Deep learning approaches to assess speech intelligibility of head and neck cancer

Sebastião Quintas

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Sebastião LEDESMA FRAZAO DE BARROS QUINTAS

Le 30 novembre 2022

Deep Learning Approaches to Assess Speech Intelligibility of Head and Neck Cancer

---

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Julien PINQUIER

Jury
M. Nicholas CUMMINS, Rapporteur
M. Mathew MAGIMAI DOSS, Rapporteur
Mme Isabel TRANCOSO, Examinatrice
Mme Julie MAUCLAIR, Examinatrice
M. Julien PINQUIER, Directeur de thèse
Mme Corinne FREDOUILLE, Présidente

UNIVERSITÉ TOULOUSE III PAUL SABATIER — Université de Toulouse

École Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Deep Learning Approaches to Assess Speech Intelligibility of Head and Neck Cancer

PhD Thesis

Presented and defended on the 30 November 2022

to obtain the title of

PhD of Science of the University of Toulouse

(Specialty: Computer Science)

by

Sebastião Quintas

Jury:

President:	Corinne Fredouille
Reviewers:	Nicholas Cummins Mathew Magimai Doss
Examinators:	Corinne Fredouille Isabel Trancoso
Advisors:	Julie Mauclair Julien Pinquier

Institut de Recherche en Informatique de Toulouse

Mis en page avec la classe thesul.

I would like to start this manuscript by thanking all those who were indispensable to the realization of this thesis. Without them, the research work done in the past three years would not have been possible. Given the international nature of the research project in which this thesis was inserted, the acknowledgments will be addressed in the person's native tongue.

Since this thesis was inserted within the TAPAS project, European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287, I would like to thank all of those who made part of this project and that greatly help me grow as a person and researcher. Out of all the early stage researchers, I would like to particularly thank Timothy and Thomas, since they were the ones that convinced me into to the project and therefore without them this thesis would have never come to fruition.

I would like to thank the members of the jury for their interest on the topic and eagerness to participate and collaborate in an insightful dialogue. In my opinion, a PhD defense day is one of the most memorable days in someone's life, and given this, I could not have asked for a better discussion.

* * *

Si je vais faire une partie des remerciements en Français, c'est impératif de commencer par les deux personnes que je tiens à remercier le plus, Julie et julien. Je pouvais pas demander une meilleure direction de thèse. Depuis le début, vous avez cru en moi et m'avez donné une chance en me faisant participer au projet. Je vous en serai toujours reconnaissant. Le chercheur que je suis aujourd'hui est le produit de votre encadrement. Merci infiniment.

Je tiens à remercier toute l'équipe SAMOVA pour ces trois belles années. Je pouvais pas demander une meilleure et plus agréable équipe qui m'a accueilli pendant cette thèse. Si tout était tellement bien passé, c'est surtout grâce à ces personnes : Verdiana, Lucile, Robin, Lila, Estelle, Étienne, Vincent, Mathieu, Jim, Leo, Alexis, Benjamin, Jérôme, Isabelle, Hervé, Thomas, Christine, Régine, Julie, Julien et tous les autres membres qui sont passés par l'équipe.

Cette thèse s'est insérée dans le projet RUGBI, qui a été déterminant pour le développement des méthodologies qui seront bientôt présentées. À ce titre, je tiens à remercier Jérôme, Corinne, Sondes, Muriel, Alain, Virginie, Marie, Corine, Christine et tous ceux qui ont contribué aux séances de brainstorming pertinentes, aux discussions passionnées et aux rencontres chaleureuses, n'importe où !

* * *

Gostaria muito de agradecer aos meus amigos e colegas do INESC-ID (Portugal), especialmente o Francisco, Catarina, Thomas, Mariana, Alberto e Isabel por todas as discussões amigáveis, interessantes e extremamente relevantes, e também por me incluírem numa grande variedade de projectos extracurrículares dentro do domínio da fala.

Queria também agradecer a todos os meus amigos e amigas que me têm acompanhado desde há muitos anos, desde o Sagrado, passando pelo Técnico até mais recentemente Toulouse. Apesar da distância, sempre me apoiaram e me motivaram a seguir os meus sonhos e ambições, e são também uma forte razão pela conclusão deste projeto. Gostaria de agradecer em particular às minhas amigas Beatriz e Rita por terem estado lá desde o princípio e por terem partilhado esta jornada de três anos comigo.

Finalmente, sem um suporte familiar sólido, esta tese não teria sido possível. Gostaria de terminar os agradecimentos por agradecer a toda a minha família, especialmente os meus pais Isabel e Pedro e os meus três irmãos Salvador, Santiago e Simão. Apesar de parecerem bastante felizes por não os chatear tanto dada a distância física que nos separa, acredito que têm tantas saudades minhas como eu deles. Gostaria também de agradecer muito aos meus avós Joaquim, Helena e Alice, e também aos meus tios e primos Rita, Ricardo, Margarida, Catarina, Luís, Madalena, José Maria, Leonor e Teresa.

A tua voz fala amorosa...
Tão meiga fala que me esquece
Que é falsa a sua branda prosa.
Meu coração desentristece.

Sim, como a música sugere
O que na música não está,
Meu coração nada mais quer
Que a melodia que em ti há...

Amar-me? Quem o crera? Fala
Na mesma voz que nada diz
Se és uma música que embala.
Eu ouço, ignoro, e sou feliz.

Nem há felicidade falsa,
Enquanto dura é verdadeira.
Que importa o que a verdade exalça
Se sou feliz desta maneira?

-Fernando Pessoa, 1929

Contents

List of Figures

Chapter 1

Introduction

A variety of pathologies that target the vocal tract, inner ear or nervous system are likely to affect the patients' communication ability [Kent, 1992]. This happens notably because of the anatomical and functional impairment of speech articulators, which is the case in **head and neck cancers**. The treatment of this set of cancers ranges from surgery to radiotherapy and/or chemotherapy, which, depending on the severity of the condition, can lead to severe speech impairments [Meyer et al., 2004]. Given this, in a clinical setting, the assessment of speech intelligibility is the key measure to evaluate functional deficit [Kent et al., 1994], and clinicians can make use of this measure in a variety of ways and for distinct purposes. An initial assessment evaluates communication disability and can set up the foundations for the therapeutic project and identify major or minor articulatory alterations. On the other hand, posterior assessments can measure the effects of treatment and give targeted hints on how to adapt it when not performing as expected. Despite this, the functional assessment of a patient with any **speech disorder** includes a variety of steps, which is also dependent on the underlying condition. Previous medical history, self-assessment questionnaires and examination of the articulatory organs are common practices for this type of clinical evaluation. Within these, speech intelligibility can be seen as an important indicator, which is usually evaluated perceptually. Given to the variety of speech-affecting disorders, it becomes relevant to emphasize that different diseases can affect speech production in distinct ways. This aspect can mean that acoustic biomarkers characteristic of patients that suffer from neurological diseases (e.g. Parkinson's) [Critchley, 1981], may not necessarily be the same for head and neck cancers. While the speech impairments experienced in Parkinson's cases result from the combination of motor and non-motor deficits, in head and neck cancers, the speech impairments result mainly from structural changes on the vocal tract, depending mostly on the tumor location [Bressmann, 2021]. Despite speech intelligibility being a relevant measure for both cases, the perceptual assessment taken in consideration for both cases differs due to the different types of speech impairments experienced.

Due to the fact that this thesis will be centered around the automatic prediction of speech intelligibility, it becomes crucial to have a proper definition of what speech intelligibility actually is. [Kent et al., 1989] consider that there are two distinct approaches used to assess this clinical measure. **The first approach is based on a subjective assessment made by clinicians. This could take the form of a score on a standardized scale after listening to the patient**. The second approach is **a perceptual objective measurement, usually obtained through the percentage of items that are accurately recognized by a listener**. These two approaches show that the definitions of speech intelligibility can be volatile, and change depending on the author or school of thought. There are a variety of other perceptual measures that are also used for the assessment of pathological speech. While these measures can relate back to intelligibility and sometimes seem interchangeable depending on the definition used, they serve distinct purposes, and therefore should be clearly disambiguated. This is the case for comprehensibility and for speech disorder severity. In the case of speech intelligibility and comprehensibility, the work of [Pommée et al., 2022] performed a consensus study on the scientific community on how should we differentiate these two terms. While both concepts are linked and contribute to functional human communication, they relate to two different reconstruction levels of the transmitted speech material.

Intelligibility refers to the acoustic-phonetic decoding of the utterance, while comprehensibility relates to the reconstruction of the meaning of the message. We can see that this definition of speech intelligibility relates to the more objective part of the two assessments proposed by [Kent et al., 1989], given it is more related to the number of recognized items by a listener. The disambiguation between speech intelligibility and speech disorder severity also becomes of high interest. While intelligibility relates more to comprehensibility and what is being perceived by the listener, speech disorder severity can be seen as a more global measure. In this case, various elements of the vocal signal are taken into account, such as the quality of the speech rate, acoustic phonetic decoding and other prosodic parameters relating to the perceived speech impairment [Kent et al., 1989, Yorkston et al., 1996].

The set of definitions previously introduced sheds some light on how subjective these measures can be. Subjective estimations are widespread in clinical practices essentially because they are easy to conduct. However, the listener subjectivity introduces a variable that is very difficult to control. The different judges used can be conditioned on a variety of aspects, such as the task or the patient being assessed, or even on previously assessed patients. The different yet similar definitions used for the distinct perceptual measures can also generate lack of consensus on a given set of judges. These aspects contribute to making the assessment of speech intelligibility a difficult task to reproduce, which can also be highly biased and variable. Given this, the **automatic prediction of speech intelligibility** becomes a hard problem to tackle not only due to the available amounts of data characteristic of pathological speech, but also due to the variability associated to the reference labels used. On the other hand, we can start to see why an automatic approach to speech intelligibility is appealing. The modeling of these perceptual measures allows the output of a standardized score that does not change based on previous assessments or on previous knowledge of the patient. It becomes then clear to see that an automatic model can be seen as a more objective, unbiased and non-variant way to predict speech intelligibility, which becomes highly relevant [Fex, 1992, Middag, 2012]. While the goal of medical technologies is never to replace a doctor or therapist, these models can provide a second opinion or even free the practitioner to perform other more relevant tasks.

During the course of this thesis, I intend to explore the automatic prediction of **speech intelligibility** for head and neck cancers using the artificial intelligence subfield known as **deep learning**. From this topic, a variety of research questions can be posed, that will jump-start the research developed.

The first research question that appears is centered around the main area for this thesis: *Can deep learning be reliably used to predict speech intelligibility?* The vast amount of advances in deep learning witnessed in the past decade shows that it is an evolving domain that has a variety of applications, in a variety of distinct fields. Nevertheless, deep learning is known to require extensive amounts of data, which will always be a constraint when handling the limited amounts of data that pathological speech is characteristic for. Techniques such as data augmentation and transfer learning become crucial when dealing with smaller amounts of data. Furthermore, the proposed systems should also scale easily with the advent of possible future data acquisition campaigns. While the quantity of data is highly relevant for the development of this type of system, the quality of the reference labels used is similarly important. The subjectivity around intelligibility measures becomes an interesting research problem to tackle that generally pose issues to the development of objective systems. Given this, the present research problem can be divided in two distinct domains. The first focuses on the small **quantity** of data available and the methods used to mitigate this aspect. The second domain deals with the **quality** of the reference measures used, where judge variance and occasional lack of inter-rater agreement should not be neglected.

The second research question to be explored is: *How can we build trust in these methodologies in order for them to be adopted in a clinical context?* Despite the recent advances in deep learning, the subject of interpretability is often neglected, mainly since some of the application domains do not require a thorough justification of the scores obtained. While the results on some of these domains can speak for themselves and do not require extensive comprehension, in the case of pathological speech that does not necessarily hold. A certain degree of interpretability should always be promoted in order to better validate the automatic intelligibility measures regressed, this aspect not only helps build trust but it could also contribute to turning a subjective measure more objective. The degree to which a score becomes explainable becomes an interesting problem to tackle, that will always be in consideration amidst the

methodologies developed in the present work. In order to better address the subject of interpretable scores, interpretability/explainability will be defined as the degree to which a human can understand the cause of a decision [Miller, 2017]. From this definition, it becomes relevant to emphasize that interpretability should be seen more a spectrum instead of a simple category (interpretable *vs.* non-interpretable). Given the intricate nature of deep learning systems, the usage of intermediate prediction scores or meaningful features that are clinically valid can provide this extra layer of explainability. These approaches can be seen as more interpretable when compared to, for example, fully end-to-end systems.

In an automatic context, it becomes relevant not to be limited to a single approach or measure, specially since there is more than one approach used to assess speech intelligibility perceptually. The process of acoustic-phonetic decoding previously introduced, known as the process of decoding speech units and posterior conversion to graphemes, is one of the main approaches to assess intelligibility, that can be seen as more objective than the subjective assessment. The speech units of this process can take different shapes depending on the test, ranging from phonemes and syllables to words or even full sentences. Since each one of these different levels is relevant in its own way (e.g. phonemic analysis can target key mispronunciations while sentence analysis can assess day-to-day communication), a **granular approach** that targets these distinct levels becomes highly interesting in the context of an automatic analysis. On the other hand, it is also known that parameters such as voice quality and prosody can also play a relevant role for speech intelligibility besides the acoustic-phonetic decoding process. This aspect is more prominent in the subjective assessment (see [Kent et al., 1989]), and therefore should not be neglected either. The relationship between these parameters and speech intelligibility also poses an interesting research question of whether different judge profiles exist, that share different opinions concerning the relevance of these parameters. The possible added benefits of modeling individual schools of thought arises as a follow-up to this question, and it becomes an interesting and relevant research field. By doing this modeling, a better understanding of what makes intelligibility subjective could be achieved, which could possibly lead to better automatic predictions.

Given the aforementioned facts, during the course of the present work, we will explore the **automatic prediction of speech intelligibility using the distinct levels: sentence, word and phoneme**. Since deep learning methodologies will be the central point of the present work, and that some degree of explainability should always be considered, the first two research questions become part of a bigger and more relevant research question: *Can a granular analysis work when predicting speech intelligibility? If so what are the added benefits?* This final research question can be seen as the most relevant one, and the resulting granular approach will serve as the backbone of the entire thesis. Figure 1.1 illustrates an overview of the proposed approach, with the three distinct levels marked, followed by a potential unified score.

In order to start this thesis, the upcoming chapter will present the scientific and application context in which the present work was developed. Here I will give an introduction to the broad field of deep learning, and also introduce the corpus that will be used throughout: the French Corpus of Head and Neck Cancer. Concerning the experiments, starting at chapter 3, we will explore the automatic prediction of speech intelligibility at the first granular level, also known as the sentence-level. Chapter 4 will be focused on the word-level granularity systems developed, while chapter 5 will be centered around the final level of granularity: the phoneme. Furthermore, a study on the relationship between parameters such as voice, resonance, prosody and phonemic distortions and the measure of speech intelligibility will be explored in chapter 6. This study will be conducted by modeling a set of perceptual judges in order to understand to what extent these parameters are related, and also to identify the different schools of thought. Finally, a unified approach that will merge all the different granular systems will be explored in chapter 7. This chapter can be seen as the culmination of all the different systems proposed and how a unified model based on the granular predictions could be seen as a better approach than the sum of its parts. To close this thesis, I will present the main conclusions, followed by a variety of perspectives and future work suggestions.

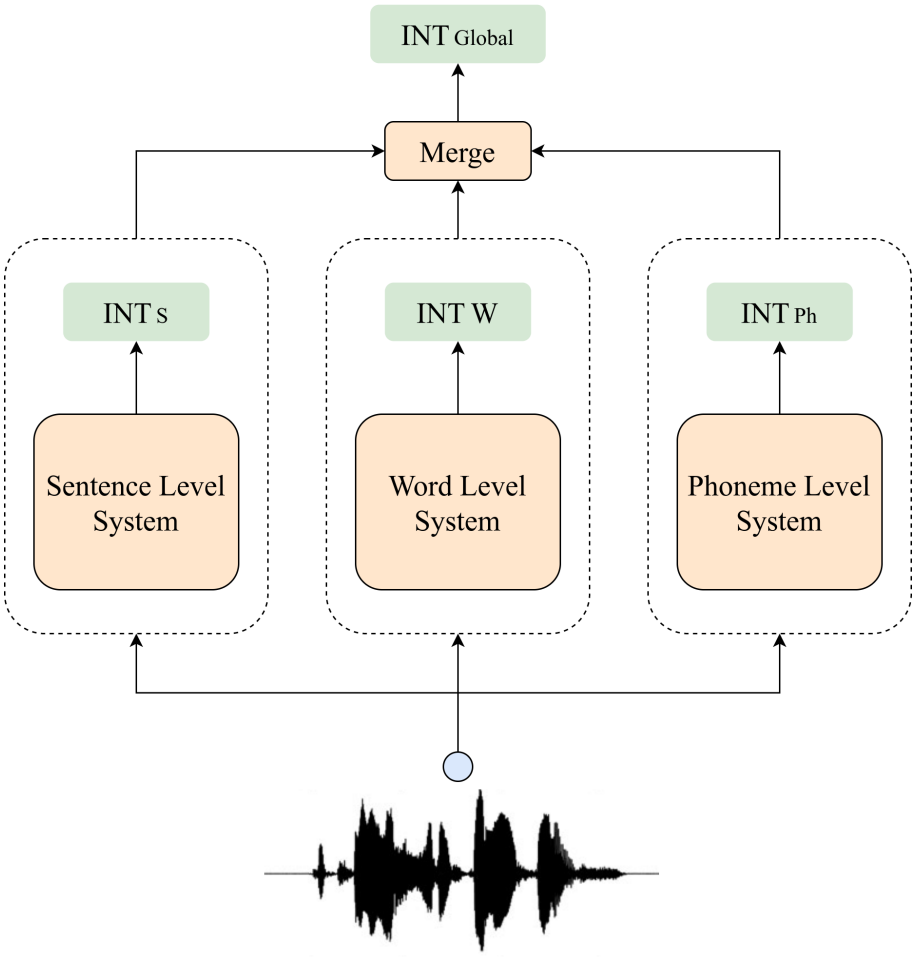


Figure 1.1: Overview of the global methodology proposed for the automatic prediction of speech intelligibility of head and neck cancers.

Chapter 2

Scientific and Application Context

In this chapter, we will present an introduction to the scientific and application context in which this thesis was written. This chapter will be organized in three distinct parts, that are complementary to each other. We will start by providing an overview on the methodologies used to predict speech intelligibility, introducing the state-of-the-art in a general way without focusing too much on technical details. The second part will be focused on the domain of choice for this work: deep learning. Here a targeted introduction to the domain will be presented, with a particular emphasis on techniques that are useful for speech processing. These two parts will mainly introduce the scientific context in which this thesis was written. The third and final part will introduce the French head and neck cancer speech corpus, which is the dataset that will be used throughout this thesis and the main application context as well.

2.1 Methods Used to Define Predictive Intelligibility Measures

From the literature, there are a variety of methodologies used to automatically predict subjective measures such as intelligibility. In this section, we aim to introduce the general approaches used recently to predict speech intelligibility. Moreover, this section will not serve as an extensive list since a briefer and more contextualized state-of-the-art will be given at the beginning of each chapter, that will target the intelligibility prediction at the different granular levels proposed. By doing this, the reader will hopefully be provided with a more targeted view on each specific topic, that will enhance the comprehension of this manuscript. Due to the latest advances in machine learning, and more specifically deep learning, the majority of the recent approaches tend to touch on these two topics. While some approaches may be more tilted towards classic signal processing instead of statistical/heuristic methods, all of them share a numerical and algorithmic basis. Given this, there is an unavoidable need for data in order to better develop these measures. Since pathological speech is a domain with characteristic scarce amounts of data, it becomes an interesting research field to explore, specially when concerning subjective measures like intelligibility. On a general view, for the automatic prediction of speech intelligibility two main schools of thought can be identified.

The first one is based on the extraction of a score as the result of the word error rate achieved by automatic speech recognition [Christensen et al., 2012, Arai et al., 2019]. This type of approach is highly common and straight-forward to implement, since there is only the need for a performant ASR system. These approaches, however, are not suited for every type of speech task. In the case of speech tasks such as pseudo-words and pseudo-sentences or even semantically unpredictable sentences, widely used in clinical contexts, the ASR systems tend to malfunction for not having tailored language models to that type of data. Still within the context of ASR models, the usage of phonological features is also a valuable method used to predict speech intelligibility [Middag et al., 2009a]. These features, obtained through methods such as forced alignment (see annex B for more details), can be used to derive an intelligibility score. While interesting results can be achieved, these approaches are also known to underperform on severe speech impairments.

The second type of approach aims to extract relevant features from pathological speech by using au-

tomatic speech processing technologies, and then output a predicted intelligibility score [Xue et al., 2019, Middag et al., 2009b]. This type of approach encompasses a large variety of systems, from purely based on signal processing [Janbakhshi et al., 2019b] to data-driven approaches [Díaz and Antolín, 2020]. It is highly common for these approaches to use traditional speech processing features [Alim and Rashid, 2018]. Some classic speech processing features are the Mel-Frequency Cepstral Coefficients (MFCC) and their respective delta coefficients, Mel filterbanks, Linear Prediction Coefficients (LPC) and more recently phonetic posteriors [Vasquez-Correa et al., 2019]. The majority of these features have been around for quite some time and are known to also hold well on fields such as linguistics and speech therapy. Moreover, recent systems such as Wav2Vec [Schneider et al., 2019] allow deep learning algorithms to operate directly on raw audio waveforms and learn features and representations themselves. These representations are quite useful to reduce the data required to train high-performance acoustic models. Wav2Vec 2.0 [Baevski et al., 2020] builds on this notion but with an added degree of facility to adapt to smaller amounts of data. The working principle behind Wav2Vec can be split in two distinct parts: an encoder and a contextual network. The encoder, through the means of a convolutional block, operates directly on the raw waveform signal to generate a latent representation. The contextual network uses these generated latent representations to create a contextualized representation, followed by a linear projection to the output. The system's pre-training is the most relevant part. Here, the output of the encoder is masked and passed on to the contextual network that reconstructs the masked parts (without the final linear projection). The resulting reconstructions are then compared to the ground truth through the means of a contrastive loss, that distinguishes real targets among distractors. Furthermore, the resulting pre-trained model can be adapted to tasks such as emotion recognition or automatic speech recognition. Other models, such as PASE+ [Ravanelli et al., 2020], can extract relevant speaker information, including speaker voice-print and phonemes, from unlabelled data. This model can also be interesting for the automatic prediction of intelligibility measures [Roger et al., 2022]. These approaches can also be applied to speech intelligibility prediction [Hernandez et al., 2022]. The main advantage lies in the usage of an already learned system to be adapted to the particular case of speech intelligibility. The feature engineering aspect of this school of thought may also affect interpretability, either due to the "black-box" nature of some heuristic approaches, or to the extensive feature analysis required to prove the correlation between some speech impairments and key features.

From the two previous general approaches presented, we can see that it becomes difficult to find a one-fits-all approach to predict intelligibility. Similarly to the different intelligibility measures that can be used for different clinical purposes, different automatic models can be used to assess different aspects of speech intelligibility. While a single model prediction may be unattainable not due to performance, but mainly due to the different facets of speech intelligibility, a combined approach could mitigate the drawbacks of each individual system (or perceptual measure used as reference). The duality between the advantages and drawbacks of some automatic approaches, as well as the combination of different systems/measures are interesting research questions that will be explored during the course of this work.

In the context of speech intelligibility, it also becomes relevant to make the distinction between intelligibility for pathological speech and intelligibility for speech in noise. Despite both topics being centered around the understanding of spoken language, the latter is more related to speech enhancement and denoising, and hence the end application of both approaches greatly differ. While the definition of intelligibility may be similar to both, the perceptual decoding also differs between the two. Despite this, some common ground can be found, and some approaches for speech in noise can also provide valuable insight for the pathological speech counterpart. Objective speech intelligibility predictors used for speech in noise, like STOI [Taal et al., 2011], ESTOI [Jensen and Taal, 2016], SIIB [Kuyk et al., 2018] and HASPI [M.Kates and H.Arehart, 2014] can also be applied towards a pathological speech context [Huang et al., 2022a]. These predictors, however, require the test signal to be time-aligned with a reference intelligible signal to obtain a divergence measure. While the usage of utterance-dependent reference signals (acquired from several healthy speakers) helps achieving the said divergence [Janbakhshi et al., 2019a], in a clinical context it becomes unfeasible to perform, specially for tasks such as spontaneous speech or picture description. Recent works, such as [Andersen et al., 2018] and [Pedersen et al., 2020], present end-to-end approaches applied to the speech in noise paradigm. It is important to state that these predictors (STOI, ESTOI, etc.) are based on psychoacoustic models and heuristics, and validated empirically using small datasets. On the other hand, due to the nature of deep

learning, during the course of the present work we will dive into the domain of data-driven approaches to estimate the intelligibility measures.

2.2 Artificial Intelligence and Deep Learning

In this section we invite the reader to dive into the topic of deep learning. In the following sections, the main goal is to provide a general introduction to the broad field of deep learning, and its current architectures that can be applied to automatic speech processing. By doing this, we will hopefully provide the reader the proper scientific basis to better understand the work developed during this thesis.

In order to answer what deep learning is, it becomes relevant to introduce it within the broader context of **artificial intelligence** (AI) [Dick, 2019]. This field encompasses every single technique that allows machines to simulate human intelligence, specially in the context of computer systems. It is clear that this domain remains vast, as a variety of techniques can be implemented to mimic human behavior. A subset of these techniques can allow, for example, a software application to become more accurate at predicting certain outcomes without being explicitly programmed to do so. The set of techniques that encapsulate this type of methodology can be inserted within the subset of AI called **machine learning** [Naqa and Murphy, 2015]. These (normally) data-driven approaches tend to use a variety of statistical methods to enable machines to improve with experience. By going deeper inside machine learning, we can finally find **deep learning** [Goodfellow et al., 2016]. Similarly to machine learning, the decision-making behind this methodology is also based on data that was previously seen by the system. What sets the two apart is the usage of artificial neural networks, a computing system inspired by the biological neural networks that constitute animal brains. These specific networks are what deep learning is all about, and will be the main focus of attention during the course of the present work. Figure 2.1 presents the relationship between artificial intelligence and the sub-branches that will be useful for this thesis.

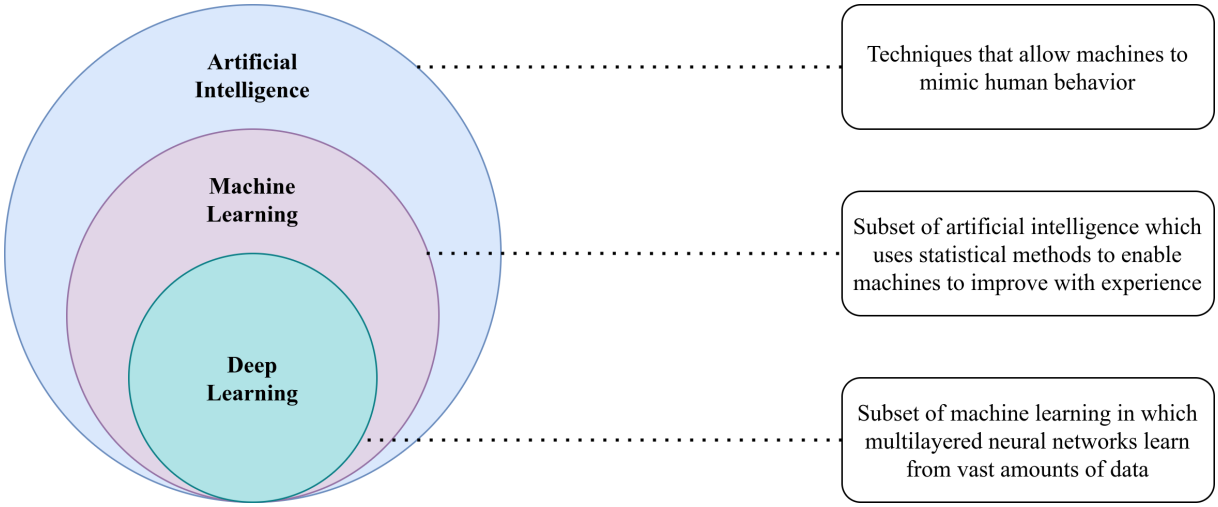


Figure 2.1: Relationship between artificial intelligence, machine learning and deep learning.

In the past few years, deep learning has been a growing methodology used to solve a variety of problems. Generally, we can group the type of problems that deep learning tackles in two categories: classification and regression problems. Classification models aim to predict discrete class labels, some examples can range from Alzheimer's detection to sentiment analysis, while regression models aim to predict a continuous quantity, such is the case for speech synthesis and the task of automatic prediction of speech intelligibility. Machine learning and deep learning models learn by seeing vast amounts of data. Through an iterative process called training, different subsets of data are seen by the deep learning algorithm, that ends up conditioning the entire system on the task at hand. Given this, it becomes

relevant to make a distinction between different types of learning. Supervised learning is known as a class of systems that learns from labelled amounts of data. This is also the type of learning that will be implemented and experimented throughout this thesis. On the other hand, in unsupervised learning no data labelling is used, and therefore the system is left to do logical connections based solely on the data seen (e.g. pattern recognition). The topic of semi-supervised learning has also seen a growing interest in the past few years. In this case, only a subset of the original data is labelled, and the system falls between supervised and unsupervised approaches.

2.2.1 Artificial Neural Networks

As it was previously stated, neural networks are not only the core subject of deep learning, but also what sets it apart from other machine learning methodologies. Given this, it becomes relevant to introduce what a neural network is within the context of deep learning. An Artificial Neural Network (ANN) is a type of computing system that uses a collection of layered units, called nodes or neurons. Similar to the synapses seen in animal brains, each connection between neurons can transmit a signal to the posterior connected neurons. While in the brain, the synapses are conveyed by electrical pulses, the signal at a connection is a real number, and the output of each neuron is computed by a non-linear function of the sum of its inputs. Neurons are typically aggregated in layers that perform different types of transformations, and normally each connection has an associated weight that adjusts as learning proceeds. The larger the weight, the strongest the signal at the connection is. Figure 2.2 presents an example of an artificial neural network, with two hidden layers, also known as the layers were the neurons are placed. A close-up of the structure behind a neuron is also presented in the same figure. For the sake of clarity, during the course of this thesis we will refer to artificial neural networks as all types of neural networks presented, Deep Neural Networks (DNN) as artificial neural networks with more than two hidden layers and shallow neural networks (SNN) as networks with two or less.

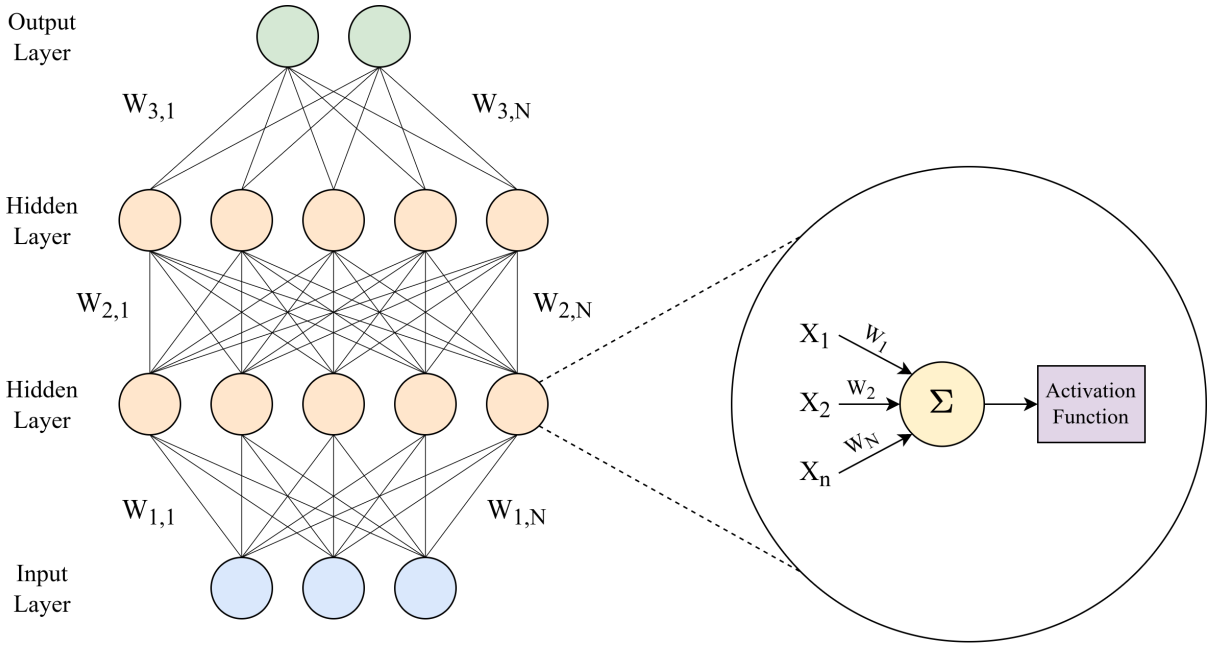


Figure 2.2: Example of a simple artificial neural network and the structure of a neuron.

Given this smaller introduction to neurons and Deep Neural Networks (ANNs), it becomes relevant to explain how these systems can learn from data seen during the process called training. As it was previously mentioned, each connection between neurons has a weight associated to it (see figure 2.2), which is typically randomly initialized. The sum of these weights is fed to the neuron's activation

function, that will define which neurons are activated, and at which intensity. The entire pipeline of activated neurons and their respective values will transport a given input to a solution, where the neurons of a given layer are conditioned by the response of the neurons located in the previous layers. There are different activation functions that can be used, each with a given set of advantages and disadvantages. The most common ones are respectively the sigmoid, hyperbolic tangent (tanh) and the Rectified Linear Unit (ReLU) [Nair and Hinton, 2010], illustrated by the following equations:

$$s(z) = \frac{1}{1 + e^{-z}} \qquad \text{(sigmoid)} \tag{2.1}$$

$$t(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \qquad \text{(tanh)} \tag{2.2}$$

$$r(z) = \max(0, z) \qquad \text{(ReLU)} \tag{2.3}$$

Once all the network is traversed, the final layer will issue an output, whether a class, sequence or set of values depending on the type of problem. In the case of supervised learning, this output will be compared to the label or ground truth associated to the input seen at the beginning of the network. The output and ground truth are then compared through the means of a loss function, or a function that calculated the distance/difference between the two. This function allows the system to know how close or far away from target the output really is. Once that difference (or loss) is known, the network can be updated through a process called **backpropagation** [Shun-ichiAmari, 1993]. This process allows the computation of the gradient of the loss function with respect to the weights of the network for a single input–output example. The main goal of backpropagation is to minimize the loss function, and hence achieving more accurate outputs. The training process of a neural network encompasses several backpropagation iterations, that generally see a large variety of data to help the generalization ability of the system. The iterated gradients are then updated according to a parameter known as **learning rate**, which controls the degree of change seen in the weights: a larger learning rate is equivalent to a larger change in the weights, which will have a greater impact on the network's outputs. There are a variety of algorithms that apply backpropagation, all of them generally building on the topic of the gradient descent. The Adaptive Movement Estimation algorithm (Adam) [Kingma and Ba, 2015], for example, is commonly known as one of the best and most efficient algorithms to perform gradient descent in regression problems. The continuous search for optimal deep learning optimization algorithms is an active and ongoing research field [Reddi et al., 2018]. Since deep learning normally makes use of larger amounts of data, it becomes unfeasible to process each input one by one, as it would already increase the already large processing time required. In order to tackle this issue, a concept known as **batch learning** is introduced. This concept makes use of a batch of randomly sampled inputs from the training corpus and performs the gradient descent based on the mean loss of each batch, instead of doing it individually. By performing this type of learning, the processing time is shortened and the generalization ability tends to improve. Batch sizes are typically powers of 2 (e.g. 16, 32, 64, etc.). There is no linear relationship between batch size and system performance. Therefore it is a parameter that should be fine-tuned manually. In order to further help the generalization ability of a given network, the normalization of the elements in a batch (**batch normalization**) is a method commonly implemented. This technique allows the stabilization of the learning process by standardizing the inputs, and therefore providing some regularization which helps reducing generalization error. Another common generalization technique used is the so-called **dropout**. In this technique, a set of random neurons are intentionally left out of a given training iteration in order to make the network more robust. The dropout parameter [Hinton et al., 2012] can be used on any layer, and can be fine-tuned as well.

In the previous paragraphs, a small introduction to the world of deep learning was done, where we have covered the basic methodologies and procedures used to train neural networks. On the other hand, due to the sheer size of ongoing developments and new methodologies that appear, it becomes hard to cover all possible topics that deep learning encompasses. Given this, in the next paragraphs we will shift our focus of attention into deep learning methodologies that are more characteristic of continuous signal processing, which can be applied directly to the case of speech processing.

2.2.2 Recurrent Neural Networks

In the previous paragraphs, a basic neural network (see figure 2.2) was used as an example to explain how it can learn from the data seen during training. On the other hand, the presented neural network has the limitation of having to work with a fixed-sized input, since the input layer has a predefined number of neurons. While fixed-sized input layers may not pose any issue to systems whose inputs always have the same dimensions (e.g. data spreadsheets, image processing after resizing, etc.), when handling continuous signals it becomes a problem. It is then easy to see why continuous signals with varying dimensions, such as speech, may pose problems to these networks. In order to tackle this issue, a new type of neural network needs to be used, which will be introduced in this section.

Recurrent neural networks (RNNs) [Liao et al., 2002, DiPietro and Hager, 2020] are a class of neural networks that are naturally suited to processing time-series data and other sequential data. This type of network allows not only the processing of variable-length sequences, but also the complex learning of inter-dependencies between the input's time steps. This process is achieved through the internal memory that the composing cells of these networks possess. By remembering different parts of the input sequence, the system can learn the task at hand by extracting context from different parts of the signal, which enhances the system's performance. Figure 2.3 illustrates a simple RNN example. The snippet of the last hidden cell displays the working principle behind these networks. At each hidden block, two inputs are given, the first corresponds to the time step to analyze (X_T) and the second to the activation results from the previous hidden cell (X_{T-1}). Each input has its respective weight. The combined dot product of the two inputs is then fed to an activation function (hyperbolic tangent), which produces not only the output of that cell, but also the hidden state to be fed to the posterior hidden cell. The passing of successive hidden states allows the network to build on the previous cells context, which permits the modelling of temporal dependencies. RNNs can also be bidirectional, in this case a second hidden layer is added that extracts context from future time steps. Similarly to the artificial neural networks previously introduced, RNNs are trained and optimized through backpropagation. RNNs can also be stacked, in this case the input of the posterior hidden layer corresponds to sequential output of the previous one. Intermediate representations (or embeddings) of the network can also be used to connect the network with regular feedforward networks. In this case, the weights of the final recurrent layer are extracted (fixed dimensions) and fed as an input to the rest of the system. These embeddings are known to encapsulate sequence properties and allow the rest of the system to work on fixed length representations instead of a sequential input. The number of hidden layers and cells are the main parameters to be chosen when planning a recurrent architecture.

In the previous paragraphs, a brief introduction to the topic of RNNs was given. Despite being performant systems, these networks face some issues. One of the biggest issues that the vanilla RNN presented faces is the problem of vanishing gradients during backpropagation. This problem arises when handling longer sequences, in which the gradients get progressively smaller and therefore the weight optimization becomes almost non-existent for those same gradients, hurting the learning capabilities of the system when handling earlier time steps. The opposite problem of exploding gradients is also common in RNNs. In this case the gradient gets increasingly high, resulting in dangerously high weight updates. In order to solve these two issues, different hidden cells are proposed, that are usually adopted instead of the vanilla RNN cell. Two of the most famous cells are the Long Short-Term Memory cell (LSTM) [Hochreiter and Schmidhuber, 1997] and the Gated Recurrent Unit (GRU) [Cho et al., 2014]. These cells can be found illustrated on figure 2.4.

The aspect that sets apart an LSTM from a vanilla cell is the introduction of memory in the system. As a consequence, the cell remembers values over arbitrary time intervals. In this specific case, the flow of information in and out of the cell is regulated by three gates, which can also be found illustrated on figure 2.4. Besides the hidden state, the LSTM introduces a new input and output when compared to the vanilla cell (and later to the GRU): the cell state (C_T and C_{T-1}). On a high level, the cell state is given by the top horizontal line present in the same figure. This line works as a conveyor belt that is conditioned by the three gates. The first gate, named **forget gate**, is responsible for removing information from the cell state, which makes the decision between information to retain and information that is no longer required for the network to understand things. The **input gate** is responsible for the addition of information to the cell state. This gate regulates the information that needs to be added to the

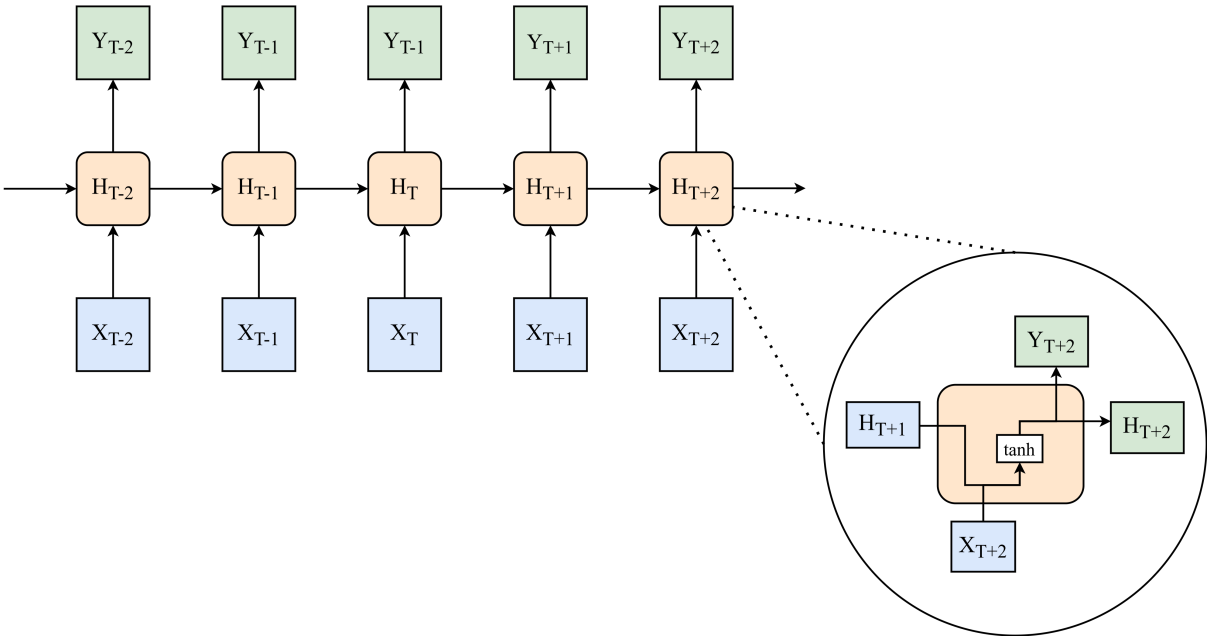


Figure 2.3: Illustration of a RNN with a snippet on a simple hidden cell. Each hidden cell takes in consideration the respective time step as well as the output from the previous time steps, which provides context for the entire sequence.

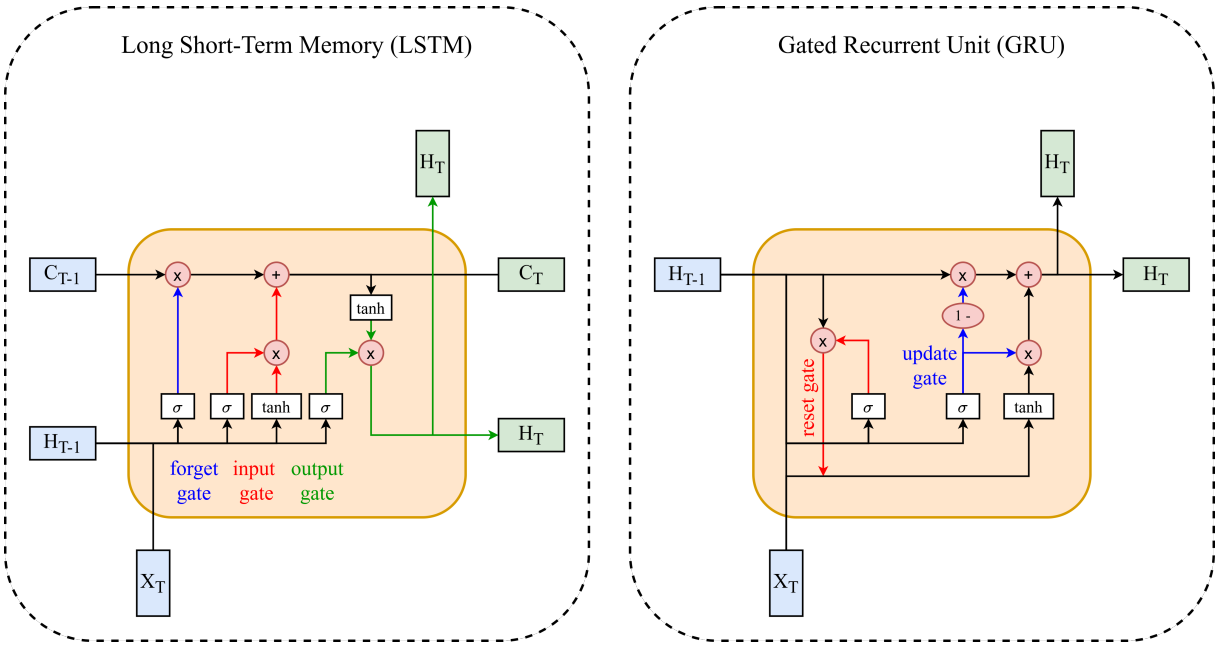


Figure 2.4: Illustration of the Long Short-Term memory (LSTM) and Gated Recurrent unit (GRU) cells, widely adopted to prevent vanishing/exploding gradients. The characteristic gates of each cell are also illustrated.

cell state, and discards unimportant or redundant information. Finally, the **output gate** is responsible for deciding the next hidden state (see that the output of this gate does not affect the cell state). The

conditioning and filtering promoted by the three gates allows the LSTM to develop long-term memory, which encapsulates important context while neglecting less relevant bits of information.

The second cell proposed to counter vanishing/exploding gradients also works in a gated fashion, however, without a cell state. The GRU (also illustrated on figure 2.4) is a variation of LSTM cell, and the two composing gates help to ensure that its memory is not taken over by tracking short-term dependencies. The network learns how to use both gates essentially to protect the memory, and therefore make longer-term predictions. The main role of the **reset gate** is to combine the input with the previous hidden state, while the **update gate** defines how much of the previous memory is to be kept around. Should the reset gate be set to all 1's and the update gate to all 0's we arrive at our vanilla RNN cell. Both LSTMs and GRUs share a common working principle, however, with some key differences: the lack of cell memory and the smaller number of gates on the GRU.

