^{-} = \frac{E^{-}}{N^{+} + N^{-}} \tag{3.2}$$

where N and E has been decomposed into the contributions of the supervised $(+)$ and unsupervised parts $(-)$. E^{+} has been assumed to be zero.

The exact values of E^{-} and N^{-} are indeed unknown during the process. A simple estimation can be done assuming an uniform distribution as on supervised counterparts:

$$\hat{N}^{-} = N^{+}\frac{R^{-}}{R^{+}} \tag{3.3}$$

$$\hat{E}^{-} = E^{+}\frac{\hat{N}^{-}}{N^{+}} \tag{3.4}$$

where R^{+} and R^{-} are the number of recognised words which have been supervised and non-supervised, respectively, up to current utterance. And E^{+} is the accumulated edition cost of the supervised parts (up to current utterance) before corrections are made.

However, a better estimation of W^{-} can be achieved by classifying recognised words into C groups depending on its confidence measure [SSJ10]. Let the groups from 1 to $C - 1$ refer to the word with lowest confidence, second lowest, and so on respectively, in a utterance. And let Group C refer to the, high-confidence, rest of words not in the previous groups. With this modification (3.3) and (3.4) are expressed as follows:

$$\hat{N}^{c-} = N^{c+} \frac{R^{c-}}{R^{c+}} \tag{3.5}$$

$$\hat{E}^{-} = \sum_{c}^{C} E^{c+} \frac{\hat{N}^{c-}}{N^{c+}} \tag{3.6}$$

Finally, WER of the unsupervised parts (W^{-}) can be obtained using the estimations (3.5) and (3.6) for the unknown WER of the unsupervised parts (3.2) :

$$\hat{W}^{-} = \frac{\sum_{c}^{C} E^{c+} \frac{R^{c-}}{R^{c+}}}{\sum_{c}^{C} N^{c+} \left(1 + \frac{R^{c-}}{R^{c+}}\right)} \tag{3.7}$$

3.3 Experimentation

3.3.1 Experimental Setup

The proposed method was exhaustively evaluated for the task of transcribing the whole WSJ0 training set of the Wall Street Journal (WSJ) speech database [PFFG93], as well as the WSJ1 plus WSJ0 training sets of summing up 80 hours. A thorough battery of experiments to evaluate the impact of the number of blocks, the incremental learning versus external the language model, the number of CM ranks, the behaviour of the CM, etc. was conducted.

However, for the sake of brevity in this chapter only the WSJ0 experimentation for the optimal number of blocks and CM ranks is presented. Also it is considered only the case in which LM is externally provided and do, it remains unchanged during the IST process. This experimental setup is motivated to evaluate the method within the usual scenario in which LM training material is easily collected. On the contrary, acoustic models are incrementally built starting from empty models.

The WSJ0 training set was split into 12 blocks, each one containing utterances from 7 different speakers. Acoustic models were completely retrained every time the semi-supervision of a block was fulfilled. These Hidden Markov Models (HMMs) consisted of clustered word-internal 3-state triphones with a number of 16 Gaussian-mixtures per state.

Several tolerance values were tested: $W^* = 1\%$, 2%, 5%, 10%, 20% and 50%. On this corpus, these values correspond to roughly allow from one error every 3 utterances ($W^* = 1\%$) up to 9 errors per utterance ($W^* = 50\%$). Additionally, as a comparison baseline, $W^* = 0\%$, which is equivalent to perform a full supervision was also tested.

For each one of the tolerance values W^*, two different experiments were conducted using the WSJ standard trigram language models with $5k$ (LM-5k) and $20k$ (LM-20k) vocabulary sizes, respectively. Human effort to perform the task of transcription has been evaluated by measuring the percentage of supervised words relative to the recognised words. Also, the benchmark tests of the Nov'92 ARPA evaluations [PFFG93] were used to assess the incremental trained HMMs with the $5k$ (4.986) word closed vocabulary and $20k$ (19.979) word open vocabulary WSJ benchmark test tasks. The overall characteristics of the WSJ speech database are shown in Table 3.1. In addition, the overall difficulty of train and test sets in terms of the perplexity is shown in Table 3.2 for both language models (LM-5k and

Table 3.1: Main characteristics of WSJ0 speech database.

	Train	Nov'92-5k	Nov'92-20k
Duration (h)	15	0.7	0.7
Speakers	84	8	8
Utterances	$7k$	330	333
Words/Line	18±8	16±6	16±6
Run. words	$129k$	$5.4k$	$5.6k$

Table 3.2: Perplexity on train and test sets depending on the LM.

	Train	Nov'92-5k	Nov'92-20k
LM-5K	115	53	101
LM-20K	170	54	142

LM-20k).

Preliminary HMMs are needed to built the initial ASR system. Also, certain ASR parameters (i.e. Grammar Scale Factor and Word Insertion Penalty) have to be properly estimated. With this purpose, the first block was considered as fully transcribed manually and used to train initial HMMs and optimise ASR parameters. In addition, the optimal confidence threshold (C_τ) to select the high-confident parts was estimated by minimising the confidence error rate (CER) [WSMN01b] in this block.

On the other hand, the balancing method updates the values of its parameters after each user interaction. Nevertheless, it still requires a good initialisation ($\{\hat{E}_0^{c+}\}$, $\{N_0^{c+}\}$, $\{R_0^{c+}\}$, $\{R_0^{c-}\}$) in order to perform properly from the very beginning. These initial values were those resulting of applying the method itself on the first block.

Once a block was recognised, corrections were performed automatically by means of a simulation of a real user. This simulation made use of the time alignment of every reference word. The alignment was found by means of forced recognition with a ASR model trained using the whole training WSJ corpus. Given a recognised word (w) asked for supervision, the simulation consisted in guessing which sequence of correct words (or no words in case of a deletion operation) matched w.

3.3.2 Results

For clarity, only the results of the supervision effort and residual accumulated WER after the transcription of each block of WSJ0 training data using a tolerance error of $W^* = 20\%$ are depicted in Fig. 3.1. Results are shown using both LM-5k and LM-20k. The rest of the lines corresponding to other tolerance values had an identical tendency to $W^* = 20\%$, from fully supervision ($W^* = 0\%$) to almost no supervision ($W^* = 50\%$). As supplement, Table 3.3 summarizes results for some significant tolerance errors after the completion of the whole task: the relative number of supervisions (%Sup.); the final residual WER of semi-supervised WSJ0 training set (Wtrain), and the WER on the corresponding external test using the final resulting ASR system (Wtest). Thus, as expected, the more allowed error, the less user interaction was needed.

Both experiments, using LM-5k or LM-20k, yielded qualitatively the same behaviour. However, for LM-20k the reduction in user effort was greater. This was because LM-20k yields better recognition accuracy than LM-5k in WSJ0 training data.

The resulting quality in the transcriptions after semi-supervision is depicted in terms of the WER in Table 3.3 (column Wtrain). All experiments

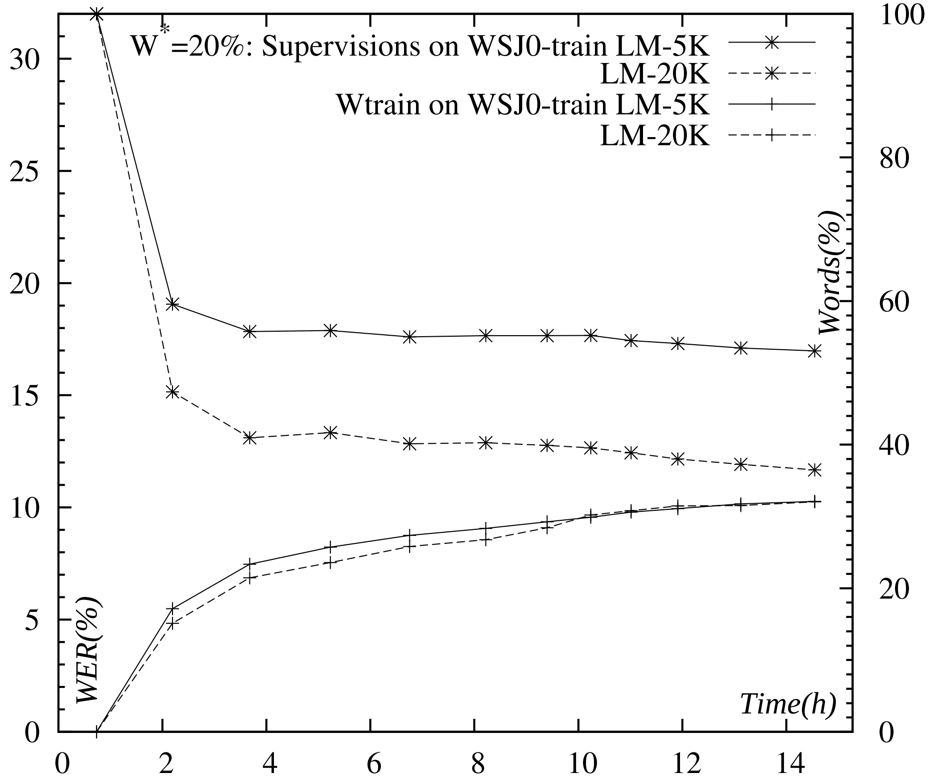


Figure 3.1: Transcription of WSJ0 training data using LM-5k or LM-20k for a tolerance of $W^* = 20\%$. Reduction of the user effort in terms of the number of supervisions relative to the recognized words in a block (at top). Residual accumulated WER after semi-supervision of each one of the blocks (at bottom).

Table 3.3: Final results for different W^* tolerances. %Sup. is the relative number of supervisions; Wtrain is the residual WER of the semi-supervised WSJ0 training set; and Wtest is the WER of the corresponding external test using the final resulting ASR system.

	LM-5k			LM-20k		
W^*	%Sup.	Wtrain	Wtest	%Sup.	Wtrain	Wtest
0	100.0	0.00	6.6	100.0	0.0	14.6
5	85.2	3.1	6.6	76.0	2.9	14.6
10	72.5	6.2	7.0	59.5	5.8	14.8
20	53.1	10.3	7.6	36.1	11.4	16.1
50	0.00	39.2	10.1	0.0	26.4	18.7

resulted in much better transcriptions than requested (WER close to the half to the requested W^* for all experiments). Thus, the system behaved in pessimistic way by requiring more user effort than a flawless predictor would. However, this is preferable rather than systems yielding poorer results than requested.

It should be highlighted that for a moderate error tolerance (20%), a great decrease in the user effort is achieved for both benchmarks (36% of supervisions for LM-20k and 53% for LM-5k). While, final output resulted of 10% of WER (less than 2 words per utterance) for the task of transcribing the whole WSJ0 training set.

The improving performance of the trained HMMs after each step is shown in Fig. 3.2 in terms of the WER on an external test (i.e. the Nov'92 benchmark tests). Here, results for 0%, 20% and 50% have been plotted. As supplement, Table 3.3 summarizes the WER on the corresponding external test using the final resulting HMMs for the rest of tolerance errors W^* (column Wtest). Tolerances between 20% and 0% do follow the same tendency.

[ht]

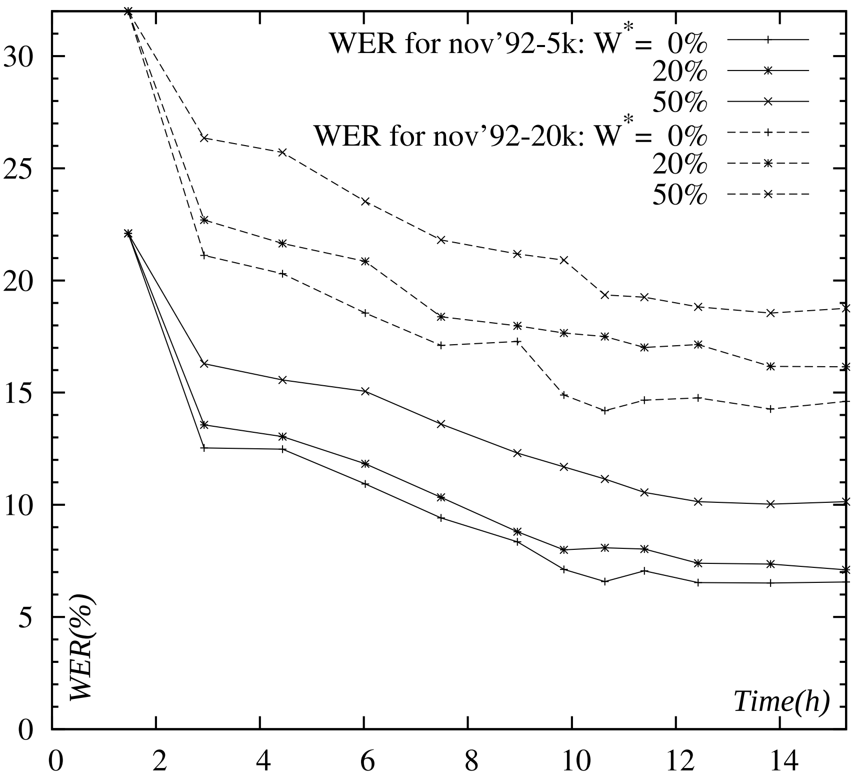


Figure 3.2: Improvement of the HMMs in terms of the WER of an external test for tolerances $W^* = 0\%, 20\%$ and 50%. Recognition of Nov'92-5k test using the HMMs learnt during semi-supervision of WSJ0 using LM-5k (Solid lines). Recognition of Nov'92-20k, using models learnt with LM-20k (Dashed lines).

Then, it can be stated that for low tolerances, the method results in HMMs almost indistinguishable to the model built using the full WSJ0 training set for both benchmarks. It should be noted that both $W^* = 50\%$ experiments also show a progressive improvement. This is still consistent with the mentioned fact that no user interaction was required for this tolerance, because the models are still improved using the high confidence unsupervised parts, and also thanks to the use of external language models suited for these tests.

3.4 Conclusions

In this chapter a simple method for interactive speech transcription has been introduced. Empirical results prove that the method is effective in finding an optimal balance between the resulting transcription quality and the supervision effort performed by an user. The desired quality of the resulting transcripts can be controlled by means of a tolerance error set by the user, and so the effort.

On the other hand, it should be highlighted that a very useful functionality arises from this method: ASR models can be improved with little supervision effort compared to the cost of manually transcribing full corpora. In fact, the incrementally learned models by means of this IST method yielded similar performance to those generated with full manual transcriptions.

3.4.1 General remarks on the behaviour of the IST method

- The estimation of the residual WER ($\hat{W}^-$) is always pessimistic because the estimation is based solely in the supervised low-confidence parts. Fortunately, this is preferable.

- Better performance of the CM would improve $\hat{W}^{-}$.
- The initialisation affects the performance, but it is easy to find.
- The update of the parameters of the recogniser and the confidence measure classifier yield no significant improvement, but it seems that in a general scenario it would.
- In practice, words incorrectly segmented pose a difficulty for the user. Although it might be alleviate enlarging the margins by just 15ms, as in the prototype on chap. 4.

3.4.2 Remarks on the large task WSJ1+0

The WSJ1+0 task has been elided from the results section for the sake of brevity. The following commentaries apply to that experimentation:

- As the task is larger, the effort is greatly reduced as it learns, especially for the incrementally-learned LM case.
- The recognition performance of the learned models is just slightly worse than using ASR models trained with fully manual transcribed data.

CHAPTER 4

IST PROTOTYPE

This chapter describes a prototype developed for IST on the basis of the framework proposed on the previous chapter. The prototype is oriented to off-line transcription of the speech.

The first section summarises the IST method including details which are specific to the implementation. Section 4.2 depicts the user interface and discusses its usability. Next, section 4.3 briefly discusses the demonstrations of the prototype in congress. Then, section 4.4 discusses about the usability. Section 4.5 arises the conclusions and proposes possible improvements for the IST framework and tool.

4.1 IST prototype balancing error and user effort

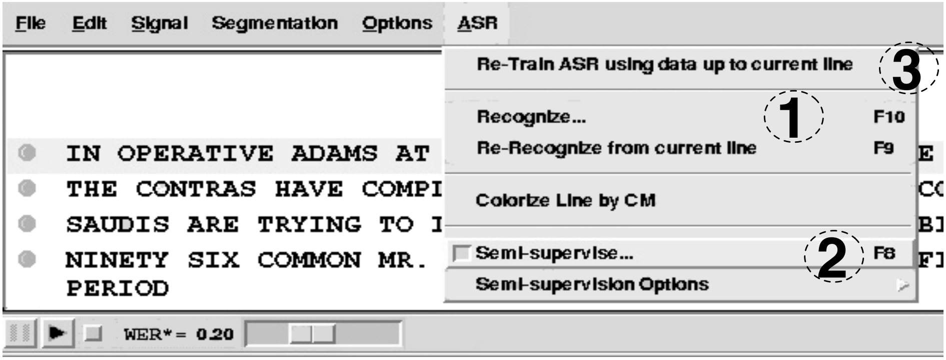
This section details how to implement the proposed IST framework as an extension for existing speech transcription tools. The procedure to semi-supervise the transcription must deal iteratively with three steps explained in 3.2: "Recognise", "Semi-supervise" and "Re-train". In summary, the user will first load a block of the speech. Then, the user will activate "Recognise". Once recognised a part of the speech, the minimum acceptable quality of the final transcript will be established by a maximum tolerated Word Error Rate (WER), W^*. After that, the user will start "Semi-supervise" in order to be assisted by the system. Additionally, at anytime, the user can make the system to "Re-train" the ASR model in order to improve subsequent recognitions.