Due to the capacity of handling sequences that RNNs possess, they become highly attractive for speech signal processing. Recent venues suggest that these types of networks can be applied to speech tasks such as speech enhancement [Zhao et al., 2018], speech emotion recognition [Mirsamadi et al., 2018] and automatic speech recognition [Amberkar et al., 2018]. In the previous paragraphs, we merely touched the tip of the iceberg concerning different recurrent structures. For more complex problem solving, different methodologies can be built around the concept of recurrent neural networks, such as the attention mechanism and the embedding paradigm. These topics will be introduced and explored later in the manuscript, and some theoretical fundamentals will be given more contextually beforehand.

2.2.3 Convolutional and Time-Delayed Neural Networks

Within the broad context of deep learning, there is also the type of network called convolutional neural network (CNN) [LeCun and Bengio, 1995, Abdel-Hamid et al., 2014]. These systems build on the concept of ANNs, however, what sets them apart is the usage of the mathematical operation called convolution, in place of general matrix multiplication in at least one of their layers. A convolution is a mathematical operation on two functions (f and g) that produces a third function ($f \circledast g$) that expresses how the shape of one is modified by the other. This operation is an integral part of signal processing [Burrus, 1985]. In the context of CNNs, the learnable parameters are convolution kernels, also known as transformation matrices, that traverse the input signal and output a transformed, convoluted output. This convoluted output is a feature map that is fed to the posterior layers of the network. Convolutional layers are usually connected with pooling layers, that down sample feature maps by summarizing the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling that summarize the average presence of a feature and the most activated presence of a feature respectively.

Due to the automatic learning of convolution kernels, which output feature maps, CNNs typically do not require a lot of pre-processing, hence making them highly attractive and useful for image processing tasks. While a large part of the CNN applications are set towards image or computer vision tasks, they still remain a highly interesting and useful architecture for speech processing. Figure 2.5 illustrates a CNN applied to a speech signal (spectrogram). The vocoder model known as wavenet [van den Oordan et al., 2016], used for speech synthesis and music generation, makes use of a fully convolutional architecture, for example. Within the context of speech intelligibility, recent works also show promising work towards interpretability and pathological speech modelling through the means of CNNs [Abderrazek et al., 2020]. Some other works also show that CNNs can be used for the prediction of speech intelligibility in noisy environments [Andersen et al., 2018, Pedersen et al., 2020].

A good understanding of the working principle behind a CNN allows the comprehension of yet another class of deep learning models: time-delayed neural networks (TDNN) [Waibel, 1989]. This class of neural networks is widely adopted in speech processing, specially in acoustic models for speech recognition [Peddinti et al., 2015]. The working principle is quite similar to CNN, the main differences lie in the kernel size of the convolution kernel, which is always one-dimensional for TDNNs, and also the lack of pooling layers.

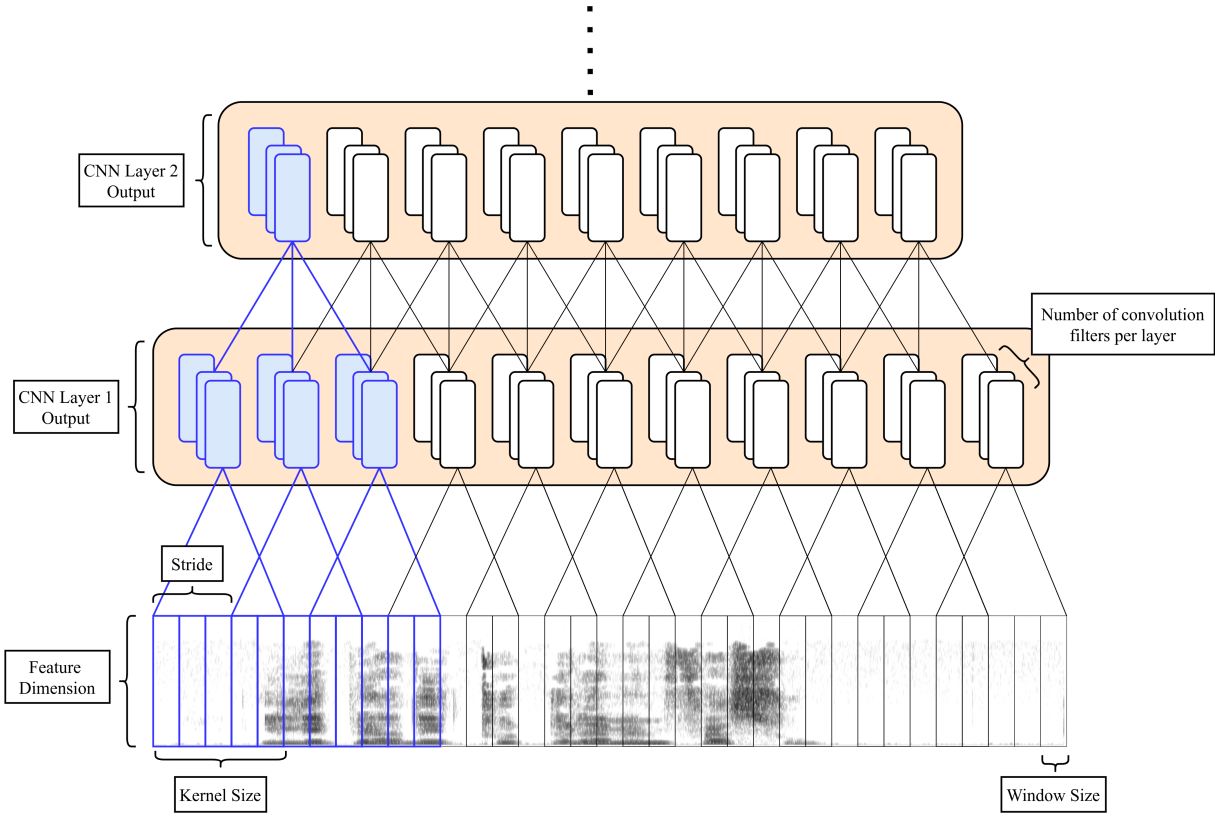


Figure 2.5: Illustration of the working principle behind a CNN. The most relevant parameters, such as kernel and window size, stride, feature dimension and number of convolution filters can be found marked in the figure as well.

2.2.4 Interpretability & Explainability

When developing systems that have a clinical application in mind, such as the automatic prediction of speech intelligibility, besides being unbiased and reproducible, it is also highly relevant that an automatic approach maintains some degree of explainability, which typically lacks in the automatic systems based on deep learning. On the other hand, there is a school of thought concerning "black box systems", stating that data-driven approaches should be believed regardless of their explainability [G.McCoy et al., 2021]. While this may indicate that there is no intrinsic value in the explainability of deep learning algorithms, and that performance and trust should be followed instead of an uncompromising search for interpretability, it is important to realize that performance systems tend to require large amounts of data to obtain good empirical results. These large amounts of data are scarce for rare and underrepresented diseases, such as head and neck cancer. Due to this, explainability becomes relevant, as it could cover possible liabilities of the system. An explainable system could provide more relevant cues in a clinical setting and promote more objective measures. The interpretability of the results, that normally lacks in machine learning systems, can also be used to build trust in the implementation of automatic approaches [Diprose et al., 2020], specially given small amounts of data. We believe that explainability should be followed up to a certain point instead of a relentless search for optimal results, as in our particular case we are dealing with the automatic prediction of a measure that is subjective by nature (speech intelligibility). Since consensus between perceptual judges can be difficult to achieve in some cases, unexplained results do not hold well in a clinical context and therefore some degree of interpretability should always be promoted. The field of explainability and interpretability has seen a recent growing tendency, specially when concerning medical or clinical data [Chakraborty et al., 2017, Ghoshal and Tucker, 2020, Schutte et al., 2020].

In the previous sections, a brief introduction to deep learning and its methodologies typically used for speech processing was presented. Since we believe there is no need for excruciating technical details about every single deep learning methodology used during the course of this work, each following chapter will have a specific state-of-the-art and methodology section that will introduce the relevant deep learning techniques used. By doing this, a more contextual approach to each chapter is given which will hopefully help comprehension.

2.3 The French Head and Neck Cancer Speech Corpus

In this section, we will introduce the corpus used during the course of this thesis, the **French head and neck cancer speech corpus (C2SI)** [Woisard et al., 2020]. The corpus contains a large database of French speech recordings aiming at validating disorder severity indexes. The database was recorded for the purpose of measuring the impact of oral and pharyngeal cavity cancer on speech production, as well as to assess patients' quality of life (QoL) after treatment. A declaration was made concerning data processing to the French national data protection authority (Commission nationale de l'informatique et des libertés) (number 1876994v 0 July 24, 2015), and a favorable decision was obtained from the research ethics committee of Toulouse Hospitals on May 17, 2016.

The introduction to the C2SI corpus will start by a brief corpus description, showcasing the speaker's distribution as well as some statistics, followed by the recorded speech tasks and ending on the perceptual measures that will be used as reference throughout the present work.

2.3.1 Corpus Description

The corpus has a total of 127 speakers, out of which 87 are patients that suffer oral cavity or oropharyngeal cancer and also 40 healthy speakers. Seven patients were recorded twice, which comprises a total of 134 recordings. All cancer patients have undergone at least one cancer treatment, such as surgery, radiotherapy and/or chemotherapy. The applied cancer treatment lasts at least six months, after which the disorders are considered stable. The criteria of non-inclusion in the corpus were to present another source of speech disorders (e.g. stuttering) or to present cognitive or visual problems that are incompatible with the assessment protocol design. The same non-inclusion criteria were applied to the controls. Among the patients, 59% were men and the remaining 41% women, with a mean age of 65.8 years old, comprised in the range of 36-87 years. For healthy speakers (controls), 45% were men and the remaining 55% women, with a mean age of 56.9 years old comprised in a range of 35-79 years.

2.3.2 Recorded Speech Tasks

Within the C2SI corpus, different speech tasks were recorded [Woisard et al., 2020]. Due to the large variety of recorded tasks, we will aim to give a simple introduction to each one in this section. This will give some context on the different tasks that are available to use, as well as lay the foundations for a more in-depth explanation of each task. Later on, during the course of this thesis, we will elaborate further on the speech tasks used. It is important to note that all patients did not complete all listed speech tasks.

- **Sustained Vowels:** These recordings consist in the production of three times the vowel /a/ in a sustained way. A sustained vowel gives information about voice level, phonation time, stability, harmonics contents, noise, unvoiced segments, etc.
- **Pseudo-words:** Each speaker had to pronounce 52 pseudo-words. The pseudo-words have the following phonotactic structure: $C(C)_1V_1C(C)_2V_2$, where $C(C)_i$ is an isolated consonant or a consonant cluster.
- **True/False Sentences:** A set of 50 sentences selected from the list of 300 sentences was produced by each speaker. These sentences have a specific syntactic-semantic structure, whereby the true or false property can be checked only when the last lexical unit was produced (e.g. "Paris is the capital of France" *vs.* "Paris is the capital of Germany"). Consequently, it is necessary to decode and understand the whole sentence before coming up with the answer.

- **Reading Passage:** The first paragraph of "La chèvre de M. Seguin", a tale by Alphonse Daudet, was read by the speakers. This text was chosen because it is long enough and it encompasses all French phonemes. The full passage is as follows: *"Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon. Un beau matin, elles cassaient leur corde, s'en allaient dans la montagne, et là-haut le loup les mangeait. Ni les caresses de leur maître ni la peur du loup rien ne les retenait. C'était paraît-il des chèvres indépendantes voulant à tout prix le grand air et la liberté."*

- **Prosodic tasks:** Different prosodic tasks were recorded, focusing on modality functions (assertion, question and injunction), focus functions (emphasis to keywords in a sentence) and syntactic functions (used for prosodic disambiguation).

- **Picture Description:** The subject was asked to choose one among several pictures representing a similar scenery (the sea with boats). Each subject had to describe the picture to the examiner so that the latter could redraw it just on the basis of the oral explanations.

- **Spontaneous speech:** The patient had to give his/her opinion on the questionnaire that he/she has to fill out before the recording session. He/she had to speak for at least three minutes. This task allows us to collect spontaneous speech recordings with no constraint on the sentences.

2.3.3 Human Perception Evaluation

In the context of the C2SI corpus, the recorded material was processed with the main goal of producing perceptual indexes. The different speech tasks were played to different sets of judges that rated the data accordingly. A large variety of perceptual evaluations was taken after the recording sessions. However for the sake of clarity, we will introduce the most relevant tasks for the present work:

- **Degree of Alteration - Voice quality, Resonance, Prosody and Phonemic Distortions:** Before the perceptual assessment of intelligibility and severity, a set of six experts (five speech and language pathologists and one phoniatrician) were asked to evaluate the degree of alteration of the four perceptual parameters of voice quality analysis (V), resonance (R), prosody (P) and phonemic distortions (PD). No definition of these concepts was given to the experts. The alteration index is comprised between 0 (normal) and 3 (severe impairment).

- **Speech Intelligibility and Speech Disorder Severity - Picture Description (Des.):** Indices of severity (SEV) and intelligibility (INT) were produced by a set of six experts. Intelligibility is defined as the comprehensibility of the message conveyed by the signal, while severity is defined as the degree of overall alteration of the sound signal. This assessment took place after the analysis of the aforementioned perceptual parameters. The indices are comprised between 0, which corresponds to the strongest alteration and unintelligible or severe speech, and 10, which corresponds to perfect speech. In order to assess for inter-judge reliability, an Interclass Correlation Coefficient (ICC) was calculated. The degree of concordance between the jury ratings is therefore good ($r > 0.69$) for the set of tasks.

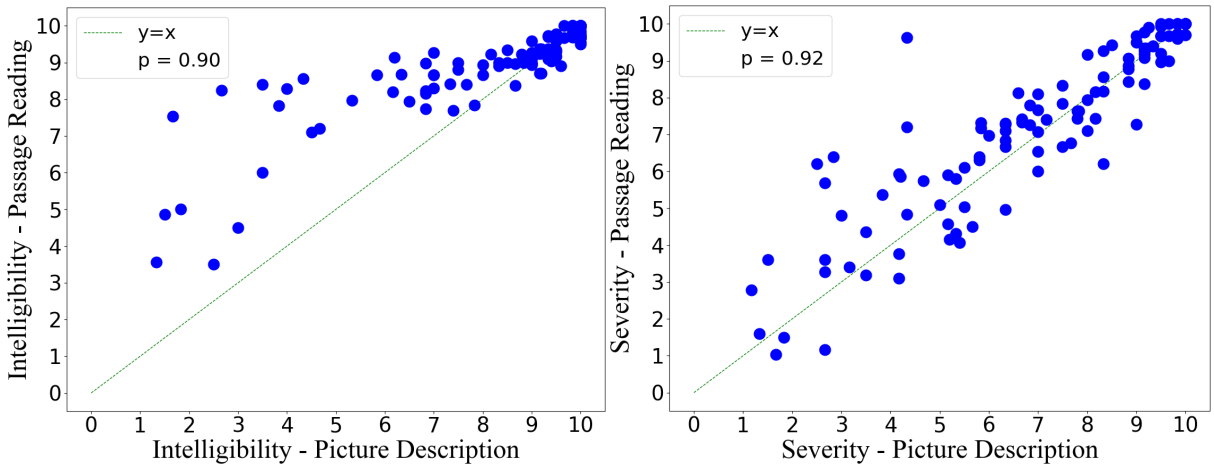
- **Speech Intelligibility and Speech Disorder Severity - Passage Reading (Lec.):** Similarly to the previous evaluation of intelligibility and severity on picture description, the same set of judges performed an analysis on the recordings of the reading passage task.

- **Perceived Phonological Deviation (PPD):** All the 52 pseudo-words pronounced by every speaker of the database were transcribed 3 times. Forty naive listeners were involved in order to transcribe the entire set of pseudo-words from all speakers. Listeners were confronted with a task that can be considered as acoustic-phonetic decoding followed by a written transcription. The mean distance between the transcribed and expected response is considered as a score of (un)intelligibility. The distance is obtained through a Wagner-Fischer algorithm [Navarro, 2001] that integrates the phenomena of insertion, elision and substitution of units (phonemes).

The correlations obtained between the different perceptual measures aforementioned can be found on table 2.1. The high correlations found (see bold values) can be interpreted as measures with higher levels of similarity. These correlations also aim to show that these measures can be biased and are inherently subjective, since the assessment of different speech tasks on the same set of patients can yield different results. The correlation between the two intelligibility and severity measures (Picture Description *vs.* Passage Reading) on the same set of patients can be seen as an example, which can be found illustrated on figure 2.6 ($\rho = 0.90$ for both intelligibilities and $\rho = 0.92$ for both severities).

Table 2.1: Correlations (Spearman) obtained between the different perceptual measures of the C2SI corpus.

	V	**R**	**P**	**PD**	**INT (Lec.)**	**INT (Des.)**	**SEV (Lec.)**	**SEV (Des.)**	**PPD**
V	1.00	0.49	0.65	0.52	-0.58	-0.60	-0.66	-0.65	0.53
R	—	1.00	0.70	0.83	-0.88	-0.85	-0.88	-0.87	0.80
P	—	—	1.00	0.72	-0.77	-0.79	-0.75	-0.75	0.69
PD	—	—	—	1.00	-0.88	**-0.93**	-0.87	**-0.91**	0.84
INT (Lec.)	—	—	—	—	1.00	**0.90**	**0.94**	0.89	-0.82
INT (Des.)	—	—	—	—	—	1.00	0.88	**0.92**	0.88
SEV (Lec.)	—	—	—	—	—	—	1.00	**0.92**	-0.82
SEV (Des.)	—	—	—	—	—	—	—	1.00	-0.85
PPD	—	—	—	—	—	—	—	—	1.00



(a) Comparison of *Lec.* and *Des.*: Intelligibility

(b) Comparison of *Lec.* and *Des.*: Severity

Figure 2.6: Comparison between the assessments of speech intelligibility and speech disorder severity on the two tasks of picture description and passage reading. While the severity assessment seems more evenly distributed, a larger skewness between the two tasks can be seen specially on the low intelligibility patients.

2.4 Conclusions

In the present chapter, an introduction to the topic of deep learning, with special emphasis on the current methodologies that are widely adopted in speech processing, was presented. This topic will serve as the main scientific context of the ongoing work. Moreover, the French corpus of head and neck cancer was also presented as the main application context. The introduction of these two domains becomes of crucial importance for a better understanding of the upcoming chapters and the present thesis as a whole.

As it was stated in the previous chapter, during the course of this thesis I will make use of deep learning methodologies for the automatic prediction of speech intelligibility in a pathological speech context, more specifically for head and neck cancer. Given that the C2SI corpus has a variety of recorded tasks and perceptual intelligibility evaluations, it becomes relevant to assess speech intelligibility in multimodal and multi-task approaches. By doing this, not only more robust scores could be obtained, but also relevant insight on the different modelizations could be found. In order to do so, the prediction of speech intelligibility will take place at the different granular levels of **sentence**, **word** and **phoneme**. This aspect foreshadows the three upcoming chapters, where different automatic models will be tailored to fit each one of the individual granularity levels. Table 2.2 illustrates the distribution of the speech tasks and reference intelligibility scores used for each granularity level. Furthermore, a unifying method that encompasses all these different predictions will be promoted, that becomes highly relevant when aiming for a high-performance general system.

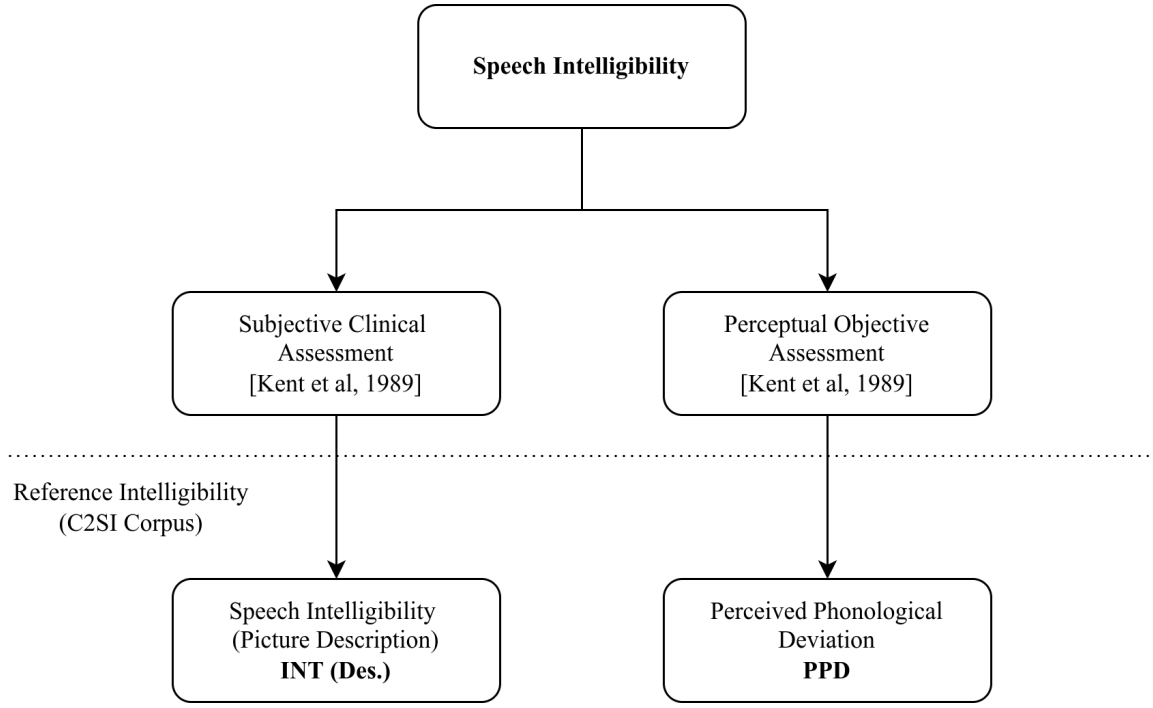


Figure 2.7: Relationship between the intelligibility approaches proposed by [Kent et al., 1989] and the intelligibility measures of the C2SI corpus to be used during this work.

Table 2.2: Tasks used at each granularity level, used during the course of this thesis.

Granularity Level	**Speech Task**	**Reference Intelligibility**
Sentence	Passage Reading (Monsieur Seguin)	Speech Intelligibility (Pic. Description)
Word	Pseudo-Words	Perceived Phonological Deviation
Phoneme	Pseudo-Words	Speech Intelligibility (Pic. Description)

Speech intelligibility is a measure of how comprehensible speech is in given conditions, that refers to the acoustic-phonetic decoding of the utterance, and therefore the reference intelligibility measures to be used during this thesis should be adapted to this definition. It is clear to see that speech intelligibility obtained from the passage reading task can be highly biased on the text itself, since previous knowledge on the text can condition the decoding process. Given this, throughout this thesis the perceived phonological deviation and the speech intelligibility obtained from picture description will be preferred as reference

measures. Later on, the voice, resonance, prosody and phonemic distortions parameters will also be automatically explored. By doing this, the present thesis is also in agreement with the two approaches to speech intelligibility proposed by [Kent et al., 1989], introduced in the previous chapter. The intelligibility on the picture description task corresponds to the "subjective clinical assessment", while the perceived phonological deviation correlates to the "perceptual objective assessment". This relationship can be found illustrated on figure 2.7.

Chapter 3

First Granularity Level: *The Sentence-XVec System*

3.1 Working Context and Hypotheses

As announced previously, a speech intelligibility measure obtained at **sentence** level will be explored. It is clear that this level of granularity plays an important role on the automatic prediction of speech intelligibility, since it is the one that more closely mimics the perceptual speech assessments performed in a clinical context. It is also the granularity level that resembles conversational level speech, which is directly correlated to day-to-day communication.

First, it becomes relevant to define what a sentence is. According to the oxford languages dictionary[1], a sentence is a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses. Given the previous definition, we can see that a sentence is a versatile construct, which we will use to our advantage during the course of this chapter. For the sake of clarity, we will also use the definition of text as any sequence of sentences that belong together.

To define any automatic supervised prediction system, which takes a speech signal as input, two important points must be reflected:

1. Starting with the input type, what is the signal processing pipeline?
2. What is the reference output for each input sentence?

Some propositions issued from the State of the Art can be found written below.

3.1.1 Choice of the Reference Intelligibility Score

As it was previously stated, it is considered a fairly common practice to use sentences in the perceptual evaluation of pathological speech. Passage reading tasks [Leea et al., 2018], picture description [Mueller et al., 2018] and spontaneous speech [Luz et al., 2020] are some of the common approaches used within a clinical setting to evaluate speech intelligibility, and they all rely on sentences or small texts to do so. On the other hand, the perceptual evaluations across different speech tasks, namely picture description and passage reading, can generate highly variant values, since the person assessing the patient can be either biased on the text (fairly common on passage reading tasks) or on the patient, pointing out again the subjective nature of these assessments [Balaguer et al., 2019]. Due to the variant nature of the intelligibility score issued by the same set of raters on different tasks, it becomes relevant to think of which set of labels would be better used in the context of an automatic assessment at sentence level. These labels will be used as a reference in the training and validation of our proposed system at this level of granularity.

[1] `https://languages.oup.com/google-dictionary-en/`

In the present chapter, we will make use of the reading passage task recorded within the context of the C2SI corpus [Woisard et al., 2020] therefore we have two perceptual intelligibility tasks at our disposal to be used as labels: the perceptual evaluation obtained by 6 judges on the **passage reading task** and the perceptual evaluation also obtained by the same 6 judges on the **picture description task**. At first, one could see that since we are making use of a passage reading task to train the system, one should use the corresponding perceptual labels, however, by looking at the distribution (section 2.3), we can clearly see that the intelligibility values are highly skewed closer to the higher values, leaving the already underrepresented set of patients with an intelligibility score below 5 (0-10 scale) even more underrepresented. This aspect poses an interesting question: Should we use a different set of perceptual labels to train a system on a different task? In our context we believe that it could produce a more robust and realistic system, since the objective of applying an automatic evaluation in a clinical setting is not only to obtain highly accurate scores, but also to generally discriminate between different sorts of speech impairment that may affect the perceptual score differently, and produce an automatic intelligibility score accordingly. Since we are dealing with smaller amounts of data, and generally deep learning or automatic systems underperform on underrepresented classes, we have decided to use the perceptual labels obtained from the picture description task, as in our eyes they provide a more realistic and unbiased perceptual intelligibility score, as well as more training data for low-intelligibility patients. Studies such as [Ben et al., 1996, Lalain et al., 2020] suggest that the usage of semantically unpredictable sentences or pseudo-words is indeed more accurate indicator of speech intelligibility, that tend to avoid the overnoting that happens when the reading task is known before (e.g. passage reading).

During the rest of this chapter, we will introduce the more specific state of the art used at this level, as well as our own methodologies, experiments, results and perspectives. We will close this chapter with some key takeaways to better add up to our previous conclusions as well as to provide a solid foundation for the following chapters.

3.1.2 Speaker Embeddings: State of the Art

The sentence granularity-level requires shifting our attention towards the speaker embedding paradigm. These fixed-length representations aim to represent speaker characteristics in a low-dimensional vector, and can have many applications for automatic processing of pathological speech.

A speaker embedding is often referred to as a single low-dimensional vector representation of the speaker characteristics from a speech signal [Bhattacharya1 et al., 2017]. These vector representations can be extracted using different means, such as by modelling speakers using Gaussian Mixture Models (GMM) via the adaptation of an Universal Background Model (GMM-UBM)[Verma and Das, 2015], or more recently, Neural Networks (NN) in the case of deep speaker embeddings [Bhattacharya1 et al., 2017]. These embeddings have recently seen a growing use in specific tasks such as speaker verification and speaker recognition and also highly valuable in the context of automatic speech recognition (ASR).

Speaker embedding representations, such as *i-vectors*, have proven to represent well speaker characteristics [Verma and Das, 2015]. Recent studies show that *i-vectors*-based approaches predict dysarthric speech evaluation metrics like intelligibility, severity and articulation impairment [Laaridh et al., 2017]. This speaker embedding (*i-vectors*) paradigm was also applied to the specific case of intelligibility in the same HNC corpus [Laaridh et al., 2018]. This work made use of those embeddings paired with a support vector regression (SVR) [Drucker et al., 1996] to predict speech intelligibility based on pseudo-words.

Deep Neural Network Speaker Embeddings have seen a growing use when compared to the GMM-UBM approach. Thus, *x-vectors* [Snyder et al., 2018b] are discriminative DNN speaker embeddings that have outperformed other speaker embeddings such as *i-vectors* in tasks such as speaker and language recognition [Snyder et al., 2017, Snyder et al., 2018a]. *X-vectors* have been successfully applied to paralinguistic tasks such as emotion recognition [Pappagari et al., 2020], and to the detection of diseases like Obstructive Sleep Apnoea [Codosero et al., 2019] and Alzheimer's [Zargarbashi and Babaali, 2019], making them highly valuable in the context of automatic assessment of pathological speech. Following the line of research present in [Laaridh et al., 2017] and [Laaridh et al., 2018], we investigate the reliability of using *x-vector* speaker embeddings as features for automatic intelligibility prediction in the context of HNC.

3.2 The Fundamentals of a Sentence-Based System

The proposed methodology at this granularity level mainly relies on two steps, illustrated in the figure 3.1. The first one corresponds to the extraction of the *x-vector* speaker embeddings, in order to obtain a fixed-length representation of every speaker's utterance, illustrated on subsection 3.2.1. We use the reading passage present in the C2SI corpus, which we will develop on section 3.3. The second step relies on automatically regressing an intelligibility score, based on the speaker embeddings previously extracted. In order to do so, we performed our experiments with shallow neural networks, further explained in subsection 3.2.2. The junction of these two steps creates the **Sentence-XVec system**, the system that will be used and explained in detail during the course of this chapter.

We make use of the reading passage of the C2SI corpus, previously introduced in section 2.3, and further elaborated on section 3.3.

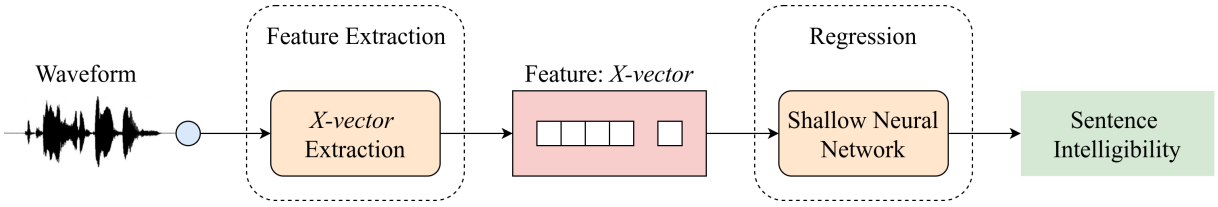


Figure 3.1: Global Overview of the proposed system. The *x-vectors* are extracted from the sentences that compose a reading passage text (LEC), and then fed to a shallow neural network that regresses an intelligibility score at sentence level.

3.2.1 *X-vector* extraction

As previously mentioned in subsection 3.1.2, *x-vectors* are speaker embeddings extracted from Deep Neural Networks that have seen a growing use in speaker recognition and paralinguistic tasks. Contrary to other speaker embeddings, such as *i-vectors*, that are extracted through a GMM-UBM approach, *x-vectors* aim to model the variability subspace by obtaining the fixed-length representations from the last layers of a Deep-Learning system, trained on the task of speaker verification. In order to extract the *x-vectors*, we used the open source implementation present in the Kaldi toolkit[2]. The system proposed by [Snyder et al., 2018b] is trained and validated on an extensive multilingual dataset. The training of a French *x-vector* extraction system was previously attempted in the context of this work. However it displayed a poor generalisation ability when compared to the multilingual approach, proving that the system remains language independent. Given this, the multilingual system was used instead. The features used are 24 dimensional filterbanks with a frame length of 25 ms, mean-normalised over a sliding window of up to 3 seconds. An energy-based Vocal Activity Detection (VAD) system selects features corresponding to speech frames, which is used in order to filter non-speech frames. The system uses Time-Delayed Neural Networks (TDNN) [Waibel, 1989] (previously introduced in section 2.2.3), that work similarly to convolutional neural networks, and fully connected layers.

The Deep Neural Network used to extract the *x-vectors* can be found outlined on the table 3.1. The first five layers operate directly on speech frames, with a small temporal context centred at the current frame *t*. This means that, for example, the input to the layer "TDNN Layer 3" is the spliced output of "TDNN Layer 2" at frames t-3 and t+3, which in turn is the spliced output of "TDNN Layer 1" respectively, at frames t-2 and t+2. This builds on the temporal context of the earlier layers, so that the layer "TDNN Layer 3" sees a total context of 15 frames.

A diagram that illustrates the part of the *x-vector* extraction system that operates at frame level can be found on figure 3.2. Afterwards, the statistics pooling layer aggregates all temporal frame-level outputs from the previous layer ("TDNN Layer 5") and computes its respective mean and standard deviation. The computed statistics are 1500 dimensional vectors, computed once for each input segment. What this process does is the consequent aggregation of information across the time dimension, so that

[2] `https://kaldi-asr.org/models/m3`

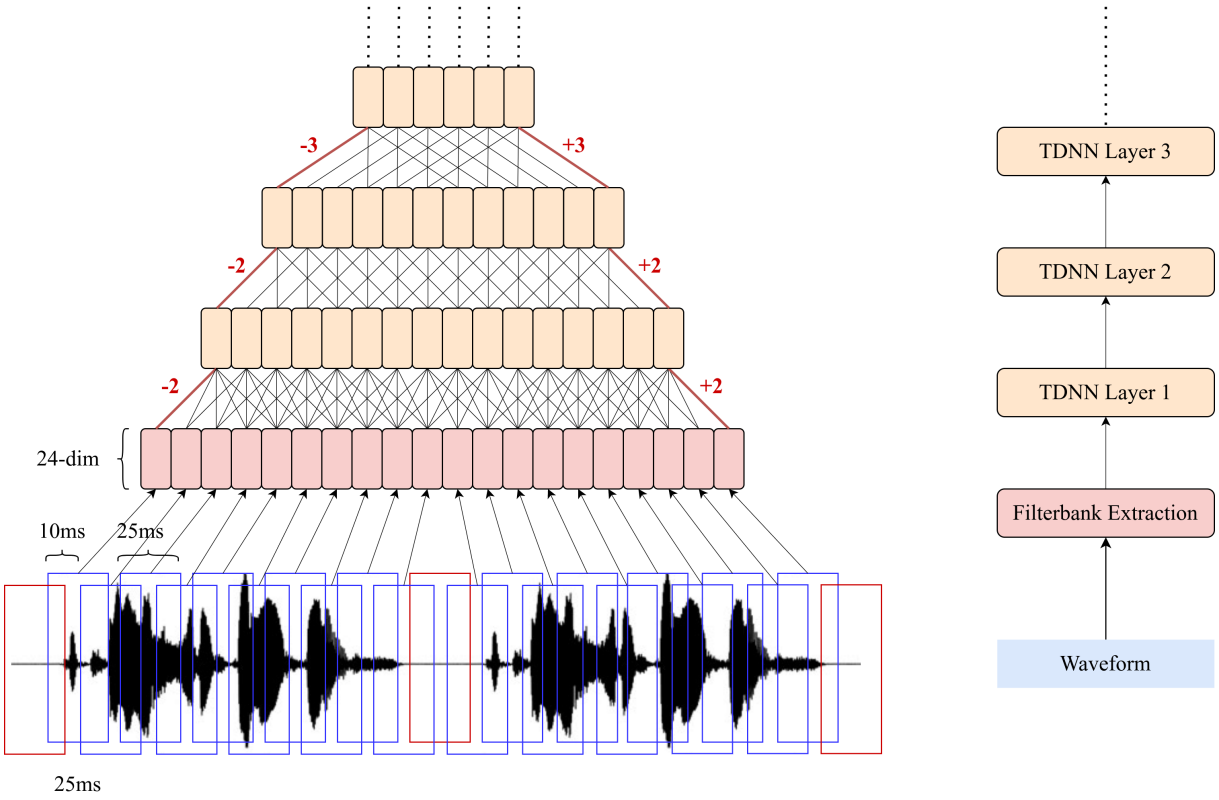


Figure 3.2: Snippet of the TDNN part of the *x-vector* extraction system. The three red frames correspond to the non-speech frames detected by the VAD system.

the subsequent layers can operate directly on the entire segment instead of a specific time instance (or frame). In table 3.1, this is denoted by the columns "Layer Context" and "Total Context", therefore starting with a Total Context of 5 time frames after "TDNN Layer 1" and finishing with a Total Context of T, which corresponds to the whole utterance. The "Input x Output" column illustrates the size of the layers used. The output mean and standard deviation from the stats pooling layer are concatenated and propagated through fully-connected layers, reaching the final softmax output layer. The nonlinearities present in the system (in between layers) are all rectified linear units (ReLUs).

Table 3.1: *X-vector* extraction DNN outline. N corresponds to the amount of speakers used to train the system (extracted from [Snyder et al., 2018b]).

Layer	**Layer Context**	**Total Context**	**Input x Output**
TDNN Layer 1	$[t-2, t+2]$	5	120×512
TDNN Layer 2	$\{t-2, t, t+2\}$	9	1536×512
TDNN Layer 3	$\{t-3, t, t+3\}$	15	1536×512
TDNN Layer 4	t	15	512×512
TDNN Layer 5	t	15	512×1500
Stats pooling	$[0, T]$	T	$1500T \times 3000$
Fully-Connected Layer 6	0	T	3000×512
Fully-Connected Layer 7	0	T	512×512
Softmax	0	T	$512 \times N$

As the *x-vector* speaker embeddings were originally designed for speaker verification, it's no wonder that in order to train the extractor, the DNN system is trained to classify the N speakers found in

the training data. The network is trained by being fed chunks of speech features (24-dim filterbanks), around 3 seconds each, and the corresponding speaker label. Generally, the average length of these chunks corresponds to the length of the type of audio files that we want to assess in this chapter (sentences).The embeddings are extracted from the affine component of layer "Fully-Connected Layer 6", therefore having a total dimension of 512 each. The system has a total of 4.2 million parameters, excluding the Softmax and fully-connected layers (see figure 3.3).

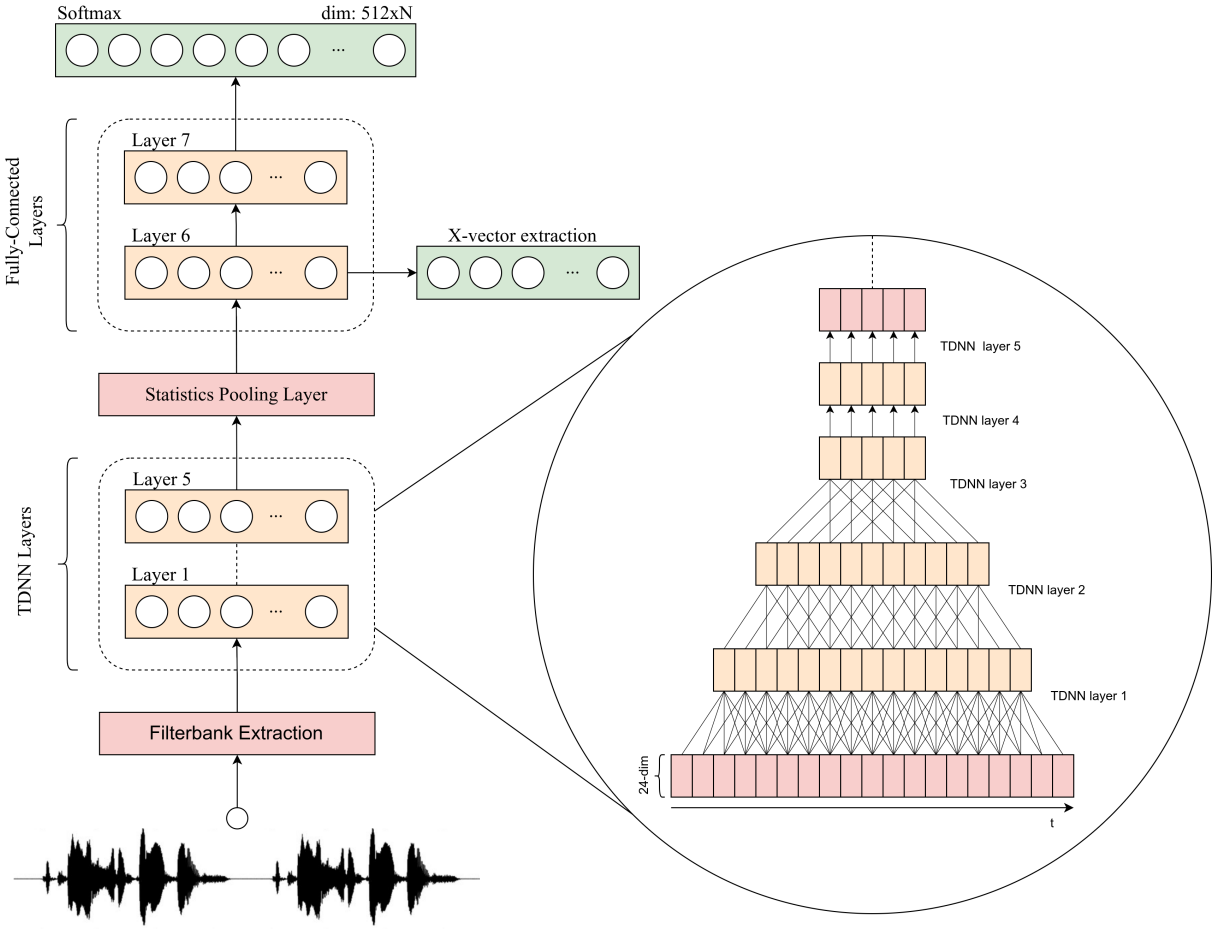


Figure 3.3: Pipeline of the *x-vector* extraction system. TDNN stands for time-delayed neural network. Frame-level layers operate at speech frame level with added neighbouring context (see table 3.1). The whole segment layers operate in the context of the whole utterance, extracted from the frame-level layers.

The data used to train the speaker embedding extractor is comprised of a variety of multilingual corpus both telephone and microphone speech, the bulk of which is in English. The corpus used (SWBD[3], SRE[4] and VoxCeleb[5]) can be found described on table 3.2.

Several data augmentations were also proposed by [Snyder et al., 2018b]. Since we've made use of the proposed Kaldi recipe to train the *x-vector* extractor system, we also employed the recommended augmentations. A 3-fold augmentation was used, that combined the original "clean" training files with two augmented copies. The augmented copies chose one of the four following data augmentations.

1. Babble Augmentation: Three to seven speakers are randomly picked from the MUSAN speech corpus [Snyder et al., 2015], summed together, then added to the original signal (13-20dB SNR).

[3] `https://catalog.ldc.upenn.edu/LDC97S62`
[4] `https://sre.nist.gov/`
[5] `https://www.robots.ox.ac.uk/~vgg/data/voxceleb/`

Table 3.2: Training data used for the *x-vector* extractor system.

Corpus Name	**Number of Speakers**	**Number of Recordings**
SWBD	2.6k	28k
SRE	4.4k	63k
VoxCeleb	1.91k	20k

2. Music Augmentation: A single music file is randomly selected from MUSAN, trimmed or repeated as necessary to match duration, and added to the original signal (5-15dB SNR).

3. Noise Augmentation: MUSAN noises are added at one second intervals throughout the recording (0-15dB SNR).

4. Reverb Augmentation: The training recording is artificially reverberated via convolution with simulated room impulse responses (RIR).

3.2.2 Shallow Neural Network

In order to regress an intelligibility score at this level of granularity, we made use of a shallow neural network. There are many different ways to regress a score based on the proposed type of features (speaker embeddings). In the same applicable context of head and neck cancer, previous works made use of Support Vector Regression (SVR) [Drucker et al., 1996], which lead to a fine generalisation. However they tend to lack in generalization ability when compared to the state of the art neural networks. Deep Neural Networks, on the other hand, can easily overfit the training data, especially in cases where the initial amount of data is scarce, which tends to be the case for pathological speech. Due to the aforementioned reasons, we have decided to model a shallow neural network, in order to better attempt to regress an intelligibility score given the small amounts of data present in the C2SI and without overfitting our model. Table 3.3 presents the proposed dimensions for the shallow neural network used. Only fully connected layers were used. ReLUs were used for the non-linearities between layers.

Table 3.3: Proposed shallow neural network outline.

Layer	**Input x Output**
Input	512×128
Fully-Connected Layer 1	128×64
fully-Connected Layer 2	64×1

3.2.3 Data Augmentation

Due to the small amounts of data present in the C2SI corpus, for example in the passage reading task used during this chapter, the overfitting of the proposed deep learning models can pose a serious concern. One of the ways to tackle this issue is to devise data augmentation schemes, in order to increase the number of files seen during training. Within the context of automatic speech processing, speed and tempo distortions are reliable data augmentations used in automatic speech recognition [Vachhani et al., 2018], which can also be employed in our particular case of automatic prediction of speech intelligibility. Since the *x-vector* extraction system aforementioned in subsection 3.2.1 was trained on small chunks of speech, and that reading passage tasks can take up to a few minutes of length, a further data segmentation into smaller chunks of speech can also be seen as a valuable data augmentation tool, which will be further discussed in section 3.3, in the context of the reading passage text used.

Tempo Distortion

Temporal distortions are valuable data augmentation tools that change the duration of an audio file according to a given coefficient. By using these augmentations, training data can be easily increased in a fast way, by promoting a variation of the original data without changing fundamental signal characteristics besides length/speed. Tempo distortion [Vachhani et al., 2018] is fairly similar to a speed distortion. However it ensures that the pitch and spectral envelope of the signal remain the same. This aspect becomes relevant since we are dealing with speech intelligibility, and therefore we want to minimise the external changes to the input signal. In the present work, we suggest this type of data augmentation to be used only on training files, given a small factor of 0.9 and 1.1, where a factor of 1 corresponds to the original signal. This in turn promotes a 3-fold data augmentation, where the same training file is distorted both ways. A subset of the augmented files was given to a specialised speech therapist, in order to validate whether there were perceptual changes in speech intelligibility in the distorted files. No perceptual changes were noted; therefore the same target values were used for both clean and mildly distorted files.

3.3 Implementation of the Sentence-XVec System

The present section illustrates the experiments performed at this level of granularity. As previously stated, we made use of the subset of speakers from the C2SI corpus that had recorded a reading passage task. A total number of 107 speakers, where 85 were patients with varying levels of speech intelligibility and the remaining 22 healthy controls. The chosen evaluation metrics used were Spearman's correlation (ρ), due to the non-normal distribution of the intelligibility scores used as reference, and the root mean squared error (RMSE).

3.3.1 The Experimental Dataset

Due to the fact that the *x-vector* extractor system was mostly trained with smaller chunks of recorded speech (around 3 seconds each), we have decided to prepare the entire reading passage text used at this granular level by splitting it in individual sentences. Given that the sentences were obtained from a text, some of them may not correspond directly to the conservative definition of sentence proposed in the beginning of this chapter. In this case, we will slightly abuse the notation and refer to these "pseudosentences" as sentences throughout. By splitting the reading passage task, we believe that this aspect can benefit the development of the automatic system proposed due to several reasons, such as:

1. A better correspondence with the studied granularity-level.
2. An increase in the training data.
3. The duration of the files being fed to the *x-vector* extractor system which matches the ones seen during training more closely.
4. More representations of the same speaker in the variability subspace, due to the increase in the number of files used.

The sentences were chosen according to natural breaks present both orally and in the text. The proposed segmented reading task (LEC) is as follows, where S_n stands for the corresponding sentences:

(S_1) *"Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres.* (S_2) *Il les perdait toutes de la même façon.* (S_3) *Un beau matin, elles cassaient leur corde,* (S_4) *s'en allaient dans la montagne, et là-haut le loup les mangeait.* (S_5) *Ni les caresses de leur maître* (S_6) *ni la peur du loup rien ne les retenait.* (S_7) *C'était paraît-il des chèvres indépendantes* (S_8) *voulant à tout prix le grand air et la liberté."*

The corresponding International Phonetic Alphabet (IPA) transcription of the reading passage task, obtained via an automatic transcriber [6], can be found on the table 3.4. The number of phonemes present in each sentence can also be found described on the same table.

[6] `https://www.openipa.org/transcription/french`

Table 3.4: International Phonetic Alphabet transcription of the reading passage task used in this chapter. The number of phonemes per sentence is also displayed.

ID	Sentence	Phoneme Quantity
S1	*məsjø səgɛ̃ nave ʒamɛz‿y də bɔnœr avɛk se ʃɛvrə.*	37
S2	*il le pɛrde tutə də la mɛmə fasõ*	25
S3	*œ̃ bo matɛ̃, ɛlə kase lœr kɔrdə*	22
S4	*sã nale dã la mõtaɲə, e la-o lə lu le mãʒɛe.*	31
S5	*ni le karɛsə də lœr mɛtrə*	20
S6	*ni la pœr dy lu ʁjɛ̃ nə le rətəne.*	24
S7	*kete par(e)t‿-ilde ʃɛvrəz‿ɛ̃depãdãtə*	27
S8	*vulã ta tu priks lə grã dɛr e la libɛrte.*	30

3.3.2 The Sentence-XVec System Training

The shallow neural network was trained, using a 5-fold cross-validation and a mean squared error (MSE) loss function. At each fold a total of 84 speakers, either patients or controls, were used for training while the remaining 21 speakers were left for testing. Figure 3.4 illustrates the train/test split of the cross-validation implemented. The same proportion of speakers in the train/test splits was kept on all folds. The tempo distortion mentioned in subsection 3.2.3 was applied to every training fold. Due to this data augmentation scheme and reading passage task segmentation, we were able to increase the training files from 84 to 2016 ($84 * 3 * 8$) and the testing files to 168 ($21 * 8$, no tempo distortion on the test set).

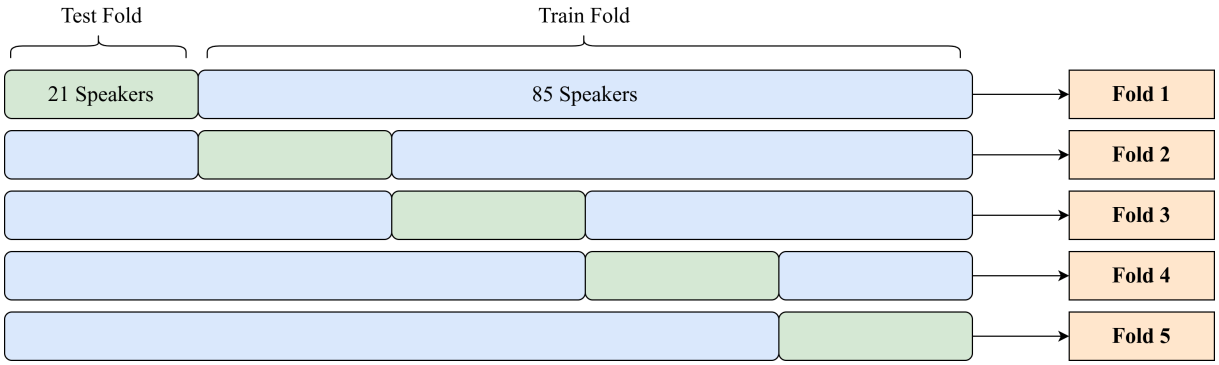


Figure 3.4: Train and Test partitions within the cross-validation context. At each fold, about 80% of the data is used as training while the rest is left for testing. No speakers are repeated across different test folds.

At each fold, the shallow neural network was trained during a total of 15 epochs using an exponential learning rate decay starting at 0.001 with a batch size of 8. Batch normalisation and a dropout rate of 25% were applied on every layer. All hyperparameters were found empirically.

3.3.3 A First Evaluation of the Sentence-XVec system

The scores were computed using the values obtained from the perceptual evaluation of speech intelligibility on the task of picture description due to the reasons aforementioned in section 3.1, issued by the six judges of the C2SI corpus (fully described in chapter 2 section 2.3). The first automatic prediction of speech intelligibility performed at this level of granularity made use of all the eight speaker's sentences. In this particular case, the *x-vectors* were extracted, fed to the shallow neural network and the output was finally paired to a predicted intelligibility value. The first possible score for each speaker was computed as the average score of each speaker's eight individual sentences. While this first score may seem simplistic, our proposed system achieved a high base correlation of 0.81 and a root mean squared error of 1.716. The

results also open the doors to further exploration, which can be found in the rest of this chapter. The error and correlation values achieved, can be found depicted on figure 3.5 along with the corresponding plot.

The results were compared to a similar methodology [Laaridh et al., 2018], which made use of the *i-vector* speaker embeddings, obtained through the pretrained Kaldi model [7]. These embeddings, that served as a baseline to our system, achieved a correlation of $\rho = 0.72$ and RMSE $= 2.121$. The same methodology, described in 3.2, was used in both the *i-vectors* and *x-vectors* case. The same shallow neural network was also applied to the *i-vectors*. However, with a dimensional change in the input since the *i-vectors* have a dimension of 400 as opposed to 512. From the comparison between the two systems, a clear improvement in performance can be seen on the *x-vectors*, both at correlation and RMSE levels.

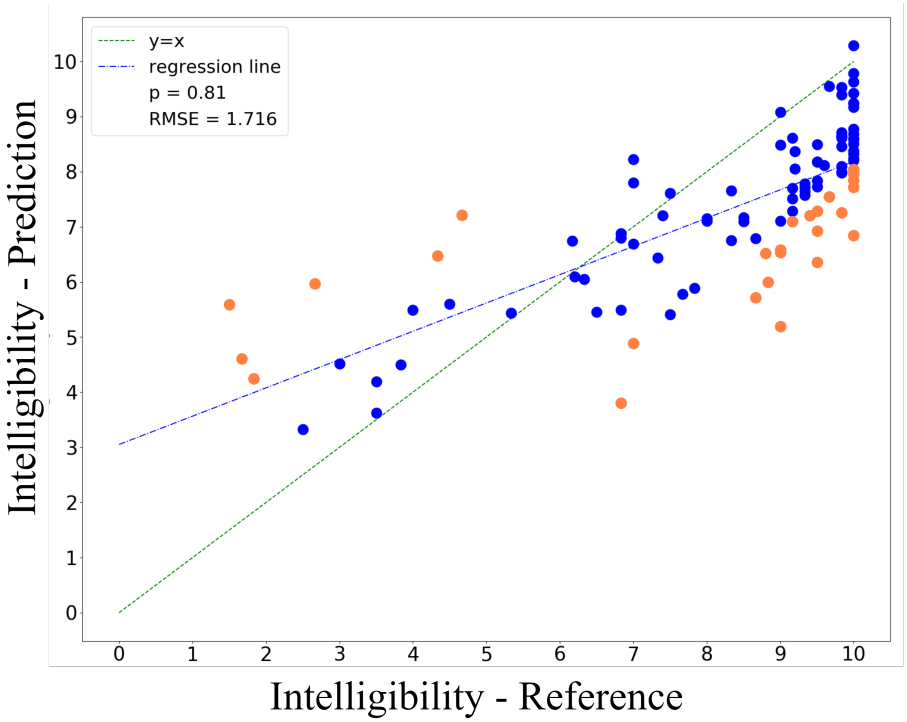


Figure 3.5: Resulting plot from the automatic prediction of speech intelligibility at sentence level, using the mean of the eight individual sentences of the reading passage task for each speaker. The figure presents the conjoined results of the five test folds, obtained from the cross-validation scheme. Controls can be found clustered between the interval [9.5,10]. Outliers are marked with orange points.

3.4 Towards a High-Performance Sentence-Based System

3.4.1 Outlier Analysis

Despite outliers being commonly found in the advent of automatic approaches, it remains highly relevant to check for relevant patterns within this subset. This could potentially help us better understand how the model behaves. The speakers that had an automatic predicted score outside a $[2, T, 2]$ boundary, where T stands for target value, were accounted as outliers. The illustration of the outliers can also be found on the figure 3.5. In this first set of experiments, from the total number of 105 speakers, a total of six outliers were found above bounds, with a higher predicted intelligibility score than the reference, and 21 outliers below bounds, with a lower predicted intelligibility score than the reference. Out of the 27 outliers found, only two of them corresponded to controls.

[7] `https://github.com/kaldi-asr/kaldi`

By looking at the figure 3.5, we can easily see that all of these six outliers above bounds corresponded to patients with a perceptual intelligibility score below 5, a class that is underrepresented in the present corpus. Given this, we can hypothesise that the system simply underperformed on these patients due to the lack of training material for this particular intelligibility level. As a matter of fact, no automatic prediction was ever below a score of 3.

The remaining 21 outliers were found below the $[2, T, 2]$ boundary (two controls and 19 patients), meaning that the system underestimated their intelligibility. In order to analyse these outliers, we can look at three different aspects that could potentially help us finding a pattern: the **tumour location**, the **reconstruction type** in the post-surgery and finally, the **standard deviation of the perceptual measures**. Both the tumour and reconstruction type, which are known characteristics of each patient, can help us to find a specific group where the system is not being accurate, while the standard deviation of the perceptual measures can point us towards a highly variant training label that does not serve as a good golden standard to begin with. From the 19 patients considered as outliers, nine patients were found to have a tumour in the amygdala region, 4 in the floor of the mouth region and the remaining ones scattered across other locations. As far as reconstruction goes, no major trend was noticed. Two patients from the set of outliers were found to have fairly high standard deviations (1.80 and 1.92 on a 0-190 scale) in the perceptual ratings, one of the controls that was an outlier as well was also the control with the highest standard deviation (0.75 respectively).

3.4.2 Analysis of the Individual Sentence Scores

From the results achieved using the average score of the eight sentences, we noticed that, on the majority of cases, there was a large variance within the individual scores of each speaker. From this we can conclude that there are sentences able to convey a much more precise intelligibility estimation when compared to the remaining ones. We further investigated this aspect by manually choosing, for each speaker, the sentences that had the predicted value closest to and furthest from the target. Similarly to subsection 3.3.3, the correlation and RMSE values were computed. The results can be found on table 3.5, paired with the results of manually choosing the worst sentence per speaker as well.

Table 3.5: Results achieved by manually choosing, for each speaker, the best and worst sentence scores. The mean sentence presents the results obtained when averaging each speaker's eight sentences. The best and worst sentences illustrate the results when manually choosing, for each speaker, the sentence closest and furthest from the target respectively.

	$\boldsymbol{p}$	RMSE
Mean Sentences	0.81	1.728
Best Sentence	0.95	0.900
Worst Sentence	0.53	2.224

The resulting values suggest that, for each speaker, there are sentences able to convey a highly precise automatic intelligibility estimation, generally displaying a very high correlation value and a low root mean squared error, and on the other hand, sentences that don't promote a good automatic intelligibility estimation. This aspect raises two questions that we will dive into:

1. Why are some sentences better at conveying an automatic intelligibility estimation than others?
2. How can we automatically find these sentences?

The resulting sets of manually chosen best and worst sentences were further assessed, in order to see if there was any particular pattern or sentence that stood out. For the 105 speakers, we counted the occurrences where each sentence was the closest and furthest from their respective reference value. The resulting distribution can be found illustrated on the histogram present on the figure 3.6. Although no clear preference was found towards a specific sentence, from the sub-list of best sentences, $S2$ was the one with larger representativeness, accounting for 23 of all cases respectively. On the other hand, $S7$ was the worst sentence with a larger presence, accounting for 24 cases respectively.

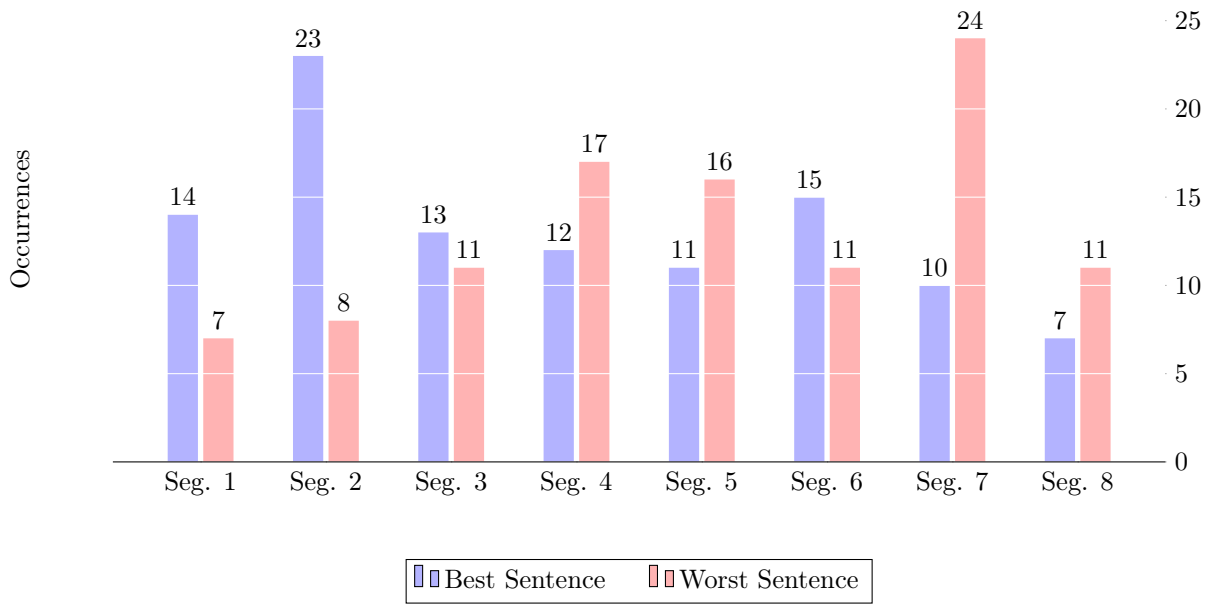


Figure 3.6: Distribution of best and worst sentences. Each bar denotes the number of speakers that have respectively the same sentence as best/worst.

Choosing the Best Sentence

As we've seen before, some of the sentences were able to convey a much more accurate intelligibility measure than others. Due to this, it becomes relevant to devise a way to detect each speaker's most relevant sentence. While the subset of speaker's best sentences displayed no clear preference towards a specific one, sentence 2 ($S2$) showed a tendency to be slightly better than the remaining ones. On the other hand, sentences 1, 3 and 6 ($S1$, $S3$ and $S6$) achieved fairly similar results, but not as good as sentence 2, accounting for 14, 13 and 15 of all CASES respectively (see figure 3.6). In order to evaluate how each individual sentence behaves, we correlated all speaker predictions from each sentence group with the target perceptual values. From this analysis, illustrated on table 3.6, we aim to evaluate the behaviour of each sentence instead of the mean of the sequence of sentences that compose the reading passage text. The results suggest that S2 has the highest correlation rate, followed closely by S6 and S1, validating the facts previously found in figure 3.6. Interestingly, by using only S2, the correlation/RMSE pair obtained (p: 0.819, RMSE: 1.434) was slightly better than the average score achieved using the mean of the eight segments.

Table 3.6: Individual correlation and RMSE values achieved when using each individual sentence instead of averaging the individual predictions based on each sentence.

	S1	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*	*S8*
p	0.764	**0.819**	0.735	0.696	0.752	0.774	0.697	0.730
RMSE	1.426	**1.434**	1.628	1.608	1.554	1.494	1.606	1.526

After finding which sentences correlate better with speech intelligibility, it becomes relevant to find an automatic way to detect them. We found that the set of speakers that obtained sentence 2 as the closest to target, achieved an above-average rate of recognized phonemes when compared to the other sets of patients. The average recognition rates, which can be found illustrated on table 3.7, were obtained using a pre-trained automatic speech recognition Kaldi model. Word and phoneme error rates have long been used as an automatic way to assess speech intelligibility [Christensen et al., 2012]. We believe that the

low recognition rate found on S6 may be due to the larger presence of nasalised vowels, which typically presents articulation issues for oropharyngeal cancer patients [Jacobi et al., 2013a].

Table 3.7: Average amount of recognized phonemes for the subset of speakers that had each respective sentence as the closest to the perceptual value. The amount of recognized phonemes was computed on the passage reading task as well. The amount of phonemes per sentence can be found on table 3.4.

	S1	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*	*S8*	S_{mean}
Average amount of recognized phonemes	175.5	**178.5**	174.5	165.5	176.5	**160.5**	160.6	178.0	168.5

For the sake of simplicity, we decided to focus only on sentences 2 and 6 for three reasons:

(i) they were the ones with larger representativeness as far as sentences with the intelligibility score closest to ground truth were concerned (see figure 3.6),

(ii) they were the two sentences with the highest individual correlations (see table 3.6),

(iii) they were the two sentences that were better discriminated by the number of recognised phonemes (see table 3.7).

Given the recognition rates found on S2 and S6, and that no other pattern was found out of the remaining attributes, a single node decision tree was implemented that chooses between sentence 2 and sentence 6 based on the phoneme recognition rate (prr). Should a speaker have a prr superior to the mean ($prr_{mean} = 168.5$, see last column of the table 3.7), S2 is chosen, while on the contrary, S6 is chosen instead. Figure 3.7 illustrates the single node decision tree implemented.

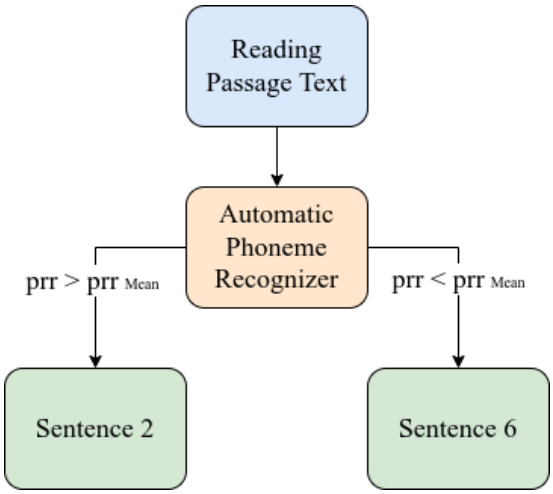


Figure 3.7: Illustration of the single node decision tree implemented, r stands for the amount of recognized phonemes.

The results of this analysis can be found on table 3.8, compared with the previous results obtained using the average of each speaker's sentences, the best individual segment (S2) and the results of the same methodology using *i-vectors*, the baseline system. The results suggest a correlation improvement of 4% when compared to the averaged scores of all sentences, and a decrement of 0.339 in error, displaying a more robust system for intelligibility prediction that also uses at each time a single sentence instead of a set of sentences.

Table 3.8: Comparison between the scores previously obtained and the single node decision tree criterion implemented.

		p	RMSE
i-vectors [Laaridh et al., 2018]	Averaged Scores	0.72	2.121
Sentence-XVec system	Averaged Scores	0.81	1.728
	Only Sentence 2	0.82	1.434
	Single Node Decision Tree	0.85	1.389

3.5 Scientific Contributions and Perspectives

3.5.1 Conclusions

In this chapter I have investigated an automatic approach for intelligibility estimation at the first granularity level: the sentence. In order to automatically estimate speech intelligibility, I have used the *x-vector* paradigm as features and a shallow neural network as our regression model. The junction of these two parts creates the Sentence-XVec system. This approach was devised for the segmented parts of a phonetically rich passage text, within the global context of our work on head and neck cancer. When using the average of all the used sentences for each speaker, a high correlation value of 0.81 was achieved, showing that the *x-vector* speaker embeddings can indeed convey reliable intelligibility predictions, surpassing the previously used *i-vectors*[Laaridh et al., 2018]. When manually choosing the sentences that are closest to the target perceptual value, a very high correlation of 0.95 was achieved, pointing out the importance of selecting the sentences used in this automatic assessment to each speaker. Moreover, I devised a single node decision tree criteria to attempt to model this sentence choice. The results suggest a correlation increase to 0.85 and a total decrease of 0.339 on the root mean squared error. While this criterion promotes a better correlation value than the average score, it displays the relevance of affecting pathology-based way to detect the best sentence for each speaker. The results obtained from the single node decision tree and on S2 show that the system can also work reliably on a single sentence as opposed to the entire reading passage text. The usage of smaller amounts of data is highly relevant in the automatic processing of pathological speech.

The results displayed in this chapter were accepted and presented at the conference Interspeech 2020, under the publication name of "Automatic Prediction of Speech Intelligibility based on X-vectors in the context of Head and Neck Cancer" [Quintas et al., 2020].

3.5.2 Perspectives

Concerning the used labels that were obtained from the perceptual assessments done in the context of the C2SI corpus, it is important to state that in some cases, high variances of up to 3 (on a 0-10 scale) were found, pinpointing the large interclass variance present in this type of clinical assessment [Plisson et al., 2017]. This aspect points out the subjectivity of the intelligibility scores used, when compared to the more objective ones found previously in the literature [Laaridh et al., 2018], which were rated by naive listeners instead of health professionals. The automatic modulation of inter-judge variance will be further described and experimented on chapter 6, that remains a highly interesting problem within the context of not only speech intelligibility, but all sorts of subjective measures.

When comparing the usage of *i-vectors* with the *x-vector* paradigm, we can conclude that the latter does not rely on larger amounts of data to output better results [Snyder et al., 2018b]. This aspect was evident by analysing the scores of only S2, present in subsection 3.4.2, which were slightly better than the averaged approach described in subsection 3.3.3. This can provide interesting cues to the development of less extensive and more precise batteries of exams, as the majority of the assessments are substantial and require much effort from both patient and therapist. A more precise and targeted assessment would strongly diminish the battery of exams required. In this particular case, we proved that the usage of a single sentence can provide similar or even better results when opposed to an entire reading passage task.

Regarding the outliers previously mentioned in subsections 3.3.3 and 3.4.1, since only 12.4% of the total number of speakers used had a target perceptual intelligibility score below 5, it was expected that the system would perform with larger margins of error in this specific context. This was the case for the 6 outliers predicted above bounds, that all had an intelligibility score below 5. Concerning the 21 outliers that were below bounds, we went searching for external attributes such as tumour location and reconstruction type. Out of the 23, we found that nearly half of them have a tumour in the amygdala region. Tumours in this location are typically associated with changes in articulation of fricatives and stop consonants [Logemann et al., 1993].

Concerning phonetic distributions, while perceptually a larger vowel presence is usually correlated to an improvement in speech understandability [Kewley-Portand et al., 2007], there is still no consensus within the scientific community as whether a larger consonant or vowel presence is correlated to a better perceptual intelligibility score [Kewley-Portand et al., 2007, Fogerty and Kewley-Port, 2009]. We introduce the hypothesis that, in the context of an automatic assessment, a larger consonant presence may be related to a more accurate intelligibility score. The reliability of a system that predicts intelligibility using only consonants will be studied in chapter 5, at the phonemic level of granularity. After using a single-node decision tree, the number of outliers decreased significantly from 23 to 5. It still leaves room for improvement, as there is still a large gap between the single node decision tree results and the cherry-picked sentences found on table 3.5. A deeper analysis on speaker individual features and speaker-specific phonetic content could provide a valuable insight to detect specific words and phonemes that are able to convey a more accurate intelligibility estimation.

Key Takeaways

1. We can reliably predict speech intelligibility at sentence level using the **Sentence-XVec system** that uses *x-vector* speaker embeddings as features.
2. The proposed system maintains a good performance when using segmented sentences, instead of an entire reading text.
3. For each speaker, there are sentences that are able to convey a much more accurate intelligibility estimation.
4. Individual speaker characteristics, such as tumour location, can affect the system.

Chapter 4

Second Granularity Level: *The Word-RNN System*

4.1 Working Context and Hypotheses

We will explore the automatic analysis of speech intelligibility at the word level, an intermediate level between sentences and phonemes. This level is particularly important since it serves as a bridge between the sentence, which can be seen as the most general and subjective level of granularity, and the phoneme, which, on the other hand, is highly precise and more objective. First of all, recall that, according to the "Oxford languages" dictionary[8], a word is a single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed. During the following lines, I tend to argue more precisely why a word intelligibility measure may be interesting and how machine learning systems may be helpful to this score evaluation.

4.1.1 Why and How to Measure Intelligibility at Word Pronunciation Level?

There are many motivations behind an intelligibility assessment at this level of granularity. While at sentence level, both perceptual and automatic assessments tend to take a more general approach to speech intelligibility (i.e. taking into consideration suprasegmental features such as prosody, voice quality, resonance, etc.), when recording smaller units as word units, evaluating these parameters becomes infeasible, especially due to the lack of temporal context present in smaller utterances. Hence we can see that an evaluation of speech intelligibility at word level tends to be more adapted to evaluate the correct/incorrect articulation of smaller sequences of phonemes, which in many ways can be seen as more objective.The utterance length has also been studied as a limiting factor for speech intelligibility, where longer utterances (e.g. sentences) can negatively influence speech intelligibility in people with speech disorders [dos Santos Barreto and Ortiz, 2019]. In this context, isolated words achieved a higher intelligibility score when compared to words inserted within a sentence, showing that depending on the assessment performed, there are always shortcomings.

In the domain of speech intelligibility, it is difficult to find a one-fits-all measure. Given this, finding and adapting different measures to different speech tasks and assessments can provide valuable clinical information. This becomes specially relevant during the post-treatment and consequent patient-monitoring. While normally the perceptual evaluation of speech intelligibility, such as the measures used in the previous chapter as reference, provides a broader view on the speech and phonation ability of a given speaker, it lacks in providing context for the day-to-day communication ability of the same speaker, which also is directly correlated to the quality of life [de Graeff et al., 2000]. In this particular case, a more naive measure based on word decoding can be seen not only as a more objective measure, but also as less biased. On the other hand, this measure fails to provide a more in-depth clinical view, to which the usage of these two types of intelligibility measure (clinical assessment *vs.* naive word decoding) can be seen

[8] `https://languages.oup.com/google-dictionary-en/`

as highly relevant. Needless is to say that these two measures relate back to the definition proposed by [Kent et al., 1989].

An idea to obtain this **naive measure of word intelligibility** emerges from the perceived phonological deviation (PPD), tailored for the pseudo-word task recorded in the context of the C2SI corpus (see section 2.3). This process globally consists of comparing the perceived phonetic pronunciation to the canonical pronunciation. We will explore this notion and give more details in section 4.2.

4.1.2 What to Expect from an Automatic Word-Recognition System?

When considering automatic approaches to assess words or acoustic-phonetic decoding tasks, probably the first class of models that comes to mind are automatic speech recognition based systems. Despite being the type of systems that more closely mimics the human process of perceived phonological deviation (i.e. transcription and consequent comparison to the ground truth), it is known that these systems tend to require large quantities of data [Ravì et al., 2017]. Similarly, these systems also tend to underperform when applied to pathological speech, especially for severe patients [Christensen et al., 2012]. The adaptation of acoustic models and state-of-the-art ASR systems to pathological speech is an active and ongoing research area [Liu et al., 2021, Calvo et al., 2021]. However in the particular case of the present work, we aim to achieve a good discrimination ability between different levels of speech intelligibility, and not necessarily a high-performance ASR system. In the case of pathological speech assessment, a perfect recognition of what was said may not be the end goal, specially for clinical applications. In this case, ASR systems can be used as a golden standard and the result deviation obtained between the controls and the pathological patients can also be used as a clinical measure. It is also important to state that patients fatigue is a well-known clinical issue, and a common reason to leave some of the speech tasks incomplete. This aspect is more recurrent in patients with severe speech impairments as well. Due to this, it becomes relevant to devise not only reliable automatic systems, but also more targeted and less extensive batteries of exams, that can also perform well in automatic systems.

Recent improvements in deep learning suggest that end-to-end models, such as the transformer, sequence-to-sequence or encoder/decoder models, outperform previous architectures namely in tasks such as ASR [Karita et al., 2019]. These models possess the advantage of not requiring pre-processing steps, nor alignments to yield competitive results, and come normally paired with attention mechanisms [Vaswani et al., 2017]. These mechanisms link the encoder and decoder through the means of sophisticated matrix operations, allowing the system to focus on more relevant parts of the sequence. The interpretation of the plots that result from the attention mechanism can provide interesting insight on specific sequence parts that may have relevant roles for sequence modelling [Gelin et al., 2021]. From this, we can directly see how relevant this aspect could be towards obtaining explainable automatic clinical measures. On the other hand, these plots are known to remain hard to decipher and highly dependent on the task and amount of data used [Serrano and Smith, 2019].

Given the fact that, besides the subjectivity associated to speech intelligibility, patients fatigue is also a common clinical problem, and that deep-learning-based systems tend to require substantial amounts of data to operate properly, we set ourselves two challenges:

- Present an automatic system that is capable of regressing objective and reproducible intelligibility measures at word level. The system should achieve a high correlation with the corresponding perceptual intelligibility scores.
- Prove that the same system can still achieve high correlations and accurate predictions when using significantly smaller amounts of data.

The rest of this chapter is organized as follows. Section 4.2 presents the word intelligibility score that will be used throughout. Subsection 4.3.2 presents the contextual state-of-the-art that will be relevant during this part of the present work. Section 4.3.3 illustrates the main modules of our system, where a recurrent model with self-attention is developed for the prediction of the speech intelligibility at word level. Section 4.4 showcases our experiments and results, with emphasis on the reference measure used: the perceived phonological deviation, as well as the system training and validation. Section 4.5 displays an analysis of the results, as well as the procedures implemented to decrease the quantity of data used

at inference time, used to combat patients fatigue. A note on the automatic modelling of the variability present in the reference scores is also presented on subsection 4.5.3. Finally, section 4.6 displays our main conclusions and perspectives.

4.2 The Word Intelligibility Score

As it was stated before, we will dive into the evaluation of speech intelligibility at word level. In order to do so, we need to use a good reference intelligibility measure to train a reliable system, which may not necessarily be the same that was used at sentence level. In the previous chapter, we discussed the pros and cons of using a different label for the automatic assessment of another speech task (as in passage reading task with picture description labels).

The perception of speech can be seen as a complex process that takes in consideration at times an ascending flux ("bottom-up") of information from the vocal signal, but also a descending flow ("top-down") based on the high-level information held by the listener. This ascending flux is essentially an acoustic-phonetic decoding operation [Fredouille et al., 2019], where a listener identifies the phonemes from the speech signal. The correct identification of phonemes, which are the base elements of speech intelligibility, can also give the precision with which a message is perceived by a listener. Given this, we can easily see that the correct identification of phonemes is a fundamental process to perceptually measure speech intelligibility [Ghio et al., 2018], especially in the context of the ascending flow part of speech perception. Since phoneme identification is a crucial part for speech intelligibility, according to the previous construct, it becomes relevant to devise ways to evaluate this ability.

Given this, we will make use of a reference intelligibility measure obtained from the perceived phonological deviation, also known as PPD [Ghio et al., 2020]. This measure was tailored for the task of pseudo-words, a speech task recorded in the context of the C2SI corpus that will be used throughout this chapter.

4.2.1 The Perceived Phonological Deviation

The PPD can be seen as a process in which a listener performs the association between perceived phonemes and the corresponding graphemes (transcription), and afterwards a distance measure is computed between the transcribed utterance and ground truth. Words and sentences can be a valuable tool for these assessments. However in both cases it is easy to extract the entirety of the word or sentence out of semantic or situational context. Due to this, a pseudo-word assessment becomes relevant, since the perceived phonological deviation derived from this task is focused solely on the identification of the composing phonemes.

Introduced in the context of the C2SI corpus, the perceived phonological deviation was obtained via a Wagner-Fischer algorithm, a dynamic programming algorithm that computes the distance between two strings of characters [Navarro, 2001]: between the ground truth pseudo-word transcription and the corresponding perceived one. The accumulated distance obtained from the algorithm, which can be found exemplified in figure 4.1, corresponds to the intelligibility measure for each word, that was used for reference during the course of this chapter. In order to compute the distance, the algorithm assumes a measure between the string characters, which can vary depending on the type of comparison performed. Since the phonetic distance between phonemes is variant, as in occlusive consonants sounding more similar between each other than fricatives for example, a distance matrix for vowels and consonants is introduced, obtained from [Ghio et al., 2018]. The matrices can be found on annex A.1.

As the perceptual transcription may be dependent of the listener performance and ability, we will use the multiple evaluations available in the C2SI corpus to propose a more robust reference measure. For that, recall that we made use of a pseudo-word task recorded within the context of the C2SI corpus. Each speaker was asked to record a set of 52 pseudo-words, that respect French phonotactic, orthography and pronunciation [Lalain et al., 2020]. Each set of pseudo-words was different, and only on rare occasions there were repeated words across different speakers. Each pseudo-word follows the structure $C(C)_1V_1C(C)_2V_2$, where $C(C)_i$ is either a single consonant or a consonant group and V_i a single vowel (see table 4.1). Each set of 52 words has a subset of 16 words with an occurrence of a double consonant at

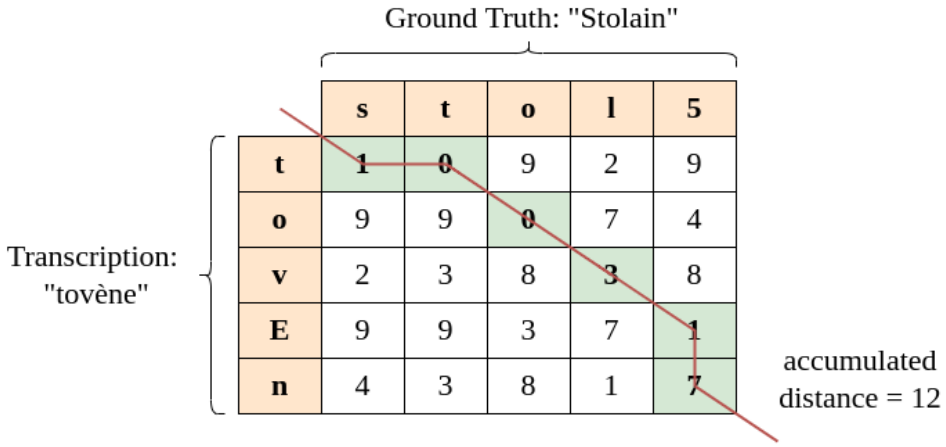


Figure 4.1: Example of the Wagner-Fischer algorithm applied to the distance between a ground truth pseudo-word and the corresponding perceptual transcription.

the beginning of the word, 16 words with double consonant in the middle and at least 26 words without double consonant. The pseudo-words can have both occurrences of the double consonant, either at the beginning or at the middle, however, in a much smaller quantity.

Table 4.1: Example of a set of 52 pseudo-words. Blue and violet represent the double consonant, that could be located at the beginning and/or middle of the pseudo-word respectively.

banfou **bl**eja bou**ct**i **br**im**pl**i chessant choniou
clifant cogu **cr**impin daillu dinrant **dr**edi
fan**rs**i **fl**in**rp**u fouma **fr**avi gabi **gl**unou **gr**o**rv**o
guchin joutu juro lanvin le**rd**a messo mouco
nianlo niejo no**ks**a nouillou pa**st**u pidant
ploniou **pr**ipin **ps**ila quiga rinta ru**rn**u
san**vr**in **sc**una souquin **sp**a**cl**ant **st**icho tan**gr**i
tou**gz**u **tr**a**dr**ou vi**rj**ant vumou yainzi
ya**lt**in zebou zouzant

4.2.2 A More Robust PPD

The robust perceptual intelligibility measure used in this study was obtained by averaging the individual transcription score of each pseudo-word by three naive listeners, also known as PPD [Ghio et al., 2020], which was previously introduced. This measure was used as reference throughout this chapter, and was calculated as a function of the distance between the transcribed word and the original one, taking into consideration the vowel and consonant matrix cost [Ghio et al., 2018] (annex A.1). The perceptual scores were comprised between 0, corresponding to the words perfectly pronounced, and 5, corresponding to unintelligible words. The final intelligibility score for each speaker was obtained by averaging the scores of the 52 pseudo-words. A set of 126 speakers was used, comprised by 40 controls and 86 patients. This set corresponds to the set of speakers that recorded the pseudo-word task within the C2SI corpus.

4.3 The Word-RNN System

4.3.1 Motivations

As argued in section 4.1.1, ASR systems may be possible. In spite of the inconveniences that these systems generally pose, the transformers appear as good candidates, given the variety of applications in which they can excel and their ability to map input sequences to different sequential outputs. Within the speech and language domain, tasks such as automatic translation [Vaswani et al., 2017], text-to-speech [Elias et al., 2021] and automatic speech recognition [Gelin et al., 2021] have seen recent interesting improvements when making use of the transformer methodology. Moreover, due to the versatility of the same methodology, transformers can be applied to a large variety of data and scenarios, ranging from image processing to natural language processing. Despite normally requiring larger amounts of data to operate properly, some works suggest that the transformer methodology can be successfully applied to the specific case of pathological speech, which is well known for the data scarcity associated with it. The work of [Lin and Tseng, 2021] made use of the transformer methodology applied to different sets of features to classify the intelligibility of children's speech, but the amount of data still remained a limiting factor in order to obtain better results. The work of [Lin et al., 2021] made use of the same methodology for the task of dementia detection, while [Huang et al., 2022b] used transformers to devise a data augmentation scheme for dysarthric speech recognition, by converting regular to dysarthric speech. Recent advances in automatic speech recognition have also made use of the transformer methodology [Dong et al., 2018].

Given the versatile nature of the transformer methodology, and the interesting results it displays, we aim to apply this type of methodology to the specific case of speech intelligibility at word level. To better understand the implementation of such class of systems, we describe in the present section the fundamental principles of transformers and their association with an attention process.

4.3.2 The Transformer Bases and the Associated Self-Attention Mechanism

Transformers, also known as sequence-to-sequence (seq2seq) models, are a type of neural network that transforms a sequence of elements into another sequence. This type of model normally consists of an encoder and decoder. The encoder takes the input data and transforms it into a higher-dimensional space in the form of a vectorial representation, also known as an embedding. The decoder takes this representation and converts it into the final output value of the system. Transformers, however, can be adapted to model different types of inputs besides sequences. Many-to-one and one-to-many approaches (see figure 4.2) can also be fit using the same methodology. Due to the nature of speech intelligibility, the many-to-one case will be particularly relevant during this part of the present work, in which we aim to extract an automatic measure from an audio sequence.

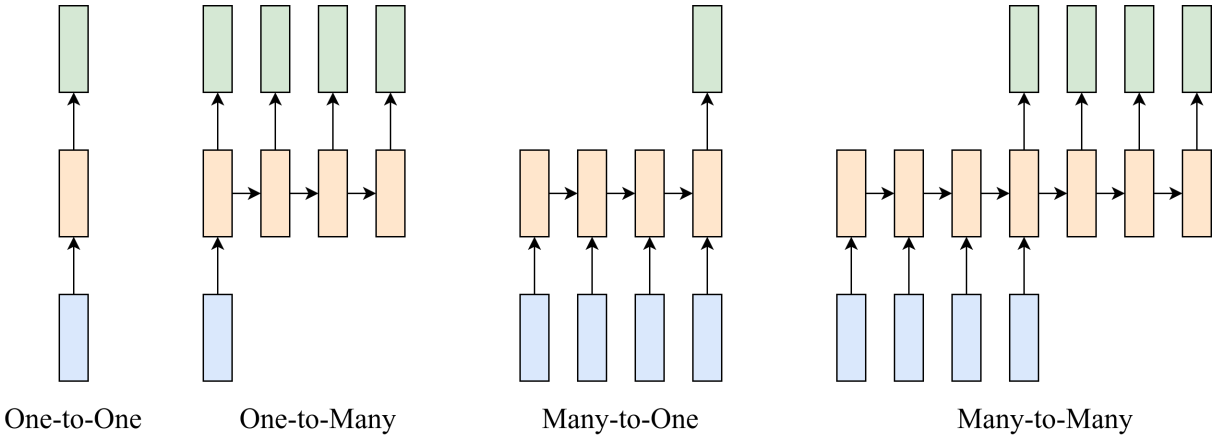


Figure 4.2: Different types of systems that can be created using a transformer methodology. Blue stands for inputs, orange for the system and green for outputs.

Given the versatility of the transformer methodology, in the way that it can map different types of inputs to different types of outputs, the composing parts of both the encoder and the decoder can also change accordingly. Should an input and output both have a fixed size, such as in the case of image processing, artificial neural networks or convolutional neural networks could easily be employed, while in the case of varying length inputs and outputs, recurrent blocks such as LSTM or GRU are normally adopted. Figure 4.3 displays a general structure used by the transformer methodology.

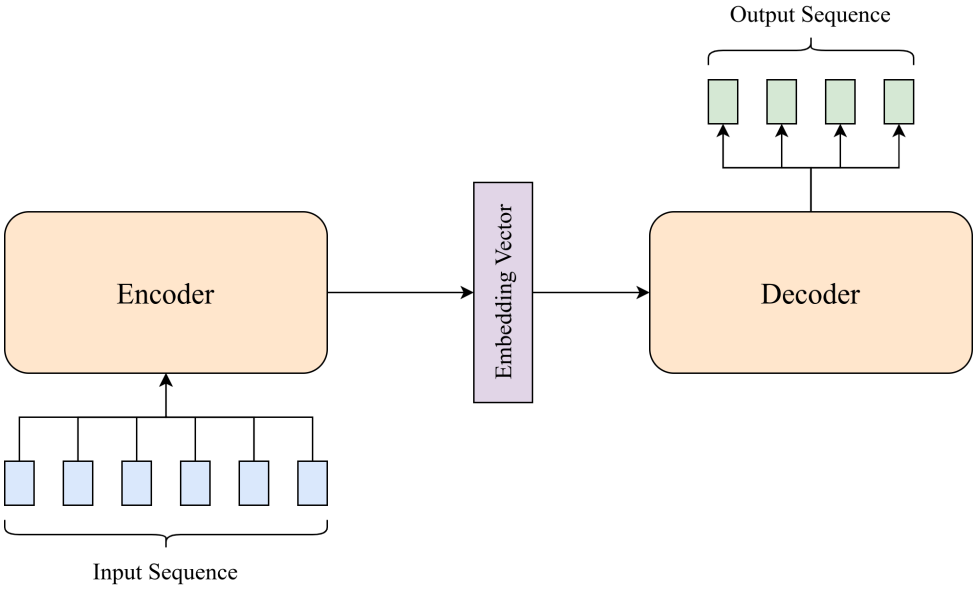


Figure 4.3: Global structure of an encoder/decoder model.

Attention

In the context of encoder-decoder models, the learning of complex sequence representations by the decoder from the embedding representation can be seen as the bottleneck problem, since the decoder would only have limited access to information provided by the encoder. This aspect becomes problematic specially in the case of longer sequences, since it is normally difficult for a system to remember earlier parts of the input sequence while processing the whole sequence. Due to this, it becomes relevant to devise a way for the decoder to be flexible and focus only on the most relevant parts of the input sequence, while giving less emphasis to parts that can be seen as irrelevant.

The attention mechanism [Bahdanau et al., 2015], an integral part of the transformer methodology [Vaswani et al., 2017], is a type of mechanism that can be used within the context of encoder-decoder models to provide an added focus to specific relevant parts of the input sequence, while neglecting the remaining non-relevant parts. This approach can be seen as a solution to mitigate the bottleneck problem of longer sequences, since the global system only takes in consideration what it considers relevant. The attention mechanism is also one of the most groundbreaking ideas in modern deep learning, and despite being introduced in the case of machine translation [Vaswani et al., 2017, Bahdanau et al., 2015], the applications quickly spread to a variety of other tasks such as speech recognition [Chorowski et al., 2015, Gelin et al., 2021], speech emotion recognition [Lieskovská et al., 2021] and also in general image processing and computer vision [Ghaffarian et al., 2021].

There are a variety of ways to implement an attention mechanism, since it is highly dependent on the application of the main system used. In the context of the present work, we will dive into the type of attention known as **self-attention**, since it will be relevant for the experiments performed in this chapter, but also to explain the foundations of more complex attention systems, such as the multi-head attention. The self-attention mechanism starts as a dot product operation executed between the input

(in this case, the input corresponds to our embedding or context vector) and the transposed input. This operation results in a square matrix that can also be known as the attention weights (see equation 4.1 and figure 4.4), where all elements of the embedding vector are multiplied by all the elements of the same embedding vector.

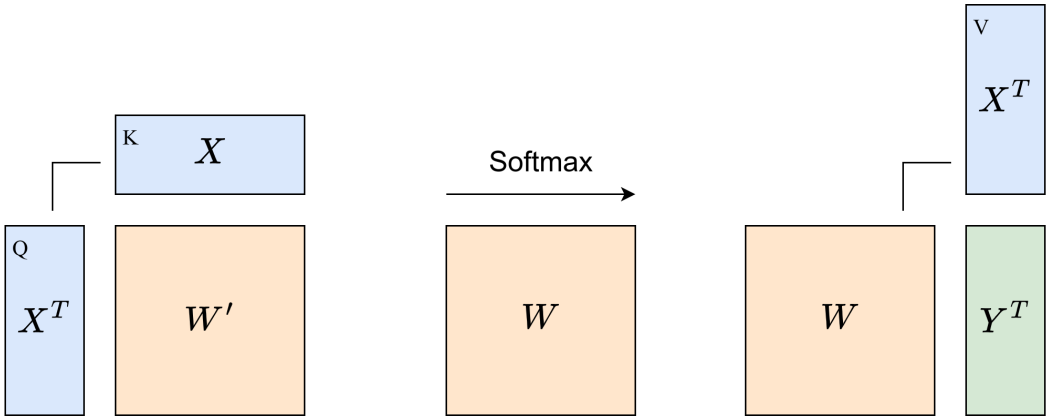


Figure 4.4: Illustration of the mechanism of self-attention. The K, Q and V characters stand for key, query and value. While in a simple self-attention mechanism, the key, query and value are analogous, in more complex attention mechanisms, such as the multi-head attention, the key and query can correspond to matrix transformations of the input while the value can be changed to the target sequence of the global model.