The "semi-supervision" procedure is that detailed on 3.2. However, let us recall here the procedure again to outline some additional details specific to the implementation:

Sequentially in a word by word basis for each recognised segment:
The error, $\hat{W}^{-}$, of the unsupervised parts of the transcriptions up to the current utterance under revision is estimated If $\hat{W}^{-} > W^{*}$:

- The system asks the next lowest-confidence word in the utterance:
  - The system catches the attention of the user over the new word. This is important, because the supervision is not performed sequentially.
  - The region to supervised is highlighted in the text-box, as well as in the audio graph. The audio portion is enlarged by a fixed margin of 15 ms to avoid chopping.
  - The corresponding audio is played once by default.
- The user can replay the audio, validate the word or type the proper correction. However, the user is also let to correct the surrounding words to the current under supervision. This is explained below in section 4.4.
- $\hat{W}^{-}$ is updated accordingly. If $\hat{W}^{-} < W^{*}$ it proceeds to next utterance.

Add new samples to the training data using the supervised and high-confidence parts of the utterance.



IN OPERATIVE ADAMS AT

THE CONTRAS HAVE COMPI

SAUDIS ARE TRYING TO I

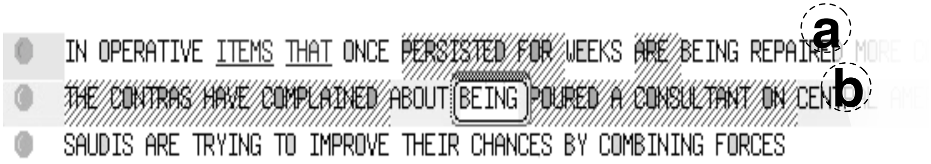
NINETY SIX COMMON MR.
PERIOD

Figure 4.1: Screenshot of the ASR menu addition. 1) Recognition. 2) Interaction. 3) Re-training.

4.2 User interface of the prototype

The prototype is implemented as an extension to the *Transcriber tool* [BGWL98]. This extension is incorporated as an additional menu entry called "ASR" . This menu lets the user to follow the proposed interaction paradigm (see fig. 4.1):
(1) First, recognise speech automatically ,
(2) Start/stop the assisting process using a method that allows up to some degree of errors in the final transcription.
(3) Finally, Re-training the ASR models with the semi-supervised utterances.

To start the recognition, the user must press the "F10" key or to select the "Recognise..." entry in the ASR menu. Then, a dialog for choosing a configuration file will appear. This configuration file defines the paths to the ASR model files; optimal configuration for the recogniser; word classifier



IN OPERATIVE ITEMS THAT ONCE PERSISTED FOR WEEKS ARE BEING REPAIRED MORE C

THE CONTRAS HAVE COMPLAINED ABOUT BEING POURED A CONSULTANT ON CEN L AME

SAUDIS ARE TRYING TO IMPROVE THEIR CHANCES BY COMBINING FORCES

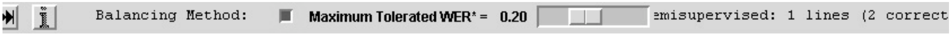
Figure 4.2: Outline of an screenshot of the textbox with the transcriptions. Shaded regions: Words recognized with low confidence measure values. Circled word: Word under supervision.

and WER prediction method parameters; and the list of the audio files to recognise. These samples should be segmented at sentence level. After all, the recognised utterances will then appear in the application.

The "F8" key starts the correction guidance (interactive transcription): The text cursor is placed to the right of the word the system has decided that it deserves to be supervised ; The word background flashes in a striking colour ; And the audio graph shows the corresponding region of the utterance corresponding to the selected word . After that, the word remains highlighted (see fig. 4.2). It should be noted that has been published that highlighting low confidence words helps user in the detection of the errors when the confidence estimation are correct ([VK08])

The selected audio is automatically played once. It should be noted that the enlargement of the piece of audio to be played influences the user decisions, especially for the number of insertions to be done.

Then, the user can enter the proper correction. At the same time, the user can re-play the piece of the audio as many times as required by pressing the tabulator key. Once finished, the user should press the enter key. The correction is underlined, and it assigned the highest possible confidence.



Balancing Method: ■ Maximum Tolerated WER* = 0.20 emisupervised: 1 lines (2 correct

Figure 4.3: Screenshot of the textbox with the transcriptions. Shaded regions: Words recognized with low confidence measure values. Circled word: Word under supervision.

After each correction, the estimator of the residual WER (W^{-}) is updated. Then the system will ask the user to correct the next lowest confidence word in the sample, or it will jump to the next utterance if the estimator is below the tolerance error threshold (W^{*}). It should be noted that the required quality on the transcription was specified in the configuration file in term of W^{*}. This threshold can also be set by the user using a slider at any time (fig. 4.3).

Finally, in order to improve the recognition of remaining utterances of the task, the speech recogniser model can be improved by retraining models. New data pairs of audio an text will be added to the training set. These pairs are the built from the contiguous subsegments not in maroon background (i.e., only those supervised and high-confidence parts): Once a new model is available, the remaining utterances can be re-recognised. This should help in reducing the number of corrections the system will ask to the user.

4.3 Demonstration

The prototype was showcased and demonstrated on a special session during the IUI 2012 congress. A small speech corpus allowed the attendants to test our interactive speech transcription prototype. The corpus was intended to be automatically recognised in a laptop in very short time. It posed some typical recognition errors. Users were able to perform the corrections while being assisted by the system, to change the tolerance error, and to re-train the models to check how the recognition has improved and it whether it has learned new vocabulary.

4.4 Usability

No formal tests on the usability of the interface have been conducted. However, the overall impression of the authors and other non-professional transcripts who tested the demo at the congress was satisfactory. The main concerns came from a small percentage of the words asked for supervision which segmentation was wrong. This mainly happened for the extra inserted words (i.e. when a deletion operation should be performed). Even so, although non performed deletions increase the resulting WER, they are preferable for a real user from the language understanding point of view. Also, incorrectly placed extra words are preferable for indexing and term search applications that would use the resulting transcriptions. Instead, the omission of words is a critical issue.

In order to avoid the most of the chopped words an enlargement of the portion of the audio to be played of 15 ms was found to be the best option. Nevertheless, it would be desirable a cleverer way to modify the margins in order to increase the chances of understanding.

As stated on the evaluation section, a user should attach strictly to some conventions. For instance, a user should deleted a word whenever less about the half of the word is uttered even if it can be figured out. However, while this servers properly for the evaluation purposes, allowing corrections more than those for the asked words to be supervised greatly improves the user experience and the final quality of the transcriptions. This behaviour was implemented by simply performing an additional estimation of which parts of the introduced corrections corresponded to which words in the utterance under supervision, instead of assuming the correction corresponds strictly to the word under supervision. The estimation is performed by means of the of the Levenshtein algorithm . The update of the $\hat{W}^{-}$ estimator is performed as if the extra corrected corresponding recognised words would have been asked for supervision, and that the corresponding corrections were introduced. It should be noted that this extra behaviour is usually useful because the user can rapidly figure out some corrections just by listening the word under supervision and reading the recognised surrounding words. This way, in case that more than one recognised word was corrected, the system will skip the supervision of those recognised words if they would have been later selected for supervision. Thus, in those cases no increase of the user effort would have happened.

Finally, it should be noted that for better comprehension of the speech, the more audio context the better and the more sequentially performed corrections , the better. Thus, this is an important issue that is left as a future work.

4.5 Conclusions

This chapter has introduced a prototype for IST on the basis of the framework proposed on the previous chapter. The prototype is built upon a manual transcription software which is still widely-used by professional transcribers. This tool showed that a great reduction in the effort can be achieved with this simple method.

Nevertheless, it should be highlighted that some aspects can be further improved: the issue of the isolated words; the estimation of the error; and the confidence measures. Exhaustive research on these and another issues was later conducted under the transLectures project. The solutions to address these aspects are introduced throughout the following chapters in this work.

CHAPTER 5

TRANSLECTURES: TRANSCRIPTION OF MASSIVE REPOSITORIES OF LECTURES

This chapter briefly summarises some of the main goals achieved within the European project transLectures to develop cost-effective IST solutions for online repositories with a massive number of educational lectures.

Following a short introduction, the rest of the chapter is organised as follows: the deployed architecture for IST in transLectures is explained over sec. 5.2. Section 5.3 remarks the main treats of the different repositories addressed with the transLectures solution. The evaluation sets used on chapters 6 and 7 were extracted from these repositories. Section 5.4 summarises the initial ASR results on poliMedia evaluation sets. The ASR models and transcripts were steadily improved later during the project as described on the next chapter. Finally, the conclusions are raised on 5.5.

5.1 Introduction

As stated on the introductory chapter, online educational repositories of video lectures are rapidly growing. As so, these sites are becoming a capital resource for high education reference. The value of the educational videos is increased even more when transcriptions and/or translations are available. In particular, the transcriptions not only make more accessible to speakers of foreign languages and to people with disabilities; but also they yield a better understanding for native speaker and they allow for further automatic analysis such as indexing of topics for later retrieval, classification, translation, summarisation, plagiarism detection, etc. Unfortunately, most lectures are neither transcribed nor translated because of the lack of efficient solutions to obtain them at a reasonable level of accuracy and cost.

For the purpose of yielding a cost-effective solution to the *Transcription and translation of video Lectures*, the European project transLectures[1] was established [JCAF$^+$12, dAGS$^+$14, SCdAG$^+$12, UXJK$^+$12, WIT$^+$14]. This project is relevant to this thesis work because of the contribution to the IST scheme, which helped to provide the massive number of transcriptions of sufficient quality to two huge online repositories: videoLectures.net and poliMedia. Both repositories will be detailed below on sec. 5.3. However, it should be outlined already the relevance of the videoLectures.net repository, since it is a free and open access web portal with more than 17882 lectures from 13273 authors at the time of the writing.

5.2 Deployed architecture for cost-effective transcription of online repositories

Given the conclusions reached for the IST prototype presented in this work (see 3.4), the three following aspects were addressed for the sake of the accomplishment of the so-mentioned goal of transLectures:

[1] The transLectures was an international cooperation project involving several research institutions and big companies on the field: Universitat Politècnica de València (Spain), Xerox S.A.S., (France), Jozef Stefan Institute, (Slovenia), Rheinisch-Westfaelische Technische Hochschule (Germany), European Media Laboratory GmbH (Germany) and Deluxe Digital Studios Ltd. (UK).

1. *Improvement of the ASR and* Statistical Machine Translation *(SMT) results by massive adaptation.*

   For this purpose, firstly specific tools were developed resulting in an state-of-the-art ASR system including deep neural networks, by the name of tLK ([dAGS$^+$14]). Secondly, and more importantly, a pipeline to adapt the ASR generic models to speaker and topic of the talk was implemented. One peculiarity exploited to achieve a great level of adaptation was the availability of time-aligned slides, which allows the successful recognition of topic-specific vocabulary.

2. *Improvement of the ASR transcripts by intelligent interaction.*

   Despite the huge improvement of the ASR technology, the only way to achieve transcripts of an acceptable quality is by means of the manual supervision/edition, as it was exhaustively justified on chapter 3. Consequently, the transLectures project deployed intelligent interaction with dedicated mechanisms oriented to scarce-collaborative non-professional users. This framework of interaction is much better suited to online scenarios rather than the conventional batch-supervision approaches, as demonstrated theoretically and empirically throughout the previous chapters.

3. *Integration into real-life online platform.*

   The described mechanism was integrated into the *matterhorn* platform; which is a well-known open source project for delivering online media. Matterhorn is a very complex platform comprising from the very "ingestion" of massive quantities of media files, media format conversions, etc. up to the delivery ([Ope12]). Being so, the IST mechanism was evaluated empirically within a real-life challenging context. A good overview of the superb product can be quickly grasped playing with the real application at `https://www.translectures.eu/player/random.php`.

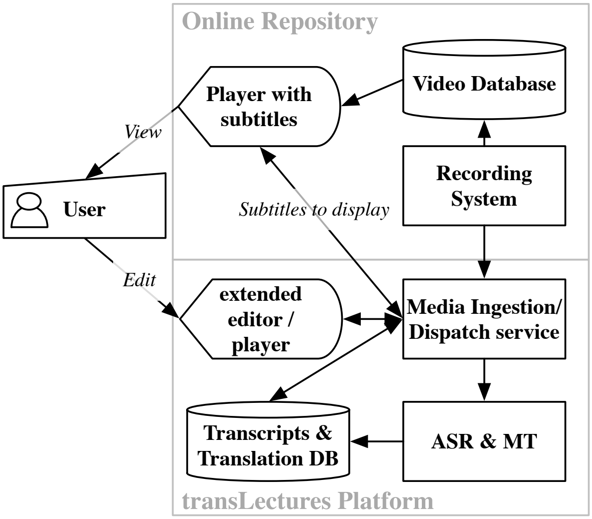


Figure 5.1: TransLectures architecture overview.

The resulting architecture implementing the aspects cited above is depicted on figure 5.1 ([SCPJ+13] for further detail). Amongst other advantages, this scheme is thought to work as a distributed architecture. That is, each component may be deployed on a different machine.

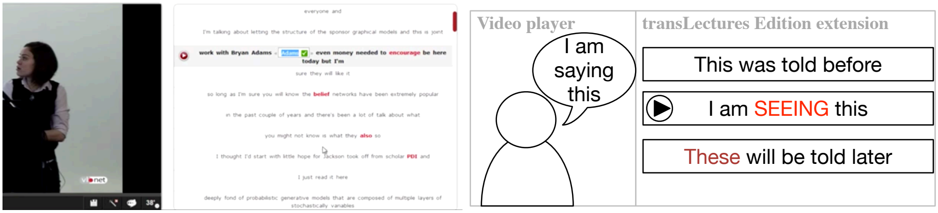
In brief, the system allows two main operational modes: just viewing (which can be performed using the default repository web player) or editing the automatic captioned transcripts. Within any of the two modes, the captions for the selected language are provided by the transLectures ingest/dispatch system, as well as the list of the available languages.

While watching a lecture in editing mode, the user may decide to introduce corrections if they estimate so. The capabilities of the editing mode are presented on sec. 5.2.1. The manual corrections are then sent back to the service which will append them to the original automatic transcript or translation stored in the transLectures database. For this purpose, the *Distribution Format Exchange Profile* (DFXP) is used. file format is used in order to be able to track the history of modifications made by users and the automatic systems, allowing the player to show the best captions available for every segment. Finally, the corrections are also committed to the ASR and MT systems which can use this modifications to improve the underlying models (sec. 5.2.2).

On the other hand, the transLectures system needs to be permanently synchronised with the video repository in order to provide transcriptions and translations for any newly recorded videos. For this purpose, the transLectures web service provides a lecture upload service, known as *ingest*, which is used by the recording system. Then, once a new lecture has been uploaded to the transLecturesweb service via the ingest interface, the transcription of this lecture, and its subsequent translation into different languages, is carried out by the ASR and SMT systems.

5.2.1 the transLectures Player

The transLectures player is an PHP/HTML5 video player and caption editor carefully designed to easing the task of manual correction with minimum for any voluntary user which watches the video. Figure 5.2 shows an actual screenshot and an outline of the implementation of the player. As it can be seen on the figure, the approach is very similar to the prototype for a standalone user introduced in chapter 4. However, this interface is more mature, as a result of previous experience, experimentation and direct feedback of hundreds of real users (mainly teachers from UPV).



everyone and

I'm talking about letting the structure of the sponsor graphical models and this is joint

work with Bryan Adams Adams ✓ **even money needed to encourage be here today but I'm**

sure they will like it

so long as I'm sure you will know the **belief** networks have been extremely popular

in the past couple of years and there's been a lot of talk about what

you might not know is what they **also** so

I thought I'd start with little hope for Jackson took off from scholar **PDI** and

I just read it here

deeply fond of probabilistic generative models that are composed of multiple layers of stochastically variables

Figure 5.2: (Left) screenshot of the tLplayer with editing capabilities. (Right) Outline: The currently played segment of the speech is highlited, and the user can type directly corrections. Low confidence recognised words are highlited in red.

Amongst other commodities, the transLectures player presents the captions with an investigated optimal length of the segments in words and duration, making more likely for the user to spot an error and introduce it without having to stop the play nor re-listening. Also, three alternative editing layouts are offered to fit user preference. And of course, two different interactions schemes are possible much as it was within the presented extension for *transcriber*: batch interaction, in which the user can freely supervise any segment of the video; and intelligent interaction, in which the player asks the user to correct only those words considered most likely to be incorrect by the system. Additionally, a complete set of key shortcuts facilitate the user control over the system.

5.2.2 ASR & SMT Systems

The ASR system used for the automatic transcription and model retraining is the state-of-the-art transLectures-UPV open source toolkit, tLK [dAGS+14]. On the other hand, the SRILM toolkit [Sto02a] is used to estimate the LM part of the ASR system.

As commented before, one of the successful ideas of the transLectures approach is the improved ASR performance due to the adaptation performed using information about the speaker, as well as LM tuned to specific vocabulary found within the slides and other related textual resources.

On the other hand, the SMT engine is the widely-used open source Moses toolkit [KHB$^+$07]. Of course, the automatic translation of the lectures into another language are generated from its automatic transcript. For what, the translation accuracy may be very low unless the source transcript has enough quality.

Finally, it should be noted that both, the transcriptions and translations of video lectures in videoLectures.net and poliMedia, are automatically regenerated every time a major upgrades of the ASR or SMT system is achieved. For this reason, the repository's overall transcription and translation quality is constantly improving. The upgrades might be the result of better acoustic, translation or language models, or of new ASR and SMT techniques.

5.3 Overview of the videoLectures.net and poliMedia repositories:

This section describes the main repositories which were the focus of the transLectures project: the videoLectures.net and the poliMedia repositories.

The videoLectures.net repository is a free and open access repository of video lectures founded on 2001 which is being used as an educational platform for several EU-funded research project, as well as other various open education resource organisations such as the OpenCourse-

	English	Slovenian	Spanish	Total
Authors	6900	1347	734	8981
Lectures	9720	1103	5056	15879
Avg. lecture duration (min)	45	45	9	34
Transcribed lectures	111	0	7	118
Lectures with slides	7013	0	734	0

Table 5.1: Basic statistics on the video lectures considered in transLectures.

Ware Consortium, MIT OpenCourseWare and Open Yale Courses; as well as other scientific institutions like CERN. The videoLectures.net delivers high-quality-educational videos, as well as other accompanying documents such as time-aligned presentation slides. This site can be visited at `http://videolectures.net`.

On the other hand, poliMedia is a more recent project aimed at the distribution of multimedia educational content at the UPV. It is designed primarily to allow UPV teachers to record lessons of short duration (up to 10 minutes). And, in contrast to videoLectures.net, the sound of the video recordings are of exquisite quality since they are always filmed at specialised studios. Also as a requirement, the videos are always accompanied with time-aligned slides. This site can be visited at `https://media.upv.es`.

The transLectures project focused the experimentation on English and Slovenian video lectures in the videoLectures.net repository and on lectures in Spanish from the poliMedia repository. For an idea as to the complexity of the task, some basic statistics have been provided in Table 6.1.

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

Table 5.2: Statistics for the evaluation partitions from poliMedia.

5.4 First results on poliMedia

The UPV obtained its first results for poliMedia working from a set of 115 hours of video lectures manually transcribed using the Transcriber tool [BGWL00]. From this set, a standard partition was defined with three speaker-independent sets: training, development and test. This will allow ongoing scientific evaluation throughout the project. Table 5.2 shows the basic statistics for this standard partition.

The baseline UPV ASR system is based on the RWTH ASR system [LGH+07, RGH+09] for acoustic modelling and the SRILM toolkit [Sto02b] for language modelling. The RWTH ASR system includes state-of-the-art speech recognition technology for acoustic model training and recognition. It also includes speaker adaptation, speaker adaptive training, feature extraction for audio files, unsupervised training, a finite state automata library and an efficient tree search recogniser. For its part, the SRILM toolkit is a well-known language modelling toolkit used across different natural language applications.

The audio data was extracted and preprocessed from the videos, in order to then extract the mel-frequency cepstral coefficients (MFCCs) [SS12]. Then, monophoneme and triphoneme acoustic models were trained by adjusting different parameters such as the number of states, Gaussian components, leaves of the triphoneme CART, etc. in the development set. The lexicon model was obtained by applying phonetic transliteration to all of the training vocabulary words. An n-gram language model was trained on the transcribed text after filtering out functional symbols such as punctuation marks, silence annotations, etc. Meanwhile, external resources were used to enrich the in-domain language model. Specifically, we considered the linear combination of our in-domain language model with a large, external out-of-domain language model computed from the Google n-gram corpus [**?**]. A single parameter λ governs the linear combination of the poliMedia language model and the Google n-gram model, which is optimized in terms of perplexity on the development set.

The entire system, including the acoustic, lexicon and language models, was trained on the poliMedia training set. The ASR system parameters were optimized in terms of word error rate (WER) on the development set. A significant improvement of more than 5 WER points was observed when moving from monophoneme to triphoneme acoustic models. Triphoneme models were inferred using the conventional CART model using 800 leaves. In addition, other parameters obtained to train the best acoustic model included 512 components per Gaussian mixture, 4 iterations per mixture, and 5 states per phoneme. The in-domain language model was an interpolated trigram model with Kneser-Ney discount. Higher and lower order n-gram models were also assessed, but no better performance was observed.

From the triphoneme ASR system, several refinements to the language

System	WER	OOV
Monophoneme Model	44.6	5.6%
Triphoneme Model	39.4	5.6%
+LM Interpolation	34.6	5.6%
+Google 50K Vocab	33.7	3.5%

Table 5.3: Test-set WER for several ASR system refinements.

model were evaluated. The in-domain language model trained on the poliMedia corpus was interpolated with the out-of-domain Google n-gram corpus [**?**]. These two language models were interpolated in order to minimize perplexity in the development set, using an approximate λ value of 0.65 for the in-domain language model and of 0.35 for the out-of-domain language model. Two interpolations were performed using different vocabulary sets, the first containing only vocabulary matching poliMedia ("LM Interpolation") and a second made up of the poliMedia vocabulary plus the 50,000 most frequent words in the Google n-gram corpus ("Google 50K"). The final experimental results in terms of WER in the test set are shown in Table 5.3.

As shown in Table 5.3, there is a significant improvement by 5.7 WER points over the triphoneme system when the language model interpolated with the "Google 50K Vocab" vocabulary set is applied. As expected, the decrease in WER is directly correlated with the number of out-of-vocabulary words (OOVs) in the test set, since the Google n-gram corpus provides a better vocabulary coverage. A similar trend is observed when comparing perplexity figures for the triphoneme system with those observed for the "LM Interpolation" system. Specifically, perplexity drops significantly from 235 to 176 simply by interpolating our in-domain language model with the Google n-gram language model containing poliMedia vocabulary alone.

5.5 Conclusions

This chapter has introduced the approach to the architecture deployed for the ambitious international transLectures project. This approach solved the huge leap between the offline IST approach on the previous chapter and the cost-effective IST platform integrated into online video repositories.

Also, the two massive repositories to be transcribed within the project are detailed. The intense research conducted within the transLectures-UPV team led to the improvements for intelligent interaction and confidence measures presented on the following chapters 6 and 7.

CHAPTER 6

INTELLIGENT INTERACTION WITH LIMITED USER EFFORT

This chapter addresses the interactive transcription over online repositories with a massive number of videos. For this purpose, refined interaction methods and additional automatic improvement are proposed and are evaluated on real well-known repositories.

Following the introduction motivating the urge for improved interaction methods, the intelligent interaction approach is introduced in sec. 6.2. Sec. 6.3 formulates the method to make an ASR system to match manual corrections. After that, experimental evaluation is presented on sec. 6.4. And finally, conclusions are raised on sec. 6.5.

6.1 Introduction

Online multimedia repositories are rapidly growing and becoming established as fundamental knowledge assets. This is particularly true in the area of open education, where large repositories of video lectures are being built on the back of increasingly available and standardised infrastructure. For instance, Superlectures [sup], Videolectures [Vid], Polimedia [dV12] and Coursera [cou] are good widely-known examples of online repositories. Nonetheless, despite most of them being freely available, the access to this knowledge is hindered by language barriers and for people with disabilities ([PP12b]). This issue can be addressed with subtitles using the textual transcripts. Unfortunately, manual transcription of video lectures is usually unaffordable in terms of time and money; while *Automatic Speech Recogntion* does not produce accurate transcripts enough yet. For this reason hybrid automatic-manual approach known as *Interactive Speech Transcription* (IST) has raised as a very cost-effective plausible solution, as thoroughly justified throughout this thesis.

Chapter 3 presented a novel and effective IST solution, but oriented to a stand-alone use of a devoted user on the role of "supervisor". However, for repositories of video lectures with a massive number of media, this approach might be neither affordable, despite of the effective decrease of the time needed to supervise the automatic transcripts up to a certain degree of quality. On the contrary, open online repositories can count instead with a huge number of non-expert user to collaborate with the transcription and translation. As their help is altruist, the amount of effort and interaction is usually scarce. For this reason, in this chapter a new approach based on limited user effort is exploited to produce the best possible transcriptions is studied.

The proposed *Intelligent Interaction with Limited User Effort* (IILUE) in this chapter was developed within the European project transLectures [trab], which provided a cost-efficient solution to the transcription of live-repositories (see chapter 5). As so, the proposed approach was tested on two challenging tasks consisting on transcribing lectures in English (videoLectures.net) and Spanish (poliMedia), counting each with 16 000 and 9 000 lectures respectively.

6.2 IILUE approach

0. **Initialization**:

a) The task must be split into a number of blocks B.
(*Preferably, similar lectures and/or expounded by the same speaker should be evenly distributed over each blocks*).

b) The user can set during how much time T will collaborate.

From the first up to the BTh. block, the IILUE takes then place as follows:

1. **Recognition** The block is transcribed using the last trained ASR.
(*For production systems, all the non-supervised blocks might be recognised to yield improved transcriptions, instead of waiting to some collaborative user to supervise it.*)

2. **Interaction**: The user is asked to supervise a total time of T/B.
Thus no WER estimation is needed to stop the interaction phase. The parts to be supervised can be selected by means of:
-**Active Learning approach**: The least confident parts or words.

3. **Learning**: New samples from the supervised and high-confidence parts, are used to improve/adapt the current ASR system.

Automatic post improvement after the IILUE is completed:

4. **Constrained Search**: All blocks are re-transcribed with the last ASR system and using a technique to constrain the produced transcript to forcedly match the user interactions.

The procedure proposed above constitutes the limited user interaction approach. The differences between the balancing user effort method, proposed on the previous chapter, are underlined. In summary as a result of the whole process, automatic transcriptions are improved by distributing user effort efficiently for the supervision of lower confidence parts of transcriptions in case of adopting the active learning method. Moreover, automatic transcriptions have been also improved by using adapted models based on user corrections.

The new presented procedure introduces several advantages:

- First, the user can set in advance how much time wills to dedicate, which it makes it suitable for collaborative non-professional users, as commented before.
- Second, the WER estimation is not necessary anymore, decreasing the system complexity and assumptions.
- Third, instead of an active learning approach such as in the chapter before, simple Batch Interaction (BI) is also possible, i.e.. supervising a sequential portion of the speech. Of course, as it will be proved on the experimental section below, yields of course worse results than the active learning approach using the confidence measures. However, it can be viewed as an additional commodity for those users who do not feel comfortable with supervising isolated phrases or words in the speech.

On the other hand, it should be noted that the gained performance with this method can be optimal depending on the choice of an appropriate number of blocks. And also, how the similar content is evenly distributed amongst them to facilitate ASR adaptation to speakers and topics on the subsequent blocks. Within real-life online repositories the optimal distribution should be easy, since they are usually accompanied with relevant metadata such as the speaker, topic, course schedule, etc.

For further details on the procedure the reading of sec. 3.2 is recommended. The novel inclusion of user interactions-constrained ASR is detailed on the following section.

6.3 Constrained search for speech recognition

In order to constrain the ASR system to use the user interactions, the Viterbi decoding process must be modified (sec. 2.1). Let be M the number of corrections a user has entered for a given segment of speech and $C = \{\vec{c}^1 \cdots \vec{c}^M\}$ a list with each of the individual corrections. In turn, each correction is represented by a vector $\vec{c}^m$ that includes information on the types of possible corrections, to be explained below. The corrections can be included into the standard MAP-generative approach to the ASR as:

$$\hat{\vec{w}} = P(\vec{w} \mid X, C) = \arg\max_{\forall \vec{w}} P(X \mid \vec{w}, C) \cdot P(\vec{w}) \tag{6.1}$$

Where it has been assumed that the language model $p(\vec{w})$ does not depend on the user interactions.

Each constrain c^m is a cuatruplet indicating the start (S^m), end (E^m) in the feature matrix X, the correct word (v^m) and the type of edit operation o^m. The acoustic model can be estimated as that which yields the maximum probability for the hypothesised word segmentation $\vec{K} =$

$(S'^1, E'^1, \cdots, S'^M, E'^M)$ (Viterbi approach):

$$P(X \mid \vec{w}, C) = \max_{\vec{K}} P(X_1^{S'^1-1} \mid \vec{w}^0) \cdot \prod_{m=1..M} P(X_{E'^m+1}^{S'^{m+1}-1} \mid \vec{w}^m) \cdot P(X_{S'^m}^{E'^m} \mid \vec{w}', c^m) \quad (6.2)$$

Where $\vec{x}_{S_{\prime M+1}} = T$ refers to the last feature in the X input vector. And $\vec{w}^m$ is the hypothesised sequences of words between the user input segments m and $m+1$.

Regarding the constrained regions of the model, it can be computed depending on whether it is a delete operation or not. In the case of o^m =deletion, v^m does not indicate the correct word, but on the contrary, the originally recognised by the previous system and that is incorrect:

$$\begin{aligned} P(X_{S'^m}^{E'^m} \mid \vec{w}', \{S^m, E^m, v^m, o^m = \text{del}\}) &= P(X_{S'^m}^{E'^m} \mid \vec{w}') \\ \text{unless } [S'^m, E'^m] \cap [S^m, E^m] &\neq \emptyset \text{ and } \vec{w}' \neq v^m. \\ P(X_{S'^m}^{E'^m} \mid \vec{w}', \{S^m, E^m, v^m, o^m \neq \text{del}\}) &= P(X_{S'^m}^{E'^m} \mid v^m) \\ \text{if } [S'^m, E'^m] \subseteq [S^m, E^m] &\text{ and } \vec{w}' = v^m. \\ P(X_{S'^m}^{E'^m} \mid \vec{w}', \{S^m, E^m, v^m, o^m\}) &= 0 \text{ otherwise.} \end{aligned} \quad (6.3)$$

For further alleviation of the computational cost[1], it can be assumed that $P(X_{S'^m}^{E'^m} \mid v^m) = 1$ or a value depending on the distance between the hypothesised and the interaction segment. Moreover, this approximation solves the case in which $P(X_{S'^m}^{E'^m} \mid v^m)$ cannot be computed accurately because the introduced word v^m is not transliterated into the proper sub-units, such as with foreign words.

[1] The $P(X_{S'^m}^{E'^m} \mid v^m) = 1$ approx. has been considered at the time of the writing of this work. Thus, it was not applied on the constrained search during the evaluation.

6.4 Evaluation

This section describes the evaluation of the *Intelligent Interaction with limited User Effort* (IILUE) approach undertaken over videoLectures.net [Vid] and poliMedia [dV12] real-life repositories.

6.4.1 Experimental setup

The videoLectures.net repository is a free and open access educational video lectures repository. The recorded lectures are mostly given by distinguished scholars and scientists at important conferences, summer schools, workshops, etc. Currently, videoLectures.net hosts more than 16.000 lectures mainly spoken in English. On the other hand, poliMedia is an innovative service for the creation and distribution of multimedia educational content at *Universitat Politècnica de València* (UPV). poliMedia is designed primarily to allow UPV professors to record their courses in video blocks lasting up to 10 minutes, accompanied by time-aligned slides. Currently, poliMedia hosts more than 9.000 lectures (mainly in Spanish) from 1.300 speakers with a duration of 2.100 hours. Table 6.1 summarises the main statistics of the evaluation tasks used in the experiments.

The IILUE approach has been tested using different English and Spanish ASR systems which were increasingly improved throughout transLectures. Consequently, the II approach has been evaluated in different settings in which the ASR performance was even better. All these systems were built using the tLK toolkit [dAGS$^+$14].

The proposed IILUE approach has been empirically tested on VL and PM during a three-year European research project, using ASR systems of

Table 6.1: Statistics of the videoLectures.net and poliMedia evaluation tasks.

	VL	PM
Full repositories:		
Language	English	Spanish
Lectures	17.4k	14.3k
Videos	20k	14.3k
Duration (h)	13.3k	2.7k
Test sets:		
Videos	4	23
Speakers	25	5
Duration	3.4h	3.4h
Running words	34k	31k

increasing accuracy. In VL, two systems were tried:

VL1 The first one was an hybrid system combining a context-dependent deep neural network plus hidden Markov models (CD-DNN-HMMs) [DMY+12a] trained on 439 hours of speech data from several datasets (mainly VL, EPPS, TED-LIUM and VoxForge). Language modelling was implemented as a linear interpolation of n-gram models trained on different corpora (Google ngrams v1, VL, TED-LIUM, Wikipedia, COSMAT, Hal, WIT3 and PM). The system used 16 MFCCs plus derivatives and fCMLLR features [DRN95]. The DNN consisted of four hidden layers with 3000 units per layer. Complete statistics for each corpus can be found in [traa]).