The attention weights are then passed through a softmax function (equation 4.2) so they are all positive and add up to 1. Finally, the output of our attention system corresponds to the multiplication of the attention weights by the transposed input vector (equation 4.3). From equations 4.1 and 4.2, we can see that the attention weights (W) provide a context matrix that encapsulates the relationships between all elements of the input sequence. The subsequent usage of this matrix when multiplied by the transposed input embedding (X^T) allows the system to provide an added degree of relevance to specific parts of the input sequence, which is the main goal behind the attention mechanism.

$$W' = X^T X \qquad (4.1) \qquad W = \text{softmax}(W') \qquad (4.2) \qquad Y^T = W X^T \qquad (4.3)$$

The relationship between the input and output of the self-attention mechanism is linear, and there are no learnable parameters in between. The self-attention module serves as the basis for more complex attention mechanisms, such as the multi-head attention. For this particular methodology, the whole attention module can be seen as a concatenation of several attention heads, whose results are concatenated into a single output vector. While in the self-attention module the weights (W) are a direct result of the multiplication of the input and transposed input, in the multi-head approach both inputs (key and query, see figure 4.4) can be transformed by learnable parameters. The final vector representation can also be a result of the attention weights multiplied by the target sequence instead of the transposed input. Given this, it becomes clear to see how this could provide extra context for the system to focus on, and it displays how versatile an attention module can be.

4.3.3 The Heart of the Word-RNN System

The proposed word-level intelligibility prediction system is based on the usage of a transformer's encoder part with a self-attention mechanism, applied to a pseudo-word task recorded in the context of the C2SI corpus. While the transformer models normally involve an encoder/decoder and sequence-to-sequence modeling, for the sake of clarity, we will refer to our proposed model as the **Word-RNN System**.

So our system can be divided in three distinct parts. After the feature extraction, we made use of a recurrent model with self-attention, previously introduced in section 4.3.2, in order to obtain automatic scores for each pseudo-word. The second part corresponds to the regression of an intelligibility score per speaker, based on the individual scores of each one's respective pseudo-words. Figure 4.5 illustrates our proposed method.

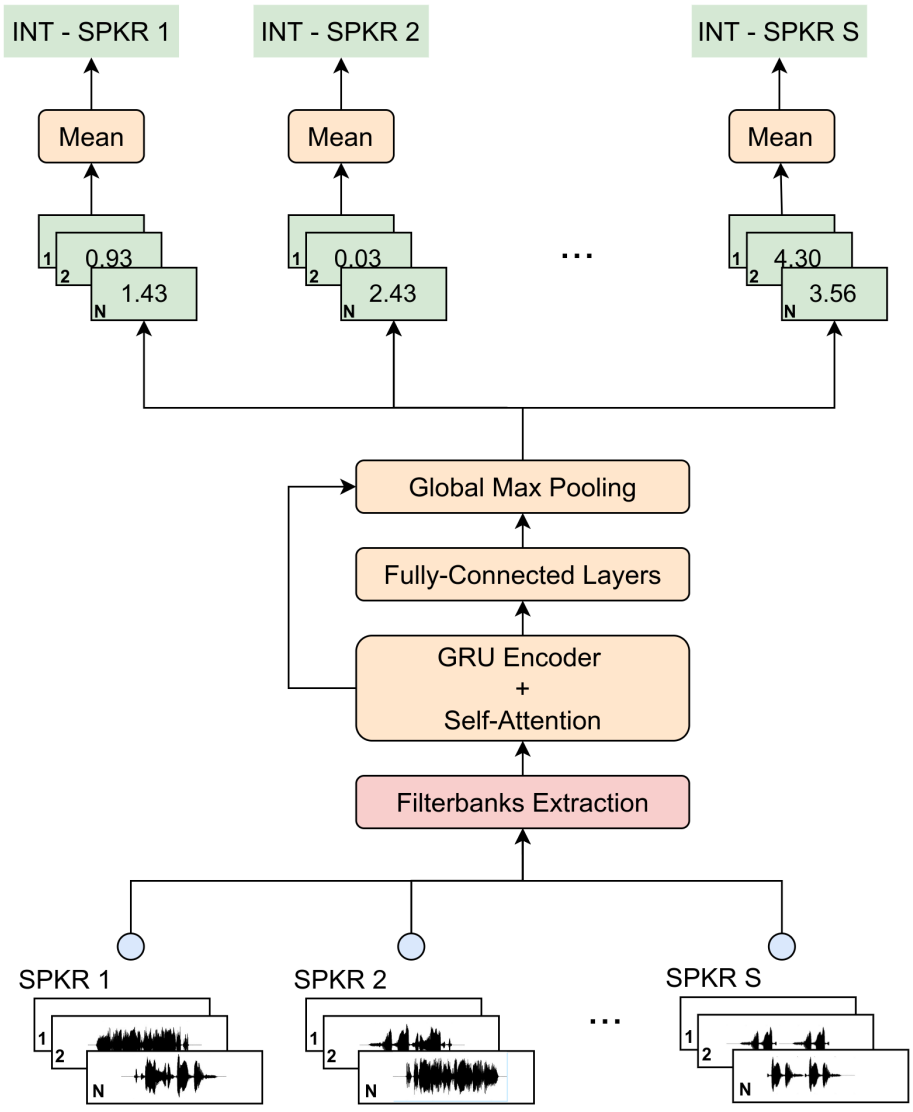


Figure 4.5: Global overview of the proposed methodology. SPKR stands for a speaker and INT stands for the intelligibility score. Here, each speaker has a set of 52 pseudo-words, to which the average of the automatic score of each speaker's words corresponds to the final intelligibility score for the same speaker.

4.3.4 Modelization

In this first part, we aim to present the proposed Word-RNN system for the automatic prediction of speech intelligibility, based on a recurrent model with self-attention, previously introduced in section 4.3.2.

Our system receives as input the individual audio files corresponding to each unique pseudo-word. From these audio files, we calculate on each window of 25 ms (with a 10 ms stride) 40 Mel filterbanks, to

be used as input features to our system. For each pseudo-word, we use as reference the intelligibility score obtained from its perceptual transcription as the target, obtained from the robust perceived phonological deviation (PPD), fully described in section 4.2.

Our proposed model uses a bidirectional recurrent encoder, with three Gated Recurrent units (GRU) layers with an input dimension of 40 and a hidden dimension of 100. At the output of the encoder, a self-attention mechanism can be found. This mechanism allows the system to focus more on particular parts of the input file, while ignoring less relevant parts. Given this, we hope the system will automatically learn interesting interdependencies between consecutive phonemes, that will result in a more robust intelligibility prediction at word level. After the attention mechanism, the fixed-length vector obtained is passed to a set of three fully-connected layers, with a dimension of 100 units each and Rectified Linear Units (ReLUs) non-linearity. Finally, at the end of the system we have a Global Max Pooling layer used to obtain the individual score for each pseudo-word, PW_n. Skip connections are added between the output of the self-attention mechanism and the Global Max Pooling layer.

Automatic Regression of Speech Intelligibility

The second part of our proposed model corresponds to the regression of an intelligibility score for each speaker, as a function of the individual score of each pseudo-word. Instead of having the proposed model directly predicting the intelligibility score at speaker level, we opted to do the average of the set of each speaker's pseudo-words at the output of the system. This aspect promotes a more explainable measure, as a speaker's score can be traced down to the individual score of his/her pseudo-words. More interpretable systems are highly relevant in a clinical context [Diprose et al., 2020]. The usage of the average set of pseudo-words also gives us more flexibility to evaluate the importance of specific words in the automatic score, which will be highly relevant later on in this work. Equation 4.4 describes the way the intelligibility score is regressed for each speaker (S_a). PW stands for pseudo-word. N stands for the number of pseudo-words used for each score.

$$Int(S_a) = \frac{\sum_{n=1}^{N} PW_n}{N} \tag{4.4}$$

4.3.5 Dataset and Training

The dataset is composed of the recordings of the C2SI corpus (see section 2.3). In order to train the system proposed in section 4.3.3, a 10-fold cross validation scheme was employed. At each fold, a set of 113 speakers (patients and controls) are used for training, and the remaining 13 speakers for evaluation. For each fold, the system was trained during 200 epochs. A learning rate of 0.001 was used, with a polynomial decay until 0.0001 during the first 50 epochs. A batch size of 16, the Adam optimizer and the MSE loss function were used during training as well. All hyperparameters were found empirically.

4.4 Performance Analysis and Explicability

In this section, I'll compare the performance of the developed system with other reference systems used within the same context, followed by a note on the analysis of the attention plots obtained.

4.4.1 System Evaluation

Spearman's Correlation (ρ) and the Root Mean Squared Error ($RMSE$) were chosen to evaluate our system. The target scores used were the perceptual intelligibility measures aforementioned in subsection 2.3. The results were compared to two other systems:

- an ASR-based system (a baseline [Fredouille et al., 2019]) i.e. an automatic transcription of pseudo-words followed by the same Wagner-Fischer methodology used for perceptual transcriptions. Due to the nature of the pseudo-words, the usage of a Language Model (LM) becomes unfeasible. Hence,

the automatic transcription took place in two parts. Firstly, a text-constrained alignment was produced by taking as input the original word, sequence of phonemes and the speech signal produced. Secondly, a semi-constrained acoustic-model decoding was used. Given the phoneme segmentation obtained in the first step, the goal now is to reconsider the phoneme labels and to search for the most appropriate ones among a set of 36 French phones. Each word was automatically transcribed in this fashion, afterwards the distance between the transcription and the ground truth was calculated. The results were correlated with the distances obtained from the perceptual (human) annotations.

- The intelligibility regression system based on *x-vectors* from our previous chapter, used at sentence granularity level. Given the short nature of the isolated word files, *x-vectors* were not extracted per individual file since they typically underperform in this type of context. Here, the embedding vector is extracted from the utterance that contains the full set of pseudo-words that each speaker recorded instead of a single embedding for each word. The signal processing pipeline is identical to the one developed in the previous chapter: a shallow neural network was modeled to fit the training data, a similar 10-fold cross-validation scheme was implemented.

The results, presented on table 4.2 and illustrated in figure 4.6, suggest a significant increase in correlation from 0.72 and 0.80 to 0.87, when compared to both previous systems, and a drastic reduction of the RMSE values: more than 50 %, when compared to the ASR baseline.

Table 4.2: Comparison between two reference systems and our proposed approach.

	ρ	$RMSE$
ASR-based system (baseline)	0.72	0.792
X-vector Speaker Embeddings	0.80	0.447
Word-RNN System	**0.87**	**0.370**

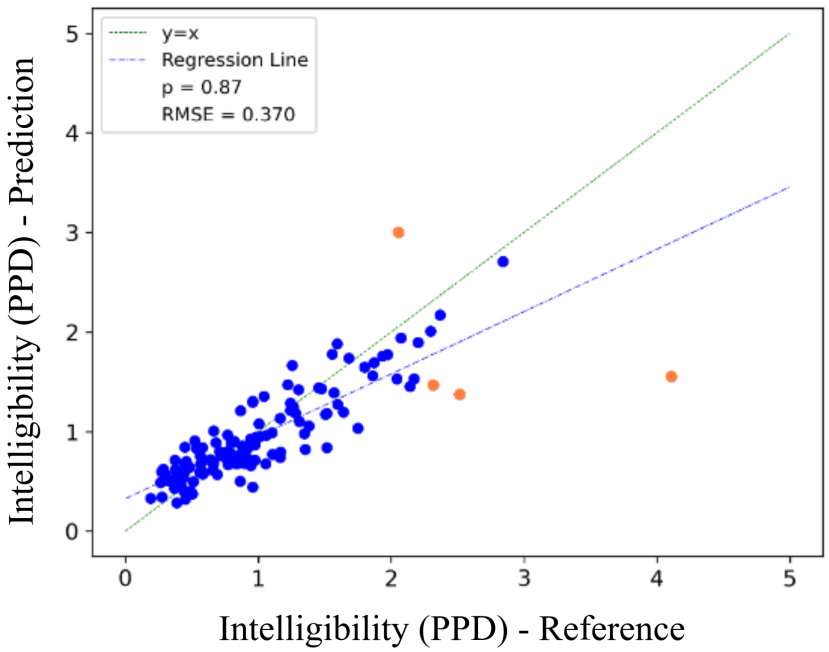


Figure 4.6: Results of the automatic prediction of speech intelligibility using the Perceived Phonological Deviation (PPD). Outliers are marked as orange points.

4.4.2 A Note on the Attention Plots

The resulting attention plots from the developed system were also analyzed. While trying to interpret these plots may not be the best approach towards justifiable interpretability [Serrano and Smith, 2019], they could still provide some interesting insight. Figure 4.7 presents four self-attention plots, two obtained from control speakers and two from patients. From the analysis of both diagonals, we can see that a straighter line appears on the control speakers, as opposed to the more distorted lines presented on both patients. Ideally, the attention weights (illustrated by the diagonal present in both figures) display parts of the input sequence that may be more relevant for the predicted output. A larger weight should present a more important part than parts with lower weights. From both plots, we can see that on the control speaker the relevance seems more uniform, while on the low intelligibility patient was slightly less visible. This could potentially mean that based on these plots, we could find key co-articulations that are more responsible for a given intelligibility score, which adds an extra layer of explainability. On the other hand, since this aspect was not evident on all pseudo-words from all speakers, it becomes difficult to derive more than some future perspectives on the interpretation of these plots.

4.5 Results Analysis

4.5.1 Effects of the Pseudo-Word Number Reduction

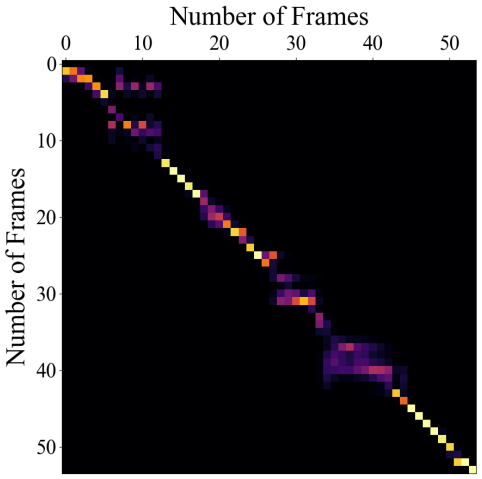
In the present study, we had at our disposal several sets of 52 pseudo-words that were used in order to obtain an automatic intelligibility estimation. On the other hand, in a clinical context, the recording of 52 pseudo-words is not only time-consuming, but also highly wearing for the patients, especially the ones with more severe speech impairments. Given that patient fatigue is not uncommon to happen in these recording sessions, which often translates into leaving the recording session incomplete, it becomes thoroughly relevant to evaluate how our proposed automatic approach behaves when predicting intelligibility with smaller subsets of data per speaker.

Different subsets of pseudo-words have been investigated depending on the position and number of occurrences of the double consonant (see section 2.3) as shown in table 4.3. The usage of the different subsets based on the pseudo-word structure and consonants was based on the work of [Marczyk et al., 2020]. The same work concluded that perceptually, words with a double consonant are better conductors of speech intelligibility. Given this, the same train of thought was implemented, this time for the automatic measures.

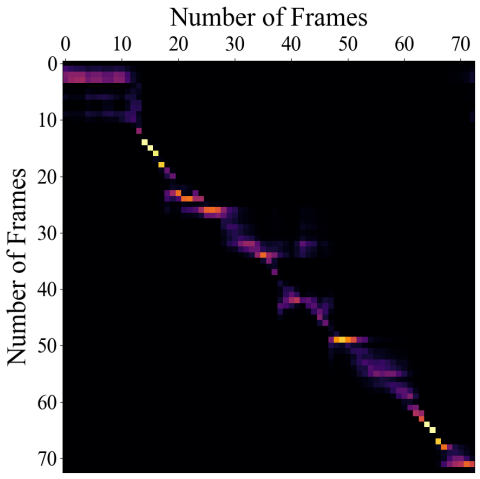
Table 4.3: Comparison between the scores previously obtained on the complete pseudo-word list and those of the reduced lists. The acronym *d.c.* stands for double consonant. The plots from the reduced subsets of pseudo-words (sets with 5, 16 and 26 pseudo-words respectively) can be found on annex A.

Model	**Amount of pseudo-words used**	ρ	$RMSE$
ASR Approach	52 (total)	0.72	0.792
X-vector Approach		0.80	0.447
Word-RNN System	52 (total)	0.87	0.370
	16 with *d.c.* at the beginning	**0.85**	**0.370**
	16 with *d.c.* at the middle	**0.85**	**0.375**
	26 without *d.c.*	0.84	0.398
	5 with *d.c.* at the beginning and middle	**0.79**	**0.413**

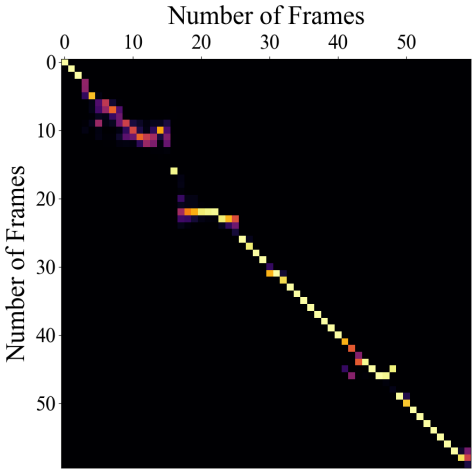
For each reduced subset of pseudo-words, the final automatic score of every subject is computed as the mean of the scores of the corresponding pseudo-words. On the other hand, the reference score still corresponds to the mean value of the 52 perceptual scores. The pseudo-word reduction only takes place during prediction and no new models were trained. Given the results obtained with the two 16 words sets, with a double consonant at the beginning and middle respectively, we can only see minor changes both at correlation and at RMSE. These results corroborate the fact found in [Marczyk et al., 2020] for the



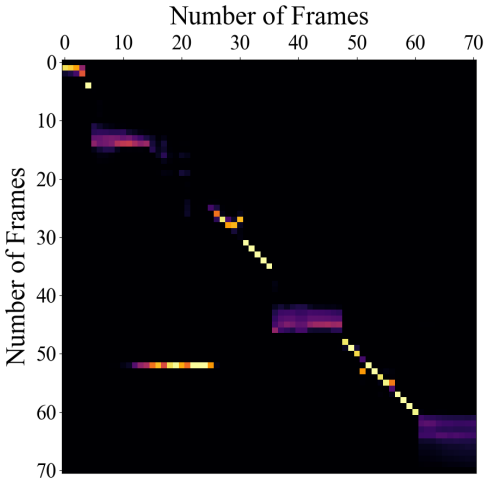
(a) Pseudo-word *"duma"*, issued by a control speaker with an INT (Des.) of 10.0.



(b) Pseudo-word *"duma"*, issued by a patient with an INT (Des.) of 6.83.



(c) Pseudo-word *"crancou"*, issued by control speaker with an INT (Des.) of 9.83.



(d) Pseudo-word *"crancou"*, issued by a patient with an INT (Des.) of 7.50.

Figure 4.7: Illustration of four plots issued by the self-attention module. Two plots correspond to the word *"damu"*, that has no double consonant, and the word *"crancou"*, that has an occurrence of double-consonant. Each pseudo-word has two plots, one corresponding to a control speaker and the other to a patient. The patient that issued the word

perceptual measures, and therefore we can say that it is possible to use significantly smaller amounts of data whilst maintaining the reliability of the automatic measures promoted. For the remaining subsets of words, the results were similar, but they tend to accentuate the difference between pseudo-words with and without double consonant. This aspect was particularly evident in the set without a double consonant (26 words), where besides having 10 more words than the two sets with double consonant, the correlation remains similar and the error increases.

Moreover, a continuous analysis on how the correlation and error change as a function of the quantity of pseudo-words used at inference time was performed. This analysis was performed by random sampling subsets of words that follow a certain criteria, in our particular case, the occurrence of double consonant. The results from this analysis are illustrated on figures 4.8 for correlation and 4.9 for RMSE.

For the correlation values, the results suggest traits of a logarithmic curve ($log(x)$, note the reverse scale on the x-axis) in the case of words with either double consonant at the beginning, middle and

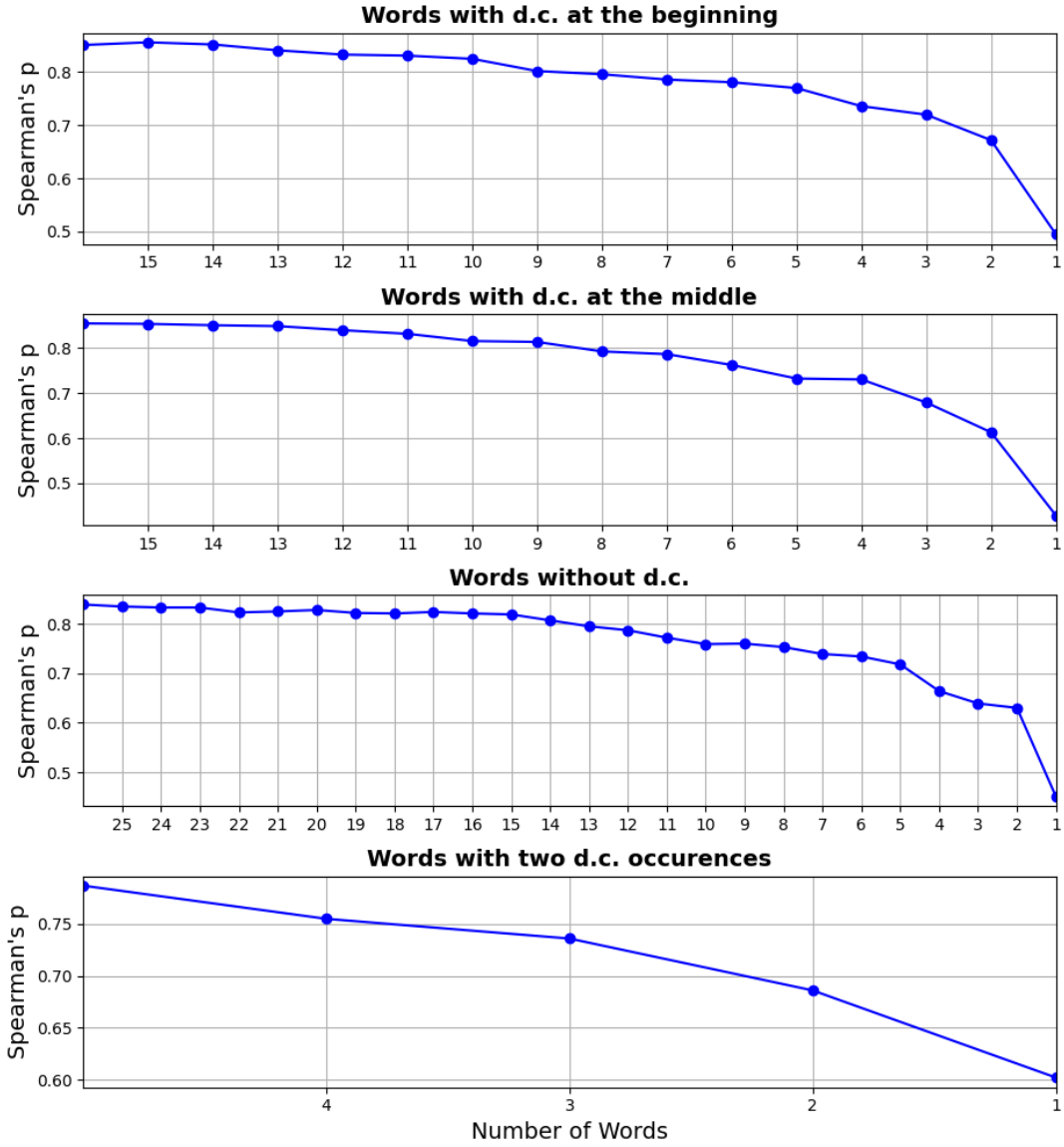


Figure 4.8: Line plots corresponding to the correlation values found when varying the quantity of pseudo-words used at inference time. Four types of words were assessed: words with double consonant (d.c.) at the beginning, middle, beginning and middle and finally without double consonant.

without d.c., showing a larger increase in correlation between the smaller sets of words when compared to the larger sets. Interestingly, the words with the dual occurrences of d.c. displayed a quasi-linear shape, not showing big correlation gains among the smallest subsets of words when compared, for example, with the words without any occurrence of double consonant. This aspect could indicate that these words can indeed serve as a better indicator of speech intelligibility, since the results obtained with only one of these words display a correlation of around $\rho = 0.60$ when compared to the three other types of words, whose singular results with only one word are always below $\rho = 0.50$. The results found on the error values mirror the ones found on the correlation. In this particular context, the error behaves similarly, displaying traits of an exponential (e^{-x}, note the reversed scale on the x-axis) on all types of words. The largest correlation gain and error decrease over all cases are always found between the singular sets of only one word and the sets with two words.

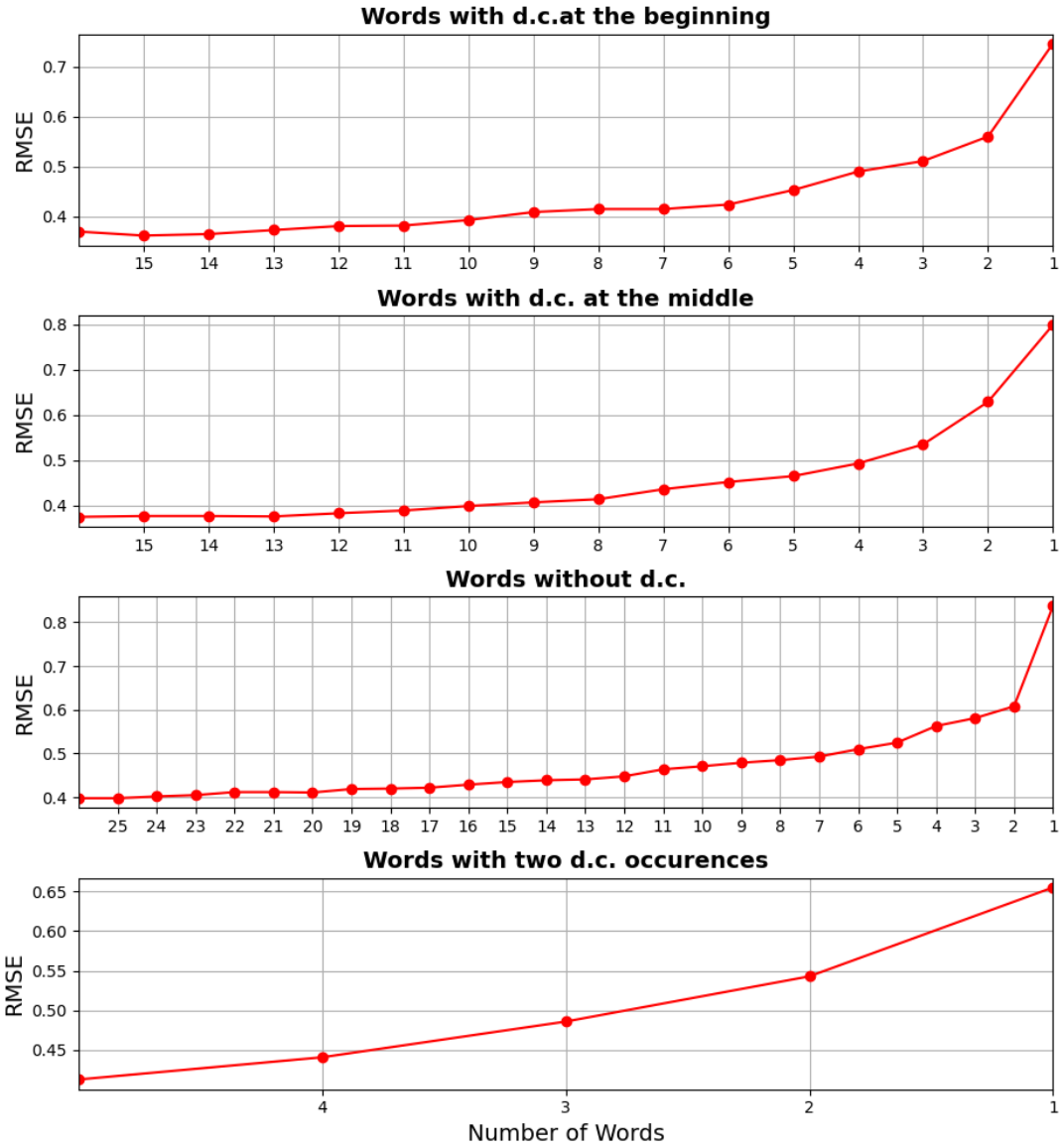


Figure 4.9: Line plots corresponding to the RMSE values found when varying the quantity of pseudo-words used at inference time. Four types of words were assessed: words with double consonant (d.c.) at the beginning, middle, beginning and middle and finally without double consonant.

This analysis shows that similarly to the perceptual evaluations, pseudo-words with at least one occurrence of a double consonant are more relevant to obtain robust and reliable automatic intelligibility measures. This aspect was also illustrated by the results obtained in the last line of table 4.3, which displayed the results obtained with pseudo-words containing the two types of double consonant, in which a fairly decent correlation and RMSE values were obtained considering the small quantity of only five pseudo-words. The results obtained with the subset of five pseudo-words with both occurrences of a double consonant are comparable with both previous approaches (Wagner-Fischer baseline and *X-vectors* approach), showing that a similar correlation and smaller error can be obtained by using less than 10% of the original quantity of words. We hypothesize that a larger subset of these words with dual occurrence of a double consonant would outperform the other reduced lists, but we were limited by the quantity of these pseudo-words in the corpus.

4.5.2 Discussion

The results obtained in the present work suggest that it is possible to obtain high correlation values between the perceptual evaluations and automatic predictions. In addition, not only the correlation values increased greatly, but also the RMSE values drastically reduced to more than half compared to a previous approach [Fredouille et al., 2019], based on the usage of the Wagner-Fischer algorithm between the ground truth and automatic transcriptions obtained via an automatic speech recognizer. Given this, we can safely conclude that we can reliably predict speech intelligibility measures automatically at word-level.

Moreover, the analysis performed in subsection 4.5.1 showed that it is possible to remove words from the global score and barely affect the reliability of the same automatic measures. This aspect becomes crucial in a clinical context, where not only we can save precious time by recording fewer data, but also avoid the dreaded patient's fatigue. The usage of pseudo-words with a double consonant was crucial to achieve these results in the present work, since it displayed a clear direction towards the best type of words to be used in these assessments. We believe that, similarly to the perceptual case [Marczyk et al., 2020], these words tend to outperform their counterpart due to the larger phonetic content and key co-articulations between consonant and vowels and consecutive consonants. These co-articulations are an important marker for the post-operative assessment of head and neck cancers [Saravanan et al., 2016]. A larger presence of consonants is also hypothesized to be a good indicator of speech intelligibility [Crevier-Buchman et al., 2002, Fort et al., 2015].

While keeping fewer pseudo-words, the results obtained with the double-consonant subsets were encouraging, leaving the possibility that we can further reduce the quantity of words used without affecting much correlation and error. The usage of a small number of words also becomes relevant in a clinical context when compared to other automatic approaches that operate, for example, at sentence level [Quintas et al., 2020], showing that a similar correlation can be achieved while using significantly smaller amounts of data. The creation of a reduced list of words, containing phonetically rich and diverse content, that could work both at perceptual and automatic level remains an interesting lead for future work.

4.5.3 A Note on Modelling judge Rating Variability

It is true that sometimes on machine learning tasks, the actual ground truth that we assume may not always be the best, especially when we rely on human labelling on subjective tasks. In this context, sometimes human agreement can be low, which directly impacts the quality of the measures used as reference, and subsequently the quality of the machine learning models. It is one of the reasons why we used a robust version of the PPD, composed by the mean of deviation score of three judges.

Nevertheless the usage of the variability associated with the reference perceptual measures is an interesting paradigm, which could lead to more robust models, or a more direct and explainable way to train an automatic model. There are many different ways to model the raters variability automatically, one can choose to individually model the opinion of several judges, or to automatically predict an uncertainty measure or even use the variability as a weight during training, to give more emphasis to certain inputs. Since in the context of the present chapter, the naive perceptual judges used were not the same for every speaker, and therefore training individual models that reflect the opinion of a single judge becomes unfeasible, we intend to explore how the automatic modeling of the intelligibility score's variability behaves during training.

While on the previous sections we assumed the mean as the gold standard (ground truth), recent works suggest that we can use variance while training deep learning systems in order to promote more robust and better models [Rizos and Schuller, 2020, Han et al., 2017, Han et al., 2020]. From the perceived phonological deviation measures, the quantifiable source of variability comes from the standard deviation issued out of the three different scores given by the experts, which we intend to use during the training of another class of models in order to investigate whether there are notifiable changes in the scores.

To introduce an automatic modeling of the variability associated with the speech intelligibility measure, three different systems were experimented, based on different loss functions applied to the proposed

system. The first loss function, which could be referred to as **Weighted Loss Function** illustrated in equation 4.6. In this loss function, the standard deviation of each pseudo-word is used as a penalty factor during training, meaning that words that have a larger variance will have less relevance during training, forcing the system to pay closer attention to pseudo-words that have a larger level of agreement. The second loss function proposed, named **Multi-Task Learning Loss** and illustrated in equation 4.7, as the name indicates, derives from the multi-task learning methodology. This approach to training deep learning models hypothesizes that the simultaneous learning of a similar task may have a positive effect when learning the main task, improving system performance [Zhang and Yang, 2018]. In this context, we adapt the system to have an extra final head (layers) that predict the standard deviation associated with each pseudo-word. This loss function is regulated by two fine-tunable weights (ω_1 and ω_2), which normally tend to give more relevance to the main task as opposed to the secondary or even tertiary tasks. Finally, the third and final loss function experimented, called **Multi-Task Learning Loss with Dynamic Tuning**, is fairly similar to the multi-task learning loss function. However in this case there is an appended term that forces the intelligibility prediction to be partially controlled by the predicted variability [Han et al., 2020], illustrated in equation 4.8. The loss function used in the baseline system is also added on equation 4.5 as a reference.

$$L(y, \hat{y}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y - \hat{y})^2} \tag{4.5}$$

$$L(y, \hat{y}, \sigma) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+\sigma} * (y - \hat{y})^2} \tag{4.6}$$

$$L(y, \hat{y}, \sigma, \hat{\sigma}) = \omega_1\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y - \hat{y})^2} + \omega_2\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\sigma - \hat{\sigma})^2} \tag{4.7}$$

$$L(y, \hat{y}, \sigma, \hat{\sigma}) = \omega_1\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y - \hat{y})^2} + \omega_2\sqrt{\frac{1}{N}\sum_{n=1}^{N}[(\hat{y} - \hat{y}\hat{\sigma}) - y]^2} + \omega_3\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\sigma - \hat{\sigma})^2} \tag{4.8}$$

The same system proposed in section 4.3.3 was trained, however, this time with the loss functions proposed for the variability modelization (equations 4.6, 4.7 and 4.8). The weights for the multi-task learning loss (MTL) and multi-task learning loss with dynamic tuning (MTL + DT) were manually fine-tuned, the resulting values were used during both system's training: $\omega_1 = 1$, $\omega_2 = 0.01$ for MTL and $\omega_1 = 1$, $\omega_2 = \omega_3 = 0.1$ for MTL + DT. The results from these experiments can be found on table 4.4.

Table 4.4: Results from the experiments on the automatic modelization of the variability present in the intelligibility scores at word level.

Model	ρ	$RMSE$
Word-RNN System (Baseline)	0.87	0.370
Weighted Loss Function	0.87	0.392
Multi-Task Learning		0.339
Multi-Task Learning with Dynamic Tuning		0.364

The results suggest the same correlation values and small error changes, compared to our previous system. Given this, we cannot clearly say that the automatic modeling of the rating's variability helps to obtain better intelligibility predictions in our particular context.

There are a number of possibilities that can cover why there were no noticeable improvements when modelling the variability in the system. The first can be seen as a non-optimal calibration of the hyper-parameters used during training. Despite the system being fairly similar to the baseline reference, with the exception of the loss functions and corresponding changes that come with it, the hyperparameters

used can usually be improved upon. This process, however is very expensive time-wise, as the training of attention-based models is highly time-consuming. An added layer of complexity is also added due to the weights used in the loss functions (ω_n), that need to be calibrated manually, increasing the overall difficulty of finding the optimal set of parameters. The second reason can be due to the variability measures themselves. The standard deviation used was obtained from the mean of three different scores, which is clearly not a large sample to begin with. Furthermore, the judges used to obtain the PPD measures were not the same for each word of the corpus, which also contributes to an added layer of variability or uncertainty. This aspect also turned the modeling of each individual judges unfeasible in the present context.

Despite the non-improvements, the automatic modeling of the variability present in the scores remains a highly interesting subject, that is highly relevant whenever modeling measures that are subjective by nature.

4.6 Scientific Contributions and Perspectives

4.6.1 Conclusions

In the present chapter, I have proposed the Word-RNN system, a new way to automatically predict speech intelligibility at word level based on a recurrent model with self-attention. Instead of using the perceptual intelligibility given by the evaluation of six independent health professionals, in this particular case I made use of the intelligibility rating given by the perceived phonological deviation, issued by naive listeners. The results suggested not only a high correlation of $\rho = 0.87$, but also a drastic reduction of more than 50% on the RMSE values (from 0.792 to 0.370) when compared to a previous approach based on the automatic transcription of the same pseudo-words. The results also suggested a significant correlation gain (from 0.80 to 0.87) when compared to our *x-vector* approach proposed in the previous chapter.

Moreover, I have also studied the reliability of the system, when using smaller amounts of data at inference time. The results suggested that similarly to a study made with the perceptual evaluations, the quantity of pseudo-words used can be significantly reduced while following a certain criterion and still maintain accurate automatic predictions. The usage of pseudo-words with double consonant was crucial for this reduction, showing that these words can indeed serve as a viable indicator of speech intelligibility. This aspect becomes highly relevant in a clinical practice, since it can not only help counter the variance and subjectivity associated to perceptual measures, but also alleviate patient fatigue by recording smaller quantities of data. An analysis of the modeling of the variability associated to the perceived phonological deviation measures was also executed during the course of this chapter. The results, however, only displayed no significant gains when compared to the (baseline) self-attention approach promoted. An in-depth study on whether it is possible to successfully model the variability of perceptual speech intelligibility measures remains an interesting lead for future work on this matter, and an ongoing research topic.

The results displayed in this chapter were accepted and presented at the conference Journées des Études sur la parole (JEP 2022), under the publication name of "Utilisation de modèles transformers pour la prédiction de l'intelligibilité de la parole de patients atteints de cancers des voies aérodigestives supérieures" [Quintas et al., 2022a].

4.6.2 Perspectives

The results obtained from the pseudo-word reduction clearly show that pseudo-words with at least one occurrence of a double consonant are a better conductor of speech intelligibility than words without. This aspect, that was first proved perceptually and now automatically, shows that an emphasis should be given to these words in the advent of the creation of new clinical evaluations that target coarticulation between not only consonants and vowels, but also consonants and consonants. The usage of longer pseudo-words with even more intricate coarticulations is an interesting lead for future work, that could provide even further discrimination ability across the different levels of speech intelligibility. Despite the pseudo-word

task having a larger emphasis towards the coarticulation of consonants and consonants and vowels, we can also see that the consonants in this context play a more important role for speech intelligibility. This aspect can serve as a validation for the hypothesis made at the end of the previous chapter, which stated that a larger presence of consonants may be correlated to a more accurate speech intelligibility prediction. This aspect will also be investigated in the next chapter at the final granularity level: the phoneme.

Despite the results obtained from the automatic modeling of the rating's variability showing no clear gain when compared to the baseline, it showed that we could use a variability measure that otherwise would be left untouched, without greatly increasing the complexity of the proposed system. This modeling could be useful when applied to either different perceptual measures, such as voice quality, non-PPD based intelligibility and speech disorder severity, that are obtained through a set of professional judges. Even if the results show no clear gain, the automatic prediction of the score variability associated to a single word can serve as a reliability measure for that same word, which provides further important clinical information, as opposed to a single gold standard. The modeling of perceptual judges will be further investigated in the course of the present work on chapter 6, however, with a different speech task and set of judges. A further analysis of the attention plots could also provide some interesting insight. On the other hand, a more in-depth study should be devised, with the main goal to find clinical value and valid interpretations of these plots. An interesting follow-up would be to find a connection between distorted bits of an attention plot and specific mispronunciations.

Key Takeaways

1. We are able to predict speech intelligibility at word level using the **Word-RNN System**, a recurrent model with self-attention, through the means of the pseudo-word task of the C2SI corpus.
2. We can obtain a high correlation and low error values when predicting speech intelligibility at word level by using the aforementioned methodology.
3. It is possible to drastically reduce the quantity of pseudo-words used in the system without affecting much the reliability of the automatic intelligibility score.
4. Pseudo-words with occurrences of double consonant are a better indicator of speech intelligibility when compared to pseudo-words without.
5. The relevance of pseudo-words with double consonant is valid for both the perceptual and automatic case.
6. In the context of the perceived phonological deviation measures used, training the proposed system with the variability measures of the scores used as reference only promoted marginal gains.

Chapter 5

Final Granularity Level: *The Phoneme-SN System*

5.1 The Phoneme: the Heart of Speech Intelligibility Measures

In this chapter, we would like to introduce to the reader the final level of granularity studied in the present work: **the phoneme**. A deeper analysis at this level becomes relevant since phonemes can be seen as the foundations of spoken communication, and therefore an automatic approach at this level can provide interesting queues for speech intelligibility. By shifting our focus of attention towards isolated phonemes, this approach can be seen as more objective, since an intelligibility score obtained directly from phonemes is more interpretable. It is also known that a phonetic analysis is highly relevant for speech intelligibility [Nuffelen et al., 2008], making individual phonemes a relevant and valuable tool.

In the present chapter we aim to study how we can devise a system that produces an intelligibility estimation at phoneme level, that is able to achieve high correlations with the perceptual intelligibility values. Similarly to chapter 3, we will use as reference the intelligibility measure obtained from picture description, in the context of the C2SI corpus.

5.1.1 Phoneme and Phone Definitions

Similarly to the previous chapters, we will start by defining the adequate necessary terminology. According to the Oxford language dictionary[9], a **phoneme** is the smallest sound unit in a language. Meaningless in themselves, phonemes can be used sequentially to construct words, and therefore are the building blocks of language. Changing one for another changes the meaning of a word. A **phone**, on the other hand, is considered in phonetics and linguistics as any distinct speech sound or gesture, regardless of whether the exact sound is critical to the meanings of words. While both definitions may sound oddly similar, there is a clear distinction between the two. A phoneme is generally regarded as an abstraction of a set (or equivalence class) of speech sounds (phones) that are perceived as equivalent to each other in a given language. In practical terms, a phone is a speech sound and a phoneme is a mental representation of that speech sound. For example, in the case of a segmented word, in the transcription we will find phoneme representations, that denominate the sound being issued, while on the segmented audio we can find the separated phones. An **allophone** is one of a set of multiple possible spoken sounds (phones) or signs used to pronounce a single phoneme in a particular language, or essentially, allophones correspond to the different ways that we can pronounce the same phoneme, depending on the context within the word. Since in the present chapter we are dealing with phonemes, it becomes relevant to define a notation to represent the same phonemes in a textual manner. For the sake of simplicity and readability, the Speech Assessment Methods Phonetic Alphabet (SAMPA[10]) is adopted as the main phonemic notation, although

[9] https://languages.oup.com/google-dictionary-en/
[10] https://www.phon.ucl.ac.uk/home/sampa/

sporadically the corresponding International Phonetic alphabet (IPA) [Lodefoged, 1990] symbols will be added to provide further context.

5.1.2 The Most Revealing Phonemes in Terms of intelligibility

In order to perform a proper analysis at phonetic level, one should first obtain the phonemes to work with. In a pathological speech context, phonemes can be extracted from a variety of recorded tasks, however the number of occurrences of each phoneme should be taken in consideration. Underrepresented phonetic classes may provide a biased idea of a given patient's phonation ability, and should be addressed accordingly. A passage reading task, for example, can be seen as a good starting point, however the majority of these tasks may promote fewer occurrences of specific phonemes, since they are normally dimensioned to include at least one occurrence of each phoneme, and not a large variety. Picture description tasks may be interesting as well, however, since there are no scripts to follow, there is no uniformity across different speakers, which can promote great discrepancies on phonetic frequencies. In this particular context, a pseudo-word task can be seen as the more reliable alternative, since these normally have several occurrences of the same phoneme in different contexts, as in the beginning, middle and end of the word. The key co-articulations between consecutive phonemes, normally present in these tasks, can provide exactly the best data to be used in our granular analysis. The pseudo-words used in this type of recording sessions can follow different structures, such as $V_1C(C)_1V_2$, $C(C)_1V_1C(C)_2V_2$ or even $C_1V_1C_2V_2C_3V_3$, where $C(C)_i$ is either a single consonant or a consonant group and V_i a single vowel. It is highly important to understand which phonetic classes are targeted in this type of task, since some of them may be focused more on the consonants than vowels (vowels can be assessed in a sustained vowel task, while consonants cannot). Different studies support the claim that consonants are better at conveying word information, while vowels, on the other hand, excel at conveying speaker identity and grammar [Bonatti et al., 2005, Owren and Cardillo, 2006, Fort et al., 2015]. Consonants have also been proved to be useful in the post-assessment of laryngectomies [Crevier-Buchman et al., 2002], a typical surgery for head and neck cancers. Given this aspect and the fact that word information decoding is highly correlated to speech intelligibility [Lalain et al., 2020], in the present work we hypothesize that consonants can indeed be a viable indicator for the automatic prediction of speech intelligibility. This aspect will be the central focus of attention throughout this chapter.

5.1.3 Similarity Estimation Systems

Since we are evaluating a speaker at the phonetic level, we can use different types of systems to extract an intelligibility score. The first approach that usually comes to mind is to regress a score based on the amount of recognized phonemes by an automatic speech/phoneme recognizer. Despite this approach being already implemented in the literature, it is known that it underperforms on patients with severe speech impairments [Christensen et al., 2012], which are also the patients that would benefit more from a continuous and precise clinical speech assessment. It is important to state that the automatic speech recognition of pathological speech is a difficult and ongoing research area [Yu1 et al., 2018, Xiong et al., 2019]. Given this, it becomes relevant to search for a reliable alternative to the usage of ASR. Similarity estimation systems, such as siamese networks, can be employed to assess the quality of phonemes when compared to well-articulated references. Considering that we can deal with pairwise comparisons of isolated phones instead of an approach based on acoustic models (e.g. ASR), these systems become an interesting and relevant alternative.