VL2 The second system was derived from the first by replacing its (monolingual) DNN by a combination of a multilingual DNN [TSN13] and a convolutional NN [AHDY13]. This multilingual extension allowed to enlarge the training data by including non-English (Spanish Agora [SF09] and

Catalan Glissando [GEA$^+$13] corpora) speech from other databases, up to 620 hours. The DNN consisted of six hidden layers with 3000 units per layer. The language model (LM) in both systems was a 200k linear interpolation of 4-gram models trained on different corpora (i.e. Google ngrams v1, VideoLectures.NET training part, TED-LIUM, Wikipedia, COSMAT, Hal, WIT3 and poliMedia) using the SRILM toolkit [SZWA11]. The individual 4-gram models were smoothed with the modified Kneser-Ney absolute interpolation method [KN95]. The linear weights were optimised in VideoLectures.NET development set.

The LM of the VL systems was adapted to each specific video lecture by extending the base vocabulary with words in the slides and the documents of the video. This LM obtained a perplexity of 144.3 on the evaluation data. The WER obtained by both systems on the test set of videoLectures.net was 24.8% in the former and 22.9% in the latter.

In the case of PM, five systems were built:

PM1 The first system consisted in a conventional HMM system based on Gaussian mixture modelling (HMM-GMM) with 1.745 HMM tri-phones, 3.342 tied-states and up to 64 Gaussians per state. Fast Vocal Tract Length Normalisation [WKN99] and constrained MLLR (CMLLR) [GGB04a] features were computed supporting speaker adaptive training. The system implements a two-pass recognition strategy in which CMLLR adaptation is applied in the second-pass decoding. It was trained on 96 hours of speech data from PM. Its language model was similar to that used in VL1 and VL2, though in this case it was trained from PM and several out-of-domain text corpora.

PM2 The second system was derived from the first by including a cluster-based cepstral mean and variance normalisation (CMVN) at video-level.

The number of Gaussians per state was increased up to 128.

PM3 The third one was an hybrid system combining a neural network and HMMs. The NN hidden layer had 4000 distributed processing linear units and it was used to classify the senones produced by the second system. It also included the CMVN step at video-level.

PM4 This system was derived from PM3 by making its neural network deep (CD-DNN-HMMs system made up of four hidden layers with 3000 units in each layer).

PM5 The fifth system was obtained from PM4 by making its DNN multilingual [TSN13] which was combined with a multilingual CNN [AHDY13] (six hidden layers with 3000 units in each layer). As in VL2, this multilingual extension allowed us to enlarge the training speech data by using all available training data from PM (in Spanish and Catalan) as well as additional databases of speech data in Catalan.

The LM was built in an identical manner to the videoLectures.net case study. In this case, training data was composed of the manual transcriptions of the poliMedia training set along with several out-of-domain text corpora [MVdAAFJ14]. A perplexity of 153.6 was obtained in the development set using the resulting LM. The WER obtained by each system on the test set of poliMedia was: 22.9% (PM1), 22.1% (PM2), 18.7% (PM3), 14.9% (PM4) and 12.7% (PM5). Specific details about all these English and Spanish ASR systems are available in [traa].

6.4.2 Results

The IILUE approach was assessed against a baseline of no user interaction (NI). Additionally, within the proposed IILUE procedure an alternative interaction method was also assessed: the batch Interaction (BI). For BI the user must to supervise the automatic transcripts in a sequential order during a certain an amount of speech (time). Obviously, using BI the IST system cannot be considered intelligent, but it serves for a fair comparison of using the intelligent guidance and not.

Regarding the only two free parameter for the proposed method: the number of blocks B, it was used $B = 2$ for videoLectures.net because of the small number of lectures contained. In the case of poliMedia, video lectures were split into five blocks. About the user interaction time T, two different values were tested: one corresponding to the 5% of the duration of the test set and another 10% of the duration. These amounts are highly correlated with the user effort that will be devoted to supervising output as a percentage of total words. It is worth mentioning that user supervision was actually computer-simulated from the groundtruth, such as in sec. 3.3. That is, reference transcriptions were used to locate errors, if any, in speech segments selected by either the batch or CM-based selection strategies.

Figure 6.1 depicts the performance in terms of the WER of IILUE, the BI method and NI for each different ASR system described on the previous section. The WER values correspond to those over the final transcripts after the methods were applied. Table 6.2 shows the relative reduction in WER over NI.

From the table it can be spotted that IILUE clearly outperforms BI with a 3 times factor for a same level of user effort in both, videoLectures.net and poliMedia. The main reason for this superior performance of the IILUE is the because of use of a CM that spots mis-recognised words. On the contrary for the BI method, only in case that errors were uniformly distributed would the performance equal the intelligent approach. Moreover, it can be stated

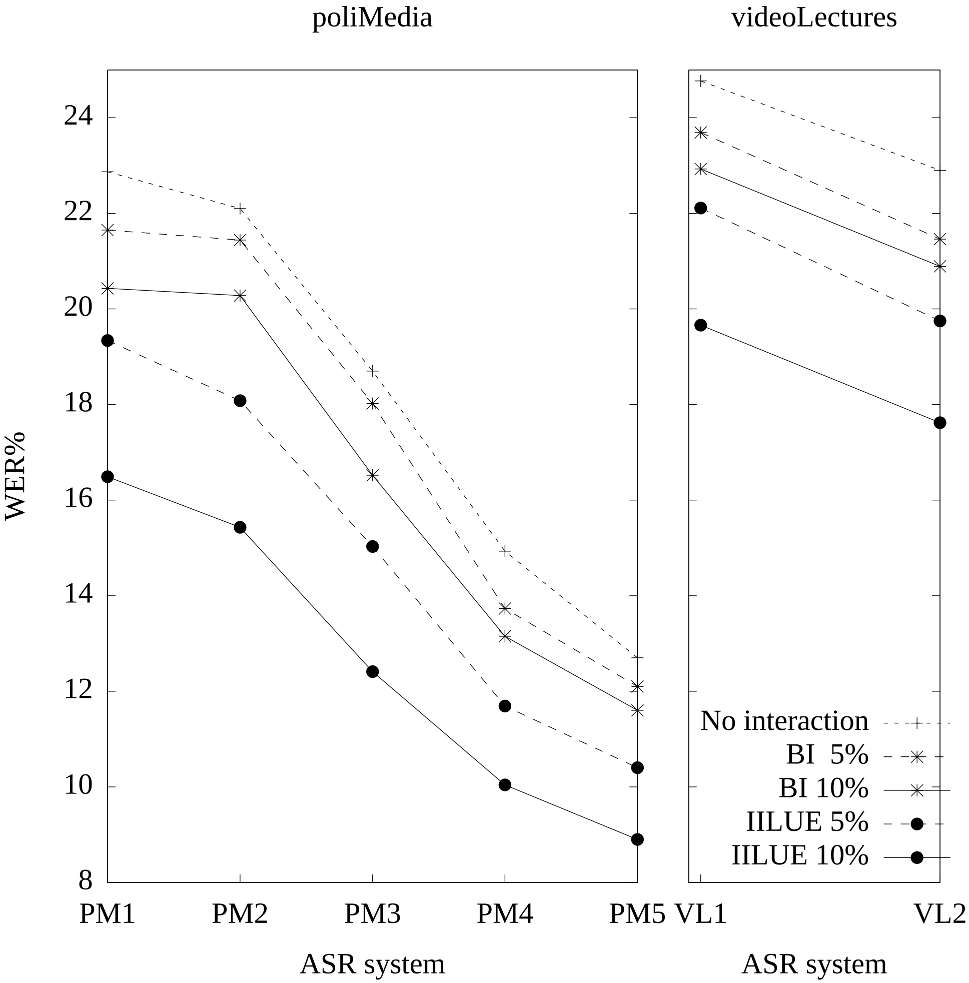


Figure 6.1: Improvement of the transcription in terms of WER for the poliMedia (left) and videoLectures.net (right) repositories over time. Intelligent Interaction (II) is compared with Batch interaction (BI) at two levels of supervision: 5% and 10%. No Interaction (NI) is the WER obtained by each ASR system.

Table 6.2: Relative decrease of the WER% depending the ASR system and repository. The improvement resulted the same in average for a given method, independently of the ASR.

Repository	ASR	BI 5%	BI 10%	IILUE 5 %	IILUE 10%
videoLectures	VL1	4	7	11	21
	VL2	6	9	14	23
poliMedia	PM1	5	11	15	28
	PM2	3	8	18	30
	PM3	4	11	20	34
	PM4	8	12	22	33
	PM5	4	9	18	30

that the behaviour of the IILUE approach is sound, since similar relative improvement is achieved independently of the ASR system and even of the repository despite their different complexity.

For further insight on how the IILUE approach behaves as a function of user effort, an additional experiment was conducted in which each video of the poliMedia test set was evaluated for each possible user effort, i.e. from 0% to 100%. In contrast to the previous experiments, neither adaptation nor CS-step was performed since supervision is carried out at video-level. Figure 6.2 shows, for each user effort in the X-axis, the percentage of the initial transcription WER which is remaining in each video after supervision (grey crosses). To better observe this trend, mean (black line) and standard deviation (red error bar) calculated using all videos is also plotted. Also, a diagonal line is plotted to simulate a random behaviour in which WER would be reduced proportionally to the user effort employed. It can be observed that II is mainly effective when the supervision effort is below 20% for which it yields a relative WER reduction from 40% and 60%. of reduction in WER. Once this point is reached, the impact of IILUE in WER reduction seems to be negligible. This confirms that the IILUE approach is very effective when the user effort is very limited (below 20%). This fact suites perfectly the assumptions over the collaborative users.

6.5 Conclusions

In this chapter,a more refined IST approach over that on previous chapters was proposed to deal with the particularities of the real-life online repositories with a massive number of videos. In summary, the introduced novelty respect to the balancing IST method is the change of objective, from a desired final WER to a desired maximum user effort. Thus, eliminating the necessity of a WER prediction. Also, the final transcripts are further improved automatically by means of a technique implemented into the ASR decoder that make the most of the user interactions. model.

The approach proves to be sound and useful no matter how good or bad is the initial underlying ASR system deployed. Empirical evaluation over two huge online repositories showed that the IILUE approach achieves relative reductions in WER that are approximately three times greater than other approaches based on post-edition effort.

However, IILUE still may suffer from other potential issues mentioned on the conclusions of the balancing methods. Thus, the improvement of the confidence measure (CM) to spot errors still have a important impact.

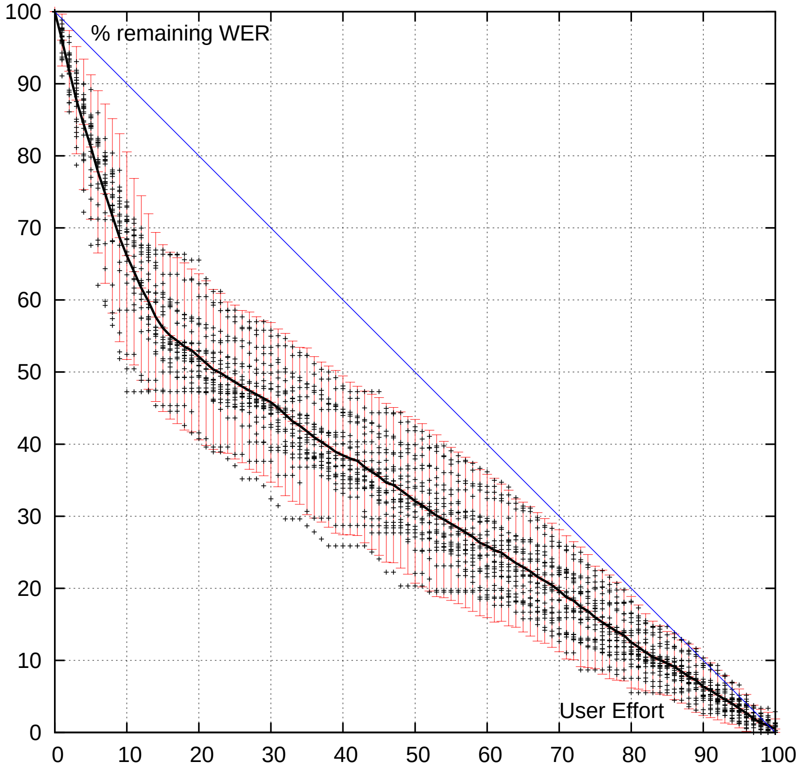


Figure 6.2: Percentage of remaining WER (from the initial transcription WER without supervision) after II for all possible user efforts on the poliMedia test set. Results are expressed for each video (gray crosses), along with the mean (black line) and standard deviation (red error bars) calculated using all videos.

CHAPTER 7

IMPROVED AND SPEAKER-ADAPTED CONFIDENCE MEASURES

This chapter details the new proposed method to compute metric called *confidence measure* (CM) to improve the interactive speech transcription system, as well as many other automatic speech recognition related application and the recognition itself.

The chapter is organised as follows: first, an introductory explanation of the problem is expounded throughout 7.1; the inclusion of speaker dependence into a baseline state-of-the-art Naive Bayes model is described on 7.2. Section 7.3 proposes a Logistic Regression model to benefit from error minimisation optimisation and formulates its corresponding speaker-dependent version. Section 7.4.2 describes the evaluation of the proposed models on two challenging tasks based on ASR transcripts from videoLectures.net and poliMedia repositories. Comparative results are presented including also Conditional Random Field (CRF) models. Section 7.5 proves that the increased CM performance results in better amended transcripts for videoLectures.net when integrated into an IST application. And finally, Section 7.6 raises the conclusions.

7.1 Introduction

Significant advances in the field of *Automatic Speech Recognition* (ASR) have been achieved over the last decades. Nowadays, automatic transcriptions of spontaneous speech in moderately noisy environments have reached an accurate enough quality ([Rou11, SNTW$^+$11a, SGR13]). This quality can be even better when ASR systems are adapted to specific scenarios ([LW95, Gal98, GL94, DRN95, WIT$^+$14, MVdAAFJ13]). Nonetheless, ASR is still far from producing error-free transcriptions and, consequently, its performance in many applications is not completely satisfactory.

To further improve the usefulness and performance of the current technology, researchers have proposed to compute a normalised score or *confidence measure* (CM) to indicate the reliability of the ASR output. This score has been computed at different levels: phoneme, word, phrase or sentence. Nevertheless, CM at the word level has been the main focus in the literature due to its usefulness for the vast majority of applications ([WSMN01a, KK05, GZXY09, BRG07, WLW$^+$10, JY11, YP13]).

One widely used word-level CM has been word posterior probability ([WSMN01a]). From then on, many works have focused on combining word posterior with additional sources of knowledge. The combination has been addressed as a classification problem in the vast majority of the works. Most well-known classifier algorithms have been tried: linear, Gaussian mixtures, neural networks, decision trees, support vector machines, etc. For further reference, a still good comprehensive survey can be found in [Jia05].

In the framework of CM as a classification problem, significant improvements were achieved by means of a combination of word-dependent (specific) and word-independent (generalised) *naïve Bayes* (NB) classifiers [SJV12].

Nonetheless, NB is learned by means of a generative criterion, the *maximum likelihood estimate* (MLE), which involves some issues. In particular, MLE overfits due to the unseen data. This issue was addressed in NB work by using a complex *backing-off* smoothing technique. But still, MLE aims at modelling the distribution underlying a given sample, which does not guarantee the solution to be the best suited for classification. Indeed, better fitted criteria may improve overall performance. For instance, the *maximum mutual information* (MMI) [HDH^{+}10] aims at better discriminating between classes without explaining the data. This criterion has been widely exploited in the literature for the *maximum entropy* (ME) models ([GS85, YLD11]).

Nevertheless, despite the success of MMI training in many applications, there is no direct relationship between maximising the MMI and minimising the probability of classification error. Instead, there are better suited criteria, which guarantee the minimisation of the *classification error rate* (CER) such as the *minimum classification error* (MCE) or the *mean squared error* (MSE). Therefore, a *logistic regression* (LR) model to be learnt by means of the MSE to surpass NB performance is proposed on this chapter.

On the other hand, speaker model adaptation has proved to be very effective for the improvement of recognition performance [LW95, Gal98, GL94, DRN95]. However, adaptation of the CMs to the speaker is nowadays unexplored. There is an increasing number of interesting scenarios in which CMs can be very useful and information about the speaker is available, such as the online lecture repositories. These repositories usually count with a large number of speeches delivered by a reduced number of speakers. Improving CM performance in these academic repositories is highly motivated since manual transcription is not affordable for such a large amount of speeches.

Moreover, ASR performance is usually poor due to the amount of technical concepts, very different native and non-native accents, etc. In this scenario, *interactive speech transcription* (IST) guided by CMs can help in massively producing acceptable transcripts for large amounts of videos with limited manual effort (see chapter 6).

Motivated by the scenario depicted above, this chapter also proposes an adaptation method of the CM models to the speaker in an attempt to improve CM classification and IST performance. The speaker adaptation is formulated for both extend both, the published NB and the proposed LR models.

7.2 Speaker-adapted naïve Bayes classifier

In this section, a speaker-adapted confidence estimator model is proposed. The model is designed to extend the *naïve Bayes* (NB) approach that was successfully applied to speech recognition [SJV12] as well as to machine translation [SJV07]. The speaker independent NB model to address CM of a word (w) as a classification problem is:

$$\hat{c} = \arg\max_{c} \, p(c|w,\vec{x}) = \arg\max_{c} \, p(c|w) \, p(\vec{x}|c,w) \tag{7.1}$$

With ($\vec{x}$) a vector of input scores. Also, it should be noted that Eq. 7.1 is obtained by applying Bayes' rule and ignoring the class-independent term.

It is worthy to note that the values of all the involved variables are assumed to be discrete. Discretisation avoids the need of explicitly modelling the probability distribution of continuous-valued features, while it renders a more flexible and data-driven model. Details on discretisation and several different approaches can be found in [Sei13]. On this work, the discretisation was computed by dividing the feature domain into a fixed number of evenly-spaced bins. The optimal number of bins was found on a development set.

The estimation of $p(\vec{x}|c, w)$ is usually biased due to the training data sparsity. More robust estimations can be obtained by simplifying the problem with the following strong independence assumption (the "naïve Bayes assumption"):

$$p(\vec{x}|c, w) = \prod_{d}^{D} p(x_d|c, w) \tag{7.2}$$

Therefore, the basic problem is to estimate $p(x_d|c, w)$ for each class-word pair and $p(c|w)$ for each target word. Given N training samples $\{(\vec{x}_n, c_n, w_n)\}_{n=1}^{N}$, these probabilities can be computed as the *maximum likelihood estimate* (MLE):

$$p(c|w) = \frac{N(c, w)}{N(w)} \qquad p(x_d|c, w) = \frac{N(c, w, x_d)}{N(c, w)} \tag{7.3}$$

where $\{N(\cdot)\}$ are suitably defined event counts on a given training data set. However, the MLE quickly overfit the training data. In order to prevent this overfitting, a backing-off smoothing method was introduced in [SJV12].

Now, the NB classifier above can be simply extended into a *naïve Bayes speaker-adapted* model (NB+spk). For that, a new variable s must introduced into Eq. 7.1 to identify the speaker:

$$\hat{c} = \arg\max_{c} p(c|w, \vec{x}, s) = \arg\max_{c} p(c|w, s) p(\vec{x}|c, w, s) \tag{7.4}$$

In order to alleviate the issue of the data sparsity, we assume mutual independence amongst the features as in Eq. 7.2. However, the conditional dependence between the speaker and the rest of the features is kept, since it may have an important impact on the classification. The MLEs are:

$$\hat{p}(c|w,s) = \frac{N(c,w,s)}{N(w,s)} \qquad \hat{p}(x_d|c,w,s) = \frac{N(c,w,x_d,s)}{N(c,w,s)} \tag{7.5}$$

Again, the issue of overfitting due to the hitherto unseen events must be addressed. For that, the following back-off following scheme is proposed:

$$p(c|w,s) = \begin{cases} \hat{p}(c|s) & \text{if } N(w) = 0 \\ p(c|w) & \text{if } N(w,s) \leq \mathcal{U}_{ws}\,,\ N(w) > 0 \\ p_s(c|w,s) & \text{if } N(w,s) > \mathcal{U}_{ws}\,,\ \exists c' : N(c',w) = 0 \\ \hat{p}(c|w,s) & \text{if } N(w,s) > \mathcal{U}_{ws}\,,\ \forall c'\, N(c',w) > 0 \end{cases} \tag{7.6}$$

$$p(x_d|c,w,s) = \begin{cases} p_s(x_d|c,w,s) & \text{if } N(c,w,s) > \mathcal{U}_{cws}, \exists x'_d : N(c,w,x'_d,s) = 0 \\ \hat{p}(x_d|c,w,s) & \text{if } N(c,w,s) > \mathcal{U}_{cws}, \forall x'_d\, N(c,w,x'_d,s) > 0 \\ p(x_d|c,w) & \text{ow.} \end{cases} \tag{7.7}$$

Thus, $p(c|w,s)$ will be estimated as in Eq. 7.5 only when the pair (w,s) has occurred above a threshold of times $\mathcal{U}_{ws}$ and the triplet (c,w,s) has occurred for each class. Whenever the second condition is not met, an absolute discounting smoothing analogous to that in the original NB classifier will be applied:

$$p_s(c|w,s) = \begin{cases} \hat{p}(c|w,s) - \frac{b}{N(w,s)} & \text{if } N(c,w,s) > 0 \\ \frac{b}{N(w,s)} & \text{ow.} \end{cases} \tag{7.8}$$

Instead, if the pair (w, s) occurs in the training, but not often enough, $p(c|w, s)$ will be estimated as $p(c|w)$. In case of an unknown word ($N(w) = 0$), $p(c|w, s)$ will be the MLE of $p(c|s) = \hat{p}(c|s) = \frac{N(c,s)}{N(s)}$.

In the case of $p(x_d|c, w, s)$, the frequency histogram in Eq. 7.5 will be used only whenever the triplet (c, w, s) is observed above $\mathcal{U}_{cws}$ times, and at least once for all possible values of x_d. However, if the latter condition is not met, the following smoothing will be used:

$$p_s(x_d|c,w,s) = \begin{cases} \hat{p}(x_d|c,w,s) - \frac{b}{N(c,w,s)} & \text{if } N(c,w,x_d,s) > 0 \\ M_{cws} \frac{p(x_d|c,w)}{\sum_{x'_d:N(c,w,x'_d,s)=0} p(x'_d|c,w)} & \text{ow.} \end{cases} \tag{7.9}$$

Where the mass $M_{cws} = \frac{b}{N(c,w,s)} \sum_{x'_d:N(c,w,x'_d,s)>0} 1$ is distributed according to the speaker independent probability $p(x_d|c, w)$. Nonetheless, if the threshold requirement $\mathcal{U}_{cws}$ is not met, the speaker independent model $p(x_d|c, w)$ will be used instead.

7.2.1 Estimation of the smoothing parameters

The smoothing parameters, b and $\{U_{(\cdot)}\}$), should be adjusted empirically on a separated data set to achieve generalisation. $b \in [0, 1]$ is discounted from every count and then distributed amongst the unseen events. $U_{(\cdot)} \in \mathbb{Z}^+$ are the thresholds which enable choosing the branch in the algorithm.

Nonetheless, unlike with the original NB model, here it is unfeasible to estimate all $U_{(\cdot)}$ by means of a grid exploration. To address this issue, two solutions were tested. The first option was to employ a common threshold per speaker, K_s, such that $U_{ws} = K_s$ and $U_{cws} = K_s/2$ for each possible class, w and s.

The second option consisted in a more sophisticated algorithm to automatically estimate all the possible thresholds: for the words seen in development and training, the $U_{(\cdot)}$ are set depending on the overall contribution of their corresponding speaker-dependent branch compared to the speaker independent:

$$\mathcal{U}^*_{cws} = \begin{cases} \infty & \text{if} \displaystyle\sum_{n:(c=c_n, w=w_n, s=s_n)} (p(\vec{x}_n \mid c_n, w_n, s_n) - p(\vec{x}_n \mid c_n, w_n)) \ > 0 \\ 0 & \text{ow.} \end{cases}$$
$$\mathcal{U}^*_{ws} = \begin{cases} \infty & \text{if} \displaystyle\sum_{n:(c=c_n, w=w_n, s=s_n)} (p(c_n \mid w_n, s_n) - p(c_n \mid w_n)) \ > 0 \\ 0 & \text{ow.} \end{cases} \tag{7.10}$$

For the training words not appearing in development, the thresholds are set similarly by considering development samples with identical $N(\cdot)$ value.

7.3 Speaker-adapted logistic regression confidence estimator

In this section, a new CM model based on *logistic regression* (LR) models is introduced. It is worth noting that, for binary classification problems such as the CM problem considered in this work, these models are equivalent to the more general *conditional random fields* (CRF). After describing the proposed speaker-dependent LR models, section 7.3.1 discusses how to discriminatively learn them using the MSE training criterion. After that, on section 7.4 this criterion is empirically compared with a similar yet different criterion that is commonly used in CRF training.

The proposed approach resembles the ones presented in [ESJV08]. However, that work formulated the classification problem as a generative model, and only the posterior of the features was attempted to be learnt in a discriminative way. Furthermore, the purpose was to mimic NB, so no improvements were obtained. Hence, in contrast to [ESJV08], here we do model the class posterior; define simpler input functions for the LR model; introduce a standard L_2 regularisation to avoid the complex set of maximum entropy constraints with cut-offs; and use the MSE learning criterion, optimised with the simple and fast *iRPROP+* [IH03] algorithm.

The assumption of a general LR distribution for the class posterior yields the following classification rule:

$$\hat{c} = \arg\max_{c} p(c|w,\vec{x}) = \arg\max_{c} \frac{\exp\left(\sum_i \lambda_i f_i(c,w,\vec{x})\right)}{\mathcal{Z}(w,\vec{x})} \tag{7.11}$$

where w is the recognised word and $\vec{x} = (x_1..x_D)$ is a D-dimensional vector of discretised input features. On the other hand, $\mathcal{Z}$ is a normalisation constant which does not affect classification; $\lambda_{(\cdot)}$ are a set of data-driven parameters; and $f_{(\cdot)}$ a set of functions which yield the model expressiveness.

As discussed before, the NB model in [SJV12] introduced several convenient assumptions: conditional independence amongst the D scores, discretisation of the continuous-valued scores, etc. Hence, here is a proposed a particular definition for the $f_{(\cdot)}$ functions to make the LR model behave similarly to the NB model in terms of classification.

Let i be the triplet of labels ($\tilde{c} \in \{0,1\}$, $\tilde{w} \in \{1..W\}$, $\tilde{x}_{\tilde{d}} \in \{1..X_{\tilde{d}}\}$) indexing the classes, the known vocabulary and the values of the score number $\tilde{d} \in \{1..D\}$ respectively. $X_{\tilde{d}}$ accounts for the total number of different possible discrete values of $x_{\tilde{d}}$. For each possible triplet, let the following function be defined as:

$$f_{\tilde{c},\tilde{w},\tilde{x}_{\tilde{d}}}(c,w,\vec{x}) = \quad \delta_{\tilde{c}}(c) \cdot \delta_{\tilde{w}}(w) \cdot \delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \tag{7.12}$$

with $\delta(\cdot)$ being the Kronecker delta and $\delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \equiv \prod_{d'}^{D} \delta_{\tilde{x}_{\tilde{d}}}(x_{d'}) \cdot \delta_{\tilde{d}}(d') = \delta_{\tilde{x}_{\tilde{d}}}(x_{\tilde{d}})$.

It becomes clear from the latter definition that the set of functions $\{f_{\tilde{c},\tilde{w},\tilde{x}_{\tilde{d}}}\}$ serves merely to activate the corresponding weights $\{\lambda_{\tilde{c},\tilde{w},\tilde{x}_{\tilde{d}}}\}$. Thus, it is the set of weights alone which will render the classification, and they are to be learned exclusively from data, as detailed in sec. 7.3.1. Also, it should be noted that each of the defined functions does not involve more than one score. This is precisely equivalent to assuming naïve Bayes over the scores, as in Eq. 7.2.

Furthermore, in order to prevent overfitting, additional weights and functions to be active independently of one or more label values are necessary:

$$\begin{aligned}
f_{\tilde{c},\emptyset,\emptyset}(c,w,\vec{x}) &= \quad \delta_{\tilde{c}}(c) \\
f_{\tilde{c},\emptyset,\tilde{x}_{\tilde{d}}}(c,w,\vec{x}) &= \delta_{\tilde{c}}(c)\cdot\delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \\
f_{\tilde{c},\tilde{w},\emptyset}(c,w,\vec{x}) &= \delta_{\tilde{c}}(c)\cdot\delta_{\tilde{w}}(w)
\end{aligned} \qquad (7.13)$$

These terms enable a behaviour similar to the smoothing in the NB model, which backs off to less specific probabilities under certain conditions.

Finally, it should be noted that the presented model typically involves a huge number of weights to be estimated, of order $\mathcal{O}$(vocabulary × number of features × mean number of values per score). Fortunately, the computation time can be halved by defining a new set of weights $\lambda_{(\cdot)} \equiv \lambda_{\tilde{c}=1,(\cdot)} - \lambda_{\tilde{c}=0,(\cdot)}$, and the corresponding activation features:

$$\begin{aligned}
f_{\tilde{w},\tilde{x}_{\tilde{d}}}(w,\vec{x}) &= \delta_{\tilde{w}}(w)\cdot\delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \\
f_{\emptyset,\emptyset}(w,\vec{x}) &= 1 \\
f_{\emptyset,\tilde{x}_{\tilde{d}}}(w,\vec{x}) &= \delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \\
f_{\tilde{w},\emptyset}(w,\vec{x}) &= \delta_{\tilde{w}}(w)
\end{aligned} \qquad (7.14)$$

in this way, Eq. (7.11) adopts the following expression:

$$p(c|w,\vec{x}) = \frac{1}{1+\exp\left((-1)^c\sum_i \lambda_i f_i(w,\vec{x})\right)} \qquad (7.15)$$

Speaker dependence can be easily introduced into Eq. (7.15), yielding a *logistic regression speaker-adapted* (LR+spk) model:

$$p(c \mid w, \vec{x}, s) = \frac{1}{1 + \exp\left((-1)^c \cdot \left(\sum_i \lambda_i f_i(w, \vec{x}) + \sum_j \lambda_j f_j(w, \vec{x}, s)\right)\right)} \tag{7.16}$$

where speaker dependence has been formulated as a separated sum over j for the sake of clarity. Now, the number of weights to be estimated is increased by S times, S being the number of known speakers. In this case, the new index j should map the triplet of labels ($\tilde{w} \in \{\emptyset, 1..W\}$, $\tilde{x}_{\tilde{d}} \in \{\emptyset, 1..X_{\tilde{d}}\}$, $\tilde{s} \in \{1..S\}$).

Thus, speaker adaptation results in the addition of the following:

$$\begin{aligned} f_{\tilde{w},\tilde{x}_{\tilde{d}},\tilde{s}}(w, \vec{x}, s) &= \delta_{\tilde{w}}(w) \cdot \delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \cdot \delta_{\tilde{s}}(s) \\ f_{\tilde{w},\emptyset,\tilde{s}}(w, \vec{x}, s) &= \delta_{\tilde{w}}(w) \cdot \delta_{\tilde{s}}(s) \\ f_{\emptyset,\tilde{x}_{\tilde{d}},\tilde{s}}(w, \vec{x}, s) &= \delta_{\tilde{x}_{\tilde{d}}}(\vec{x}) \cdot \delta_{\tilde{s}}(s) \\ f_{\emptyset,\emptyset,\tilde{s}}(w, \vec{x}, s) &= \delta_{\tilde{s}}(s) \end{aligned} \tag{7.17}$$

7.3.1 Discriminative learning

As discussed on sec. 7.1, the weights of the discriminative models can be estimated to minimise the MSE, which may be preferable for classification problems instead of the MMI criterion and the MLE criterion for generative models. Given N training samples $\{(\vec{x}_n, c_n, w_n)\}_{n=1}^{N}$, the MSE can be formulated as an optimisation problem by means of the objective:

$$F_{\text{MSE}}(\vec{\lambda}) = \sum_{n=1}^{N} (c_n - p_\lambda(c_n = 1 \mid w_n, \vec{x}_n))^2 \tag{7.18}$$

However, there is no closed form solution for the optimal $\vec{\lambda}$ under the minimum MSE constrain. Fortunately, any simple gradient descent based optimisation algorithm can succeed in finding the solution despite the MSE not being a convex criterion. In this work we opted for the simpler iRPROP+ [IH03] iterative algorithm, which provides faster convergence than other more expensive methods such as *generalised iterative scaling* (GIS) [DR72]. A recent evaluation of different optimisation algorithms on a large task can be found in [WRSN13].

Another common issue of many training criteria, including MSE, is that they easily overfit the weights to the training data. Since there is no clear way to smooth discriminatively trained models, a typical amendment is to add a L_2 regularisation term to the objective:

$$F(\vec{\lambda}) = F_{\text{MSE}}(\vec{\lambda}) - \frac{C}{2} \sum_i (\lambda_i - \lambda_i^{(0)})^2 \tag{7.19}$$

where $\vec{\lambda}^{(0)}$ can be either a reliable estimation of the weights or simply $\vec{0}$.

For our model, $\lambda_i^{(0)} = \vec{0}$ is a clever guess, since it prevents the features from having an overrated impact. During experimentation, the zero regularisation made the feature-independent term $\lambda_{\emptyset,\emptyset}$ drop quickly to zero after a few iterations. This behaviour can be interpreted as an increased generalisation of the model, since $\lambda_{\emptyset,\emptyset}$ is proportional to the logarithm of the class prior $p(c)$ from the generative point of view. Thus, for two different models yielding the same performance on a certain test, the one with $\lambda_{\emptyset,\emptyset}$ closer to zero is likely to perform better on a new test with different prior distribution.