Similarity estimation systems, as the name indicates, are a class of systems used to compute the similarity between two given inputs. Although no single definition of similarity exists, usually these measures take some form of distance between the two inputs. There are many ways to compute a similarity measure, however, the majority of the approaches used in the literature tend to transform both inputs into latent space representations (also known as embeddings or fixed-length vector representations) and then calculate a distance measure between the two. The distance measure can be obtained from a multitude of operations, namely the absolute distance, mean square displacement or cosine similarity. Non-linear approaches such as neural networks can also be used to obtain distance measure or a similarity estimation.

One of the main approaches used to obtain similarity measures, within a deep learning framework, are the so-called siamese networks[11]. A siamese network is an artificial neural network that uses the same weights while working in tandem with two inputs to compute comparable output vectors. After obtaining those vectors, a distance can be computed between these fixed-length representations. This distance can be interpreted as the similarity between the two inputs. Siamese Networks work very similarly to artificial neural networks, since the main difference lies in the usage and respective optimization of the same system for both inputs. There are many different ways to optimize a siamese network. Seeing that the main objective of the system is to push latent representations of similar inputs closer in the embedding space while moving away dissimilar pairs, multiple loss functions can be applied. The contrastive loss function [Wang and Liu, 2021] is widely used within this context. This loss function uses pairs of similar and dissimilar inputs and maximizes or minimizes the Euclidian distance between them accordingly. The triplet loss function [Dong and Shen, 2018] works similarly, however in this case, a given input is compared to a similar and dissimilar pair respectively. By using artificial neural networks after a distance calculation between latent representations, a siamese network can also be optimized to increase or decrease the Euclidian distances. Then, a binary-cross entropy function can be used in the case of binary classification (either similar or dissimilar) or a (Root) Mean squared Error in the case of the regression of an exact similarity measure [Khan et al., 2020].

Siamese networks have seen a growing use in tasks like speaker verification and sentence similarity [Wan et al., 2018, Khan et al., 2020, Mueller and Thyagarajan, 2016]. Recent works used the aforementioned methodology in a pathological speech context [Wang et al., 2019, Ng and Lee, 2020]. In both cases, the systems were developed for the detection of children's speech disorder, focusing on the binary task of detecting specific mispronunciations.

In conclusion, we introduce an intelligibility prediction system based on consonant similarity. There are multiple motivations behind the development of such a system, namely that:

(i) ASR based intelligibility prediction systems typically underperform in patients with severe speech impairments [Christensen et al., 2012];

(ii) Individual phonemes, especially consonants, are highly relevant for speech intelligibility measure [Nuffelen et al., 2008, Saravanan et al., 2016], whether in healthy [Fort et al., 2015] or pathological people [Crevier-Buchman et al., 2002];

(iii) Finally, automatic systems tend to lack explainability [G.McCoy et al., 2021], which is normally demanded by health practitioners [Diprose et al., 2020].

Given this, we propose an automatic system that predicts speech intelligibility based on consonant similarity, that is able to output not only an objective, but also a fully-explainable prediction. Since we have found previously on chapter 3 that there are sentences able to conduct a more accurate intelligibility prediction [Quintas et al., 2020], we also explore the relevance of specific phonemes in our automatic speech intelligibility score.

5.2 The Siamese Network and Phonetic Similarity

The proposed system for the automatic prediction of speech intelligibility at phoneme level relies on three components. The first one corresponds to the feature extraction which includes data preparation. In subsection 5.2.1 we will explain how we obtained the segmented phonemes as well as the features used. The second component is based on a recurrent siamese network in order to compute the phonetic similarity between the consonant-phone sets: the set pronounced by the test speaker and the reference-consonant set. This set will be further explained in subsection 5.3.1. In subsection 5.2.2, we will introduce the prediction module followed by a detailed explanation of the working principle. Finally, the third and final step corresponds to the calculation of the intelligibility score, based on the phonetic similarity scores previously achieved. This can be found described in subsection 5.2.3. Figure 5.1 presents a global overview of how the proposed system works, highlighting the three components previously mentioned.

[11] `https://www.mygreatlearning.com/blog/siamese-networks/`

Given the importance of the Siamese Network (SN) in our system, we will subsequently call our system the Phoneme-SN system.

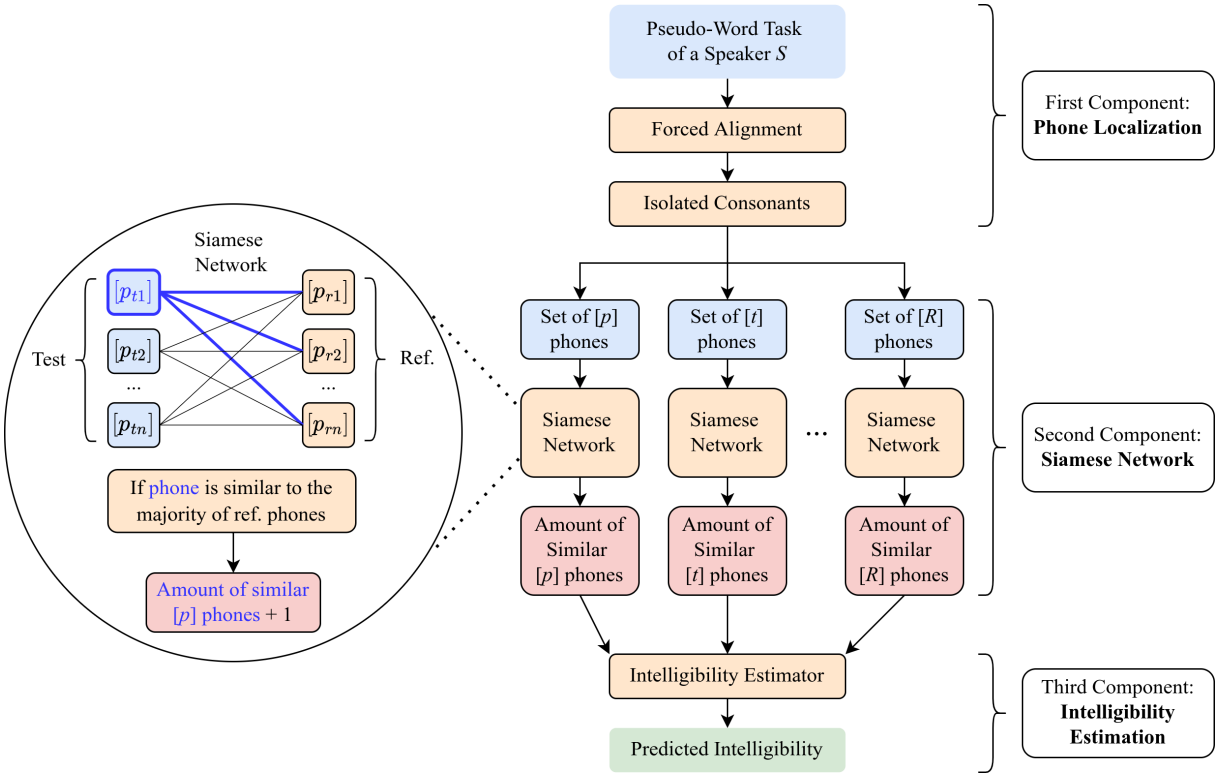


Figure 5.1: General overview of the proposed system. The pseudo-words from a given speaker are force aligned in order to obtain the isolated consonants. Afterwards, each phone of a given class is compared to the reference canonical phones (obtained from the control speakers) through the means of a siamese network. From this comparison, the amount of similar phones of a given class is obtained. Finally, the predicted intelligibility is computed based on the amount of similar/dissimilar phones.

5.2.1 Phone Localization

We use individual phones as input to our system. In order to isolate the phones present in the input utterance, we made use of a forced alignment system. Forced alignment is a technique that takes an orthographic transcription of an audio file and generates a time-aligned version using a pronunciation dictionary to look up phones for words. Figure 5.2 illustrates an example of a forced alignment process, where we have the audio file and the corresponding timestamps for isolated words and phones. These systems have a crucial importance in tasks such as automatic speech recognition, grapheme-to-phoneme conversion and diverse segmentation tasks. In the context of the present work, we made use of the Montreal Forced Aligner (MFA) toolkit [McAuliffe et al., 2017], which has several trained models for several languages. Annex B presents an overview of the working principle and training behind forced alignment, more specifically the MFA. The software was used without any additional changes, the inputs correspond to the canonical written pseudo-word and the utterance file, the output corresponds to the aligned file with the marked time-stamps.

Feature Extraction

For each file, 13 Mel Frequency Cepstral Coefficients (MFCCs) were extracted. Filterbank features, MFCCs + delta coefficients and phonetic posteriors [Vasquez-Correa et al., 2019], which are also relevant

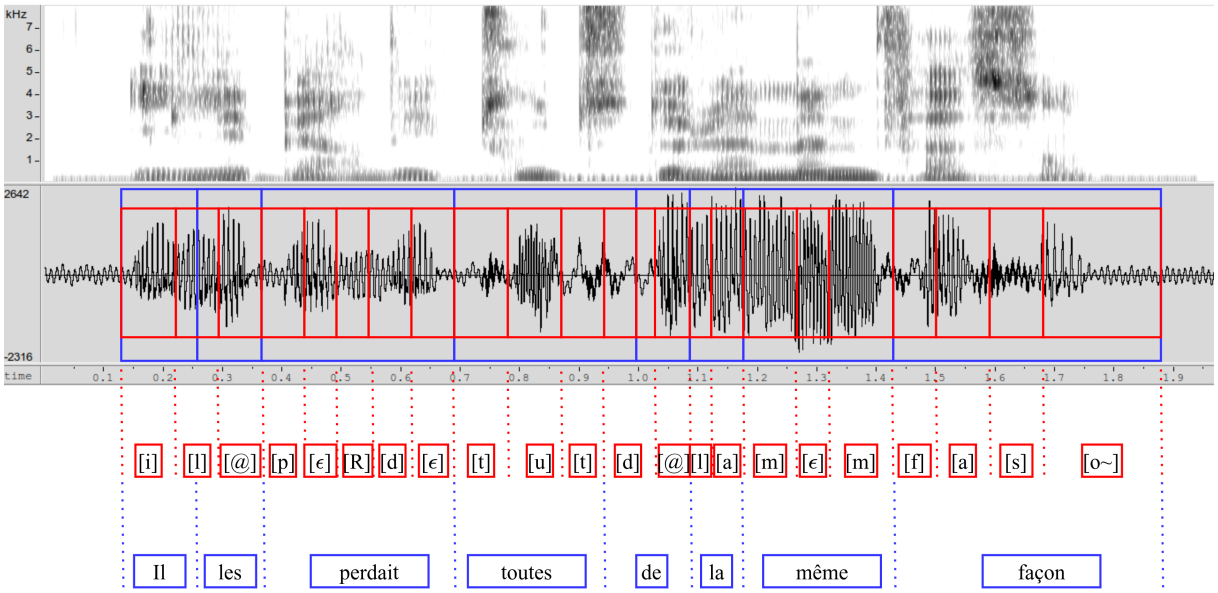


Figure 5.2: Example of a force-aligned audio file. The image presents an audio waveform, the corresponding spectrogram with traced formants, and the respective timestamps for isolated phonemes words.

features for the automatic processing of pathological speech, were also experimented, however, in our specific context they conducted to a poorer generalization ability of the system. Interestingly, the work of [Lin and Tseng, 2021] also came across a similar finding regarding the features used. The authors concluded that on the automatic classification of children's speech intelligibility, the MFCCs were more useful than Mel Spectrogram and acoustic-phonetic features.

5.2.2 The Siamese Network Components

Inspired by the usage of similarity estimation systems for the detection of phonetic mispronunciations [Wang et al., 2019], the proposed system uses a siamese network to evaluate phonetic similarities. The system receives as input two phones, one to be used as a reference and the other as a test, and computes their similarity. A mispronunciation is detected whenever the test phone is found to be dissimilar from the reference one. The level of similarity is given by a threshold, should a phoneme pair have a similarity score (given by the sigmoid between 0.0 and 1.0) above that threshold, the phonemes are considered similar.

Our Siamese Network[12] uses two bilateral Gated Recurrent Units (GRU) as encoders with shared weights, in which the two outputs form embedding representations (see figure 5.3). Both encoders share the same structure, and are comprised of two hidden layers, with a hidden dimension of 100.

The embedding representation corresponds to the concatenation of the last two hidden states, of forward and backward context on the bilateral GRU used. At the output of each encoder, we obtain a fixed-length representation of size 200, where half of it corresponds to the forward and the other half to the backward context respectively. The absolute difference between the two fixed-length representations is then computed. As it was mentioned before, the main goal of the proposed system is to learn the similarity between two input phones. Given this, it is expected that the system models the input phonemes so that same-pair phonemes are closer in the embedding space than different-pair phonemes. Afterwards, a Deep-Neural-Network block is appended. This block is composed of 3 fully-connected layers of size 200. The Rectified Linear Unit function is used as an activation function in all of the layers except for the final one, which used a sigmoid. A dropout rate of 0.25 and batch normalization are applied to every fully-connected layer. The final output gives a phonetic similarity score. Finally, the level of similarity is given

[12] `https://github.com/Elquintas/siamese_intel`

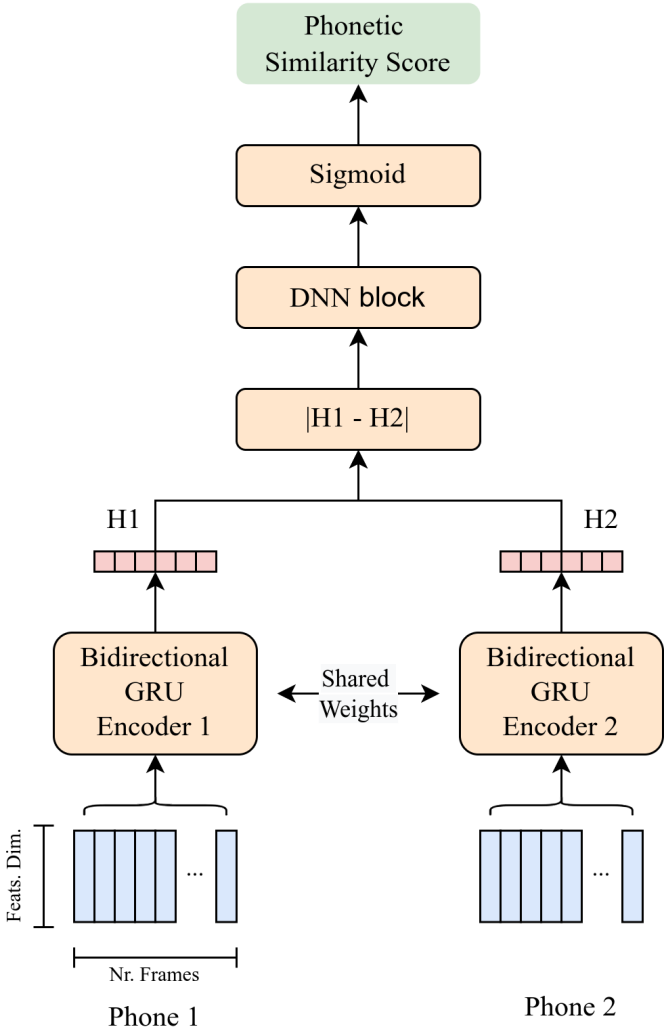


Figure 5.3: Schematic diagram of the proposed siamese network. If a pair of phonemes has a similarity score above a certain threshold (that will be further explained), those phonemes are considered similar, otherwise they are dissimilar.

by a threshold, should a phoneme pair have a similarity score above that threshold, the phonemes are considered similar. The phonetic similarity score is either 1 or 0 to convey the similarity or dissimilarity.

5.2.3 Intelligibility Estimation

Based on the similarity scores obtained by the Siamese Network, we compute an intelligibility score for each speaker:

$$I(S_a) = \frac{\sum_{n=1}^{16} \frac{\mathrm{Sim}_a(n)}{\mathrm{Tot}_a(n)}}{16} * 10 \tag{5.1}$$

This scoring function corresponds to the arithmetic mean of each patient's individual consonant score.

If n is a consonant and S_a the evaluated speaker, $\mathrm{Sim}_a(n)$ refers to the number of occurences of the consonant n pronounced by S_a considered similar to the reference-set. while $\mathrm{Tot}_a(n)$ is the total number of representations of that consonant issued by a speaker. By using this score function, we are able to regress an automatic intelligibility prediction that is a direct function of the number of similar

phones, consonants in this case, that a given speaker has. This, in turn, promotes a more explainable prediction than the dreaded "black box" approach, normally associated to deep learning methodologies. This aspect becomes highly relevant in a clinical context. Figure 5.4 presents a global overview of the proposed methodology, from the individual phone similarity to the regression of the intelligibility score per speaker.

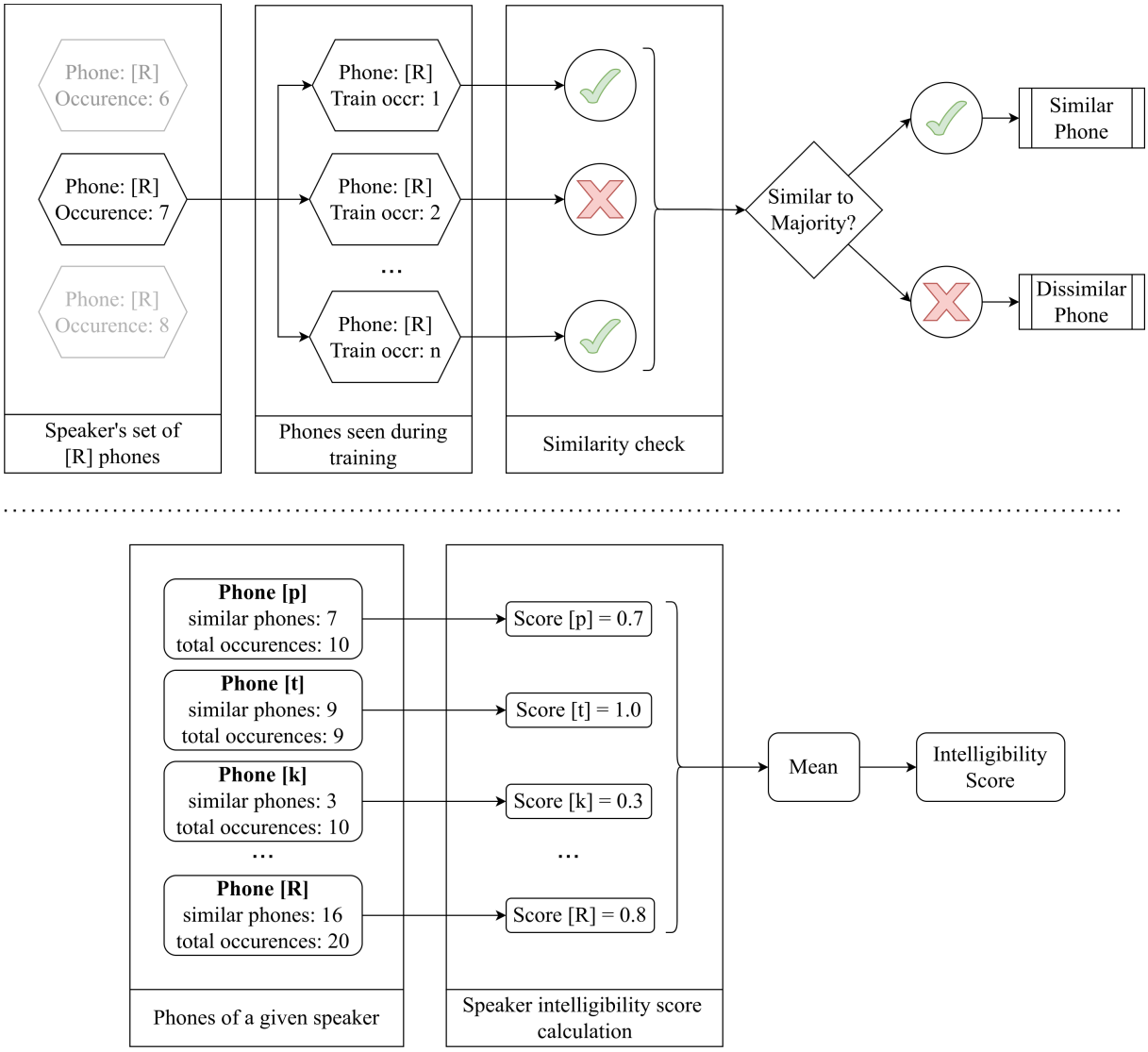


Figure 5.4: Overview of the proposed approach. Each new phone is compared to all the phones of that same type seen during training. A phone is considered similar if it is similar to the majority of the training phones of that same type. The phonetic score for a given phoneme corresponds to the number of similar phones divided by the total number of occurrences of that same phoneme. The intelligibility score is the mean of the individual scores of each phoneme. In our case, the phonemes correspond to the 16 French consonants.

5.3 The Phoneme-SN System Implementation

5.3.1 The Training Phase

Selection of the Training Corpus

Similarly to the previous study, we made use of the isolated pseudo-word task of the C2SI corpus (see section 4.2 for details). Each pseudo-word follows a $C(C)_1V_1C(C)_2V_2$ structure, where $C(C)_i$ is either a single consonant or a consonant group and V_j a single vowel. An example of a set of 52 recorded pseudo-words can be found in the table 5.1.

Table 5.1: Example of a set of 52 pseudo-words with the aforementioned structure of $C(C)_1V_1C(C)_2V_2$.

banfou	bleja	boucti	brimpli	chessant	choniou
clifant	cogu	crimpin	daillu	dinrant	dredi
fanrsi	flinrpu	fouma	fravi	gabi	glunou	grorvo
guchin	joutu	juro	lanvin	lerda	messo	mouco
nianlo	niejo	noksa	nouillou	pastu	pidant
ploniou	pripin	psila	quiga	rinta	rurnu
sanvrin	scuna	souquin	spaclant	sticho	tangri
tougzu	tradrou	virjant	vumou	yainzi
yaltin	zebou	zouzant			

While the majority of the French consonants were present in each set of 52 pseudo-words, there were only 8 vowel representation. As indicated in the introduction of this chapter (section 5.1), consonants can be a good perceptual indicator of regular speech intelligibility and for the intelligibility of post-operatory laryngectomies [Fort et al., 2015] [Crevier-Buchman et al., 2002], so we have decided to perform the present intelligibility estimation using only consonant similarity. Table 5.2 displays the description of the 16 consonants used. In this table, we present a given consonant by its SAMPA symbol, International Phonetic Alphabet (IPA [Lodefoged, 1990]) symbol, sonority, manner and place of articulation and finally minimum and maximum number of occurrences for the 52 pseudo-word sets. Despite the random nature of the pseudo-word generation, we observe that the number of occurences of each consonant slightly changes across different speakers. The pseudo-word task of the C2SI corpus did not contain instances of the consonant [J] (IPA symbol "ɲ", place of articulation: palatal, manner of articulation: nasal. e.g. monta**gn**e) nor semi-consonants[13], and therefore both were not contemplated in the present analysis.

The intelligibility values, used as targets, were computed based on the independent perceptual evaluation of six different therapists. Similarly to chapter 3, the intelligibility measure was obtained from the picture description task of the C2SI corpus. A score between 0 and 10 was attributed, based on those evaluations, the smaller the score is, the less intelligible the speaker is. A total of 102 speakers, 24 healthy controls (HC) and 78 patients, were used in the present study. The set of speakers corresponds to those who recorded the pseudo-word task and were also submitted to the perceptual evaluation aforementioned.

Siamese Network Training

Our proposed system relies on having input pairs of phones. In order to better prepare our data for training, the phones were paired in two categories:

- **same-phoneme pairs**: each one is composed of two different occurrences of the same phoneme,
- **different-phoneme pairs**: such pair mimics mispronounced phonemes by pairing occurrences of different phonemes.

[13] https://nathaliedoucet.live/2020/09/28/the-french-semi-consonants/

Table 5.2: Description of the 16 French consonants used. The occurrence interval states the expected number of representations of a given consonant in a random set of 52 pseudo-words.

Sampa symbol	IPA symbol	Sonority	Manner of articulation	Place of articulation	Occurrence interval [min,max]	Example in regular french
[p]	p	unvoiced	plosive	bilabial	[10, 12]	***p**artir*
[t]	t	unvoiced		dental	[9]	***t**emp*
[k]	k	unvoiced		velar	[10]	***c**ouche*
[b]	b	voiced		bilabial	[6, 9]	***b**anane*
[d]	d	voiced		dental	[5, 7]	***d**atte*
[g]	g	voiced		velar	[7, 8]	***g**randir*
[s]	s	unvoiced	fricative	alveolar	[4]	***s**ans*
[S]	∫	unvoiced		palatal	[6, 8]	***ch**amps*
[z]	z	voiced		alveolar	[4, 5]	***z**one*
[Z]	3	voiced		palatal	[4, 5]	*man**g**er*
[f]	f	unvoiced		labiodental	[6, 8]	***f**emme*
[v]	v	voiced		labiodental	[4, 6]	***v**ouloir*
[m]	m	voiced	nasal	bilabial	[4, 5]	***m**é**m**oire*
[n]	n	voiced		dental	[4, 5]	***n**ager*
[l]	l	voiced	liquide	dental	[12]	*fina**l**ement*
[R]	R	voiced		velar	[20]	***r**atatouille*

In order to learn phonetic similarity, we used same-phoneme pairs as positive and different-phoneme pairs as negative. 24 HC of the C2SI corpus were used as training data. After force aligning and extracting the features, we obtained a total of 3,323 training phones (consonants only). For each of the 16 consonants groups, an individual training set was created. Such a set is composed by all the representations of a given consonant and a random set of 650 different consonant phones, whose size is around 850 (based on the occurence interval, see table 5.2). From this set, the pairs are formed without repetition, and a random subset of 50k pairs was extracted from all the possible combinations. This subset corresponds to the final training set for that same consonant.

Sixteen different models, one for each consonant, were created and trained individually with the corresponding consonant subset. Each individual model was fine-tuned on four hyperparameters: learning rate (between 0.001 and 0.0001), epochs (between 2 and 8), batch size (either 256 or 512) and GRU dropout rate (either 0.0 or 0.25). A binary cross-entropy loss function was used, optimized by the Stochastic Gradient Descent (SGD) algorithm [Ketkar, 2017]. Similarly to the previous chapters, optimal hyperparameter searching was performed empirically.

Validation of the Phoneme-SN System

The first necessary validation of our global system concerns the Siamese Network. A set of six patients from the C2SI corpus, unseen during training, was used to validate the proposed model instead of a subset of the HC. The chosen patients had high intelligibility (over 9.8 on a 0 to 10 scale), making them virtually indistinguishable from HC. A validation set of 960 phones was deduced.

Each new phone of the validation set is paired with all the reference phones seen during training. After gathering the validation phones into the 16 different consonant groups, the median similarity is computed between each test and reference consonant groups (see figure 5.5).

We obtain a reliable distinction between all phonetic groups. This aspect was less evident in the plosive consonants, where the difference between the key consonant to be tested and the other consonants was more subtle when compared to the remaining groups. This was expected due to the nature of forced-alignment, that segments plosives in short time intervals. This, in turn, provided less contextual information, making the phonetic distinction a slightly more difficult task. This aspect, however, did not prove to be problematic as all speakers were submitted to the same forced-alignment.

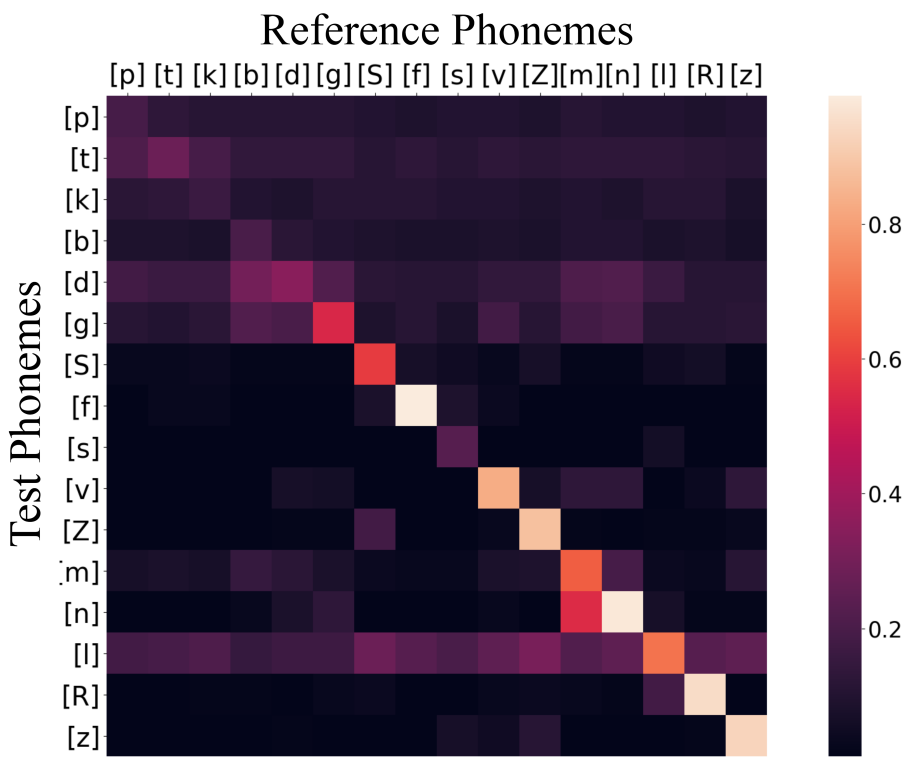


Figure 5.5: Siamese network validation heatmap.

To finalize the validation of our system, we must fix the threshold applied to the level of similarity: we used, as a threshold, the median values obtained by the **validation** patients for each consonant group, also known as the diagonal values of figure 5.5. In order to make the system more robust to the different median values, we added to the threshold the median absolute deviation (MAD) [Pham-Gia and Hung, 2001]. The classification method is as follows. If a new phone has a similarity score above the threshold ($\tilde{x_i} - MAD_i$), where $\tilde{x_i}$ is the median, it is considered a similar phone, otherwise it is considered dissimilar. The intelligibility score is computed according to equation 5.1. Figure 5.4 illustrates the phone-wise comparison and also the respective intelligibility score calculation.

5.3.2 A First Evaluation of the Phoneme-SN system

Similarly to the previous chapters, the Spearman's correlation coefficient ρ and the Root Mean Squared Error (RMSE) were used as evaluation metrics. By correlating the predicted scores with the given intelligibility reference values, we were able to achieve a Spearman's correlation coefficient of $\rho = 0.82$. The results can be found displayed on table 5.3 and illustrated on figure 5.6.

Table 5.3: Comparison between the results obtained at this level of granularity and two other approaches, one based on automatic speech recognition and the other based on *x-vector* speaker embeddings (see chapter 3).

	ρ	$RMSE$
Automatic Speech Recognition Approach	0.63	2.406
X-vector Speaker Embeddings Approach	0.74	2.488
Siamese Network	0.82	2.350

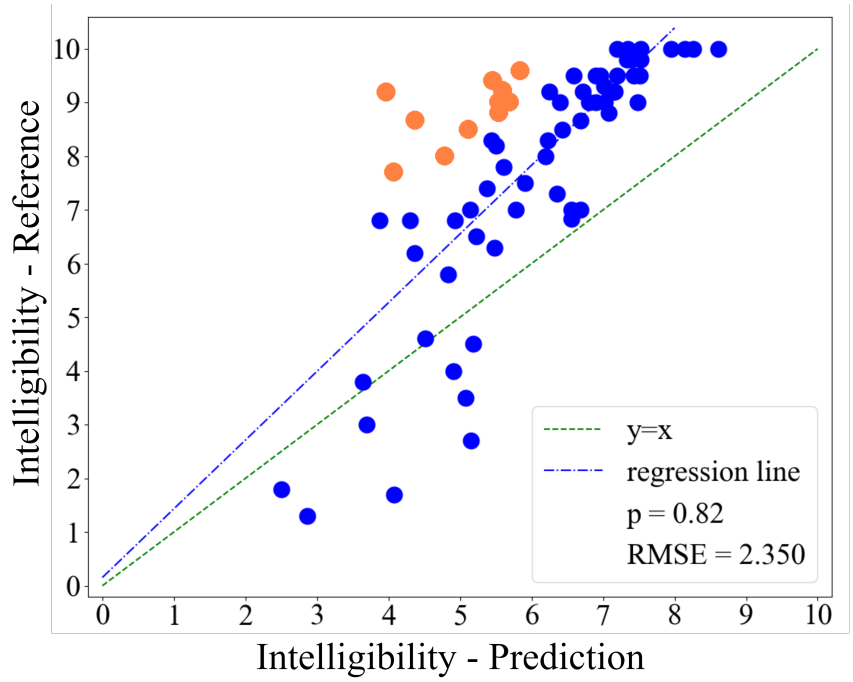


Figure 5.6: Intelligibility prediction plot from the proposed system that operates at phoneme level. Outliers are marked with orange points.

The same results were compared to two other approaches used to predict speech intelligibility. The first one is based on the distance between a pseudo-word and its corresponding automatic transcription, obtained via an automatic speech recognition system [Fredouille et al., 2019]. Similar to the previous chapter, these ASR results were obtained from the Wagner-Fischer algorithm, fully described in section 4.2 and adapted here to the number of patients used. While on the present evaluation we give more emphasis to correlation due to the system not being optimized on the reference intelligibility, we scaled the Wagner-Fischer results so the results are comprised in a $[0, 10]$ scale: this way both approaches become comparable. The second reference system is based on the *x-vector* speaker embedding paradigm, introduced in chapter 3. The same set of speakers (see subsection 5.3.1) was used in the three different systems. When compared to these two systems, our results suggest a significant correlation gain and an added level of interpretability, since the results when using the proposed approach can be traced back to the amount of similar/dissimilar phonemes. As expected, the approach based on ASR underperformed when compared to the other two. The RMSE values obtained, despite being larger than the ones obtained in the previous granularity levels (chapters 3 and 4), were smaller for the proposed system than the two reference systems used.

5.4 Result Analysis: Towards Interpretability

From the plot illustrated in figure 5.6, despite the high correlation achieved, visualized by the regression line, the trace of the line $y = x$ indicates that our intelligibility-predicted score is, very often, lower than the perceptual intelligibility reference measure. To precise this trend, it becomes relevant to analyze what can be interpreted as outliers in order to better understand this behavior and perhaps give a first interpretability of our score. Due to the high RMSE value found on the predictions (2.350), we considered the outliers as the speakers with a prediction score outside a $[-3, T, 3]$ boundary.

As it was previously said, the predicted values of these outliers had a large deviation from the target intelligibility, and a prediction below target. Interestingly, no speakers with a reference intelligibility score below 5.0 were found as outliers, contrary to what the results suggested on chapter 3. From the

perceptual evaluation of these outliers, we found that the outliers are observed in specific contexts, namely very breathy/hoarse speech and patients with nasalized plosives. All of the cases corresponded to phonetic mispronunciations and it was expected that the system would classify them as such for the obvious reason that it was neither trained nor validated with speakers that had similar mispronunciations. On the other hand, despite the phonemes being mispronounced, we noticed that in these cases the perceptual reference intelligibility values were still high, pointing out that in the aforementioned contexts, those specific mispronunciations had a little effect on this perceptual measure. This aspect points out that despite a phonemic analysis having a high correlation with speech intelligibility [Nuffelen et al., 2008, Saravanan et al., 2016], there are certain phonemes which, in specific contexts, convey complementary information. This aspect can also be visible by the high correlation ($\rho = -0.93$) found between the phonemic distortions parameter and the intelligibility obtained from picture description (see table 2.1). In other words, there are phonemes that, despite being mispronounced, do not play a relevant role for the perceptual intelligibility of the same speaker.

5.4.1 Phonetic Suppression Post-Processing

To validate this hypothesis, we assumed that, by suppressing specific consonants, we were able to obtain an intelligibility estimation closer to the reference. This suppression took place when computing the final intelligibility score (see equation 5.1), after computing the similarity score for each speaker's phonemes (see "Phonetic Similarity Score" from figure 5.3). Thus, we made use of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS): an ensemble of acoustic parameters tailored for indexing physiological changes in voice production [Eyben et al., 2015], with an added degree of theoretical significance and explainability, and we went to search for hand-crafted features used in the literature to address the key types of speech impairment aforementioned. We found a set of three acoustic parameters:

- **Slope UV0-500 (mean) -** Mean value of the linear regression slope of the logarithmic power spectrum within 0-500 Hz on unvoiced segments. Related to breathy and hoarse voice qualities [Alipour et al., 2012].
- **Loudness (percentile 20) -** Estimate of perceived signal intensity from an auditory spectrum. In our context, we hypothesize that it can help detecting nasalized plosives due to the added intensity found in this type of mispronunciation [Tjaden and Wilding, 2004].
- **LogRel F0-H1-A3 (mean) -** Ratio of energy of the first $F0$ harmonic (H1) to the energy of the highest harmonic in the third formant range (A3). Relevant feature for breathy/hoarse voice assessment as well [Narasimhan and Vishal, 2017].

These three acoustic parameters will be used to model our phonetic suppression scheme, devised as a post-processing. The **Slope UV0-500 (mean)** and **LogRel F0-H1-A3 (mean)** features are relevant to recognize speakers with a breathy/hoarse voice quality while the **Loudness (percentile 20)** is relevant to identify speakers with nasalized plosives. The thresholds for the phonetic suppression were fixed empirically. Table 5.4 presents the thresholds used for the phonetic suppression, as well as the corresponding suppressed phonemes. The results from these multiple potential post-processing can be found in table 5.5.

5.4.2 Discussion

The results suggest that we can reliably predict speech intelligibility using consonant similarity. Moreover, by conditioning the used consonants on key mispronunciations and external features, we are able to obtain an even higher level of correlation ($\rho = 0.89$), illustrated in figure 5.7. The number of outliers obtained from the post-processing, outside a $[-3, T, 3]$ boundary, was also significantly lower (see figures 5.6 and 5.7). This aspect points out that, depending on the speech impairment a speaker may have, there are mispronounced phonemes that do a little contribution to the overall intelligibility score:

- For patients with a high level of hoarseness, all voiced phonemes were classified as non-similar by the system. By suppressing the voiced phonemes, we were able to obtain a more accurate prediction

relative to the perception. To detect this hoarseness, we used the **Slope UV0-500 (mean)** and **LogRel F0-H1-A3 (mean)** features (see table 5.4). Any patient that had a feature value above a threshold had their voiced phonemes suppressed, and the remaining phonemes were used for the score.

- The suppression of the full plosive group, displayed on table 5.4, also lead to more accurate predictions, showing that those mispronunciations did not affect much of the perceptual intelligibility estimations (see correlation values on table 5.5).

As expected, the used features also isolated a few patients that did not have the specific mispronunciations aforementioned. However, the same phonetic suppression poorly affected those intelligibility scores, confirming a certain level of robustness of the chosen features.

Table 5.4: Phonemes used and suppressed in the intelligibility score function according to the GeMAPS features.

Feature name	**Threshold**	**Phonemes Suppressed**	**Phonemes Used**
Slope UV0-500 (mean)	> 2.41	[b], [d], [g], [z], [Z], [v], [m], [n], [l]	[p], [t], [k], [s], [S], [f], [R]
LogRel F0-H1-A3 (mean)	< 15.00		
Loudness (percentile 20)	> 0.31	[p], [t], [k], [b], [d], [g]	[s], [S], [z], [Z], [f], [v] [m], [n], [l], [R]

Table 5.5: Correlation and RMSE results achieved by the proposed methodology and by the phonetic suppression post-processing. The results are also compared to two baselines, an ASR-based intelligibility prediction (described on the previous chapter and adapted here to the number of patients) and the *x-vector* analysis proposed in chapter 3.

		$\boldsymbol{\rho}$	RMSE
Automatic Speech Recognition Approach		0.63	2.406
Speaker Embedding Approach (*x-vectors*)		0.74	2.488
Consonant Similarity Approach	Predicted	0.82	2.350
	Predicted + Loudness	0.84	2.266
	Predicted + LogRelF0-H1-A3	0.85	2.288
	Predicted + SlopeUV0-500	0.86	2.204
	Predicted + LogRelF0-H1-A3 + SlopeUV0-500 + Loudness	**0.89**	2.080

The assumption that different phonemes have different degrees of relevance corroborates the fact that for each speaker, there are sentences that are able to convey a better intelligibility estimation than others, concluded in chapter 3. A deeper feature analysis should be investigated in order to identify other contextual key phonemes that are less important in the intelligibility score. Further robust feature conditioning could help provide more accurate scores and also a more objective and explainable patient-specific information.

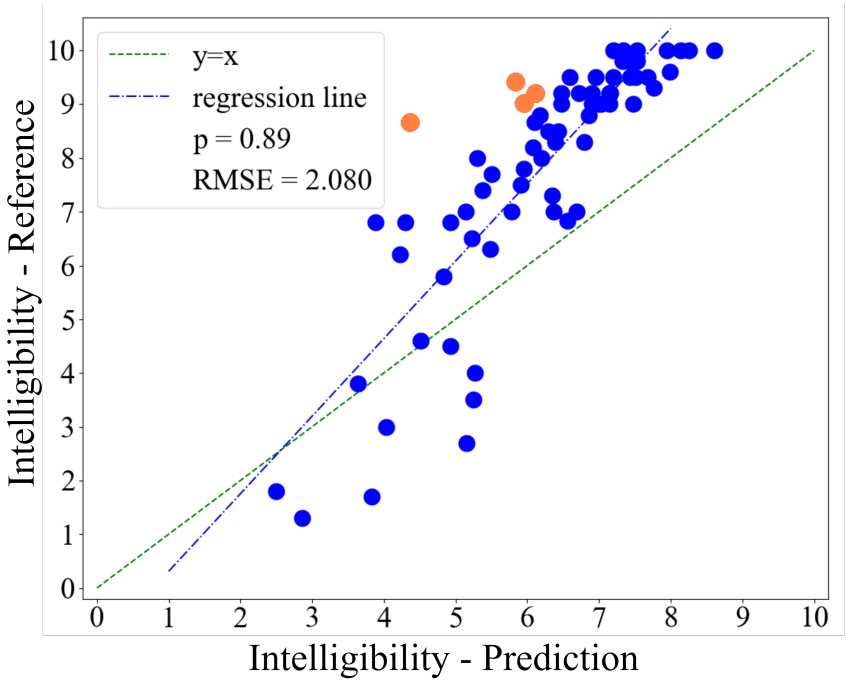


Figure 5.7: Intelligibility prediction plot obtained from the post-processing (LogRelF0-H1-A3 + SlopeUV0-500 + Loudness). Outliers are marked with orange points.

While the systems developed in the two previous chapters made use of entirely data-driven approaches to regress an intelligibility score, **the Phoneme-SN system regresses a fully objective and explainable intelligibility score without using an intelligibility measure as a training reference**. This aspect becomes highly relevant since the system can objectively predict speech intelligibility while correlating well with the perception-based measures. While the outliers show that the predicted score was deviated from the reference intelligibility, we cannot say that the system was underperforming on the aforementioned cases, specially since those suppressed phonemes (see table 5.4) were in fact being mispronounced. Given this, several questions could be raised concerning the relevance of individual phonemes for the intelligibility score, which were partially assessed during the present post-processing. The correlations found between phonemic distortions and speech intelligibility (see table 2.1) show that despite both being highly correlated, there is still no perfect connection between the two parameters. This phenomenon could be attributed to a variety of reasons, such as the relevance of other perceptual parameters or lack of agreement among the judge set. While these aspects were not analyzed just yet, they foreshadow the heart of the upcoming chapter, that will deal directly with the modeling of individual schools of thought promoted by the set of judges of the C2SI corpus.

5.5 Scientific Contributions and Perspectives

5.5.1 Conclusions

In this chapter I have presented an automatic approach that predicts speech intelligibility at phoneme level, which is also the thinnest level of granularity that was used in the present work. The proposed system uses consonant similarity to regress our intelligibility score. The consonants were obtained from the force-alignment of a pseudo-word task, present in the C2SI corpus. A base correlation of $p = 0.82$ was obtained between the predicted and reference intelligibility values, showing a correlation gain of 8% over previous approaches. Moreover, by conditioning the predictions on certain consonants, the correlation increased up to $p = 0.89$. This aspect showed that, depending on the speech impairment experienced by the speaker, there are phonemes that may have a greater or smaller importance in the intelligibility score.

The proposed system also maintains a high degree of interpretability, since the final intelligibility score is a function of the individual scores of each phoneme (consonants). All of the external features used in the conditioning have clinical relevance, and therefore the same level of interpretability is maintained in the post-processing. The high correlations displayed show that the present automatic approach can be used to promote a more interpretable and objective measure, which correlates well with the subjective perceptual measures without requiring a reference intelligibility score for training. The need for more explainable and interpretable systems is highly relevant in a clinical setting.

The results displayed in this chapter were accepted and presented at the conference Interspeech 2022, under the publication name of "Automatic Assessment of Speech Intelligibility using Consonant Similarity for Head and Neck Cancer" [Quintas et al., 2022b].

5.5.2 Perspectives

From the three different granular levels proposed, due to the regression of a score solely based on consonant similarity, we can see that the present predictions could be considered more objective than the others. The correlations obtained, despite being issued on a smaller set of speakers, showed that the amount of similar/dissimilar phonemes is highly correlated to speech intelligibility, showcasing that the phonemic distortion parameter is exceedingly relevant for speech intelligibility. This aspect comes as no surprise due to previous studies that displayed both the relevance of consonants, and also the importance of phonemic distortions to the overall perceptual intelligibility score [Fort et al., 2015, Crevier-Buchman et al., 2002, Nuffelen et al., 2008, Saravanan et al., 2016], when compared to other parameters (e.g. prosody, voice quality, resonance, speech disorder severity, etc.) [Balaguer et al., 2021].

The root mean squared error values found greatly differ from the ones found in the previous chapter, showing a larger error when compared to the other granular levels. Despite this, it is important to state that at the present level of granularity, the system developed was the only one not trained nor optimized on the perceptual intelligibility reference scores, either the clinical evaluation (INT) or the perceived phonological deviation (PPD). By not giving any reference intelligibility score to the system, it is even more surprising that we can find such high base correlations (ρ=0.82) when developing an intelligibility score based on similar/dissimilar phonemes. This aspect shows that we can automatically find an objective and explainable way to predict a highly subjective measure, avoiding the black-box paradigm associated with deep learning. In contrast, there are also other relevant factors that affect speech intelligibility, such as the perceptual parameters aforementioned (voice, prosody, etc.). The study of the different levels of relevance that these parameters have can help to demystify even more the subjectivity associated to speech intelligibility. This aspect served as an interesting lead that will be explored thoroughly in the next chapter.