7.4 Experiments

7.4.1 Experimental setup

The evaluation of the proposed models (NB+spk, LR and LR+spk) and the baseline model (NB) has been carried out over two difficult tasks from English (videoLectures.net) and Spanish (poliMedia) video lectures. These tasks have been used in the context of the EU-funded project transLectures, which had the aim of developing innovative, cost-effective tools for the automatic transcription and translation of online educational videos [SCdAG+12]. The English task has been defined over the free and open access educational video lecture repository VideoLectures.NET (VL). In VL, the recorded lectures are mostly delivered by distinguished scholars and scientists at important conferences, summer schools, workshops, etc. Currently, VL hosts more than 16.000 lectures from 12.698 speakers. The Spanish task has been defined over Polimedia (PM), which is a recent, innovative service for the creation and distribution of multimedia educational content at *Universitat Politècnica de València* (UPV). PM is designed primarily to allow UPV professors to record their courses in video blocks lasting up to 10 minutes, accompanied by time-aligned slides. PM hosts more than 9.000 lectures from 1.300 speakers with a duration of 2.100 hours.

The state-of-the-art ASR *TLK* toolkit ([dAGS+14]) has been used for the experiments. Acoustic models (AM) were learned using TLK by means of a pre-trained *deep neural network hidden Markov model* (DNN-HMM) hybrid architecture, in a similar fashion to [DMY+12b]. Speaker adaptation was implemented using constrained MLLR (CMLLR) features [SBG05, GGB04b]. The speech data to train the English AM consisted of out-of-domain corpora (TED-LIUM [RDE12], EPPS [Rou11, RSS07, TC-] and Voxforge [vox]), as well as in-domain VL speeches. In contrast, only in-domain PM speech data was used for Spanish. Additionally, it should be noted that the speakers related to the AM data are different from those selected to evaluate the CM models. The statistics of the AM train data are summarised in Table 7.1.

On the other hand, the *language model* (LM) consisted of 5-gram models computed with the SRILM toolkit ([Sto02c]). It is worth mentioning that a common LM was used for all the lectures of the VL task. However, a different LM was used for the PL task depending on the speaker who delivered the speech. Each different LM was adapted to the speaker by exploiting the textual content in the slides available for these PM lectures [MVdAAFJ13].

The evaluation of CMs has been carried out over a distinct corpus from the data used to build the ASR systems. This corpus was split into *training*, *development* and *test* partitions in a balanced way for each of the speakers (statistics are summarised in Table 7.2). As a measure of the difficulty of the task, it should be noted that about 25% of the words of each test are not found in the training sets . The *word error rates* (WER) on the automatic transcripts of the VL and PL test sets were 29.97% and 11.83%, respectively.

Table 7.1: Acoustic data statistics for the English and Spanish ASR systems.

	videoLectures.net				poliMedia			
Set	Spks	Dur.	Words	Voc.	Spks	Dur.	Words	Voc.
ASR data	4034	427h	2.8M	41K	73	107h	936K	27K

Table 7.2: Data partitions for the VL and PL evaluation tasks.

	videoLectures.net				poliMedia			
Set	Spks	Dur.	Words	Voc.	Spks	Dur.	Words	Voc.
Train	8	3.9h	34K	4K	29	20h	117K	13K
Dev.	8	1.3h	11K	2K	29	6.5h	59K	6K
Test	8	1.3h	11K	2K	29	6.7h	59K	6K

7.4.2 Evaluation of CMs

For the purpose of evaluation, the recognised words must be labelled as correct or incorrect. The labelling was computed as the tagging error between the automatic transcripts and the reference transcripts over the minimum Levenshtein cost path. Additionally, class prediction (correct, $c = 1$, or incorrect, $c = 0$) is carried out by minimising the Bayes risk as follows:

$$c^* = \begin{cases} \text{correct} & \text{if } p(c = 1 | w, \vec{x}, s) > \tau \\ \text{incorrect} & \text{ow} \end{cases} \tag{7.20}$$

Where the threshold τ must be empirically estimated on a Dev set. Though this was necessary only for the generative models; for the discriminative models, τ resulted always very close to 0.5 due to the training criteria.

The CM performance was evaluated by means of the following metrics:

- ***Classification Error Rate* (CER)**

The relative number of wrongly classified samples on an evaluation sample set, given the rule in (7.20). It is the direct natural metric to assess the performance of two classifiers: the higher the value, the worse. A simple way to estimate the goodness of a classifier is to compare the CER value to the *relative number of incorrect samples* produced by the system (usually referred as the "baseline"). Unfortunately, the CER as a metric has some flaws: results cannot be directly compared for different tests sets; and the CER is very sensitive to the test set itself, not only to the classifier.

- ***Area under the ROC curve (AROC)***
  The area under the *Receiving Operating Characteristic* (ROC) curve [Faw06]. Briefly, the ROC curve is the set of points in the *False Positive Rate (FPR)-True Positive Rate (TPR)* space, yielded by the classification for every possible different value of the classification threshold τ. The AROC is usually normalised within $[0, 100]$, 100 being a perfect classification and 50 a random classification. The AROC has been a commonly used metric to evaluate the replicability of the CER results. Nonetheless, this metric has been severely criticised since it can give potentially misleading results if ROC curves cross, and it is incoherent in terms of misclassification costs [Han09].

- ***h-measure*** [Han09]
  Normalised metric which is proportional to the overall misclassification loss incurred when using an optimal threshold (which depends on the costs) averaged by a certain function $u(c)$ over the cost ratio $c \in [0, 1]$, $c = c_0/(c_0 + c_1)$ and (c_0, c_1) being the misclassification costs. For the common case in which it cannot be derived which kind of misclassifications are preferable (false positive, or false negatives, etc.), the author proposes a normalised symmetric function $u(c) \propto \beta(c; 2, 2) \propto (c - c^2)$. This measure

was proposed to avoid the issue of the AROC metric, since it is proportional to the expectation of the overall misclassification loss weighted by a function depending on the distribution of the scores. Thus, the weight function to measure the AROC depends on the classifier to be tested.

- ***Normalised cross entropy (NCE)*** [NCE]
  Metric proportional to the cross entropy of the classified set. This metric is related to the average log distance of the score to the true class. NCE equals 1 for a perfect classification in which the predicted posteriors of the correct class score 1 for the correct samples and 0 for the incorrect. Unfortunately, the lowest value is unbounded, since it involves the sum of the logarithm of zero or arbitrarily low values for samples which scored high on the opposite class to the true one. Despite this flaw (noticed shortly after its publication [WMS98]), it is still widely used.

7.4.3 Results

Experiments have been carried out computing the set of input scores that performed the best for the NB model in [SJV12]:

- **SP**: Word acoustic log-score per time frame (10-ms).
- **D**: Duration (in ms.) of the word per phone.
- **NL**: [1] Length of the N-gram in which the word has been decoded.
- **PAvg**: Average of frame-based word posteriors [WSMN01a].
- **PMax**: Like PAvg but using the maximum instead of the average [WSMN01a].

[1] The NL score is not exactly the same as that used in [SJV12], since the length of the N-gram is used instead of the Boolean feature representing the LM back-off behaviour.

Table 7.3: Performance of the models on VL and PL tasks.

TASK	MODEL	CER%	CER% 95%CI	AROC%	h	NCE
VL	NB	17.27	[16.57, 17.97]	85.4	0.37	0.17
	CRF	16.62	[15.93, 17.31]	86.2	0.39	0.31
	LR	16.43	[15.75, 17.11]	86.4	0.40	0.32
	NB+spk	16.56	[15.87, 17.25]	86.2	0.39	0.19
	CRF+spk	14.99	[14.33, 15.65]	88.1	0.44	0.36
	LR+spk	14.82	[14.16, 15.48]	88.2	0.45	0.36
PL	NB	8.14	[7.92, 8.36]	84.9	0.30	0.07
	CRF	7.99	[7.77, 8.21]	85.9	0.31	0.29
	LR	7.89	[7.67, 8.11]	85.5	0.31	0.29
	NB+spk	8.09	[7.87, 8.31]	85.7	0.31	0.10
	CRF+spk	7.97	[7.75, 8.19]	86.9	0.33	0.30
	LR+spk	7.81	[7.59, 8.03]	86.4	0.32	0.30

Table 7.3 summarises the performance of the proposed models on the VL and PL test sets in terms of the different metrics discussed above. We also include results from additional experiments using *Conditional Random Field* (CRF) models which, as stated in recent publications, are of particular importance [Sei13, SW11, FMR$^+$10, LCY10] [2]. It is worth noting that all models have been compared under identical conditions. To assess statistical significance of results, 95% confidence intervals are included for the CER% evaluation metric.

From the results in Table 7.3, it can be stated that speaker-adapted models outperform their non-adapted counterparts. This is true, indeed,

[2] Both, CRF++ and wapiti toolkits were tested. Results presented here correspond to wapiti toolkit (https://wapiti.limsi.fr/), which in turn outperformed CRF++. The optimisation algorithm used was RPROP+ too with L2 regularisation. The optimisation criterion was Maximum log-Likelihood conditional Estimate.

for all models and all evaluation metrics, and also holds for both, VL and PL tasks. Statistically speaking, this statement is significant to a great extent, especially in the case of VL. In this case, in terms of CER%, the best results are: 14.99, with CRF+spk, and 14.82 with LR+spk. These figures are clearly below the lower limit of the 95% confidence intervals for CRF and LR, respectively. On the other hand, the results on PL are similar, though the CRF+spk result overlap the CER% confidence interval for CRF at its lower half, and the same happens with LR+spk. This might be influenced by the comparatively low values of CER% on PL for all models.

Another conclusion that can be drawn from Table 7.3 is that the NB model is clearly superseded by CRF and LR, and that this also holds for their speaker-adapted versions. Given that the LR model is designed as a discriminatively trained version of NB, this result was well expected. On the other hand, although LR(+spk) results are slightly but consistently better than those of CRF(+spk), there is no clear statistical evidence to support its superiority. Indeed, the main difference between them is the training criterion used which, from our experiments, has little effect on the results.

The ROC curves of the NB(+spk), CRF(+spk) and LR(+spk) models are depicted in Fig. 7.1 and Fig. 7.2 for VL and PL, respectively. The classification thresholds adjusted on Dev (operating points) and the optimal ones are also plotted. As can be observed, the speaker-adapted models show better performance than their non-adapted counterparts for all thresholds.

Table 7.4 shows detailed results on the VL test, at speaker level, using the CER evaluation metric. As above, the best results are achieved by LR+spk and CRF+spk. The results at speaker level using other evaluation metrics are similar and are omitted for simplicity.

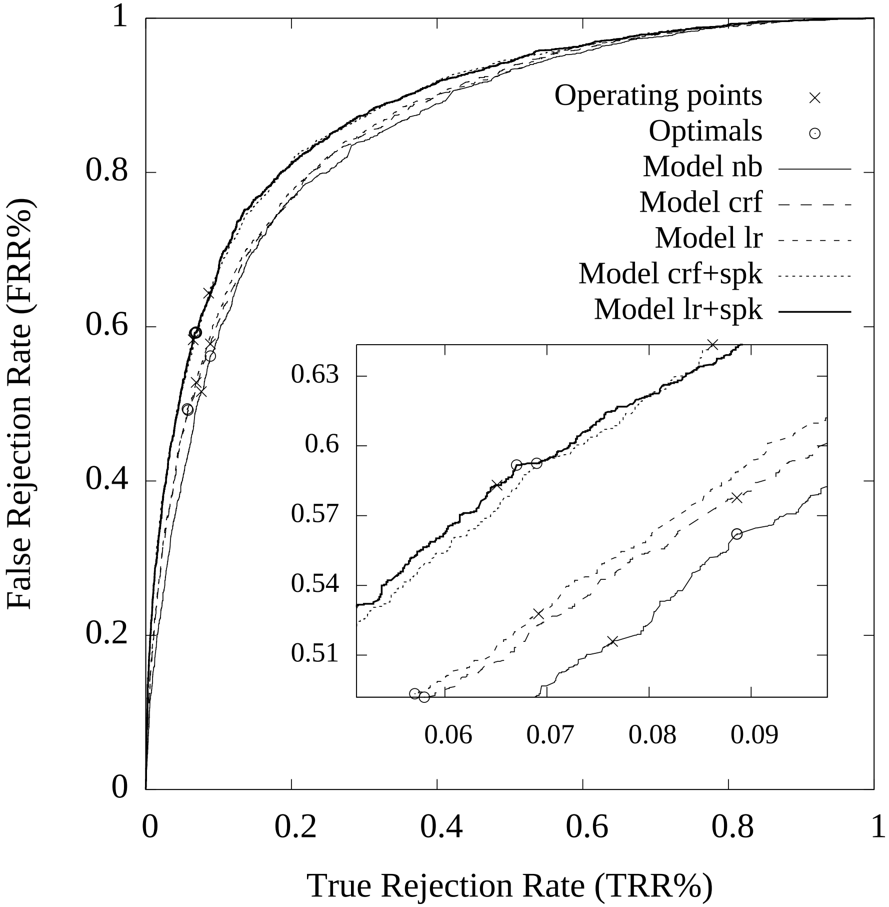


Figure 7.1: ROC curves on the videoLectures.net test set.

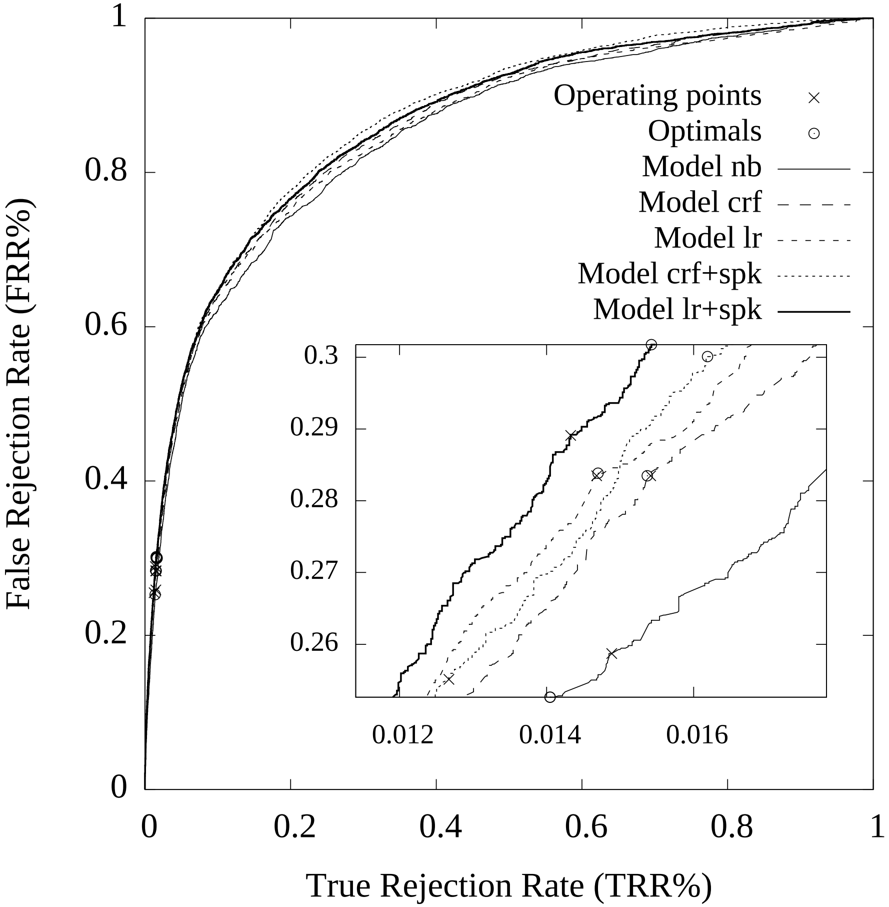


Figure 7.2: ROC curves on the poliMedia test set.

Table 7.4: CER in [%] for each speaker on the VL test set.

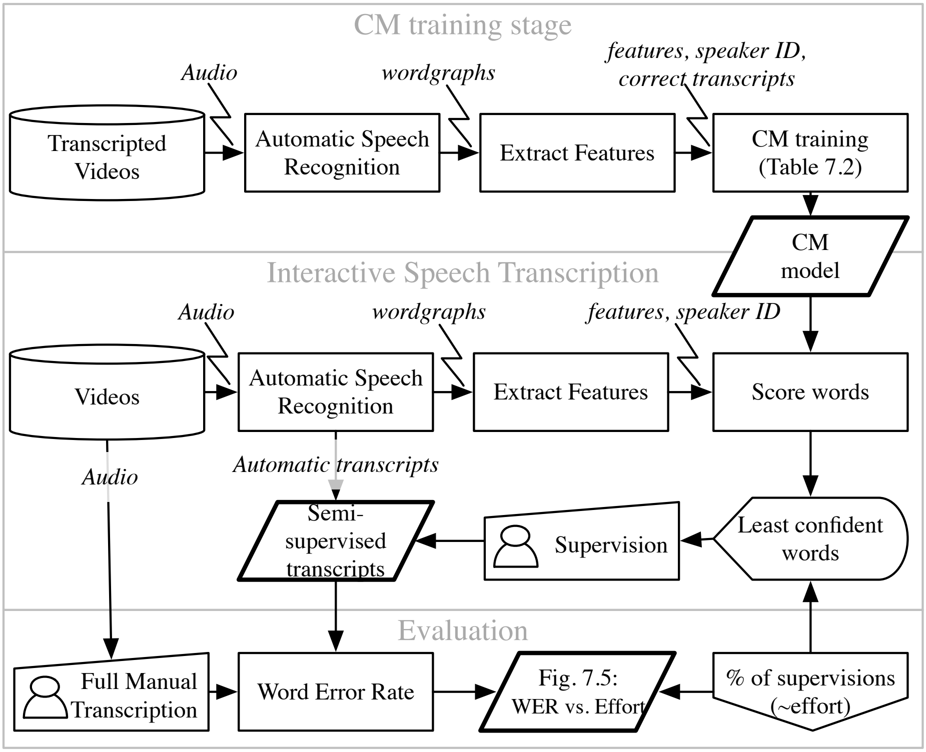
Speaker	1st	2nd	3rd	4th	5th	th 6th	7th	8th
Baseline	15.37	12.79	21.94	16.10	48.76	22.72	31.86	45.89
NB	14.01	12.72	15.91	13.10	27.24	16.15	22.43	30.78
CRF	13.37	12.37	16.20	13.88	25.11	15.96	19.71	28.49
LR	13.00	12.23	16.20	13.55	24.68	15.33	20.31	29.06
NB+spk	13.91	11.74	15.56	13.21	22.03	16.15	21.84	25.62
CRF+spk	12.64	11.81	14.41	12.83	19.56	15.52	18.86	21.99
LR+spk	12.73	11.39	14.41	12.38	19.73	14.77	18.61	23.14

7.5 Interactive Speech Transcription Application

With the aim of measuring the benefits of the LR+spk model in a practical application, we have evaluated its performance in an *interactive speech transcription* (IST) setting applied within the EU project transLectures. In this setting, users devote a limited amount of effort to supervising a given percentage of words of the automatic transcriptions. User effort is optimised by ordering the speech segments selected for supervision from lower to higher reliability based on CMs. The scheme of this IST approach is depicted on figure 7.5.

The VL test set has been used for the assessment of the NB and LR+spk models. Corrections were performed by means of a simulated user in a similar way as in chapter 6. However, there is no block iteration, neither a final constrained search automatic improvement similarly to the last experiment presented on 6.4. Also, despite here the total user time T for interaction is not directly set, the number of words selected for supervision is strongly connected to the average effort and time necessary for that manual process.

Figure 7.3: Flowchart depicting the process of the IST system used for evaluation. The rhomboids in bold indicate the product of each one of the stages.



The final quality (measured in WER) of partially supervised transcriptions resulting for different percentages of supervised words is depicted on Fig. 7.4. The figure assesses the behaviour when using the NB model or the LR+spk and CRF+spk models to compute CMs. A random strategy corresponding to a sequential supervision of the words is also depicted.

As can be observed in the figure, it can be stated that the LR+spk and CRF+spk models outperform the NB model for any level of user effort (percentage of supervised words). In particular, for the reasonable range of percentages from 10% to 20%, the LR+spk and CRF+spk produce relative WER improvements between 2.5% and 7%.

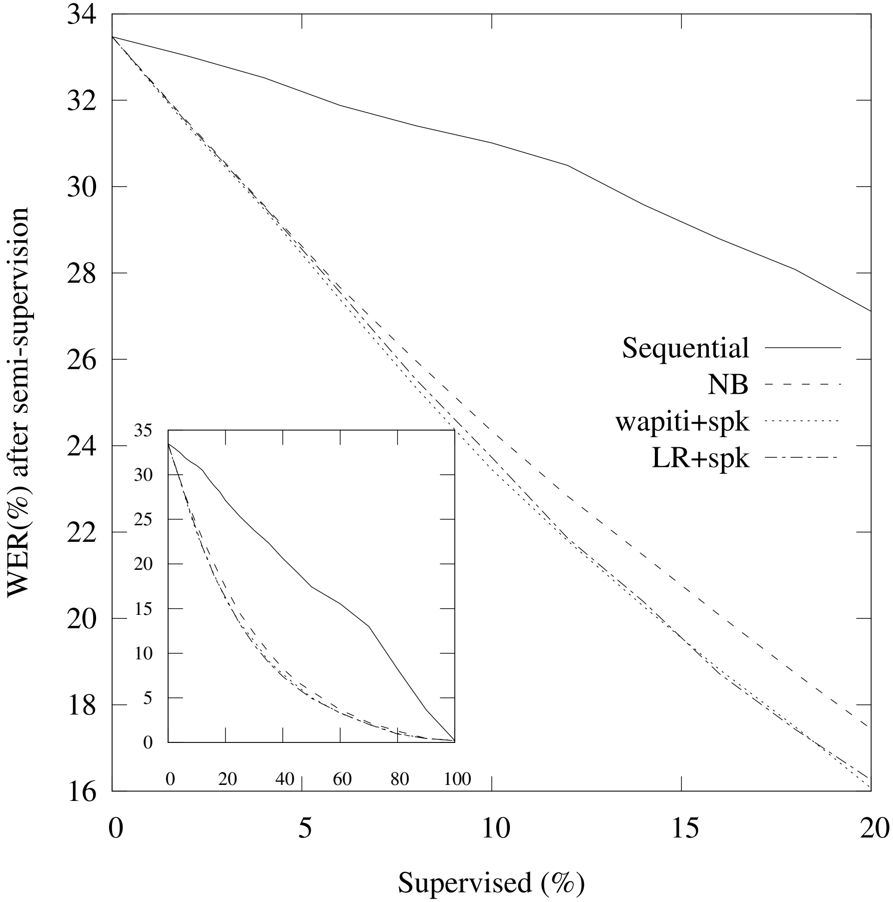


Figure 7.4: Resulting WER for the partial supervision of the VL test set in a reasonable range of effort. The sub-figure shows the behaviour under all percentages up to full supervision.

7.6 Conclusions

A new particular logistic regression model has been introduced in this chapter to improve the reliability of the confidence measures for automatic speech recognition. The conducted experimentation proved the significantly superior performance of the proposed model on two realistic challenging tasks.

Additionally, it has been introduced speaker dependence for the sake of further improvement. While this adaptation is not always possible, it is still highly motivated due to the practical applications on large repositories of lectures. The proposed logistic speaker-adapted model significantly outperforms the compared models.

Finally, a simple real application of interactive speech transcription guided by confidence measures has confirmed that the gains obtained by the proposed models translate into a noticeable improvement of the resulting semi-supervised transcriptions for an equal level of user effort. Furthermore, the conducted interactive supervisions proves that, in practice, most of the insertions can be successfully corrected by the user due to their distribution, despite of the a priori limitation of the word-level interactions to capture that kind of errors.

CHAPTER 8

CONCLUSIONS

8.1 Summary

• First, on chapter 3 a novel method for interactive speech transcription has been introduced. The other few publications on the field dealt instead the IST by means of segment/sentence-based approaches; and they sought for perfect final transcript. Thus, those approaches were not able to achieve the same reduction in terms of effort (i.e. number of necessary interactions).

On the contrary, the approach presented here has been proved experimentally to be effective in finding an optimal balance between the resulting transcription quality and the supervision effort performed by an user. The desired quality of the resulting transcripts can be controlled by means of a tolerance error set by the user, and so the effort.

Moreover, results show that a tolerance error in the transcriptions does not affect critically on the incremental learning of acoustical models. Thus, this method can be used also for producing ASR models similar in performance to those generated from full manually transcribed corpora.

• Second, in order to deal with the particularities of the real-life online repositories with a massive number of videos, another similar IST approach was investigated. This approach has a different goal from the previous: performing intelligent interaction constrained to a limit of user effort. Thus, eliminating the necessity of a WER prediction. Additionally, the final transcripts are further improved automatically by means of the constrained search: an automatic decoding technique to gain the most from the user interactions.

The approach proved empirically to be sound and useful no matter how good or bad was the initial underlying ASR system: the final transcription errors were 3 times lower in average than other approaches based on post-edition effort over challenging real-life tasks.

- Third, all the IST conducted research was put into real deployment to achieve a cost-effective solution to the *Transcription and translation of video Lectures* under the European project. The more refined version of the online prototype was evolved into a modern web player and editor which provided a massive number of subtitles for videoLectures.net and poliMedia educational repositories with more than 17882 lectures from 13273 authors at the time of the writing. These real-life repositories, and so challenging, where the source for the data used for most of the experiments on this work.

- Finally, for the sake of providing better intelligence guidance experience under the IST frameworks, but also to improve a big number of ASR applications, a new logistic regression model has been introduced. This model improves the reliability of the confidence measures for automatic speech recognition, as demonstrated on challenging tasks, when compared to state-of-the-art models such as the Naive Bayes and Conditional Random Fields.

Additionally, it has been introduced speaker dependence for the sake of further improvement. While this adaptation is not always possible, it is still highly motivated due to the practical applications on large repositories of lectures. The proposed logistic speaker-adapted model significantly outperforms the compared models.

The achieve improvement on the performance of the CM, effectively translated into further improvement on a real application of interactive speech transcription (guided of course by confidence measures). Furthermore, the conducted interactive supervisions proves that, in practice, most of the insertions can be successfully corrected by the user due to their distribution, despite of the a priori limitation of the word-level interactions to capture that kind of errors. Validating the word-level approach of the presented IST and CM methods.

8.2 Publications

- "A prototype for Interactive Speech Transcription Balancing Error and Supervision Effort" for the *"Intelligent User Interfaces 2012"* congress, CORE-A ranked.
- "A prototype for Interactive Speech Transcription Balancing the Error and Human Effort" for the *"Innovation and Applications in Speech Technology 2012"* workshop.
- "TransLectures", *"IberSpeech'12"* congress.
- "Speaker-adapted confidence measures for speech recognition of video lectures" on the journal *"Computer Speech and Language"* with 1.09 impact factor.
- "Intelligent Interaction in Automatic Transcription of Video Lectures", in preparation for *Speech Communication.*

8.3 Future work

The balancing the error and supervision effort, as well as the intelligent interaction with user limited effort (IILEU) achieved a great reduction in the effort. However, they suffered from certain aspects:

• The word-level approach induces a lack of audio context, and the words could be chopped when listened, making it harder for the user to figure up the proper correction. Although, in practice this does not critically affects the performance as revealed by the experimentation.

However, the proposed IST jumps from one place in the sentence to another until it decides to jump to the next utterance. Thus a as future work, the driver of the method should include rules and cost functions in order to group into one larger segment words that are likely to be wrongly recognised. Also, it will be better if the segments were asked from left to the right.

• The confidence measures: A bad performance of the confidence measures mislead the overall process. Although a significant improvement has been achieved with the proposed speaker adapted LR model. Still, there is free room for testing aspects, which have been already investigated on the literature, on the presented model. For instance, the how deep is the impact of using one discretisation technique or another; extension to catch insertions operations; etc.

• The estimation of the error: Although the new IILUE bypasses this problem, it will be still interesting to apply better WER estimation algorithms. For instance, using more features than the rank level of confidence measure, such as the dependence with the words themselves, the relative position in the sentence, the continues value of the confidence measure, etc.

BIBLIOGRAPHY

[AB98] Tasos Anastasakos and Sreeram V Balakrishnan. The use of confidence measures in unsupervised adaptation of speech recognizers. In *ICSLP*, 1998.

[AD77] D. Rubin A. Dempster, N. Laird. Maximum likelihood from incomplete data via the em algorithm. In *ournal of the Royal Statistical Society*, volume 39, pages pp. 1 – 38, 1977.

[AHDY13] Ossama Abdel-Hamid, Li Deng, and Dong Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Proc. of Interspeech 2013*. ISCA, August 2013.

[Bau72] L.E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *O. Shisha, editor, Inequalities, Vol. 3, pp. 1–8. Academic Press*, 3:1–8, 1972.

[BGWL98] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *Proceedings of LREC*, pages 1373–1376, 1998.

[BGWL00] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication special issue on Speech Annotation and Corpus Tools*, 33(1–2), 2000.

[BN08] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434 – 451, 2008.

[BRG07] M. Bardideh, F. Razzazi, and H. Ghassemian. An svm based confidence measure for continuous speech recognition. In *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*, pages 1015–1018, Nov 2007.

[BSR13] Peter Bell, Pawel Swietojanski, and Steve Renals. Multi-level adaptive networks in tandem and hybrid asr systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6975–6979. IEEE, 2013.

[CGM$^+$97] C Julian Chen, Ramesh A Gopinath, Michael D Monkowski, Michael A Picheny, and Katherine Shen. New methods in continuous mandarin speech recognition. In *Eurospeech*, 1997.

[Cha97] Lin L Chase. *Error-responsive feedback mechanisms for speech recognizers*. Carnegie Mellon University Pittsburgh, PA, 1997.

[cou] coursera.org: Take the World's Best Courses, Online, For Free. `http://www.coursera.org/`.

[CR96] Stephen Cox and Richard Rose. Confidence measures for the switchboard database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 511–514. IEEE, 1996.

[CZM04] Barry Y Chen, Qifeng Zhu, and Nelson Morgan. Learning long-term temporal features in lvcsr using neural networks. In *INTERSPEECH*, 2004.

[dAGS+14] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. The translectures-upv toolkit. In *Proc. of IberSpeech 2014*, Las Palmas de Gran Canaria (Spain), 2014.

[DHS12] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification.* John Wiley & Sons, 2012.

[DM80] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[DMY+12a] George E. Dahl, Student Member, Dong Yu, Senior Member, Li Deng, and Alex Acero. Context-dependent pretrained deep neural networks for large vocabulary speech recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 20, pages 30–42, 2012.

[DMY+12b] George E. Dahl, Student Member, Dong Yu, Senior Member, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[DR72] J N Darroch and D Ratcliff. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, October 1972.

[DRN95] V. Digalakis, D. Rtischev, and L. Neumeyer. Speaker adaptation using constrained reestimation of gaussian mixtures. *IEEE Trans. on Speech and Audio Processing*, 3(5):357–366, 1995.

[dV12] Universitat Politècnica de València. poliMedia: Videolectures from the “Universitat Politècnica de Valencià. http://polimedia.blogs.upv.es, May 2012.

[DYDA12] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):30–42, 2012.

[ESJV08] C. Estienne, A Sanchis, A Juan, and E. Vidaf. Maximum entropy models for speech confidence estimation. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4421–4424, 2008.

[EW00a] Gunnar Evermann and PC Woodland. Posterior probability decoding, confidence estimation and system combination. In *Proc. Speech Transcription Workshop*, volume 27. Baltimore, 2000.

[EW00b] Gunnar Evermann and Philip C Woodland. Large vocabulary decoding and confidence estimation using word posterior probabilities. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1655–1658. IEEE, 2000.

[Faw06] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[FMR+10] Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, and Patrick Gros. CRF-based combination of contextual features to improve a posteriori word-level confidence measures. In *INTERSPEECH*, pages 1942–1945, 2010.

[FWK07] Christian Fügen, Alex Waibel, and Muntsin Kolss. Simultaneous translation of lectures and speeches. *Machine translation*, 21(4):209–252, 2007.

[Gal98] M.J.F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[Gan05] Juri Ganitkevitch. Speaker adaptation using maximum likelihood linear regression. In *Rheinish-Westflesche Technische Hochschule Aachen, the course of Automatic Speech Recognition, www-i6. informatik. rwthaachen.*

de/web/Teaching/Seminars/SS05/ASR/Juri Ganitkevitch Ausarbeitung. pdf, 2005.

[GEA+13] J.M. Garrido, D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. De-La-Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, E. Estebas, M. Cabrera, and A. Bonafonte. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language resources and evaluation*, 47(4):945–971, 2013.

[GGB04a] Diego Giuliani, Matteo Gerosa, and Fabio Brugnara. Speaker normalization through constrained mllr based transforms. In *Proc. of the Eighth ICSLP*, 2004.

[GGB04b] Diego Giuliani, Matteo Gerosa, and Fabio Brugnara. Speaker normalization through constrained MLLR based transforms. In *INTERSPEECH*, 2004.

[GIY97] Larry Gillick, Yoshiko Ito, and Jonathan Young. A probabilistic approach to confidence estimation and evaluation. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 879–882. IEEE, 1997.

[GKB01] Vaibhava Goel, Shankar Kumar, and William Byrne. Confidence based lattice segmentation and minimum bayes-risk decoding. In *INTERSPEECH*, pages 2569–2572, 2001.

[GKKC07] Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocky. Probabilistic and bottle-neck features for lvcsr of meetings. In *Acoustics, Speech and Signal Processing,* *2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–757. IEEE, 2007.

[GL94] J. Gauvain and C. Lee. Maximum a posteriori estimation of multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.

[GS85] Silviu Guiasu and Abe Shenitzer. The principle of maximum entropy. *The Mathematical Intelligencer*, 7(1):42–48, 1985.