Key Takeaways

1. We can predict speech intelligibility at phoneme level using the **Phoneme-SN System**. This system can be trained without using perceptive measures as ground truth.
2. We can obtain a high base correlation when predicting speech intelligibility using the proposed methodology.
3. Our approach promotes a more interpretable measure when compared to other automatic systems, since the intelligibility score is a direct function of the amount of similar/dissimilar phonemes.
4. For each speaker, different phonemes have different levels of relevance for the perceptual and automatic speech intelligibility.
5. We can further condition the system on features that have clinical pertinence, and obtain a significant increase in correlation.

Chapter 6

Towards Modelling Perceptual Judges

6.1 Why Individual Judge Modelling?

In the previous chapters, we proposed different ways to predict speech intelligibility. These ways made use of different deep learning paradigms, as well as different speech tasks and reference labels. The approaches also aimed to tackle different problems that arise when using data-driven approaches, such as **data scarcity**, illustrated by the data augmentation/segmentation schemes proposed at sentence level (chapter 3), the **explainability** of the promoted score at phoneme level (see chapter 5), and also the **reliability** of a deep learning system when using smaller subsets of data, represented by the pseudo-word reduction (chapter 4). All of those systems worked as a single judge, that promoted a single intelligibility measure assumed to be the ground truth. This ground truth corresponds to the mean of the individual predictions of the judges used, which could also be called the gold standard. While gold standard prediction is a tremendously common practice when modelling subjective measures, since it can be seen as the best way to evaluate a system's performance, it normally leaves individual judge predictions out of the equation. These predictions can provide further context to our understanding of speech intelligibility, either perceptually or automatically obtained, and that is the topic that we will dive into during the course of this chapter: individual judge modelling.

There are several motivations behind individual judge modelling as a source of uncertainty modelling. First and foremost, this type of methodology can be helpful during clinical evaluations. By creating a variety of automatic judges instead of a general one, many different opinions can be modelled, which can provide more insight than a simple intelligibility prediction. Secondly, this type of system can help understanding how different judges behave either perceptually or automatically, which can be seen as valuable information in order to promote more objective measures. Finally, individual judge modelling can also help finding the difference between professional and novice raters, to which an added degree of relevance can be added to the professional raters, in order to compensate the lack of experience of the novice judges.

6.1.1 The Robustness of Perceptual Measures

In this chapter, we will explore judge modelling as a way to model the variability associated to the perceptual measures, always within the context of speech intelligibility. Given this, it becomes relevant to briefly talk about the concept of hard and soft labels [Galstyan and Cohen, 2007]. A hard label is a label assigned to a member of a class where membership is binary. A soft label, on the other hand, is one which has a score (probability or likelihood) attached to it. So the element is a member of the class in question with a probability/likelihood score. Given these two definitions, it becomes easy to see that these notions are usually adapted towards classification tasks, however a similar notion becomes relevant for the task of automatic prediction of speech intelligibility, specially since there is a difference between the individual prediction of each judge and the mean ensemble of intelligibility predictions. Given that intra-rater reliability is one of the many issues that arises when assessing measures that are subjective by nature, individual judge measures can be seen as less robust than the mean of the different predictions.

We will refer to this mean as the gold standard. Furthermore, the standard deviation of this same mean provides a confidence interval that can be interpreted as the level of robustness of the given measure: a smaller standard deviation is associated to a higher inter-judge agreement, and therefore a more robust measure.

Before evaluating speech intelligibility, the six judges of the C2SI corpus were asked to evaluate four perceptual parameters: voice quality, resonance, prosody and phonemic distortions, used to help them assessing intelligibility. These parameters were assessed by each judge before the assessment of speech intelligibility, the protocol can be found described in section 2.3. It is highly hypothesized that these parameters are linked to speech intelligibility, however, subjectivity comes into play on the extent of their connection. Different judges can have different opinions concerning the levels of relevance of each parameter, and therefore create agreement or lack of agreement in some cases, which is translated into high standard deviations, observable in the gold standard speech intelligibility. A study performed by [Balaguer et al., 2021] conducted an evaluation on the relationship of these four parameters with speech disorder severity. The results suggested different levels of relevance depending on the judge analyzed. Since speech intelligibility and speech disorder severity are highly correlated (see table 2.1), the same evaluation conducted this time on the intelligibility could provide an interesting insight. While on the previous chapters we were more concentrated towards achieving an optimal performance on the automatic prediction of speech intelligibility, in the present chapter we intend to look more under the hood, in order to better understand what makes speech intelligibility subjective and to find a different way to model the associated variability through individual judge modelling. This analysis results directly in the activity of modelling uncertainty, since we are no longer assuming a simple golden standard prediction.

6.1.2 The Relevance of Modelling Uncertainty

Recent works show different schools of thought concerning Machine Learning Under Subjectivity Uncertainty [Rizos and Schuller, 2020]. From assuming hard labels to learn under subjectivity, there are many ways to tackle areas where rater disagreement can be seen as noise or bad data. While the premise of using hard labels may not always be possible, especially in the case of highly subjective measures, using the characteristic uncertainty of these measures as a source of information can be seen as an interesting alternative. In the previous chapters, we have explored briefly this topic when possible (see section 4.5.3). However a deeper analysis on this subject becomes of high interest for the ongoing work. Given that in the C2SI corpus we have access to each individual rating of each judge (see section 2.3), we have the possibility to explicitly model each judge individually, and study the resulting predictions.

Individual judge modelling can be seen as an interesting approach when there is a correspondence between raters and labels. Different methods propose model parameter estimation and rater trustworthiness in a joint manner [Tschiatschek et al., 2018, Raykar et al., 2010, Chou and Lee, 2019, Yan et al., 2010, Rodrigues and Pereira, 2019, Morales-Álvarez et al., 2019, Guan et al., 2018]. Schools of thought can also be modelled when the number of raters is exceedingly high, and an individual modelling becomes unfeasible [Tian and Zhu, 2012]. Using separate models for explicitly modelling a rater has been used for machine translation [Cohn and Specia, 2013] and emotion recognition [Fayek et al., 2016]. In these cases, a hard label is predicted for each model and then all predictions are merged together to form a soft label. This aspect takes in consideration that there may be raters that are better at annotating certain types of data than others.

6.1.3 Modelling Individual Judge Profiles

Given that in the context of the C2SI corpus, more specifically the assessment of intelligibility based on picture description, we have access to the individual predictions of each perceptual judge, their respective modelling becomes a possibility. Hence, we set the following objectives:

- Study the relationship between the four aforementioned perceptual parameters and speech intelligibility at judge level: Creation of individual **perceptual judge profiles**.
- Create an array of automatic systems that model each perceptual judge separately: Creation of individual **automatic judge profiles**.

- Compare both sets of judge profiles obtained, perceptual and automatic.
- (Secondary) Introduce a system that is more objective and interpretable.

The methodology and experiments in this chapter can be found divided in two major groups. The first group presents the statistical analysis that was performed on the perceptual measures. Here we can find different statistical approaches that assess the relationship between intelligibility and the parameters of voice quality, resonance, prosody and phonemic distortions. In the second group, we present an automatic modelling based on the *x-vector* paradigm, similarly to chapter 3, however, this time we want to model each judge individually and predict the four isolated perceptual parameters instead of the intelligibility. Furthermore, similarly to the statistical analysis, we want to perform an analysis on the same relationship between intelligibility and the perceptual parameters, using this time the automatic measures. Figure 6.1 presents a high-level view of the methodology to be adopted during this chapter.

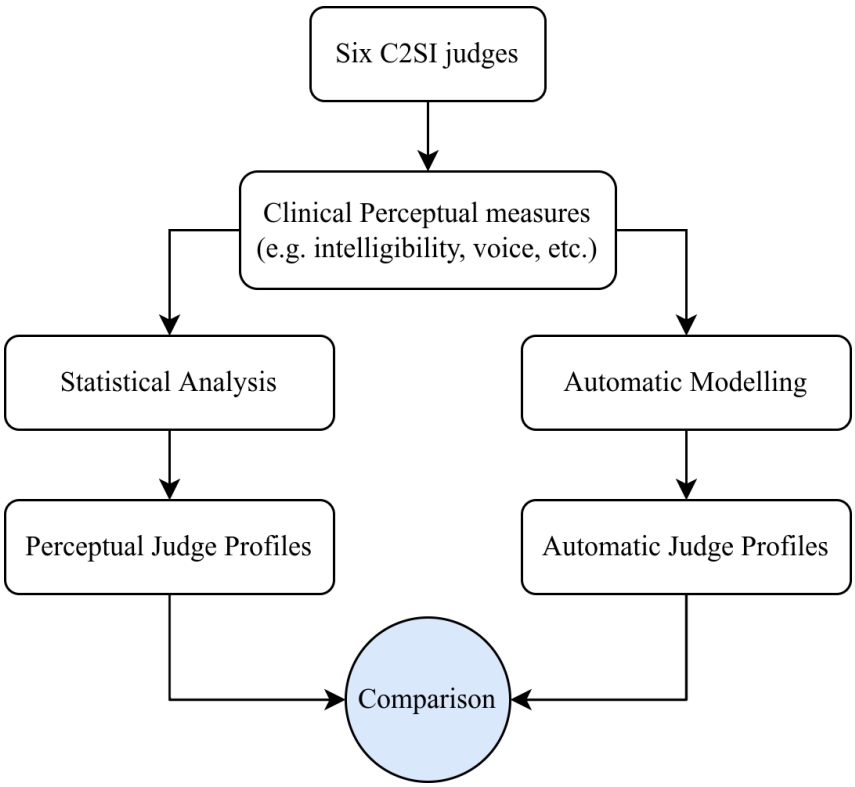


Figure 6.1: Illustration of the proposed methodology to be developed during the course of this chapter. From the clinical perceptual measures obtained from the six judges of the C2SI corpus we will perform a statistical analysis to identify perceptual judge profiles. Furthermore, an automatic model will be fit for each individual judge in order to obtain the same judge profiles, but from our automatic model. Finally, we will perform a comparison between the two. This comparison and following analysis is the main objective of the present chapter.

The rest of this chapter is organized as follows. Section 6.2 describes the aforementioned statistical analysis, with the respective methodology and experiments. Consequently, section 6.3 presents our automatic modelling applied individually to each judge of the C2SI corpus, divided in the methodology and experiments respectively. Section 6.4 proposes a discussion, where we will compare the two sets of experiments performed. Finally, section 6.5 presents our main conclusions and perspectives on the work performed during this chapter.

6.2 A Statistical Analysis Applied to the C2SI Judges

In the present work, we aim to study how we can reliably do the modelling of perceptual judges for the task of automatic prediction of speech intelligibility. Given this, it becomes relevant to understand how individual judge profiles vary across different listeners and how the perceptual evaluation of speech intelligibility is taken in consideration by different judges. Before proceeding to the automatic analysis, it becomes relevant to have an in-depth understanding of the variability behind the perceptual measures. We hope to show how this variability is manifested through the means of the following statistical analysis.

6.2.1 Methodology

In order to find the most relevant perceptual parameters for speech intelligibility, three statistical methods were employed. The first statistical method studied the inter-rater agreement on the perceptual measures and intelligibility via the interclass correlation coefficient (ICC) [Bartko, 1966]. This coefficient is one of the few measures used to evaluate intra-rater reliability [Gwet, 2008]. In the present work, we can only make use of this coefficient by applying directly to the perceptual measures, that have different observations of the same event rated by a set of judges. Since there was only one set of observations (ratings) per judge, the ICC cannot be computed at judge level. This coefficient will give us a measure that displays the level of agreement that the different judges have on different parameters. The second method, inspired by [Balaguer et al., 2021], was a multivariate analysis [Good, 2005] that uses robust linear regression analyses as a top-down variable selection approach, where the explanatory perceptual parameters were chosen based on the significance of their coefficient (using the p-value and the 95% confidence interval). Finally, a grid search system [Álvaro Barbero Jiménez et al., 2007] was also employed that searches for the optimal combination of weights given a set of constraints. Equation 6.1 illustrates the way intelligibility (**INT**) can be regressed, where each weight $(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ is associated to one of the four perceptual parameters aforementioned (**V**oice quality disorder, **R**esonance, **P**rosody and **P**honemic **D**istortions). The grid search is optimized according to the reference (perceptual) speech intelligibility.

$$\mathrm{INT}_{Jn} = 10 - \frac{5}{6} * (\sigma_1 \mathrm{V}_{Jn} + \sigma_2 \mathrm{R}_{Jn} + \sigma_3 \mathrm{P}_{Jn} + \sigma_4 \mathrm{PD}_{Jn}) \tag{6.1}$$

Due to the nature of the four perceptual parameters, that have a different and inverted scale when compared to speech intelligibility (3-0 opposed to 0-10, see section 2.3), the sum of the parameters and respective weights are multiplied by a $\frac{5}{6}$ coefficient and then subtracted by 10, so the resulting values are scaled accordingly.

6.2.2 Statistical Analysis Experiments

For the statistical analysis, no prior data preparation is required, since the proposed analysis will rely only on the perceptual measures obtained clinically. Given that we are evaluating the relationship between four perceptual parameters and speech intelligibility, no controls were used in this part of the study. In this case, the usage of controls could provide a biased look on the same relationship, since they all achieved similar values either on intelligibility and on the remaining parameters.

As it was stated before in section 6.1.3, before proceeding to the evaluation of speech intelligibility, each judge was asked to assess four perceptual parameters: voice quality, resonance, prosody and phonemic distortions. In this section we aim to study the relationship behind these four perceptual measures and speech intelligibility, that due to the subjective nature of perceptual measures, it is hypothesized that will vary across different judges. In order to study this relationship, an **inter-rater agreement** assessment, a **multivariate analysis** and a **grid search** approach were used.

We start this statistical analysis with an **inter-rater agreement** measure. This analysis will give us a relevant insight on the measures that share higher/lower levels of agreement among the different judges. An intraclass correlation (ICC) coefficient was computed [Bartko, 1966]. A low ICC (closer to 0) means a low agreement among the judges on a specific parameter, while a high agreement (closer to 1) means that tends to be agreement on that measure by the set of judges [Mukaka, 2012]. The results can be found on table 6.1. The ICCs obtained for prosody and voice quality are low, the value for resonance

is acceptable, and only the ICC for phonemic distortions is considered good. The highest ICC is found for the speech intelligibility ratings.

Table 6.1: Intraclass correlation coefficient (ICC) on the perceptual parameters analyzed. A low ICC means a low agreement between the six perceptual judges on a given parameter. The 95% confidence intervals are displayed as well.

Perceptual Parameter	ICC	95% Confidence Interval
Voice Quality	0.37	[0.27, 0.47]
Resonance	0.42	[0.32, 0.52]
Prosody	0.30	[0.20, 0.39]
Phonemic Distortions	0.67	[0.59, 0.75]
Intelligibility	0.78	[0.76, 0.80]

The second method implemented was a robust **multivariate linear regression analysis**. This approach was conducted by listener, with the speech intelligibility as the dependent variable and the four perceptual parameters as explanatory variables. The Spearman's correlation coefficient was used due to a non-parametric distribution of variables. The results from this analysis can be found on table 6.2.

From the results presented in this multivariate analysis, we can see that the voice quality parameter achieved fairly low correlations with the intelligibility on all six listeners, meaning that it was not largely taken in consideration by all judges when evaluating speech intelligibility. Two listeners (2 and 4 respectively) only took phonemic distortions in consideration. This perceptual parameter was also the one that achieved the highest correlation values among all listeners. A high correlation between speech intelligibility and phonation ability is not a surprise, and is coherent with previous studies found in the literature [Saravanan et al., 2016, Nuffelen et al., 2008]. Three judges (2, 3 and 6) consider prosody and phonemic distortions, while only the first listener considered three parameters. The most important parameters considered by each judge are marked in bold.

For the last method used to evaluate the relationship between the four perceptual parameters and speech intelligibility, a **grid search** approach was devised. A weight was multiplied by each perceptual parameter, varying between 0.0 and 1.0, with an increment of 0.05. The sum of all weights used in the grid search always adds to 1.0 (see equation 6.1). The weights were chosen according to the values that maximize the correlation between the perceptual intelligibility and the output of equation 6.1.

Table 6.2: Significant parameters explaining speech intelligibility, by listener (robust linear regression analysis and Grid Search Approach)

Judge	Considerated Parameters from the Multivariate Analysis of the Perceptual Measures				Constrained Grid Search Approach (weights between [0.0,1.0])				General Grid Search Approach (weights between [0.0,1.0])			
	V	R	P	PD	V	R	P	PD	V	R	P	PD
1	0.14	**0.22**	**0.18**	**0.52**	—	0.30	0.30	0.40	0.10	0.20	0.30	0.40
2	0.08	0.03	**0.40**	**0.61**	—	—	0.33	0.66	0.10	0.0	0.30	0.60
3	0.06	0.01	**0.29**	**0.54**	—	—	0.10	0.90	0.0	0.0	0.10	0.90
4	0.10	0.13	0.05	**0.82**	—	—	—	1.0	0.20	0.20	0.0	0.60
5	0.0	0.09	0.17	**0.56**	—	—	—	1.0	0.0	0.0	0.0	1.0
6	0.02	0.13	**0.43**	**0.38**	—	—	0.40	0.60	0.0	0.20	0.30	0.50

The results, also illustrated on table 6.2, show a similar pattern to the one found previously on the multivariate analysis. Judge 1 displayed a similar weight distribution to the correlations found in the multivariate analysis. The same aspect can be said for judges 2 and 3, whose grid search weights showed a clear tendency towards prosody and phonemic distortions, as opposed to voice quality and

resonance. For judges 4 and 5, the same tendency towards phonemic distortions as the most relevant parameter can be seen as well. However, the grid search showed that, for judge 4, the consideration of voice quality and resonance would achieve an optimal correlation. Despite the small correlation achieved in the multivariate analysis by these two parameters for the same judge, the values can be seen as non-negligible, and therefore slightly relevant. Judge 6 was the judge that gave the least emphasis to phonemic distortions. Although this aspect is not found on the grid search approach (a weight of 0.5 associated to PD), the weight distribution was similar to the multivariate analysis, with a slight shift in consideration between prosody and phonemic distortions.

A second grid search approach was devised, this time, however, conditioned by the results of the multivariate analysis. This means that at each time, a set of weights is fit only to the considered parameters (bold values on table 6.2). The main motivation behind this second analysis is to investigate how the weight distribution of the grid search analysis would behave if a set of constraints was introduced. The results (also present on table 6.2) show a clear preference for phonemic distortions as the most considerate parameter. No aberrant values were found in the weights, which proves a certain level of robustness of the two different approaches.

Relationship of the Four Perceptual Measures With Intelligibility

Given the previous analysis performed on the most relevant parameters for each judge, it becomes relevant to see the correlations promoted by each approach. Moreover, we want to assess how the different perceptual parameters as well as the intelligibility based on the multivariate analysis/grid search correlate with the perceptual intelligibility. Seven different correlations are proposed, between the assessed parameter/measure and the perceptual intelligibility given by each judge. The results from this analysis can be found on table 6.3. The first four correspond to the four perceptual parameters previously assessed (voice quality, resonance, prosody and phonemic distortions). Afterwards, the "uniform weights" approach promotes the correlation between the perceptual intelligibility and the results from equation 6.1 in the specific case of $\sigma_{1,2,3,4} = 0.25$. Finally, the last two columns correspond to the results of equation 6.1 given the optimal weights for both grid searches (see Constrained Grid Search and General Grid Search from table 6.2).

Table 6.3: Correlation results between the four perceptual parameters and the perceptual intelligibility.

Judge	Correlation Values (Spearman's ρ) - Perceptual Judges						
	V	R	P	PD	INT (equation 6.1)		
					Uniform Weights	Grid Search Constrained Weights	Grid Search Optimal Weights
1	0.492	0.561	0.432	0.748	0.828	0.839	0.851
2	0.249	0.559	0.627	0.807	0.774	0.849	0.852
3	0.533	0.617	0.738	0.836	0.836	0.879	0.879
4	0.178	0.558	0.020	0.813	0.767	0.813	0.863
5	0.212	0.353	0.323	0.784	0.648	0.784	0.784
6	0.222	0.575	0.620	0.732	0.750	0.792	0.810

6.2.3 Results Analysis - Perceptual Measures

From table 6.3, we can see that for the four perceptual parameters (V, R, P and PD) the results suggest a clear higher correlation for the phonemic distortions parameter, which is coherent with the results from the multivariate analysis and both grid searches. The voice quality parameter was also coherent, displaying the lowest correlations as expected. The resonance parameter was fairly uniform across all six judges with the exception of judge 5 that presented a lower correlation. The prosody parameter was the most erratic one, displaying correlation values as high as 0.738 (judge 3) and as low as 0.020 (judge 4). Consequently, we can conclude that the prosody parameter was the one that achieved the

least consensus among all judges in relation to the perceptual intelligibility, followed closely by the voice quality parameter. These conclusions are consistent with the ICC values found in table 6.1.

In the case of the uniform weights, for judges 1, 3 and 6 this score promoted a better correlation to speech intelligibility than any one of the individual four perceptual parameters. In the case of the remaining judges, the phonemic distortion parameter displayed a higher correlation than the score based on the uniform weights, which is coherent with the weights of the grid search and the results of the multivariate analysis (table 6.2) that gave a larger emphasis on this same perceptual parameter. All the remaining analysis (constrained and general grid searches) displayed similar correlations in an increasing way. Since the general grid search approach was optimized on the correlation with the perceptual intelligibility, it was expected that the correlation values will be the highest for that particular approach, with the second grid search (multivariate analysis with optimal weights) being a close second. The correlation gains between these two analyses were relatively small (around 1-2%), with the exception of judge 4 that displayed a correlation gain of 5%. This was expected since the grid search weights for this judge were the ones that changed the most when compared to the multivariate analysis. This aspect points out the importance of performing more than a single analysis on the relationship between the four perceptual parameters and the intelligibility. By doing this, we can search for key patterns and relationships that can help promote a better explainability of the reasoning behind the assessment of speech intelligibility.

From the analysis promoted in this section, on the relationship between the different perceptual parameters evaluated by a set of judges, a profile of each judge can be traced. These profiles not only showcase how variable speech intelligibility ratings can be, but also give us a powerful insight on how we can automatically model a judge. This aspect becomes crucial for the development of an explainable automatic system that operates more according to a human evaluation instead of the black-box paradigm, typically associated to deep learning approaches. The added degree of explainability becomes highly relevant in a clinical context [G.McCoy et al., 2021]. Despite the different levels of relevance found among the judges for each perceptual parameter, the ICC value for speech intelligibility was the highest of all parameters, showing a certain level of consensus despite the different approaches.

6.3 Automatic Modelling of the C2SI Judges

The second set of experiments present in this chapter is centered around the automatic prediction of speech intelligibility, and the four linked perceptual parameters. Similarly to subsection 4.5.3, we will investigate how our automatic intelligibility prediction works when modelling uncertainty, however, this time by modelling each judge individually instead of using uncertainty during training.

6.3.1 Methodology

In order to model the behavior of each perceptual judge, an ensemble of automatic models was created. This set predicts speech intelligibility at judge level and later on merges the individual predictions of each judge to obtain the final intelligibility score for each speaker. Figure 6.2 displays a global overview of the proposed system.

The system was based on the work developed during chapter 3, the Sentence-XVec system, that made use of the *x-vector* speaker embedding paradigm as features [Snyder et al., 2018b, Quintas et al., 2020], extracted using the Kaldi toolkit [Povey et al., 2011]. A shallow neural network was modelled to each judge, the dimensions can be found on table 6.4. Figure 6.3 presents an illustration of the previous table, that corresponds to each individual judge model.

The loss function used takes in consideration all of the four perceptual parameters of each judge equally (V, R, P and PD), containing a total of 24 parameters to optimize (six judges times four perceptual parameters), following a multi-task learning methodology [Zhang and Yang, 2018]. Equation 6.2 illustrates the proposed loss function.

$$L_{total}(y_v, ..., \hat{y_{pd}}) = L_v(y_v, \hat{y_v}) + L_r(y_r, \hat{y_r}) + L_p(y_p, \hat{y_p}) + L_{pd}(y_{pd}, \hat{y_{pd}}) \tag{6.2}$$

By using a multi-task learning methodology, we hope to model the four perceptual parameters instead of simply regressing the intelligibility score. Moreover, by also modelling each judge individually, we

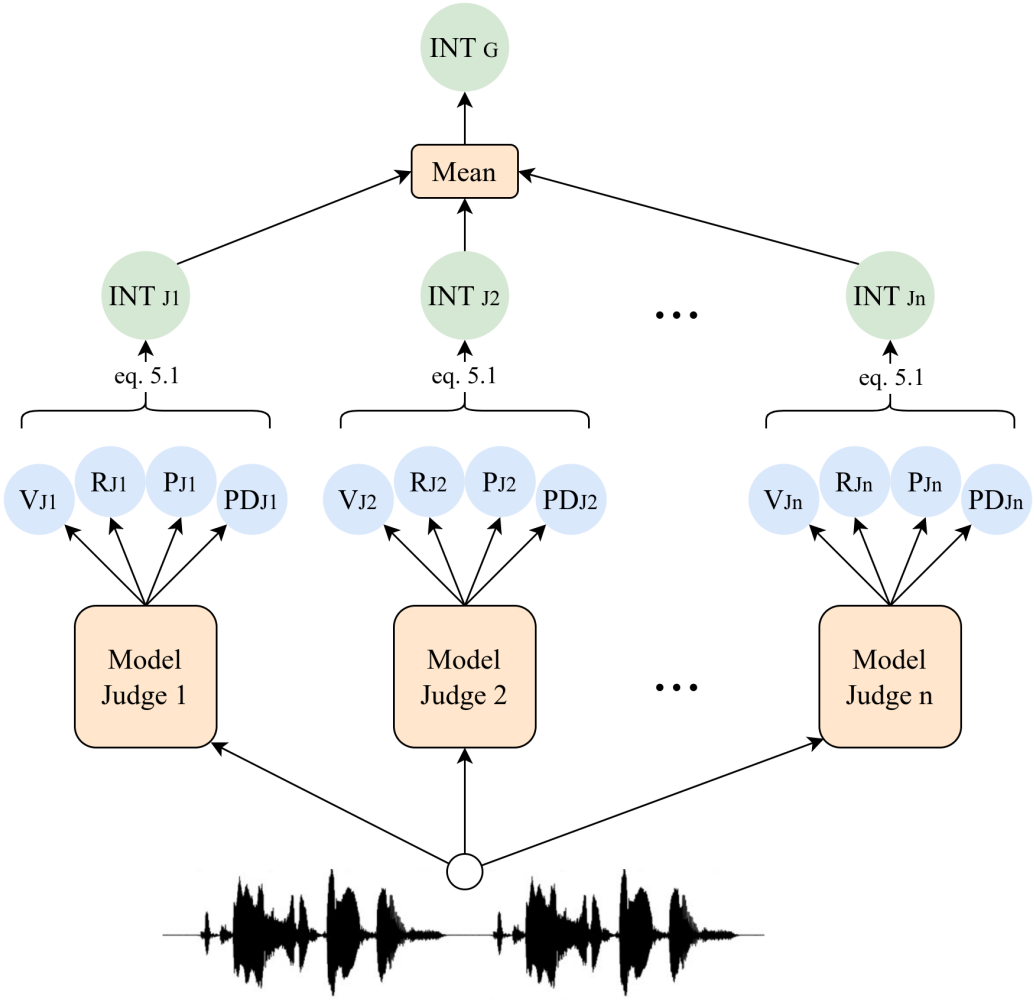


Figure 6.2: Global Overview of the proposed system. The x-vectors are extracted from the segmented parts of a reading passage task (LEC), and then fed to a individual shallow neural networks that model each perceptual judge. Equation 1, previously introduced, merges the individual prediction of the four parameters.

Table 6.4: Shallow neural network outline used for each automatic judge, "fc" stands for fully-connected. The first two layers correspond to the shared layers of each automatic judge while the remaining four correspond to the individual prediction of each one of the four perceptual parameters.

	Layer	**Input x Output**
shared layers	*fc-1*	512×128
	fc-2	128×64
prediction layers	*fc: Voice Quality*	64×1
	fc: Resonance	64×1
	fc: Prosody	64×1
	fc: Phonemic Distortions	64×1

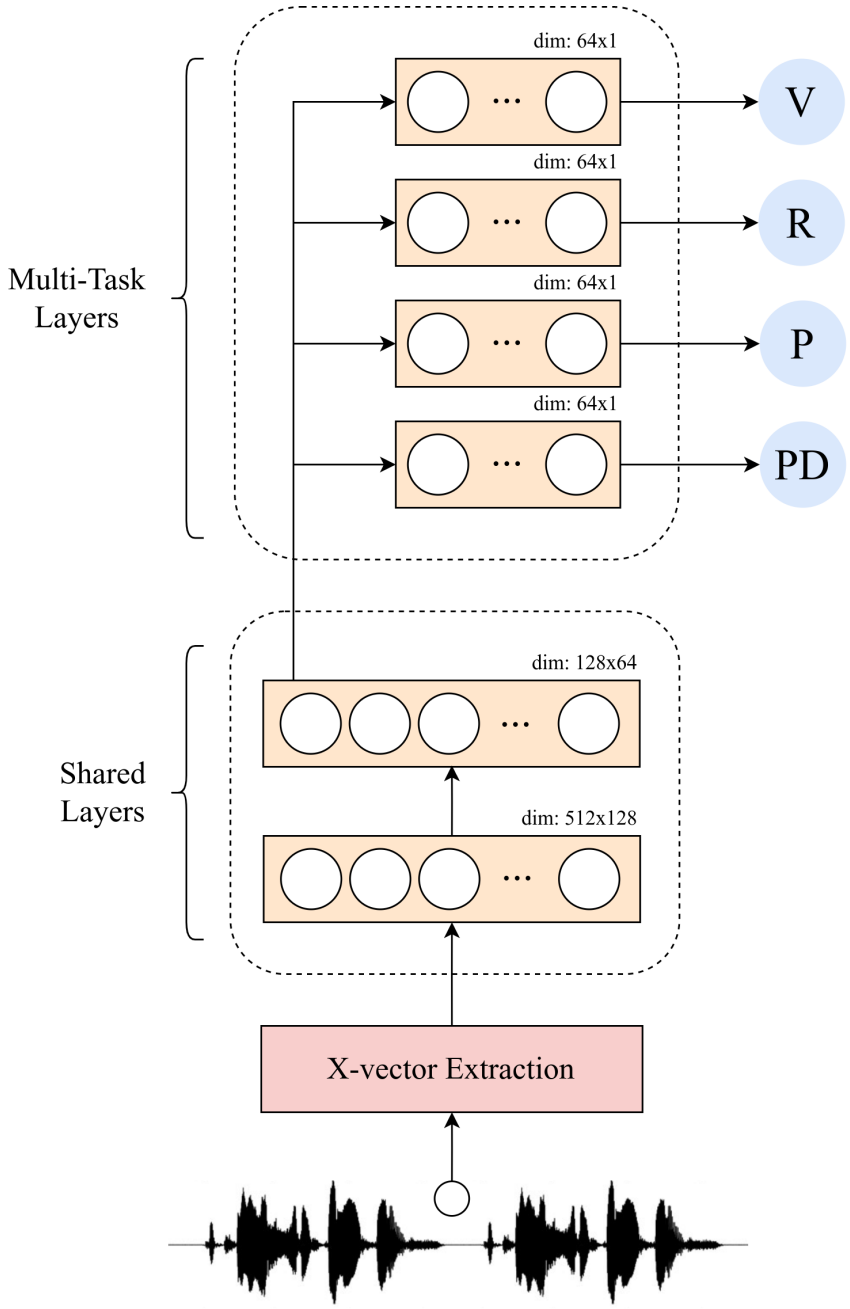


Figure 6.3: Illustration of a model corresponding to one of the modeled judges. This diagram corresponds to a "Model Judge" box displayed on figure 6.2.

can obtain a variety of measures that directly impact the gold standard, providing an added degree of explainability to the final intelligibility estimation.

The methodology present in this chapter for the automatic part is fairly similar to chapter 3, however, instead of having one general model that predicts speech intelligibility (gold standard), we have one model per judge that predicts four perceptual parameters linked to speech intelligibility.

6.3.2 Experiments with Individual Judge Modelling

The second part of our experiments corresponds to the modelling of the same set of judges used to assess speech intelligibility. In this section we aim to show how we built our automatic model as well as to demonstrate why the measures obtained from an automatic approach can be seen as more objective.

For the automatic modelling of each judge, similarly to chapter 3, we made use of the reading passage task recorded in the context of the C2SI corpus [Woisard et al., 2020]: *La chèvre de Mr. Seguin*. The passage reading task was also segmented in eight distinct sentences (see sub-section 3.3.1, and the same data augmentation scheme employed in sub-section 3.2.3 was used during training. In order to compare the results with the statistical analysis, no controls were used. A total of 85 patients were used in the present study.

System Training

As it was stated previously in section 6.3, an individual model was fit for each judge. The target scores used were the individual ratings of each judge on the four perceptual parameters of voice quality, resonance, prosody and phonemic distortions. The set of models was optimized simultaneously. In order to train the proposed system, a 10-fold cross validation scheme was implemented, similarly to chapter 3. The system was trained during 13 epochs, with a batch size of 32. A learning rate of 0.001 was used and the system was optimized with the Adam optimizer algorithm. A dropout of 0.1 was added between consecutive layers.

System Evaluation

Similarly to the previous chapters, the Spearman's correlation coefficient (ρ) and the root mean squared error (RMSE) were chosen to evaluate our system. The system was compared to the approach promoted in chapter 3, which also made use of the *x-vector* speaker embeddings. This approach directly predicted the global speech intelligibility instead of doing an individual judge modelling and subsequent average of the resulting predictions. Controls were used exclusively to be compared to the previous system. The results, present on table 6.5, suggest a high correlation of 0.845 and an error of 1.623. Our model also suggests a correlation gain of near 4% when compared to the previous approach. The corresponding plot can be found on figure 6.4.

Table 6.5: Results achieved by the proposed system and comparison to a baseline. Controls were used here so the system becomes comparable to the previous system developed in chapter 3.

	$\boldsymbol{\rho}$	RMSE	Controls
Single General Model (chapter 3)	0.81	1.728	yes
Individual Judge Modelling	0.85	1.623	
	0.78	1.733	no

Since the system was trained, using the four perceptual measures as a target for each judge, it becomes relevant to assess how the corresponding automatic model of each perceptual judge performs on these perceptual measures. Table 6.6 presents the correlations between each judge's perceptual parameter and the corresponding automatic prediction of that same parameter. From these results, we can see a clear tendency towards the phonemic distortions being the perceptual parameter that achieves higher correlations for each speaker. This can also be interpreted as the easiest parameter to model since during the system's training, all parameters had the same weight in the loss function. From table 6.6, we can also see that the voice quality and prosody parameters were the ones that achieved the lowest correlations. Contrary to the phonemic distortions, these two parameters appear to be the hardest to model in an automatic context. The high gap found between the prosody values of judge 3 and judge 4 can serve as an example of either the variability associated to this particular perceptual measure, or to the low *vs.* high intra-judge reliability on this particular parameter. The same aspect can be concluded for the voice quality parameter of judge 5.

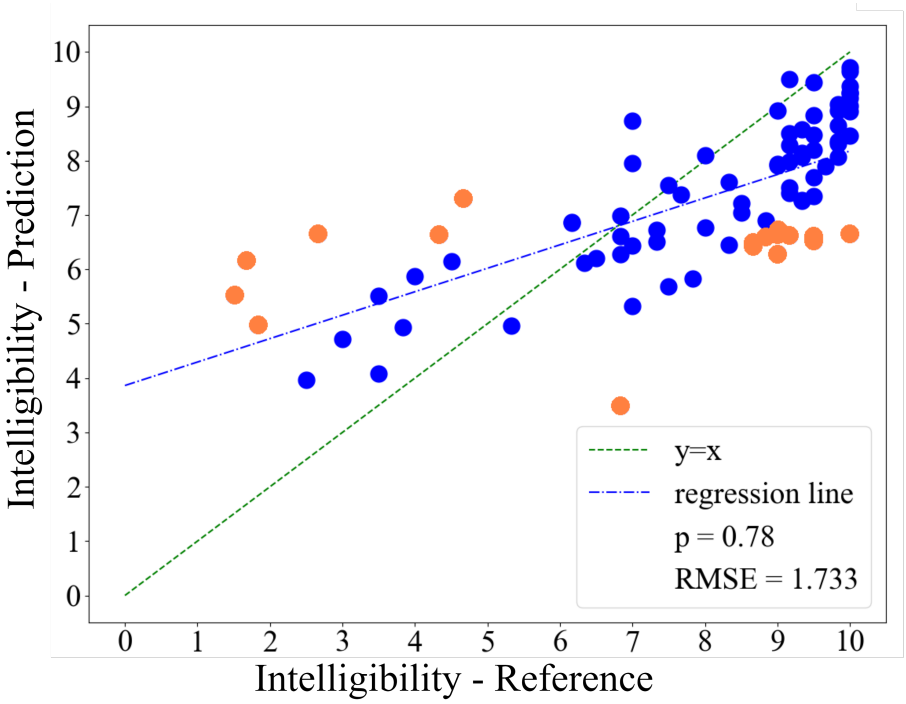


Figure 6.4: Results of the automatic prediction of speech intelligibility using the proposed methodology. Outliers are marked as orange dots.

Table 6.6: Correlation values of each judge's perceptual parameters when compared to the reference values of those same individual parameters.

Judge	Correlation Values (Spearman's ρ)			
	V	R	P	PD
1	0.447	0.639	0.387	0.662
2	0.653	0.379	0.362	0.708
3	0.404	0.622	0.161	0.748
4	0.405	0.612	0.654	0.714
5	0.299	0.563	0.443	0.770
6	0.443	0.533	0.407	0.637
mean	0.441	0.558	0.402	0.707

Relationship of the Four Automatic Measures With Intelligibility

Similarly to subsection 6.2.2, an analysis of the most relevant parameters for speech intelligibility was performed, however, in this case using our set of automatic judges. The results from the multivariate analysis of the perceptual parameters (see table 6.2) were used as a constraint on the automatic measures, in order to see if the same analysis would translate into the automatic domain. The results of the two grid search experiments can be found on table 6.7, along with the results from the multivariate analysis of the perceptual measures.

The correlation results from the automatic prediction of speech intelligibility (see equation 6.1) can be found on table 6.8, where each value corresponds to the correlation between the respective parameter or scoring approach and the reference intelligibility score of each judge.

Table 6.7: Significant parameters explaining speech intelligibility, by listener (robust linear regression analysis and Grid Search Approach).

Judge	Considerated Parameters from the Multivariate Analysis of the Perceptual Measures				Constrained Grid Search Approach (weights between [0.0,1.0])				General Grid Search Approach (weights between [0.0,1.0])			
	V	R	P	PD	V	R	P	PD	V	R	P	PD
1	X	✓	✓	✓	—	0.10	0.50	0.40	0.33	0.0	0.33	0.33
2	X	X	✓	✓	—	—	0.40	0.60	0.20	0.40	0.0	0.40
3	X	X	✓	✓	—	—	0.33	0.66	0.0	0.30	0.50	0.20
4	X	X	X	✓	—	—	—	1.0	0.20	0.10	0.30	0.40
5	X	X	X	✓	—	—	—	1.0	0.0	0.30	0.60	0.10
6	X	X	✓	✓	—	—	0.50	0.50	0.20	0.10	0.40	0.30

Table 6.8: Correlation results from the automatic analysis.

Judge	Correlation Values (Spearman's ρ) - Perceptual Judges						
	V	R	P	PD	INT (eq. 6.1)		
					Uniform Weights	Grid Search Constrained Weights	Grid Search Optimal Weights
1	0.467	0.471	0.451	0.559	0.613	0.616	0.630
2	0.269	0.678	0.575	0.692	0.699	0.700	0.728
3	0.439	0.604	0.555	0.609	0.613	0.623	0.633
4	0.291	0.599	0.349	0.679	0.726	0.679	0.742
5	0.279	0.476	0.387	0.452	0.490	0.452	0.539
6	0.269	0.574	0.513	0.585	0.637	0.627	0.649

6.3.3 An Analysis of the Automatic Measures Obtained

From the weight distribution present on table 6.7, we want to evaluate the weight difference between the approach conditioned by the multivariate analysis and the grid search. Moreover, we want to assess how those different weights hold in terms of correlation results (table 6.8). If, for example, there is only a minor correlation gain between the grid search for the constrained parameters and the general grid search, it could mean that the modelling of a certain judge was behaving very closely to the perceptual case. The results were also compared to a "uniform weight" intelligibility prediction, where all perceptual parameters shared the same weights.

Out of the four perceptual parameters correlated with the perceptual intelligibility (first 4 columns of table 6.8), the voice quality parameter was the one that had the lowest correlation on all six judges, except for judge 1. However, despite being the parameter with the lowest values, the correlations were slightly higher than the ones found on the perceptual results (compared with table 6.3). A similar aspect can also be seen on the prosody parameter, where in the automatic case is more normalized on all judges, when compared to the skewed results of the perceptual case. On the resonance parameter, the results stayed fairly uniform, despite this, judges 2 and 5 displayed larger correlations to speech intelligibility on the automatic setting than on the perceptual case. As expected, the phonemic distortion parameter was the one that showed the highest correlation values with the perceptual intelligibility, however, not as high as the ones witnessed on the perceptual evaluation, and not on judge 5. Given the correlation on all the four perceptual parameters, we can see that, on the automatic analysis, the correlations between the four automatically predicted perceptual parameters are more normalized and less skewed across all judges and parameters when compared to the simple perceptual parameters.

The correlation results from equation 6.1 applied to the automatic case (columns 5, 6 and 7 of table 6.8) behaved differently when compared to the perceptual approach (table 6.3). In the automatic case, a more even distribution of the optimal grid search weights can be found (table 6.7), where all judges

take in consideration at least three parameters. We can also witness a slight shift in the most relevant parameters, while on the perceptual case there was a clear tendency towards phonemic distortions, in the automatic case we can see more emphasis being given to prosody than phonemic distortions on all judges (except judges 2 and 4). The correlation results on the constrained grid search weights (based on the multivariate analysis of the perceptual case) showed a decrease in correlation on judges 1, 4, 5 and 6 when compared to the uniform approach. The correlation gains between the uniform approach and the optimal weights from the grid search are relatively small, except for judge 5, which is also the judge that presented the largest weight for prosody. The small correlation gain between the uniform and optimal weights showcases that in an automatic setting, a more even distribution tends to work better when compared to a more cherry-picking approach of choosing the weights for the parameters whose impact was larger on the perceptual case. Figure 6.5 illustrates a visual comparison between the two sets of judge profiles found (perceptual vs automatic).

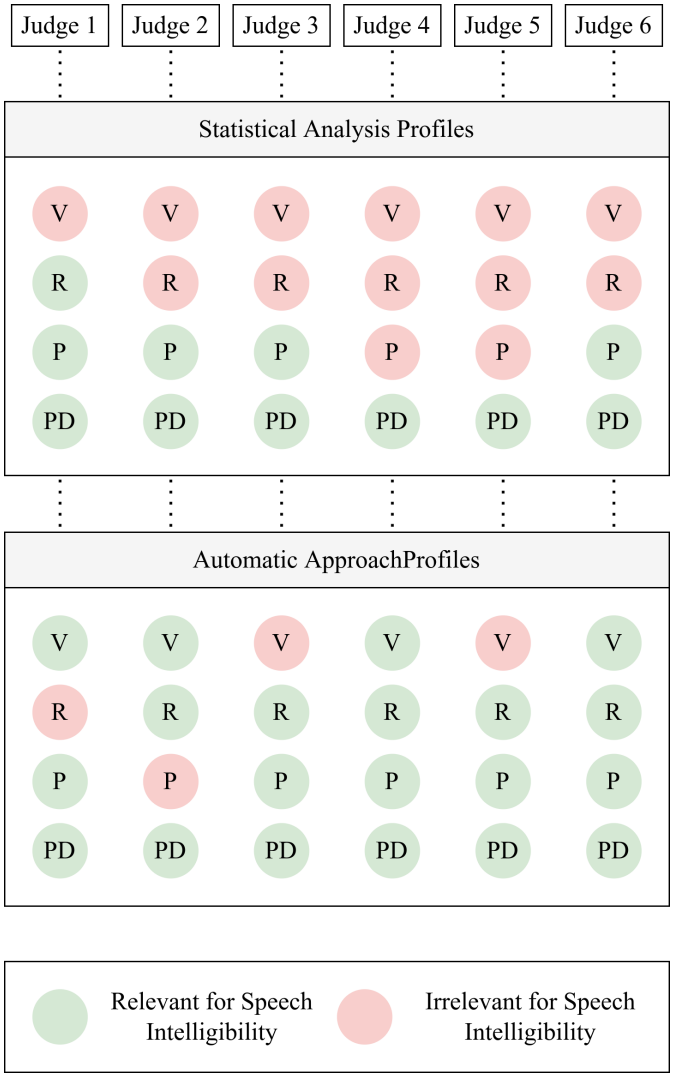


Figure 6.5: Illustration of the judge profiles found, showcasing the relationship between the four perceptual parameters and speech intelligibility on the perceptual case (multivariate analysis) and on the automatic approach (general grid search approach).

6.4 Discussing the Different Judge Profiles Found

In this section we present a detailed discussion of the results obtained. Given the multitude of experiments performed in the present chapter, we will organize it as follows. First we start by discussing the results of the perceptual and automatic analysis of the four perceptual parameters. Second, we will discuss the results of the approaches used to predict speech intelligibility as well as to model each perceptual judge.

6.4.1 Four Parameters: Perceptual *vs.* Automatic

From the results obtained for the parameters of voice quality, resonance, prosody and phonemic distortions in both analyses (perceptual and automatic), we were able to notice interesting differences among these parameters. Starting with the ICC evaluation on the perceptual case, table 6.1 displayed that voice quality and prosody parameters shared a low inter-rater agreement. This could mean a variety of things such as differences in the judge's opinion concerning the affected parameters (subjectivity) or a misinterpretation of the measure before the judges are being asked to assess the speakers. On the other hand, the parameters of resonance and phonemic distortions achieved better levels of agreement among the judges. This could mean that those two measures can be seen as more objective or easier to interpret in a clinical context, which provides an added degree of explainability. The correlations between those four parameters and the intelligibility (table 6.3) show that the phonemic distortion parameter is the one more closely related to speech intelligibility, on the other hand, the second-closest parameter was prosody, which achieved a fairly low level of agreement among judges. This aspect clearly showcases that despite the high ICC values obtained for speech intelligibility, it still remains a measure that is subjective by nature, and that can take several other parameters in consideration depending on the judge.