[GVB03] David Grangier, Alessandro Vinciarelli, and Hervé Bourlard. Information retrieval on noisy text. Technical report, IDIAP, 2003.

[GZXY09] Jie Gao, Qingwei Zhao, Ran Xu, and Yonghong Yan. Improved lattice-based confidence measure for speech recognition via a lattice cutoff procedure. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, volume 4, pages 473–476, Aug 2009.

[Han09] DavidJ. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, 2009.

[Haz06] Timothy J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *In Proc. Interspeech*, 2006.

[HDH$^+$10] Georg Heigold, Philippe Dreuw, Stefan Hahn, Ralf Schüter, and Hermann Ney. Margin-Based Discriminative Training for String Recognition. pages 1–10, February 2010.

[Hei10] Georg Heigold. *A Log-Linear Discriminative Modeling Framework for Speech Recognition.* PhD thesis, RWTH Aachen University, Computer Science Department, RWTH Aachen University, Aachen, Germany, June 2010.

[HES00] Hynek Hermansky, Daniel W Ellis, and Shantanu Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1635–1638. IEEE, 2000.

[HN94] R. Kneser H. Ney, U. Essen. On structuring probabilistic dependencies in language modeling. *Computer Speech and Language*, 2(1–38), 1994.

[HSP02] Timothy J Hazen, Stephanie Seneff, and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech & Language*, 16(1):49–67, 2002.

[HTRT06a] D. Hakkani-Tür, G. Riccardi, and G. Tur. An active approach to spoken language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(3):1–31, 2006.

[HTRT06b] Dilek Hakkani-Tur, Giuseppe Riccardi, and Gokkan Tur. An active approach to spoken language processing. *ACM Transactions on Speech and Language Processing*, 3:1–31, 2006.

[IH03] Christian Igel and Michael Hüsken. Empirical evaluation of the improved RPROP learning algorithms. *Neurocomputing*, 50:105–123, 2003.

[JCAF$^+$12] Alfons Juan, Jorge Civera, Jesús Andrés-Ferrer, Nicola Cancedda, Mitja Jermol, Davor Orlic, John Shawe-Taylor, Hermann Ney, Marion Mast, and Yota Georgakopoulou. translectures: Transcription and translation of video lectures. 2012.

[JD01] Hui Jiang and Li Deng. A bayesian approach to the verification problem: Applications to speaker verification. *Speech and Audio Processing, IEEE Transactions on*, 9(8):874–884, 2001.

[Jel76] F. Jelinek. Continuous speech recognition by statistical methods. In *Proceedings of the IEEE*, volume 64, pages 532–556, 1976.

[Jia05] H Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

[Jia10] Hui Jiang. Discriminative training of hmms for automatic speech recognition: A survey. *Computer Speech & Language*, 24(4):589–608, 2010.

[JY11] Zhao Junfeng and Zhu Yeping. A multi-confidence feature combination rejection method for robust speech recognition. In *Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on*, pages 2556–2559, Dec 2011.

[Kat87] Slava M Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):400–401, 1987.

[KHB$^+$07] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christie Moran, Richard Zens, Chris Dyer, Ontraj Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, pages 177–180, 2007.

[KHS99] Simo O. Kamppari, J. Hazen, and Arthur C. Smith. Word and phone level acoustic confidence scoring, 1999.

[KK05] Tae-Yoon Kim and Hanseok Ko. Bayesian fusion of confidence measures for speech recognition. *IEEE Signal Processing Letters*, 12:871–874, December 2005.

[KN95] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proc. of the ICASSP*, volume 1, pages 181–184, 1995.

[KS97] T Kemp and T Schaaf. Estimating confidence using word lattices. *Proc Eurospeech*, 1997.

[LABG05] Lori Lamel, G Adda, E Bilinski, and Jean-Luc Gauvain. Transcribing lectures and seminars. In *Proc. of Interspeech 2005*, pages 1657–1660, 2005.

[LCY10] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proceedings the 48th Annual*

Meeting of the Association for Computational Linguistics (ACL), pages 504–513. Association for Computational Linguistics, July 2010.

[LGH+07] J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney. The rwth 2007 tc-star evaluation system for european english and spanish. In *Proc. of Interspeech*, pages 2145–2148, 2007.

[LMR08] Saturnino Luz, Masood Masoodian, and Bill Rogers. Interactive Visualisation Techniques for Dynamic Speech Transcription, Correction and Training. In Stuart Marshall, editor, *Proceedings of CHINZ 2008, The 9th ACM SIGCHI-NZ Annual Conference on Computer-Human Interaction*, pages 9–16, Wellington, New Zealand, 2008. ACM Press.

[LR96] Eduardo Lleida and Richard C Rose. Likelihood ratio decoding and confidence measures for continuous speech recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 1, pages 478–481. IEEE, 1996.

[LR00] E Lleida and R C Rose. Utterance verification in continuous speech recognition: decoding and training procedures. *Speech and Audio Processing, IEEE Transactions on*, 8(2):126–139, 2000.

[LW95] CJ Leggetter and PC Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech and language*, 9(2):171, 1995.

[MBP$^+$06] Cosmin Munteanu, Ronald Baecker, Gerald Penn, Elaine Toms, and David James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc. of the CHI*, pages 493–502, 2006.

[MBS00] Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, 2000.

[MLN97] Sven C Martin, Jörg Liermann, and Hermann Ney. Adaptive topic-dependent language modelling using word-based varigrams. In *In Proc. Eurospeech'97*, 1997.

[MLR01] P.J. Moreno, B Logan, and B. Raj. A boosting approach for confidence scoring. *Seventh European Conference on Speech Communication and Technology*, 2001.

[MS99] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[MVdAAFJ13] A. Martinez-Villaronga, M.A. del Agua, J. Andres-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8450–8454, May 2013.

[MVdAAFJ14] A. Martinez-Villaronga, M. A. del Agua, J. Andres-Ferrer, and A. Juan. Language model adaptation for lecture transcription by document retrieval. In *Proc. of IberSpeech 2014*, Las Palmas de Gran Canaria (Spain), 2014.

[NCE] NCE. `www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`.

[Ney90] H. Ney. Acoustic modeling of phoneme units for continuous speech recognition. *Signal Processing V: Theories and Applications, Fifth European Signal Processing Conference*, pages 65–72, 1990.

[NHUTO92] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder. Improvements in beam search for 10000-word continuous speech recognition. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 9–12 vol.1, Mar 1992.

[NMW97] Hermann Ney, Sven Martin, and Frank Wessel. Statistical language modeling using leaving-one-out. In *Corpus-based methods in Language and Speech processing*, pages 174–207. Springer, 1997.

[NRE97] Chalapathy V Neti, Salim Roukos, and E Eide. Word-based confidence measures as a guide for stack search in speech recognition. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 883–886. IEEE, 1997.

[OFN97] Stefan Ortmanns, Thorsten Firzlaff, and Hermann Ney. Fast likelihood computation methods for continuous mixture densities in large vocabulary speech recognition. In *EUROSPEECH*, 1997.

[ONA97] Stefan Ortmanns, Hermann Ney, and Xavier Aubert. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech & Language*, 11(1):43–72, 1997.

[Ope12] Opencast. Matterhorn. http://opencast.org/matterhorn/, May 2012.

[OVWY94] J. J. Odell, V. Valtchev, P. C. Woodland, and S. J. Young. A one pass decoder design for large vocabulary recognition. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 405–410, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[PFFG93] D.S. Pallett, J.G. Fiscus, W.M. Fisher, and J.S. Garofolo. Benchmark tests for the DARPA spoken language program. In *Proceedings of the workshop on Human Language Technology*, pages 7–18. Association for Computational Linguistics, 1993.

[PN05] Michael Pitz and Hermann Ney. Vocal tract normalization equals linear transformation in cepstral space. *Speech and Audio Processing, IEEE Transactions on*, 13(5):930–944, 2005.

[PP12a] Miltiades Papadopoulos and Elaine Pearson. Improving the accessibility of the traditional lecture: An automated tool for supporting transcription. In *Proc. of the BCS-HCI*, pages 127–136, 2012.

[PP12b] Miltiades Papadopoulos and Elaine Pearson. An intelligent system to support accurate transcription of university lectures. In *Proc. of the 11th ITS*, pages 718–719, 2012.

[Qin13] Long Qin. *Learning Out-of-Vocabulary Words in Automatic Speech Recognition.* PhD thesis, School of Computer Science. Carnage Mellon University, 2013.

[RDE12] A Rousseau, P Deléglise, and Y Estève. TED-LIUM: an Automatic Speech Recognition dedicated corpus. *LREC*, 2012.

[RGH+09] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The rwth aachen university open source speech recognition system. In *Proc. of Interspeech*, pages 2111–2114, 2009.

[Rou11] A. Rousseau. Lium's systems for the iwslt 2011 speech translation tasks. In *International Workshop on Spoken Language Translation*, San Francisco (USA), 8-9 Sept 2011.

[RRCV07] Luis Rodríguez-Ruiz, Francisco Casacuberta, and Enrique Vidal. Computer Assisted Transcription of Speech. In *Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis, Volume 4477 of LNCS*, pages 241–248, 2007.

[RSS07] B Ramabhadran, O Siohan, and A Sethy. The IBM 2007 speech transcription system for European parliamentary speeches. In *IEEE Workshop on ASRU*, pages 472–477, 2007.

[RTDG$^+$13] R. Ranchal, T. Taber-Doughty, Yiren Guo, K. Bain, H. Martin, J.P. Robinson, and B.S. Duerstock. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, 6(4):299–311, 2013.

[SA91] Richard Schwartz and Steve Austin. A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 701–704. IEEE, 1991.

[SBG05] G. Stemmer, F. Brugnara, and D. Giuliani. Adaptive training using simple target models. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 997 – 1000, 2005.

[SCdAG$^+$12] J. A. Silvestre-Cerdà, M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. translectures. In *Online Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012)*, pages 345–351, Madrid (Spain), nov 2012.

[SCPJ$^+$13] J.A. Silvestre-Cerda, A. Perez, M. Jimenez, C. Turro, A. Juan, and J. Civera. A system architecture to support cost-effective transcription and translation of large video lecture repositories. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 3994–3999, Oct 2013.

[SCSSJ12] Isaias Sanchez-Cortina, Nicolás Serrano, Alberto Sanchis, and Alfons Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *IUI '12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, February 2012.

[Sei13] Matthew Stephen Seigel. *Confidence Estimation for Automatic Speech Recognition Hypotheses*. PhD thesis, Department of Engineering, University of Cambridge, 2013.

[SF09] Henrik Schulz and José AR Fonollosa. A catalan broadcast conversational speech database. In *I Joint SIGIL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, 2009.

[SGR13] P Swietojanski, A Ghoshal, and S Renals. Revisiting hybrid and GMM-HMM system combination techniques. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6744–6748. IEEE, 2013.

[Six03] Achim Sixtus. *Across-word phoneme models for large vocabulary continuous speech recognition*. PhD thesis, Bibliothek der RWTH Aachen, 2003.

[SJV07] A. Sanchis, A Juan, and Enrique Vidal. Estimation of confidence measures for machine translation. In *Proceedings of the MT Summit XI*, 2007.

[SJV12] A. Sanchis, A. Juan, and E. Vidal. A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):565–574, 2012.

[SL83] M.M. Sondhi S.E. Levinson, L.R. Rabiner. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. In *Bell System Technical Journal*, volume 62, pages 1035–1074, April 1983.

[SLCY11] Frank Seide, Gang Li, Xie Chen, and Dong Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 24–29. IEEE, 2011.

[SNTW$^+$11a] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A.E.D. Mousa, S. Hahn, D. Nolden, R. Schlüter, and H Ney. The RWTH 2010 quaero ASR evaluation system for English, French, and German. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Prague, Czech Republic*, pages 2212–2215, 2011.

[SNTW$^+$11b] Martin Sundermeyer, Markus Nußbaum-Thom, Simon Wiesler, Christian Plahl, Amr El-Desoky Mousa, Stefan Hahn, David Nolden, Ralf Schlüter, and Hermann Ney. The

RWTH 2010 Quaero ASR evaluation system for English, French, and German. In *ICASSP*, pages 2212–2215. IEEE, 2011.

[SS12] M. Sahidullah and G. Saha. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. 54(4):543–565, 2012.

[SSHTT00] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1):127–154, 2000.

[SSJ10] N. Serrano, A. Sanchis, and A Juan. Balancing error and supervision effort in interactive-predictive handwriting recognition. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 373–376. ACM, 2010.

[SSSB10] R. Sánchez Sáez, J.A. Sánchez, and J.M. Benedí. Confidence measures for error discrimination in an interactive predictive parsing framework. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1220–1228. Association for Computational Linguistics, 2010.

[Sto02a] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *ICSLP*, 2002.

[Sto02b] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of ICSLP*, 2002.

[Sto02c] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904, Denver, USA, 2002.

[sup] SuperLectures: We take full care of your event video recordings. `http://www.superlectures.com`.

[SW11] Matthew Stephen Seigel and Philip C Woodland. Combining information sources for confidence estimation with CRF models. In *INTERSPEECH*, pages 905–908, 2011.

[SZWA11] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. Srilm at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, page 5, 2011.

[TC-] TC-STAR Evaluation Report. `www.tcstar.org/documents/D30.pdf`.

[traa] 3.1.3 final report on massive adaptation (m36). `https://www.translectures.eu/wp-content/uploads/2015/01/transLectures-D3.1.3-31Oct2014.pdf`.

[trab] transLectures: Transcription and Translation of Video Lectures. `http://www.translectures.eu/`.

[TSN13] Zoltan Tüske, Ralf Schlüter, and Hermann Ney. Multilingual hierarchical mrasta features for asr. In *Proc. of Interspeech 2013*, pages 2222–2226, 2013.

[UXJK$^+$12] UPVLC, XEROX, JSI-K4A, RWTH, EML, and DDS.

transLectures: Transcription and Translation of Video Lectures. In *Proc. of EAMT*, page 204, 2012.

[Vid] Videolectures.NET: Exchange ideas and share knowledge. `http://www.videolectures.net/`.

[VK08] Keith Vertanen and Per Ola Kristensson. On the benefits of confidence visualization in speech recognition. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. ACM Request Permissions, April 2008.

[vox] Voxforge. `www.voxforge.org/`.

[VVP+07] Fabio Valente, Jithendra Vepa, Christian Plahl, Christian Gollan, Hynek Hermansky, and Ralf Schlüter. Hierarchical neural networks feature extraction for lvcsr system. In *INTERSPEECH*, pages 42–45, 2007.

[Wal06] Mike Wald. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3(2):131–141, 2006.

[WBR+97] Mitch Weintraub, Francoise Beaufays, Ze'ev Rivlin, Yochai Konig, and Andreas Stolcke. Neural-network based measures of confidence for word recognition. In *icassp*, page 887. IEEE, 1997.

[WDAO07] Connor White, Jasha Droppo, Alex Acero, and Julian Odell. Maximum entropy confidence estimation for speech recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–809. IEEE, 2007.

[WIT+14] Simon Wiesler, Kazuki Irie, Zolt'an Tüske, Ralf Schlüter, and Hermann Ney. The rwth english lecture recognition system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3322–3326, Florence, Italy, May 2014.

[WKN99] L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. In *Proc. of the ICASSP*, volume 2, pages 761–764, 1999.

[WLW+10] Zhi-Guo Wang, Cong Liu, Hai-Kun Wang, Yu Hu, and Li-Rong Dai. Phonetic clustering based confidence measure for embedded speech recognition. *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 186–189, 2010.

[WMS98] Frank Wessel, Klaus Macherey, and Ralf Schluter. Using word probabilities as confidence measures. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 225–228. IEEE, 1998.

[WN05] Frankz Wessel and Hermann Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23 – 31, 2005.

[WRSN13] Simon Wiesler, Alexander Richard, Ralf Schluter, and Hermann Ney. A critical evaluation of stochastic algorithms for convex optimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6955–6959. IEEE, 2013.

[WSMN01a] F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 9(3):288–298, Mar 2001.

[WSMN01b] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298, 2001.

[WWR00] Frank Wallhoff, Daniel Willett, and Gerhard Rigoll. Frame-discriminative and confidence-driven adaptation for lvcsr. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1835–1838. IEEE, 2000.

[XH89] M.A. Jack X.D. Huang. Semi-continuous hidden markov models for speech signals. In *Computer Speech and Language*, volume 3, pages 329–252, 1989.

[YLD11] D Yu, J Li, and L Deng. Calibration of confidence measures in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(8):2461–2473, 2011.

[You92] S. J. Young. The general use of tying in phoneme-based hmm speech recognisers. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 569–572, Washington, DC, USA, 1992. IEEE Computer Society.

[YP13] K A Yadav and M Patil. Confidence calibration measures to improve speech recognition. In *Communications and Signal Processing (ICCSP), 2013 International Conference on*, pages 826–829. IEEE, 2013.

[YW94] Steve J Young and Philip C Woodland. State clustering in hidden markov model-based continuous speech recognition. *Computer Speech & Language*, 8(4):369–383, 1994.

[ZR01] Rong Zhang and Alexander I Rudnicky. Word level confidence annotation using combinations of features. 2001.