In the case of the automatic approach, we can also notice that the correlations obtained by each judge on the four individual parameters (table 6.6) closely mimic the results from the ICC analysis, showing that voice quality and prosody achieved the lowest average correlations, while resonance remained slightly better and the phonemic distortions remained the best out of the four. This shows that despite the four parameters of each of the six judges being optimized in the same way (no added relevance to specific parameters during training), there are parameters that are inherently easier to model, and that aspect is directly related to the quality of the measure itself. This means that a measure with a larger inter-judge agreement will be easier to model than a measure with low agreement, especially in data-driven approaches such as the proposed model. This aspect becomes relevant in situations such as the one presented in this work, where we attempt to model and understand the uncertainty of a measure that is subjective by nature.

6.4.2 Speech Intelligibility: Perceptual *vs.* Automatic

From both the multivariate and grid search analysis on the perceptual parameters, we found that both prosody and phonemic distortions played a more relevant level for speech intelligibility than the remaining parameters. When using equation 6.1 to compute the intelligibility score based on the four perceptual parameters, we can see an increment in correlation when adding more relevance to those parameters. Both grid search approaches (table 6.3) promoted correlation gains on all judges when compared to a uniform approach where all parameters have the same relevance. This aspect shows that different profiles emerge with regard to these specific parameters, showcasing the variability and bias associated to these clinical measures. Despite this, speech intelligibility remains fairly homogeneous, showing high levels of agreement between judges when compared to the other parameters.

On the other hand, when modelling those same perceptual judges via our automatic approach, we can see that the profiles that emerge from each judge are different. In this case, the weight distribution for those same four parameters appears to be more uniform for all six judges. This aspect is validated by the correlation results obtained on table 6.8, that show only minor gains between the uniform and the optimal weights. Contrastingly, by constraining the weights on the results of the previous analysis of the perceptual measures, the results tend to deteriorate, showing that the profiling made via the perceptual analysis does not necessarily hold in an automatic setting. This aspect shows that despite the variability witnessed among the different perceptual judge profiles, showed by the different levels of relevance for

each parameter, an automatic approach will promote a more uniform and objective way to predict speech intelligibility, that shares high levels of correlation with the perceptual measures (figure 6.4 and table 6.5).

6.5 Scientific Contributions and Perspectives

6.5.1 Conclusions

In the present chapter, we have conducted the individual modelling of the set of perceptual judges present in the C2SI corpus. Moreover, the relationship between speech intelligibility and the four perceptual parameters of voice quality, resonance, prosody and phonemic distortions were studied at judge level. The results suggest that the four perceptual parameters are highly related to speech intelligibility, however, with different levels of relevance. Furthermore, this aspect favored the appearance of different judge profiles that attribute different levels of relevance to the perceptual parameters. An ensemble of automatic models was proposed to do the modelling of each individual judge. The results of this automatic approach suggest an increase in correlation when compared to the previous system performed in chapter 3. The two types of judge profiles, the ones obtained from the perceptual measures and the automatic ones, were later compared. This analysis suggested that a more uniform judge profile is found among all automatic judges, without large discrepancies on the relevance of each one of the four perceptual parameters. This aspect can be seen as a more objective and less variant prediction. The results from both the perceptual and automatic analysis suggest that the four perceptual parameters are highly intertwined with speech intelligibility.

6.5.2 Perspectives

In the present work, we performed a study on the different profiles that emerge from a set of six perceptual judges, and the corresponding automatic modelling of those same judges. The main focus of attention was set towards the relationship between four perceptual parameters (voice quality, resonance, prosody and phonemic distortions) and speech intelligibility, both rated by the same set of judges. We hypothesized that those parameters were deeply connected to speech intelligibility, however, with different levels of relevance. This aspect was illustrated by the results of the statistical analysis, which display the different degrees of relevance that each judge attributed to each parameter.

The different profiles that emerge from the perceptual analysis display the high variability associated to the perceptual measures, where each judge takes in consideration different parameters before the evaluation of speech intelligibility. This variability is considered one of the main impairments behind the clinical evaluation of speech intelligibility, among with judge bias, variance and lack of agreement. Moreover, we can see that the perceptual and automatic judges follow different approaches concerning the relevance of each perceptual parameter. The results from the automatic approach proposed show that we can predict speech intelligibility in a more objective way, where each automatic judge uses a more uniform approach concerning the relevance of each perceptual parameter. The approach is not only more objective, but also more interpretable, since the intelligibility value can be traced back to each individual prediction given by each automatic judge, and each judge's prediction can be derived from the four perceptual parameters evaluated.

Despite the different levels of relevance used in the automatic approach when compared to the perceptual case, a high correlation was achieved with the proposed automatic model on the global intelligibility ($\rho = 0.845$), showing that despite using the perceptual measures as a reference, an automatic approach can follow a different weight distribution, in this case following a more uniform trend. Both perceptual and automatic analysis provided insight on how both types of judge perform the assessment of speech intelligibility, which becomes highly relevant for subjective measures in a clinical context. A study on the direct impacts that a hybrid approach can have on the prediction of speech intelligibility (a set of automatic and perceptual judges) is an interesting lead for future work, as well as exploring the relevance of more objective automatic parameters [Eyben et al., 2015] on both perceptual and automatic scores, as opposed to the four parameters studied during the course of this work.

A posterior evaluation of the reliability of each individual perceptual judge becomes of high interest,

however, in the present work it was unfeasible due to the fact that each judge only rated the corpus once. By having a single corpus rating per judge, the ICC coefficient cannot be calculated, and therefore no measure of intra-rater reliability could be extracted for each judge. Several corpus ratings by the same set of judges, despite impractical and time-consuming, could provide the tools to calculate the reliability of each individual judge. With this knowledge, the automatic system could be made accordingly, by providing more relevance towards reliable judges than non-reliable ones. On the perceptual side, more effort could be given towards understanding the schools of thought behind the ratings of reliable judges than the contrary, which could help to promote more objective and explainable measures as well.

Key Takeaways

1. Different judge profiles arise from the modelling of each perceptual judge.
2. The four perceptual parameters play a different role for speech intelligibility in each individual judge profile.
3. An automatic modelling of each judge is possible, and displays more uniform relevance levels across the four perceptual parameters.
4. The approach promoted, based on the independent evaluation of different automatic models that mimic each judge, can be seen as more objective.

Chapter 7

A Unified Automatic Intelligibility Score

7.1 Working Context and Hypotheses

In the previous chapters, we made use of different levels of granularity to automatically predict speech intelligibility. From sentence to phoneme level, high correlations were achieved and different conclusions were drawn. These conclusions led to a better understanding behind the automatic modelization of speech intelligibility, and in some cases led to interesting post-processing ideas that generally increased performance. Given this, the idea of aggregating each system's individual predictions into a single, universal automatic intelligibility score becomes hugely appealing, specially in the advent of obtaining a high-performance system.

Regardless of the high correlations achieved, each granular system had its respective shortcomings. Starting at sentence level, two methodologies were proposed, on chapters 3 and 6. Both methodologies were able to achieve high correlations and decent error values. However both systems were underperforming on patients with severe speech impairments (mainly due to the lack of data of this type). At word level, an extremely high base correlation was achieved, and almost no outliers or aberrant values were found. Despite this, the measure used was based on the perceived phonological deviation, obtained by the transcription of naive judges. Although this measure is valid and highly valuable in a clinical environment, to assess for example day-to-day communication, it serves a different purpose than the perceptual speech intelligibility evaluated by the set of professional judges. Finally, at phoneme level, with the post-processing of the original system, a high correlation was also achieved, however, this time with a smaller set of speakers and an RMSE value larger than the remaining systems. Given that each granularity level adds something to the table, it becomes highly relevant to promote an unified intelligibility score, that merges the individual predictions obtained at each individual level.

Different approaches can be found in the literature concerning merging individual systems to obtain more accurate predictions/classifications. In [der Burgh et al., 2017], the authors made use of three different automatic measures, obtained from distinct deep learning models, to predict the survival class (short, medium or long) of amyotrophic lateral sclerosis patients. Each distinct measure, based on clinical characteristics, brain morphology and structural connectivity respectively, was fed directly to a classifier that combined these three information sources. This methodology not only promoted higher accuracy, but also interesting insight on how the relationship between the information sources and the disease prognostic works. The work of [Badgeley et al., 2019] showed that a combined prediction of image processing plus patient data and hospital process features significantly increased the detection ability of hip fractures, when compared to using only image processing. This work again validates the fact that combined approaches to a specific problem can indeed promote better results. Other merging approaches were already implemented during the course of this thesis. In chapter 6, the gold standard prediction (as in global intelligibility) is based on the mean prediction of each individual automatic judge, which followed a multi-task learning [Zhang and Yang, 2018] methodology. On chapter 5, the post-processing

implemented can also be seen as a merging approach that takes in consideration a set of external features to better predict the final intelligibility score.

In the present chapter, we aim to promote an unified score based on the granular systems developed during the course of this thesis. The concept of unifying different automatic predictions comes as a logical step towards promoting more robust and precise predictions. In addition, by combining different predictions in a single system, a more interpretable and explainable measure can also be obtained. The added degree of interpretability can be seen as a combination of the different granular scores obtained. The individual score at phoneme, word and sentence can help justifying the automatic score obtained.

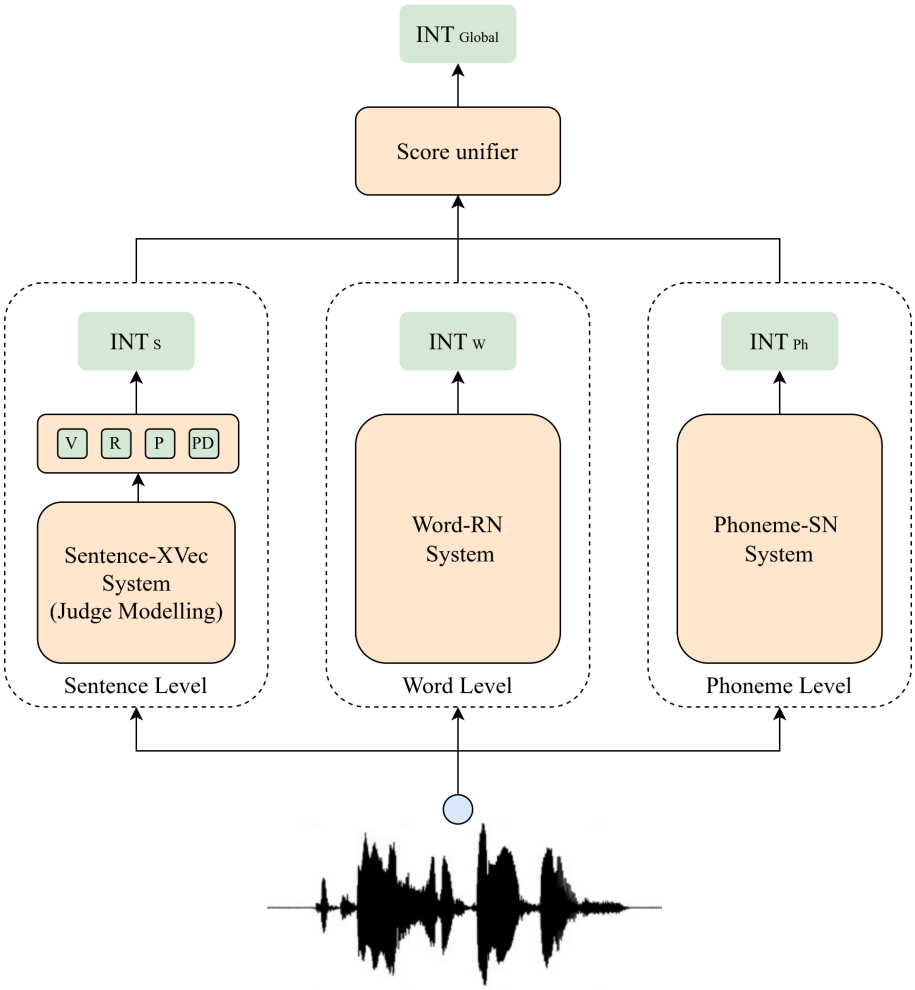


Figure 7.1: Diagram of the three granular systems proposed followed by a score unifier, which will be the topic discussed throughout the present chapter.

7.2 How to Merge the Different Developed Systems

In the present section, we will introduce the methodology used to regress an unified intelligibility score. Before diving into the unifying approaches, it becomes relevant to indicate the systems that will be used during this chapter. As far as the different levels of granularity are concerned, the systems used are as follows (see Figure 7.1):

- **Sentence:** The Sentence-XVec System, adapted for the individual judge modelization (see chapter 3 for base methodology plus chapter 6 for judge modelling improvements),

- **Word:** The Word-RNN System, a recurrent model with self-attention (see chapter 4),
- **Phoneme:** The Phoneme-SN system, the siamese network-based system with GeMAPS conditioning (see chapter 5).

The systems were chosen based on performance on both Spearman's correlation coefficient ρ and root mean squared error (RMSE), which can be found in their respective chapters. All of the chosen systems maintain at least some sort of explainability, whether that being based on the score of multiple utterances, such as the case at word and sentence level, or on a clear formula definition, like the one found at phoneme level for the phonetic similarity-based intelligibility score. Figure 7.1 presents a diagram of the proposed methodology, where the "score unifier" box corresponds to the different methods that will be experimented during the course of this chapter.

There are a variety of options that can be used to merge the different systems promoted. However for the sake of clinical explainability, simplicity should be preferred to complex merging approaches. Due to this aspect, we opted for simple yet interpretable statistical measures to be used as unifying approaches. The **mean** and **median** can be seen as the simplest and most comprehensive merging approaches, however, due to the sample size per speaker (of three different systems), it may provide a biased view, especially in the case of outliers. A **weighted mean** is also a valuable tool that we intend to implement (see equation 7.1). Similarly to the mean approach, it is also fast and easy to interpret, however, with an added margin to add different levels of relevance to different systems. In order to fine-tune the weighted mean weights, a grid search is employed [Álvaro Barbero Jiménez et al., 2007]. The statistical operators **min** and **max**, that give either the minimum and maximum value out of a set of predictions, are also simple and relevant measures that can provide insightful information. The mode is also a relevant statistical measure. However it becomes unfeasible to obtain due to the small sample size of three systems used to assess each speaker.

$$\mathrm{INT}_{global} = \omega_S \mathrm{INT}_S + \omega_W \mathrm{INT}_W + \omega_{Ph} \mathrm{INT}_{Ph} \tag{7.1}$$

7.3 Experimenting the Different Merging Methods

7.3.1 How to Handle Missing Data?

In order to unify the intelligibility scores promoted, several aspects should be taken into consideration. The first aspect that arises is the **numbers of speakers** used at each granularity level. Due to different methodologies and recorded speech tasks, the number of speakers used was not uniform across all the different granular systems. Given this, it becomes relevant to find ways to handle the missing values. From the literature, one can find a variety of ways to handle missing data [Kang, 2013, Graham et al., 2013]. Some of the best-known methods include either removing the incomplete data (removal methods) or use some form of estimation to reasonably guess the missing values (imputation methods). Since the present work dwells on pathological speech, a domain known for the smaller amounts of data used, removing speakers that do not have predictions at every granular level seems incautious, as it could provide ill-predicted intelligibility estimations, and a biased idea of the system's performance. This aspect points directly to the second well-known option: an imputation method to handle missing values. Mean, median and mode methods are widely used as an imputation method. However they need to be devised carefully as these methods can give a variation loss in the data when handling several missing values. By picking the predicted results for a set of speakers by one of the different systems, a distribution can be formed and the missing values predicted. This aspect, however, assumes a specific type of distribution (e.g. normal) that may not necessarily hold, such is the case of the reference speech intelligibility from picture description used in the present work (see chapter 2, section 2.3). This approach could also provide highly biased or skewed values for the same speaker, which is far from ideal as well. In our particular context, the most prudent method to handle these values is the imputation of the mean at speaker level instead of system level. By proceeding this way, no new-aberrant values will be added, and the posterior unified methods will be left unharmed. Figure 7.2 illustrates the process used to handle missing values in the context of the present work, using a mean imputation method.

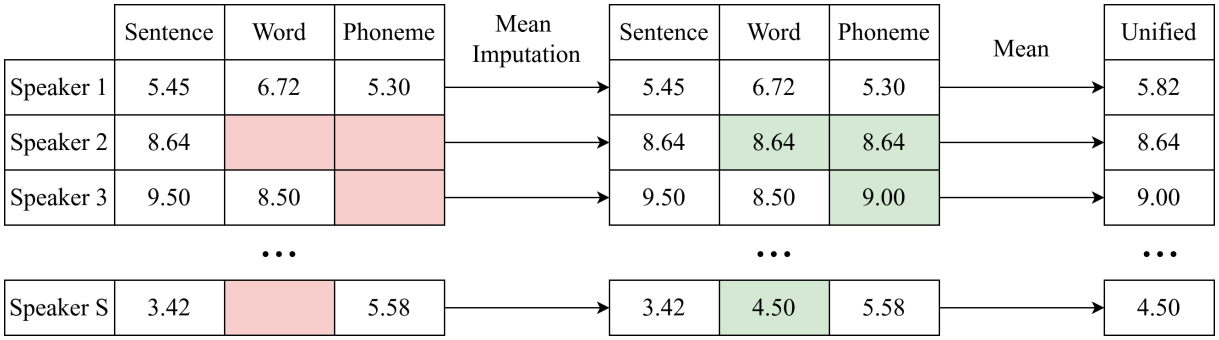


Figure 7.2: Illustration of the mean imputation method used to handle missing values.

The second aspect that needs to be taken in consideration when unifying our different models is the reference intelligibility measures. When working at sentence and phoneme levels the reference intelligibility values used were obtained from a picture description task, while at word level the perceived phonological deviation was used instead. Despite the different intelligibility measures sharing high correlations ($\rho = 0.87$ in this case, see chapter 2, section 2.3), it becomes relevant to harmonize and use a single reference intelligibility measure. Moreover, this aspect could provide some valuable insight on how the two measures relate to each other in the context of an automatic unified measure. In the present chapter, we will use the intelligibility obtained from picture description, since it was not only used by two of the preceding models (phoneme and sentence), but it was also obtained through a set of six professional judges, instead of the naive judges used for the word-level perceived phonological deviation measure.

7.3.2 Evaluating the Unifying Methods

Before merging each speaker's scores, the imputation method based on the mean (see figure 7.2) was employed on all missing values. Moreover, the unifying method was tested on the set of 108 speakers of the C2SI corpus that underwent the perceptual evaluation of speech intelligibility based on picture description. Similarly to the previous chapters, the results were evaluated using two metrics: Spearman's correlation coefficient ρ and the root mean squared error (RMSE). Table 7.1 displays the results obtained from the different methods employed. The results were also compared to the manual choice of the best system (i.e. oracle), which can be seen as the best possible value that both metrics can achieve with the given methodology.

A high correlation can be found on the majority of the methods tried, on the other hand, some fluctuations on the error were witnessed. The **mean** approach, illustrated on figure 7.3 stands out with a particularly high correlation of 0.91 and an error of 1.588. This method is only topped by the **weighted mean** approach, which promoted a 1% correlation increase and a non-significant error decrement. The results from the **median** approach are comparable to the mean and weighted mean approaches, however, slightly worse. The **max** approach showed simultaneously the lowest correlation and lowest error (RMSE) of all the merging approaches while the **min** approach showed the highest error value, surpassed only by the isolated phoneme-level model. The results of the three combinations of two systems can also be found on the same table, which can provide insightful information on the system pairs that share the best synergy. In this case, only the mean was employed as a statistical measure. The combinations of two systems show that the **sentence + phoneme** models obtained the highest correlation and error as well. On the other hand, the **word + phoneme** and **sentence + word** combinations displayed slightly smaller correlations and smaller errors, showing no clear consensus on the best combinations concerning the two evaluation metrics used. The **Oracle** approach promoted a correlation of 0.92, which is comparable to the correlation values obtained by the mean, median and weighted mean. On the other hand, despite the correlations being comparable, the error obtained by the Oracle is significantly (32%) lower. The different error values obtained point directly to an outlier analysis that could show specific speakers whose three systems share highly different predictions.

Table 7.1: Results obtained from the unifying methods implemented.

Model		Speakers	ρ	RMSE
Phoneme-SN System		71	0.89	2.080
Word-RNN System		102	0.82 †	1.548
Sentence-XVec (Adapted to Judge Modelling)		108	0.85	1.623
2 Automatic Measures (Mean Only)	Sentence + Phoneme		0.91	1.650
	Word + Phoneme		0.90	1.578
	Sentence + Word		0.88	1.508
3 Automatic Measures	Max		0.86	1.450
	Min		0.89	1.952
	Mean		**0.91**	**1.588**
	Median		0.89	1.597
	Weighted Mean (Grid Search) ††		**0.92**	**1.557**
Manually Best (Oracle)			0.92	1.075

†Correlated to INT from picture description instead of the PPD measures from chapter 4. The 108 speaker subset of the original 126 speakers was used in this chapter.

††Optimal grid search weights: $\omega_{Ph} = 0.4$, $\omega_W = 0.35$ and $\omega_S = 0.25$

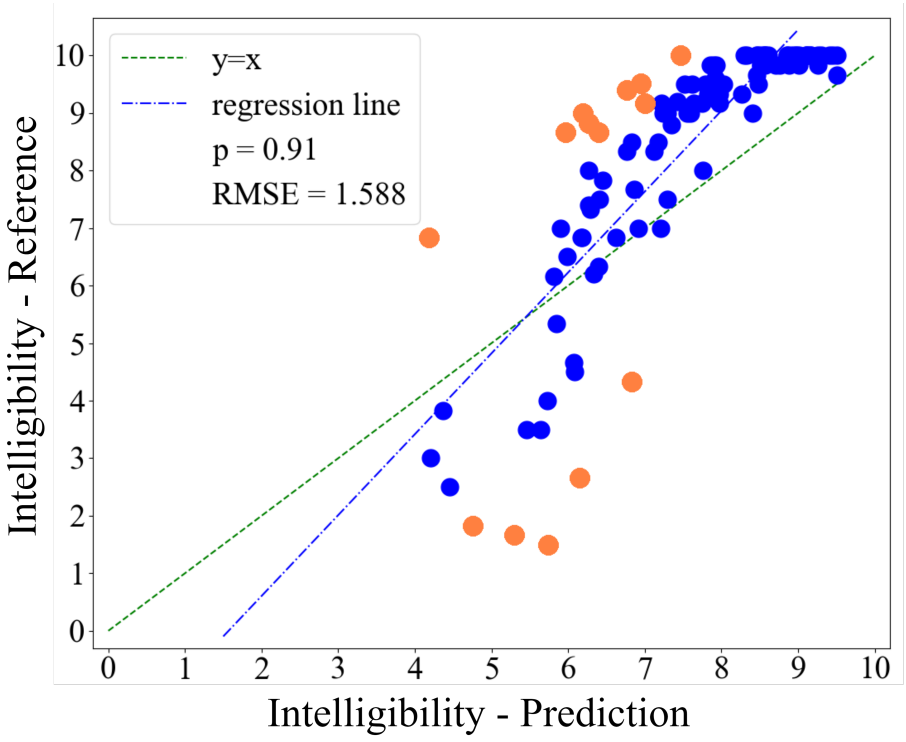


Figure 7.3: Results from the mean unifying method. Outliers are marked as orange dots.

7.4 Discussing the Merger Approaches

7.4.1 Performance Analysis

As it was stated before, no significant change was found between the mean and weighted mean approaches. Moreover, the optimal grid search weights ($\omega_{Ph} = 0.4$, $\omega_W = 0.35$, $\omega_S = 0.25$) showed a slight preference

towards the phoneme model in favor of the sentence model. Despite this, it is important to state that the phoneme model is the one with fewer scores (71 speakers) due to the training methodology used (see chapter 5), which consequently was also the system that had the largest mean inputation for the missing values.

When evaluating the results from the merge of only two systems (lines 4-6 of table 7.1), the mean between the sentence and phoneme systems was the unifying method that displayed the highest results correlation-wise ($\rho = 0.91$). This was expected since both systems were trained and optimized on the perceptual intelligibility based on picture description, while the word-level system made use of the perceived phonological deviation scores. From the three combinations of two systems, a trade-off appears to happen between error and correlation values, since the merging approaches that use the phoneme-level display a high correlation and higher error while approaches that use the word-level system display a lower correlation and error as well. Given this we can see that despite the word-level system not being trained nor validated on the perceptual intelligibility measure used as a target, it still has a benign effect on the final intelligibility estimation. This aspect can be witnessed by the error decrement found between the **sentence + phoneme** systems (2 automatic measures) and the **mean** of the three systems used (3 automatic measures).

Despite some error and correlation fluctuations, all of the merging approaches displayed clear correlation gains. This aspect states that an unified approach promotes more precise and discriminant intelligibility predictions when compared to the isolated granular models. On the other hand, the error obtained from the oracle showed that a custom selection of the best systems could further decrease the prediction error. Given this, an outlier analysis remains relevant to fully understand how different speakers and systems impact the error.

7.4.2 A Look Into the Outliers Found

Outliers were accounted for the predictions that share a Maximum Absolute Error (MAE) superior to 2.0, and therefore outside a $[-2, T, 2]$ boundary. From the experiments performed, a total of 15 outliers were found, illustrated by the orange dots on figure 7.3. The same set of outliers was found in the mean and weighted mean approaches. Out of the 15 outliers, only three had a single system prediction, while the remaining had predictions at all granular levels. Concerning the reference intelligibility scores, five had a score below 5.0, making part of an underrepresented class, moreover, concerning the standard deviations of the reference intelligibility scores, only four outliers had a standard deviation above 1.50, showing a moderate disagreement level among the judges on their reference score. As expected, on all outliers with a reference score below 5.0, the phoneme-model displayed the results closest to the ground truth. Only three outliers displayed aberrant values with an error above 3.0. We believe these outliers are partially responsible for the high gap in the RMSE values found between the unifying approaches and the Oracle. These three outliers all had a reference intelligibility score below 3.0, making them highly unintelligible to begin with. A comprehensive list of the 15 outliers can be found on annex A, table C.1. No pattern was found on the outliers concerning sex, tumor region/location or reconstruction type.

From the outlier analysis, it becomes clear that the differences in the error between the unifying approaches and the Oracle are mainly due to low-intelligibility speakers. These patients were expected to underperform simply due to the small quantities of data of this type present in the corpus when compared to other intelligibility levels (e.g. above 8.0). This aspect plus the non-normality of the reference intelligibility scores contributed to a harder modelization of this class of patients, making it harder for an automatic system to differentiate between different tiers of highly unintelligible speech.

7.5 Scientific Contributions and Perspectives

7.5.1 Conclusions

In the present chapter, the junction of three systems was performed that operate at different granularity levels. They were all developed during the course of the present thesis. The systems operate at the sentence, the word and the phoneme level respectively. A set of statistical measures (max, min, median,

mean and weighted mean) was used, at speaker level, on the individual predictions of each system. The results suggest that all unifying methods (i.e. statistical measures used) promoted correlation gains and error decrements when compared to the individual granular models. The mean and weighted mean approaches promoted the higher correlations ($\rho_{mean} = 0.91$ and $\rho_{weighted} = 0.92$), while the max and weighted mean approaches gave the lowest errors ($RMSE_{weighted} = 1.557$ and $RMSE_{Max} = 1.450$). The results were compared to an Oracle approach that chose manually, for each speaker, the best system. Despite the weighted mean correlation being the same as the Oracle, a gap in the error was found: this indicated an outlier analysis on the speakers with a larger deviance from the reference intelligibility. The same analysis revealed that, despite the unifying methods promoting very high correlations, the system underperforms on some speakers with a lower intelligibility value. This aspect was expected due to the small number of patients at this intelligibility level, an aspect that was also evident in previous chapters. From the same outlier analysis, we can draw the conclusion that the phoneme-level system, despite showing a higher RMSE value than the other granular systems, performs the best for this grade of patients. This aspect shows that the combined usage of granular systems can indeed help mitigate a system's particular weaknesses, while achieving higher correlations than the isolated models.

7.5.2 Perspectives

Despite the high correlations found when using different merge approaches, the error gap found between these methods and the oracle still leaves some room for improvement. This error gap raises an interesting question: Should we sacrifice interpretability for an uncompromising search of optimal results on the error? An automatic model could be fitted to choose, for each patient, the best model. However this approach would not contribute towards clinical interpretability. In my point of view, an automatic approach to predict speech intelligibility should be devised to mainly counter the subjectivity associated to the perceptual measures. By adding another complexity layer, which would remain non-interpretable, we would be working towards an uncompromising search for optimal results, instead of an application that can be used and explained in a clinical setting. Given this, we have opted for the unifying methods employed during this chapter to remain simple and easy to interpret. On the other hand, there is always room for improvement. Approaches that use external features that remain clinically valid and easy to interpret (e.g. GeMAPS [Eyben et al., 2015]), can be used to condition the final intelligibility score. In this particular context, the intelligibility score could even be computed as a direct function of the systems developed and the conditioning ability of those same features, provided that it would not interfere with the generalization ability of the system. This aspect was partially explored in chapters 3 and 5 during the post-processing. The further development of this conditioning remains an interesting perspective and an active lead for future work.

Given that the present chapter works as the culmination of all the previous chapters, there are a variety of other perspectives that can be drawn from the present work. Nevertheless, we will save the larger medium-term and long-term perspectives of the merging approaches promoted for the following chapter, that will address not only the main conclusions, but also the full-spectrum of perspectives on the present thesis.

Key Takeaways

1. "The whole is bigger than the sum of its parts" - The unifying methods used displayed correlation and error improvements.
2. The oracle approach showed that while the correlation is near the maximum achievable value, there is still room for improvement on the error.
3. Since the unified prediction makes use of simple statistical operators (e.g. mean, median), the predictions remain easily interpretable.

Chapter 8

Conclusions and Perspectives

8.1 Conclusions

The work developed during the course of this thesis explored the automatic prediction of speech intelligibility using deep learning methodologies. More specifically, during the present work I developed a variety of systems that predict speech intelligibility at the three distinct granular levels: sentence, word and phoneme. These systems were developed and tested on the French corpus of head and neck cancer. The granular analysis to speech intelligibility allowed a better understanding of why intelligibility measures are highly subjective by nature, and how different speech impairments relate to speech intelligibility. This analysis leads to a study of the individual judge profiles that showcased the profile variance present in the perceptual judges, and the more uniform profiles found on the automatic modelization. Moreover, the combination of the granular predictions displayed higher correlations and lower prediction errors when compared to the individual granular models.

While the granular predictions were an idea that started early on, during the development of this thesis, the main intention was to provide an answer to a variety of research questions. In the present work, the main contributions were always related to the specific case of head and neck cancer. The results found on the different chapters display performance gains on the distinct systems promoted that held well with the previous results found in the literature. The results obtained foreshadow the first research question set at the beginning: *Can deep learning be reliably used to predict speech intelligibility?* Given that this artificial intelligence domain is known for having large data requirements, and that the domain of pathological speech is known for its data scarcity, it became a relevant and central question. By looking at the different systems proposed, we could see that this aspect was mainly problematic for the sentence granularity level (see chapter 3), which had the smallest amounts of data. In order to tackle this issue, data augmentation schemes were implemented, and a pre-trained model to extract the x-vectors was employed. The proposed methodology allowed for high correlation values, that mitigated the short-data scenario. The same methodology was also used when modeling each individual perceptual judge (see chapter 6). Moreover, novel data augmentation schemes could be devised to further enhance the developed systems, which will be left as a perspective. As far as the amount of data is concerned, for the word granularity level (chapter 4) this aspect did not appear to be very problematic. The results obtained in this chapter even suggested that the number of pseudo-words can be reduced at inference time without significant changes in the scoring. Finally, at phoneme level (chapter 5), due to forced alignment, a large quantity of isolated phones was obtained, which allowed the generalization of the model. Given all these findings, we can say that deep learning can be reliably used in our specific context. However, it should be paired with suitable methodologies that address the specific issue of smaller amounts of data and interpretability.

Seeing that speech intelligibility can be automatically predicted with deep learning, the following research question that arose was: *How can we build trust in this type of systems for use in a clinical context?* Given that the black-box paradigm tends to not hold well among doctors and therapists, an understanding of how the measures were obtained becomes of crucial importance. This aspect is directly

related to the explainability of the measures obtained through the systems developed. In the present work, and at the distinct levels of granularity, I have always tried to add a certain degree of explainability to each system. Starting at sentence-level, the segmentation of the reading passage task into smaller sentences allowed the system to assess individually each chunk. This allowed us to have individual scores for each sentence that displayed interesting results as far as different types of speech impairments were concerned. Furthermore, this finding pointed us directly to the hypothesis that there are sentences that convey a better intelligibility estimation than others, depending on the type of speech impairment of the speaker. This aspect was a recurrent finding throughout the whole thesis, that was witnessed similarly at word and phoneme levels with different words and phonemes playing distinct relevance levels for different speakers. At word-level, by experimenting with different sets of pseudo-words, I was able to obtain interesting results that correlated well with the perceived phonological deviation scores. While the usage of the score variance during training and the attention plots did not display the expected improvements, they can still be considered an interesting venue towards interpretability to be pursued in future endeavours. The regression of the word-level intelligibility score was made based on the mean of different sets of pseudo-words, which similarly to the sentence-level, can be seen as more interpretable since the final score is a function of each individual word score. Finally, at phoneme-level the regression function used, based on the phonetic similarity, can be seen as the most objective and interpretable of them all. Since the score can be directly regressed from the amount of similar/dissimilar phonemes that a speaker has, it becomes easily justifiable in a clinical context, which corroborates the fact that intelligibility is directly related to acoustic-phonetic decoding. By navigating towards interpretable systems, we can start easily building trust in these approaches, which was one of the main goals set out during the course of this thesis. By being explainable, a system can also be seen as more trustworthy, and therefore these methodologies are on the road towards being added value to clinical practices.

The different levels of granularity were merged to obtain a universal prediction (see chapter 7). The high correlation values found in the different methods used show that a precise general intelligibility estimation is possible given the smaller quantities of data used. However, the error values obtained still point towards certain improvements on severe patients. This aspect was expected due to the small presence of this type of patient. The outliers found in the merged system also show that there is a synergy between the different granular systems. This synergy avoided highly aberrant or deviant predictions, since besides some exceptions on low-intelligibility speakers, the granular systems generally had slightly different sets of outliers. This detail also validates the relevance of a granular approach, that targets different aspects of speech intelligibility for each speaker, and therefore promotes a more robust and complete prediction.

While the methodologies presented during the course of this thesis displayed interesting results, they also faced some limitations. Since the main scope of research was set towards head and neck cancer, no other corpus was assessed. Given this, it becomes interesting to explore how the same (or improved) approaches would behave when predicting speech intelligibility for other diseases with dysarthric symptoms, such as cerebral palsy and amyotrophic lateral sclerosis. On the other hand, it also becomes difficult to manage and merge the sometimes even smaller quantities of recorded data that these diseases tend to have, especially when considering the fact that the perceptual intelligibility ratings used can be obtained in many different ways. Hybrid approaches that combine multi-corpus from various diseases could be an interesting perspective, in the case some common ground can be found as far as the intelligibility measures of each corpus are concerned. The present work was also performed solely on the French language, and generalizing the methodologies on different languages is also an interesting lead for future work. Despite this, the approaches promoted should be tailored to fit the target language, as different languages with different sets of phonemes pose different challenges to this type of systems. A global intelligibility model that is language and disease independent is also an interesting lead for future work. However, some specificity may be lost in the pursuit of this approach. The number of patients that presented a low intelligibility score (below 5 on a 0-10 scale) was also a limitation throughout this thesis, that hurt the generalization ability of the proposed systems on this type of patients. Since the recruitment of this class of patients is a difficult and complex task, measures such as data augmentation could further help the system's generalization ability. These possible approaches will be discussed in greater detail in the upcoming perspectives section.

The granular approach promoted addresses a gap in the literature concerning the need for reliable and

explainable automatic predictions of speech intelligibility. This thesis aimed to separate intelligibility in three distinct parts in order to construct a general, more robust and interpretable automatic prediction. This type of unified methodology was previously unseen in the literature. Despite the limitations mentioned in the previous paragraph, the results obtained throughout the present work strongly support the hypothesis that a granular approach is able to tackle different aspects of speech intelligibility, and that the posterior unified approach is able to compensate the shortcomings of each individual model. This aspect provides an answer to the final and most important research question set at the beginning of this thesis, of *whether a granular analysis can work when predicting speech intelligibility, and if so what are the added benefits*. Moreover, the study performed concerning the modeling of each individual judge also validated the hypothesis that automatic judges are more objective when compared to the perceptual judge profiles witnessed. While the automatic prediction of speech intelligibility is a vast and ever-changing field, the contributions made during this thesis can be seen as a stepping-stone towards obtaining high-performance intelligibility predictions. The automatic measures obtained are not only significantly less subjective, but also faster, unbiased and explainable when compared to the perceptual counterpart. Hence, the future implementation of these systems in a clinical context can be seen as an added value.

8.2 Perspectives

From the different studies performed within the context of the present work, a variety of perspectives can be drawn. These perspectives can encompass not only the different research questions that were addressed, but also some propositions on what a possible continuation of the present work would be. While each chapter had a dedicated outlook section used to address the future work concerning the developments promoted, in this final section we aim to present the global perspectives of the entire work. For the sake of clarity, we will group them in three distinct groups: short, medium and long-term perspectives.

Short-term perspectives would be centred around new and upcoming methodologies that could have a potential direct use to the automatic prediction of speech intelligibility. These perspectives are around new technical aspects and on possible performance boosters for the developed models. Medium-term perspectives will target the clinical usage of the developed models. Since the automatic analysis of pathological speech still remains a vastly academic domain, it is only natural that some of the research advances promoted will take some time to be adopted. Given this, these perspectives will address the question of how to build trust in these approaches, and how they would operate in a clinical scenario. Finally, the long-term perspectives will tackle some research questions in a wider and unrestrained manner, addressing questions such as the definition of different concepts used in this work, namely intelligibility and severity, and whether these measures could be combined to reduce subjectivity. Due to the multimodality of these concepts that are present not only in the automatic domain, but also in speech science, speech therapy and linguistics, these perspectives are left as a small, thought-provoking contribution to consummate this thesis.

8.2.1 Short-Term

As it was previously stated, the short-term perspectives will encompass upcoming methodologies that could be of use to the problem at hand. These perspectives will be grouped into techniques that could aid the development of posterior deep learning models, such as data augmentation, and also novel architectures that could potentially be interesting.

Data Augmentation

It is clear that data augmentation plays an important role in the development of deep learning systems. Normally due to the large data requirements of these models, data augmentation comes in handy as a training and performance booster. This is especially the case for pathological speech assessment, that is a domain typically known for its data scarcity. In the present work, typical audio data augmentation schemes such as temporal distortion [Ko et al., 2015] that can also be applied to pathological

speech [Vachhani et al., 2018], were implemented. While these methodologies performed fine and allowed the augmentation of scarce quantities of data, there is still room for improvement when thinking about dysarthric speech. While temporal distortion is a widely relevant and easy-to-implement data augmentation scheme, it does not exactly mimic pathological speech. Despite being clear that some unintelligible patients have a lower speaking rate, the slowdown witnessed is not necessarily uniform. The works of [Jacobi et al., 2013b] and [de Bruijn et al., 2013] explore the perceived vowel changes that head and neck cancer experience in post-surgery, manifested by overcompensation and a reduced vowel space. This line of research, mainly clinical, gives us interesting indices towards a tailored data augmentation, that could be based on different types of distortions of healthy speech.

Given this, a possible solution that could serve as data augmentation is the temporal distortion of key phonemes and also key silences between consecutive phonemes. This approach could also be extended towards other aspects, such as formant modulation and intensity alteration at phoneme level. The time windows for each phoneme could be obtained through forced alignment, and then those specific parts of the audio file distorted accordingly. By doing this, a healthy speaker could be distorted to more closely mimic key-speech impairments. This comes with the added advantage of not requiring computationally expensive models. However, it would require a great deal of clinical know-how to properly target key aspects of dysarthric speech, in order to reproduce them based on a healthy utterance. Figure 8.1 presents an example of the same word being uttered by a healthy and a pathological speaker, and a suggestion of phonetic distortion to more closely match the speech impairment witnessed.

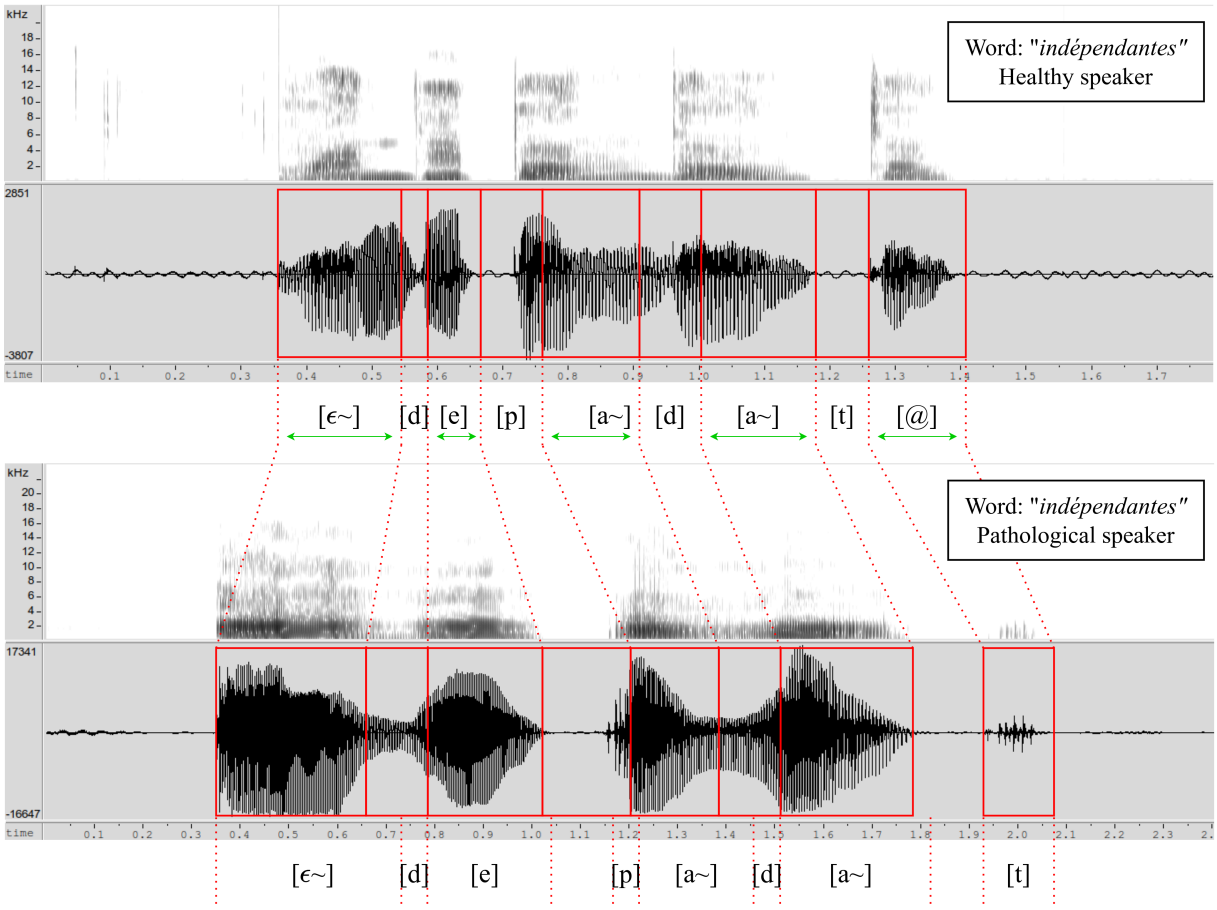


Figure 8.1: Illustration of what a temporal phonetic distortion would appear to be. From the comparison between the two instances of the word *"indépendantes"*.

Moreover, the inverse could also be possible. With a proper know-how, pathological phonemes and

silences could be shortened or extended in order to make the utterance more intelligible. This aspect, however, would require a good forced-alignment system to properly detect phoneme boundaries for pathological speech. Both types of phonetic distortions presented could be validated either perceptually or through the usage of the systems developed during the course of this work. Should an utterance be deemed more intelligible or unintelligible, these data augmentation methods would be on the road towards validation. Given that in this setting, we have utterance pairs (normal vs distorted), STOI and ESTOI intelligibility measures could also be implemented. The corresponding utterance pairs could also serve as data to train models such as autoencoders or adversarial models that will be briefly introduced in the next paragraph.

The usage of new methodologies such as generative adversarial networks (GANs) could provide interesting data augmentation schemes that could be of use for the automatic prediction of speech intelligibility. These (adversarial) models are a deep learning methodology firstly introduced in the work of [Goodfellow et al., 2014] that has shown interesting improvements in the field of generational models. This methodology aims to train two classes of models: a generator and a discriminator. The goal of the generator is to create realistic utterances while the discriminator's main objective is to differentiate between the real and forged utterances. When training models in this adversarial fashion, a high-quality generator and discriminator are achieved. For the particular context of data augmentation, the generator will be the most relevant. The work of [Jiao et al., 2018] already shed some light on this aspect, presenting an adversarial model that transformed healthy into pathological speech. On the other hand, the model was trained to generate pathological utterances used to the task of binary classification of pathological speech. An improved model that could generate different sorts of speech impairment, based on different post-surgery situations and tumour locations, for example, would be an interesting future contribution.

Despite this, it is known that adversarial models are hard to train and have large data requirements. The forged utterances previously introduced could also be helpful to better train these models. While these utterances could lack some naturalness due to the key-phonetic distortions, the GAN conditioned on both forged and real utterances could remove the sound artefacts by itself and promote high quality data augmentations.

End-to-End Models

The usage of end-to-end models, or models that operate directly on the raw audio waveform instead of extracted features, is also a relevant and interesting perspective for the automatic prediction of speech intelligibility. Models such as Wav2Vec [Schneider et al., 2019] can directly learn speech-related features from the audio signal without being explicitly told to do so, and output embedding representations that have proven to be quite useful in speech recognition. The work of [Hernandez et al., 2022] shed some light on this topic by using the Wav2Vec model to learn self-supervised speech representations to improve dysarthric speech recognition. Similarly to the *x-vector* speaker embeddings, these fixed-length representations could be used to learn interesting intelligibility models that could be language-independent and high-performance. On the other hand, for the sake of clinical usage, some interpretability should always be promoted somehow when using this sort of large-scale model.

The usage of semi-supervised learning [Hady and Schwenker, 2013] can also be an interesting perspective. This type of training methodology uses small amounts of labelled data to learn a model up to a certain quality standard, and then makes predictions on the unlabelled data based on the distance to the labelled points. While the unlabelled data provides context on the entire dataset distribution, the labelled data serves as a sanity check, that provides structure and a certain degree of guidance to the system's development. Semi-supervised learning can be interesting in the advent of having larger quantities of unlabelled data, particularly data obtained "in the wild" (e.g. WhatsApp messages, self-recorded YouTube videos, etc.) of people with speech impairments. While the applications of using this methodology applied to pathological speech assessment are still to be explored, the work of [Serrà et al., 2021] presented a semi-supervised learning approach applied to the assessment of speech quality, a problem that shares a reasonable amount of similarities with speech intelligibility.

8.2.2 Medium-Term

In the present work, we have discussed a variety of systems and their respective performances. Despite this, the effects of a direct application of the proposed systems in a clinical environment still remain an interesting and uncharted territory to explore. While the deployment of these approaches requires a big deal of effort, mainly due to privacy constraints and low confidence on automatic approaches, hybrid approaches that combines perceptual and automatic assessments could tackle some of these issues. These approaches become an interesting perspective, since the main objective behind the development of automatic approaches was never to replace a doctor or therapist, but to provide a system that could enhance clinical analysis and serve as a second opinion.

In order to shed some light on this matter, a brief study on what would happen when replacing one of the perceptual judges by his/hers corresponding automatic model is presented. The results are derived from the experiments performed on chapter 6, which conducted the modelization of each one of the six judges used in the C2SI corpus. The main goal of this study is to assess the agreement levels found among the judges. High levels could potentially mean that the perceptual judges are in agreement with the modelization of the left-out judge, which could help build trust in the automatic model. In order to do so, the correlations between each pair of judges are computed. Furthermore, in order to obtain a general agreement score for the replacement of one judge, the mean of all judge pairs is computed. The results of this analysis can be found on table 8.1, and its respective sub-tables.

From the presented results, we can see an average level of agreement of 0.747 between all the six perceptual judges (sub-table 8.1a), and an agreement level of 0.841 for the six automatic judges (sub-table 8.1b), modeled according to the intelligibility scores of each one of the perceptual judges respectively. By comparing the two, a big difference can be witnessed, showing that the automatic approach promotes a larger agreement level between the different judges. Moreover, when analyzing each individual replacement (sub-tables 8.1c to 8.1h), slightly lower agreement levels can be found, between 0.713 and 0.730, when compared to sub-table 8.1a. Despite these small decrements, the agreement levels still remain comparable to the reference one. Additionally, the mean of the results obtained from sub-tables 8.1c to 8.1h is also presented on the sub-table 8.1i. This result presents a general idea of the agreement level found when replacing one of the judges by his/her respective automatic model. The results also suggest a small change (from 0.747 to 0.722), which can be seen as a good index and an interesting perspective for the future application of these technologies in clinical situations.

In the present short study, an analysis of the replacement of one judge was performed. Furthermore, there are a variety of studies that can also be performed within the same context, namely replacing more than one judge, adding an external automatic judge that was not trained on the scores obtained by the other judges, and so on.

8.2.3 Long-Term

During the course of this thesis, different intelligibility measures from the C2SI corpus were used as reference. Since the main contributions of this thesis dwell around the development of automatic systems to predict speech intelligibility, the validity of the intelligibility measures used was not assessed, as it is a subject outside of the domain and scope of this line of research. Nevertheless, I would like to leave as a thought-provoking long-term perspective the merge of different intelligibility measures used as reference into a single general intelligibility score. In the present work, two main intelligibility measures were used. The first one made use of the perceptual ratings of six experts (INT *Des.*), while the second measure was based on the transcription of pseudo-words by three different naive listeners each, in a process known as perceived phonological deviation (PPD). Despite these two measures being used to assess speech intelligibility, they serve two distinct purposes. While the PPD measure aims to tackle intelligibility at a spoken communication level, based on phonetic acoustic decoding, the INT *Des.* measure takes in consideration different aspects of speech such as voice quality, prosody, resonance and phonemic distortions (see chapter 6). Given this, the PPD measure can be seen as more direct and naive, while the INT *Des.* measure more subjective and educated, due to the usage of a set of experts. Despite being different, the question of whether these two measures should be combined to form a unified reference intelligibility measure arises. In the previous chapter, the unifying methods promoted are already mixing the automatic modeling of

Table 8.1: Average levels of agreement when replacing one of the judges (J) by his/hers corresponding automatic model (M). All results illustrate Spearman's correlation ρ. The Average Level of Agreement is obtaining by averaging the correlations of all possible judge pairs (upper or lower triangle of each matrix). Red values mark lower agreements (<0.65).

(a) Agreement level for the six **perceptual** judges

	J1	J2	J3	J4	J5	J6
J1	1.0	0.71	0.79	0.74	0.76	0.70
J2	0.71	1.0	0.71	0.63	0.72	0.67
J3	0.79	0.71	1.0	0.80	0.82	0.81
J4	0.74	0.63	0.80	1.0	0.81	0.75
J5	0.76	0.72	0.82	0.81	1.0	0.78
J6	0.70	0.67	0.81	0.75	0.78	1.0
Average Level of Agreement: 0.747						

(b) Agreement level for the six **automatic** judges

	M1	M2	M3	M4	M5	M6
M1	1.0	0.80	0.83	0.83	0.84	0.80
M2	0.80	1.0	0.86	0.79	0.81	0.80
M3	0.83	0.86	1.0	0.87	0.88	0.88
M4	0.83	0.79	0.87	1.0	0.91	0.85
M5	0.84	0.81	0.88	0.91	1.0	0.80
M6	0.80	0.80	0.88	0.85	0.80	1.0
Average Level of Agreement: 0.841						

(c) Agreement level with replacement of J1

	M1	J2	J3	J4	J5	J6
M1	1.0	0.56	0.68	0.68	0.72	0.61
J2	0.56	1.0	0.71	0.63	0.72	0.67
J3	0.68	0.71	1.0	0.80	0.82	0.81
J4	0.68	0.63	0.80	1.0	0.81	0.75
J5	0.72	0.72	0.82	0.81	1.0	0.78
J6	0.61	0.67	0.81	0.75	0.78	1.0
Average Level of Agreement: 0.717						

(d) Agreement level with replacement of J2

	J1	**M2**	J3	J4	J5	J6
J1	1.0	0.61	0.79	0.74	0.76	0.70
M2	0.61	1.0	0.66	0.63	0.64	0.58
J3	0.79	0.66	1.0	0.80	0.82	0.81
J4	0.74	0.63	0.80	1.0	0.81	0.75
J5	0.76	0.64	0.82	0.81	1.0	0.78
J6	0.70	0.58	0.81	0.75	0.78	1.0
Average Level of Agreement: 0.725						

(e) Agreement level with replacement of J3

	J1	J2	**M3**	J4	J5	J6
J1	1.0	0.71	0.69	0.74	0.76	0.70
J2	0.71	1.0	0.56	0.63	0.72	0.67
M3	0.69	0.56	1.0	0.74	0.78	0.66
J4	0.74	0.63	0.74	1.0	0.81	0.75
J5	0.76	0.72	0.78	0.81	1.0	0.78
J6	0.70	0.67	0.66	0.75	0.78	1.0
Average Level of Agreement: 0.713						

(f) Agreement level with replacement of J4

	J1	J2	J3	**M4**	J5	J6
J1	1.0	0.71	0.79	0.69	0.76	0.70
J2	0.71	1.0	0.71	0.54	0.72	0.67
J3	0.79	0.71	1.0	0.77	0.82	0.81
M4	0.69	0.54	0.77	1.0	0.76	0.68
J5	0.76	0.72	0.82	0.76	1.0	0.78
J6	0.70	0.67	0.81	0.68	0.78	1.0
Average Level of Agreement: 0.727						

(g) Agreement level with replacement of J5

	J1	J2	J3	J4	**M5**	J6
J1	1.0	0.71	0.79	0.74	0.68	0.70
J2	0.71	1.0	0.71	0.63	0.58	0.67
J3	0.79	0.71	1.0	0.80	0.74	0.81
J4	0.74	0.63	0.80	1.0	0.75	0.75
M5	0.68	0.58	0.74	0.75	1.0	0.69
J6	0.70	0.67	0.81	0.75	0.69	1.0
Average Level of Agreement: 0.717						

(h) Agreement level with replacement of J6

	J1	J2	J3	J4	J5	**M6**
J1	1.0	0.71	0.79	0.74	0.76	0.67
J2	0.71	1.0	0.71	0.63	0.72	0.55
J3	0.79	0.71	1.0	0.80	0.82	0.77
J4	0.74	0.63	0.80	1.0	0.81	0.70
J5	0.76	0.72	0.82	0.81	1.0	0.77
M6	0.67	0.55	0.77	0.70	0.77	1.0
Average Level of Agreement: 0.730						

(i) Final average agreement level comparison

	ρ
Average Level of Agreement (6 human judges)	0.747
Average Level of Agreement (5 human judges + 1 automatic judge)	0.722

these two measures. Despite high correlations being achieved, there is still room for improvement. In this section we aim to look at the problem from a different angle, leaving the perspective of whether a global intelligibility measure that could encompass both naive and expert approaches would have any clinical or automatic utility.

In order to shed some light on this aspect, I tried merging the two different reference intelligibility measures, instead of simply using one. Similarly to table 7.1, different methods were assessed, however, this time with the merge of the INT *Des.* and PPD measures in a gradual way. This means that the two measures were combined using a weighted mean, where the relevance of each measure is given by a percentage. The results of this small study can be found on table 8.2. The patients used in the table correspond to the set of 107 patients that had both perceptual intelligibility (INT *Des.*) and perceived phonological deviation (PPD) evaluations. As expected, the results suggest a correlation increase when using a mixed intelligibility measure as opposed using only the INT *Des.* or PPD measures. The obtained correlations can achieve values up to 0.94 or 0.95 when using either the mean or weighted mean approaches from the last chapter, and a 50% contribution of each perceptual measure as the reference score. While the merge of the two measures displays an expected correlation gain when applied to the unified system, the added benefit of using a measure like this is still unknown. The two measures have distinct utilities, however, the lack of a generalized consensus on the definition of intelligibility, and the overlap with other measures (comprehensibility, severity, etc.) aggravates this issue. The work of [Pommée et al., 2022] did a survey among specialists in order to see where the separation between intelligibility and comprehensibility lies. Their results suggest that, despite the two measures being related to functional human communication, intelligibility refers to acoustic-phonetic decoding while comprehensibility relates to the reconstruction of the meaning of the message. On the other hand, this definition validates the two measures used. The INT *Des.* measure was obtained by a set of experts that took phonemic distortions as the most relevant parameter (see chapter 6) and the PPD measure relates directly to acoustic-phonetic decoding due to the pseudo-word transcription. The high correlation found between the two perceptual measures ($\rho = 0.87$) shows that they can be seen as highly similar, and therefore a unified measure could help reduce the subjectivity associated with these measures.

Table 8.2: Results obtained from using a gradual merge of the two reference intelligibility measures used during this study. The 107 patients that share both perceptual measures were used (INT and PPD). Mean replacement (see section 7.3.1) was used to handle the missing scores for the speakers at phoneme-level.

Model	Reference Intelligibility Measure									
	INT: 100 % PPD: 0%		INT: 75% PPD: 25%		INT: 50% PPD: 50%		INT: 25% PPD: 75%		INT: 0 % PPD: 100%	
	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE
PPD	0.88	—	0.93	—	0.96	—	0.99	—	1.00	—
INT *Des.*	1.00	—	0.99	—	0.96	—	0.93	—	0.88	—
Phoneme-SN	0.90	1.856	0.92	1.701	0.92	1.392	0.90	1.227	0.86	1.135
Word-RN	0.82	1.548	0.86	1.276	0.88	1.104	0.89	0.859	0.88	0.784
Sentence-XVec	0.85	1.623	0.87	1.363	0.88	1.143	0.87	0.985	0.85	0.921
Max	0.86	1.450	0.89	1.199	0.90	0.994	0.89	0.867	0.88	0.855
Min	0.89	1.952	0.92	1.701	0.92	1.480	0.90	1.306	0.87	1.197
Mean	0.91	1.588	**0.94**	1.303	**0.94**	1.044	0.93	0.836	0.91	0.721
Median	0.89	1.597	0.92	1.318	0.92	1.067	0.91	0.870	0.89	0.768
Weighted Mean	0.92	1.557	**0.94**	1.273	**0.95**	0.996	0.93	0.771	0.91	0.654

In order to finish this thesis, I would like to leave a few open-ended questions concerning the intelligibility measures used. These could be seen as longer-term perspectives since they relate directly to the definitions of speech intelligibility, and the clinical context in which they are used. Due to the fact that different speech tasks have different intelligibility ratings (with high intra-judge variability in some cases

of the C2SI corpus), should we aim for a unified measure that shares high levels of consensus among experts? Should we search for a general measure that contains a variety of other sub-measures (e.g. intelligibility, severity, comprehensibility, etc.) in order to regress a global intelligibility score? Will a general, unified perceptual intelligibility measure have any clinical utility? The correlation gains seen on table 8.2 paint the possibility of a global measure in an optimistic way. However, a deeper analysis should be performed. By walking towards consensus on clinical measures such as speech intelligibility, the further development of high-quality assistive systems will be facilitated. Given this, I strongly believe that in the long-term, the domain of speech intelligibility will walk not only towards robust high-performance automatic systems, but also to a better definition of the clinical measures used. It becomes clear to see that these two aspects walk hand-in-hand. By seeing the research, development and implementation of the type of systems developed during the present work in a hopeful way, the present domain will gravitate towards a futuristic clinical setting rich in trusted assistive technologies. These technologies hopefully will not only aid doctors and therapists in a variety of ways, but also directly lead to better diagnosis, prognosis and treatment of speech affecting diseases such as head and neck cancer.

Appendix A

Pseudo-Words Experiments

A.1 Cost Matrices

In this section of the present annex, the cost matrices that showcase the distances between the different phonemes are presented, introduced by the work of [Ghio et al., 2018]. These two matrices, one for the vowels and another for the consonants (see tables A.3 and A.4 respectively), present the distance between phonemes based on trait decomposition. The more different traits two phonemes have, the larger the distance will be. The traits used for the vowels (nasal, rounded, open, etc.) can be found on table A.2, and for the consonants (voiced, nasal, vocalic, etc.) on table A.1.

These matrices become relevant to further understand the reference scores that were used for the pseudo-words (see word-level granularity, chapter 4). Based on the distance obtained from the table, the Wagner-Fischer algorithm can be computed to obtain the distance between two given pseudo-words.

Table A.1: Trait decomposition of French consonants.

	p	t	k	b	d	g	f	s	S	v	z	Z	m	n	ŋ	l	R
vocalic	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
continuous	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1
nasal	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
voiced	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1
compact	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
acute	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1

Table A.2: Trait decomposition of French vowels.

	a	i	u	o	e	y	ϕ	ɛ	ɔ	œ	Ô	Û	Ê	ə	ã	ɛ̃	ɔ̃	œ̃	μ
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
rear	1	0	1	1	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0
high	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
rounded	0	0	1	1	0	1	1	0	1	1	1	-	1	0	0	0	1	1	-
open	1	0	0	0	0	0	0	1	1	1	-	-	-	-	1	1	1	1	1

Table A.3: Consonant cost matrix, used to obtain the perceived phonological deviation score. Imported from [Ghio et al., 2018].

p	t	k	b	d	g	f	s	S	v	z	Z	m	n	ŋ	l	R	
0	1	2	1	2	3	1	2	3	2	3	4	3	4	5	3	4	p
1	0	1	2	1	2	2	1	2	3	2	3	4	3	4	2	3	t
2	1	0	3	2	1	3	2	1	4	3	2	5	4	3	3	4	k
1	2	3	0	1	2	2	3	4	1	2	3	2	3	4	2	3	b
2	1	2	1	0	1	3	2	3	2	1	2	3	2	3	1	2	d
3	2	1	2	1	0	4	3	2	3	2	1	4	3	2	2	3	g
1	2	3	2	3	4	0	1	2	1	2	3	4	5	6	4	3	f
2	1	2	3	2	3	1	0	1	2	1	2	5	4	5	3	2	s
3	2	1	4	3	2	2	1	0	3	2	1	6	5	4	4	3	S
2	3	4	1	2	3	1	2	3	0	1	2	3	4	5	3	2	v
3	2	3	2	1	2	2	1	2	1	0	1	4	3	4	2	1	z
4	3	2	3	2	1	3	2	1	2	1	0	5	4	3	3	2	Z
3	4	5	2	3	4	4	5	6	3	4	5	0	1	2	2	3	m
4	3	4	3	2	3	5	4	5	4	3	4	1	0	1	1	2	n
5	4	3	4	3	2	6	5	4	5	4	3	2	1	0	2	3	ŋ
3	2	3	2	1	2	4	3	4	3	2	3	2	1	2	0	1	l
4	3	4	3	2	3	3	2	3	2	1	2	3	2	3	1	0	R

Table A.4: vowel cost matrix, used to obtaine the perceived phonological deviation score. Imported from [Ghio et al., 2018].

a	i	u	o	e	y	ϕ	ɛ	ɔ	œ	ã	ɛ̃	ɔ̃	œ̃	Ê	Ô	Û	μ	ə	
0	3	3	2	2	4	3	1	1	2	1	2	2	3	1	1	2	2	1	a
3	0	2	3	1	1	2	2	4	3	4	3	5	4	1	3	2	3	1	i
3	2	0	1	3	1	2	4	2	3	4	5	3	4	3	1	2	4	2	u
2	3	1	0	2	2	1	3	1	2	3	4	2	3	2	0	1	3	1	o
2	1	3	2	0	2	1	1	3	2	3	2	4	3	0	2	1	2	0	e
4	1	1	2	2	0	1	3	3	2	5	4	4	3	2	2	1	3	1	y
3	2	2	1	1	1	0	2	2	1	4	3	3	2	1	1	0	2	0	ϕ
1	2	4	3	1	3	2	0	2	1	2	1	3	2	0	2	1	1	0	ɛ
1	4	2	1	3	3	2	2	0	1	2	3	1	2	2	0	1	2	1	ɔ
2	3	3	2	2	2	1	1	1	0	3	2	2	1	1	1	0	1	0	œ
1	4	4	3	3	5	4	2	2	3	0	1	1	2	2	2	3	1	2	ã
2	3	5	4	2	4	3	1	3	2	1	0	2	1	1	3	2	0	1	ɛ̃
2	5	3	2	4	4	3	3	1	2	1	2	0	1	3	1	2	1	2	ɔ̃
3	4	4	3	3	3	2	2	2	1	2	1	1	0	2	2	1	0	1	œ̃
1	1	3	2	0	2	1	0	2	1	2	1	3	2	0	2	1	1	0	Ê
1	3	1	0	2	2	1	2	0	1	2	3	1	2	2	0	1	2	1	Ô
2	2	2	1	1	1	0	1	1	0	3	2	2	1	1	1	0	1	0	Û
2	3	4	3	2	3	2	1	2	1	1	0	1	0	1	2	1	0	1	μ
1	1	2	1	0	1	0	0	1	0	2	1	2	1	0	1	0	1	0	ə

A.2 Pseudo-Word Reduction: Extra plots

In this section of the annex, some extra plots are presented that correspond to the scores obtained from the pseudo-word reduction performed in chapter 4, section 4.5). The plots obtained from the two sets of 16 pseudo-words (double consonant in the beginning or middle), illustrated in figures A.1 and A.2

respectively, present a similar plot the one obtained with the full set of 52 pseudo-words (see figure 4.6). The same can be said for figure A.3. For the final plot (figure A.4), a more scattered score distribution can be found when compared to the others. This aspect comes as no surprise due to the small number of pseudo-words used, and the larger RMSE value found when compared to the previous ones.

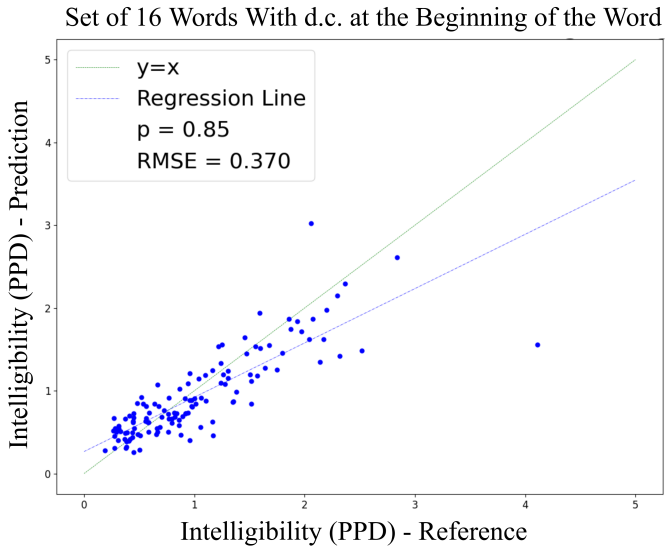


Figure A.1: Results of the automatic prediction of speech intelligibility using the Perceived Phonological Deviation (PPD) when using the subset of 16 pseudo-words with double consonant (d.c.) at the beginning of the word.

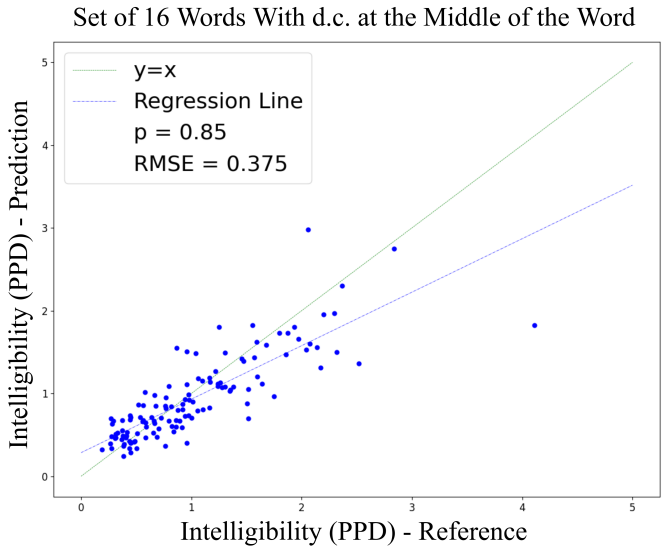


Figure A.2: Results of the automatic prediction of speech intelligibility using the Perceived Phonological Deviation (PPD) when using the subset of 16 pseudo-words with double consonant (d.c.) at the middle of the word.

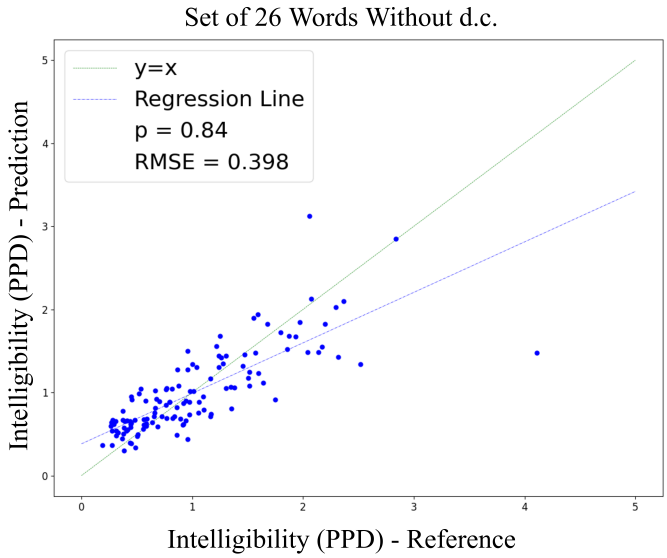


Figure A.3: Results of the automatic prediction of speech intelligibility using the Perceived Phonological Deviation (PPD) when using the subset of 26 pseudo-words without any double consonant (d.c.).

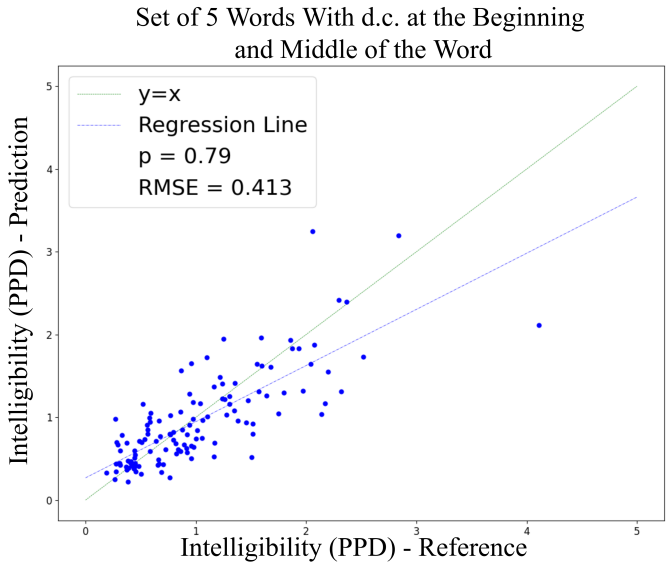


Figure A.4: Results of the automatic prediction of speech intelligibility using the Perceived Phonological Deviation (PPD) when using the subset of 5 pseudo-words with dual occurrences of double consonant (d.c.), in the beginning and middle of the word.

Appendix B

Forced Alignment

In the present annex, I will provide an overview of the working principle behind the forced alignment system used, the Montreal Forced Aligner (MFA) [McAuliffe et al., 2017]. Although not a contribution or central to the understanding of chapter 5, it can provide a simplified view behind the training of acoustic models, an essential part to not only forced alignment, but also automatic speech recognition (ASR). Despite the present thesis not making any contribution towards improving ASR, these systems were still used on multiple occasions as a baseline. Given this, this annex serves as an extra background on the root of either ASR and forced alignment: the acoustic model.

In the context of the MFA, this system passes through four primary stages of training, in order to fully develop the acoustic model. The first pass of the alignment uses monophone models. In this type of model, a single phone is modelled the same regardless of phonological context of the utterance that it is in. On the second pass, the system uses triphone models, where instead of operating on the isolated phone, the context on either side of the phone (silences included) is taken into consideration. The third pass performs a Linear Discriminant Analysis – Maximum Likelihood Linear Transform (LDA-MLLT) to learn a transform of the feature that maximizes the difference of the features between phonemes. This transformation builds Hidden Markov Model (HMM) states, but with a reduced feature space for all data. Finally, the final pass enhances the previous triphone model by taking into account speaker differences, and calculates a transformation of the Mel Frequency Cepstrum Coefficients (MFCC) features for each speaker. Seeing that the triphone model takes in phonetic context within a word, it is expected that it promotes better alignments than the monophone counterpart. The four training steps, from monophone to the triphone model with speaker differences accounted (fMLLR + SAT), can be found fully illustrated on figure B.1.

After obtaining the trained system, one can obtain the forced alignment timestamps by providing the system an audio file and a corresponding transcription. In our particular case (see chapter 5), the phonetic transcription was provided, and therefore the time intervals corresponding to each phonetic occurrence were obtained. For further reference to the MFA system see [McAuliffe et al., 2017][14].

[14] `https://kaldi-asr.org/doc/tree_externals.html`

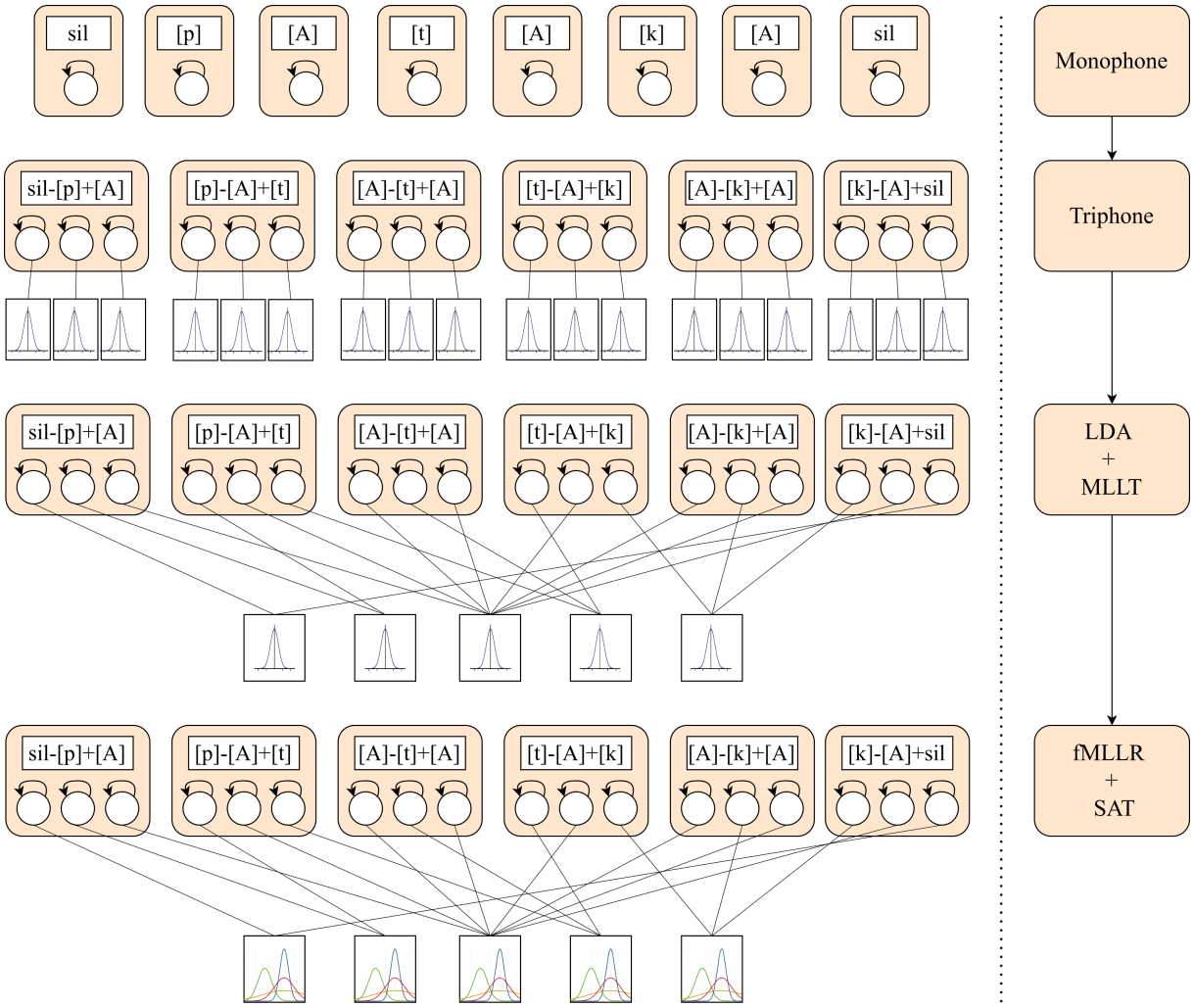


Figure B.1: Overview of the four stages of the acoustic model used to perform the forced alignment. Similarly to ASR, acoustic models are an essential part of forced alignment.

Appendix C

Global Outliers

In this annex, the outliers obtained from the experiments performed in chapter 7 are illustrated. Outliers were accounted for the speakers whose predictions had a mean absolute error superior to 2.0. A total of 15 outliers were found, that can be found presented on table C.1. Bold values mark either low reference intelligibility patients (column INT ref), higher standard deviations in the reference score (column STD ref) or a large Mean Absolute Error between the predicted and reference scores (column MAE).

C.1 Outliers - Unified Score

Table C.1: Outlier table from the unified score experiments. The same set of outliers was found in the mean and weighted mean approaches.

Name	Phoneme Score	Word Score	Sentence Score	INT ref	STD ref	INT pred	STD pred	MAE
AVG062	—	—	5.70	**3.50**	**1.98**	5.70	0.0	2.14
BOG109	—	—	6.90	**4.33**	**1.97**	6.90	0.0	2.50
BOM123	6.10	7.20	6.10	8.67	1.25	6.40	0.52	2.65
BOM094	2.50	6.90	5.00	**1.83**	1.07	4.75	1.80	2.92
CAA116	5.95	6.70	6.10	9.00	0.82	6.19	0.32	2.81
DES144	6.18	6.50	6.30	8.83	1.07	6.26	0.13	2.57
HEB114	5.83	7.40	7.30	9.40	0.80	6.78	0.72	2.63
JAR011	7.34	8.60	6.70	10.0	0.0	7.47	0.79	2.53
JUM122	4.36	7.10	6.60	8.67	0.94	5.96	1.19	2.71
MAF006	—	—	5.80	**1.50**	0.96	5.74	0.0	**4.24**
PEJ149	6.96	7.70	6.40	9.50	0.76	6.95	0.53	2.55
PRG014	3.87	5.70	3.10	6.83	0.69	4.18	1.09	2.65
REA127	6.11	8.20	6.90	9.17	0.69	7.00	0.86	2.17
SER007	3.83	6.20	6.00	**1.67**	**1.70**	5.29	1.07	**3.62**
SOM147	5.15	6.90	6.60	**2.67**	**1.70**	6.16	0.76	**3.49**

Glossary

Adam: Adaptive Movement Estimation Algorithm
AI: Artificial intelligence
ANN: Artificial Neural Network
ASR: Automatic Speech Recognition
CNN: Convolutional Neural Network
C2SI: French Corpus of Head and Neck Cancer
d.c.: Double Consonant
DES: Picture Description Task (C2SI corpus)
DNN: Depp Neural Network
GAN: Generative Adversarial Network
GeMAPS: Geneva Minimalist Acoustic Parameter Set
GMM: Gaussian Mixture Models
GMM-UBM: GMM - Universal Background Model
GRU: Gated Recurrent Unit
HC: Healthy Controls
HNC: Head and Neck Cancer
ICC: Intraclass Correlation Coefficient
INT: Speech Intelligibility
IPA: International Phonetic Alphabet
LEC: Passage Reading Task (C2SI corpus)
LSTM: Long Short-Term Memory Cell
MAD: Median Absolute Deviation
MAE: Maximum Absolute Error
MFA: Montreal Forced Aligner
MFCC: Mel Frequency Cepstral Coefficient
MTL: Multi-Task Learning Loss
MTL + DT: MTL with Dynamic tuning
NN: Neural Network
QoL: Quality of Life
ReLU: Rectified Linear unit
RNN: Recurrent Neural Network
P: Prosody
PD: Phonemic Distortions
Phoneme-SN System: Phoneme-level intelligibility prediction system
PPD: Perceived Phonological Deviation
prr: Phoneme Recognition Rate
R: Resonance
RIR: Room Impulse Responses
RMSE: Root Mean Squared Error
SAMPA: Speech Assessment Methods Phonetic Alphabet
Sentence-XVec System: Sentence-level intelligibility prediction system
SEV: Speech Disorder Severity
seq2seq: Sequence-to-Sequence

SGD: Stochastic Gradient Descent
SN: Shallow Neural Network
SNN: Shallow Neural Network
SPKR: Speaker
SVR: Support Vector Regression
TDNN: Time-Delayed Neural Network
tanh: hyperbolic tangent
V: Voice Quality
VAD: Vocal Activity Detection
Word-RNN System: Word-level intelligibility prediction system

Bibliography

[Abdel-Hamid et al., 2014] Abdel-Hamid, O., rahman Mohamed, A., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545.

[Abderrazek et al., 2020] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2020). Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders — step 1: CNN model-based phone classification. *Proceedings of Interspeech, Shanghai, China*, pages 2522–2526.

[Alim and Rashid, 2018] Alim, S. A. and Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms. *From Natural to Artificial Intelligence*.

[Alipour et al., 2012] Alipour, F., Scherer, R. C., and Finnegan, E. (2012). Measures of spectral slope using an excised larynx model. *Journal of Voice*, 26(4):403–411.

[Amberkar et al., 2018] Amberkar, A., Awasarmol, P., Deshmukh, G., and Dave, P. (2018). Speech recognition using recurrent neural networks. *Proceedings of ICCTCT, Coimbatore, India*, pages 1–4.

[Andersen et al., 2018] Andersen, A. H., de Haan, J. M., Tan, Z.-H., and Jensen, J. (2018). Nonintrusive speech intelligibility prediction using convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26:1925–1939.

[Arai et al., 2019] Arai, K., Araki, S., Ogawa, A., Kinoshita, K., Nakatani, T., Yamamoto, K., and Irino, T. (2019). Predicting speech intelligibility of enhanced speech using phone accuracy of DNN-based ASR system. *Proceedings of Interspeech, Graz, Austria*, pages 4275–4279.

[Badgeley et al., 2019] Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., and Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(31).

[Baevski et al., 2020] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of NeurIPS, Vancouver, Canada*, pages 12449–12460.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representations*.

[Balaguer et al., 2021] Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., and Woisard, V. (2021). Paramètres perceptifs expliquant la sévérité du trouble de parole mesurée automatiquement en cancérologie orl. *Rééducation orthophonique, Ortho édition*, pages 1–13.

[Balaguer et al., 2019] Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V., and Speyer, R. (2019). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review. *Journal of the Sciences and Specialities of Head and Neck*, 1:111–130.

[Bartko, 1966] Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*.

[Ben et al., 1996] Ben, C., Grice, M., and Hazan, V. (1996). The sus test ‘: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18(4):381–392.

[Bhattacharya1 et al., 2017] Bhattacharya1, G., Alam, J., and Kenny, P. (2017). Deep speaker embeddings for short-duration speaker verification. *Proceedings of Interspeech, Stockholm, Sweeden*, pages 1517–1521.

[Bonatti et al., 2005] Bonatti, L. L., Peña, M., Nespor, M., and Mehler, J. (2005). Linguistic constraints on statistical computations: the role of consonants and vowels in continuous speech processing. *Psychological Science*, 16(6):451–459.

[Bressmann, 2021] Bressmann, T. (2021). Speech disorders related to head and neck cancer. *Laryngectomy, Glossectomy, and Velopharyngeal and Maxillofacial Defects, chapter 22*.

[Burrus, 1985] Burrus, C. S. (1985). Convolution algorithms. *Citeseer: New York, NY, USA*.

[Calvo et al., 2021] Calvo, I., Tropea, P., Viganò, M., Scialla, M., Cavalcante, A., Grajzer, M., Gilardone, M., and Corbo, M. (2021). Evaluation of an automatic speech recognition platform for dysarthric speech. *Folia Phoniatrica et Logopaedica*, 73:432–441.

[Chakraborty et al., 2017] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., and Gurram, P. (2017). Interpretability of deep learning models: A survey of results. *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, San Francisco, CA, USA*, pages 1–6.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv preprint :1406.1078*.

[Chorowski et al., 2015] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*.

[Chou and Lee, 2019] Chou, H.-C. and Lee, C.-C. (2019). Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. *Proceedings of ICASSP, Brighton, united Kingdom*, pages 5886–5890.

[Christensen et al., 2012] Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Proceedings of Interspeech, Portland, OR, USA*, pages 1776–1780.

[Codosero et al., 2019] Codosero, J. M. P., Espinoza-Cuadros, F., Antón-Martín, J., Barbero-Alvarez, M. A., and Gómez, L. A. H. (2019). Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training. *IEEE Journal of Selected Topics in Signal Processing*, 14:240–250.

[Cohn and Specia, 2013] Cohn, T. and Specia, L. (2013). Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

[Crevier-Buchman et al., 2002] Crevier-Buchman, J, V., S, M., and D, B. (2002). Intelligibility of french consonants after partial supra-cricoid laryngectomy. *Revue de Laryngologie - Otologie - Rhinologie*, 123(5):307–310.

[Critchley, 1981] Critchley, E. M. (1981). Speech disorders of parkinsonism: a review. *Journal of Neurology Neurosurgery and & Psychiatry*, 44(9):751–758.

[de Bruijn et al., 2013] de Bruijn, M. J., Rinkel, R. N. P. M., Cnossen, I. C., Witte, B. I., Langendijk, J. A., Leemans, C. R., and de Leeuw, I. M. V. (2013). Associations between voice quality and swallowing function in patients treated for oral or oropharyngeal cancer. *Supportive Care in Cancer 21*, 21(7):2025–2032.

[de Graeff et al., 2000] de Graeff, A., de Leeuw, R. J., Ros, W. J., Hordijk, G.-J., Blijham, G. H., and Winnubst, J. A. (2000). Long-term quality of life of patients with head and neck cancer. *The Laryngoscope*, 110(1):98–106.

[der Burgh et al., 2017] der Burgh, H. K., Schmidt, R., Westeneng, H.-J., Reus, M. A., Leonard, den Berg, H., and den Heuvel, M. P. (2017). Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *NeuroImage: Clinical*, 13:361–369.

[Dick, 2019] Dick, S. (2019). Artificial intelligence. *Harvard Data Science Review*.

[DiPietro and Hager, 2020] DiPietro, R. and Hager, G. D. (2020). Chapter 21 - deep learning: RNNs and LSTM. *Handbook of Medical Image Computing and Computer Assisted Intervention*.

[Diprose et al., 2020] Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., and Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, 27(4):592–600.

[Dong et al., 2018] Dong, L., Xu, S., and Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

[Dong and Shen, 2018] Dong, X. and Shen, J. (2018). Triplet loss in siamese network for object tracking. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[dos Santos Barreto and Ortiz, 2019] dos Santos Barreto, S. and Ortiz, K. Z. (2019). Speech intelligibility in dysarthrias: Influence of utterance length. *Folia Phoniatrica et Logopaedica*, 72(3):202–210.

[Drucker et al., 1996] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapoik, V. (1996). Support vector regression machines. *Proceedings of Advances in Neural Information Processing Systems, Denver Colorado*, pages 155–161.

[Díaz and Antolín, 2020] Díaz, M. F. and Antolín, A. G. (2020). An attention long short-term memory based system for automatic classification of speech intelligibility. *Engineering Applications of Artificial Intelligence*, 96.

[Elias et al., 2021] Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R., and Wu, Y. (2021). Parallel tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling. *Proceedings of Interspeech, Brno, Czechia*, pages 141–145.

[Eyben et al., 2015] Eyben, F., Scherer, K., and Truong, K. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

[Fayek et al., 2016] Fayek, H. M., Lech, M., and Cavedon, L. (2016). Modeling subjectiveness in emotion recognition with deep neural networks: ensembles vs soft labels. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

[Fex, 1992] Fex, S. (1992). Perceptual evaluation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 6(2):155–158.

[Fogerty and Kewley-Port, 2009] Fogerty, D. and Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America*, 126:847–587.

[Fort et al., 2015] Fort, M., Martin, A., and Peperkamp, S. (2015). Consonants are more important than vowels in the bouba-kiki effect. *Lang Speech*, 51(2):247–266.

[Fredouille et al., 2019] Fredouille, C., Ghio, A., Laaridh, I., Lalain, M., and Woisard, V. (2019). Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. *International Congress of Phonetic Sciences (ICPhS)*.

[Galstyan and Cohen, 2007] Galstyan, A. and Cohen, P. R. (2007). Empirical comparison of "hard" and "soft" label propagation for relational classification. *Proceedings of International Conference on Inductive Logic Programming, Corvallis, Oregon, USA*, pages 98–111.

[Gelin et al., 2021] Gelin, L., Pellegrini, T., Pinquier, J., and Daniel, M. (2021). Simulating reading mistakes for child speech transformer-based phone recognition. *Proceedings of Interspeech, Brno, Czechia*, pages 3860–3864.

[Ghaffarian et al., 2021] Ghaffarian, S., Valente, J., van der Voort, M., and Tekinerdogan, B. (2021). Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. *Artificial Intelligence Algorithm for Remote Sensing Imagery Processing*, 13(15):2965.

[Ghio et al., 2020] Ghio, A., Lalain, M., Giusti, L., Fredouille, C., and Woisard, V. (2020). How to compare automatically two phonological strings: Application to intelligibility measurement in the case of atypical speech. *12th Conference on Language Resources and Evaluation (LREC), Marseille, France*, pages 1689–1694.

[Ghio et al., 2018] Ghio, A., Lalain, M., Giusti, L., Pouchoulin, G., Robert, D., Rebourg, M., Fredouille, C., Laaridh, I., and Woisard, V. (2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. *XXXIIe Journées d'Études sur la Parole, Aix-en-Provence, France*, pages 285–293.

[Ghoshal and Tucker, 2020] Ghoshal, B. and Tucker, A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv:2003.10769*.

[G.McCoy et al., 2021] G.McCoy, L., Brenna, C. T., Chen, S. S., Vold, K., and Das, S. (2021). Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, 142:252–257.

[Good, 2005] Good, P. (2005). Multivariate analysis. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, pages 169–188.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. *The MIT press*.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

[Graham et al., 2013] Graham, J. W., Cumsille, P. E., and Shevock, A. E. (2013). Methods for handling missing data. *Handbook of psychology: Research methods in psychology*.

[Guan et al., 2018] Guan, M., Gulshan, V., Dai, A., and Hinton, G. (2018). Who said what: modeling individual labelers improves classification. *Proceddings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

[Gwet, 2008] Gwet, K. L. (2008). Intrarater reliability. *Wiley encyclopedia of clinical trials*.

[Hady and Schwenker, 2013] Hady, M. F. A. and Schwenker, F. (2013). Semi-supervised learning. *Handbook on Neural Information Processing*, 49:215–239.

[Han et al., 2020] Han, J., Zhang, Z., Ren, Z., and Schuller, B. (2020). Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening. *Cognitive Computation*, 13:231–240.

[Han et al., 2017] Han, J., Zhang, Z., Schmitt, M., Pantic, M., and Schuller, B. (2017). From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. *Proceedings of ACM - International Conference on Multimedia, Mountain View California, USA*, pages 890–897.

[Hernandez et al., 2022] Hernandez, A., Perez-Toro, P. A., Noth, E., Orozco-Arroyave, J. R., Maier, A., and Yang, S. H. (2022). Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv:2204.01670*.

[Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv preprint :1207.0580*.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[Huang et al., 2022a] Huang, W.-C., Halpern, B. M., Violeta, L. P., Scharenborg, O., and Toda, T. (2022a). Towards identity preserving normal to dysarthric voice conversion. *Proceedings of ICASSP, Singapore*, pages 6672–6676.

[Huang et al., 2022b] Huang, W.-C., Halpern, B. M., Violeta, L. P., Scharenborg, O., and Toda, T. (2022b). Towards identity preserving normal to dysarthric voice conversion. *Proceedings of ICASSP, Singapore*, pages 6672–6676.

[Jacobi et al., 2013a] Jacobi, I., van Rossum, M. A., and van der Molen, L. (2013a). Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy. *The annals of Otology, Rhinology and Laryngology*, 12:754–762.

[Jacobi et al., 2013b] Jacobi, I., van Rossum, M. A., van der Molen, L., Hilgers, F. J. M., and van den Brekel, M. W. M. (2013b). Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy. *Annals of Otology, Rhinology & Laryngology*, 122(12):754–762.

[Janbakhshi et al., 2019a] Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019a). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. *Proceedings of ICASSP, Brighton, united Kingdom*, pages 6405–6409.

[Janbakhshi et al., 2019b] Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019b). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. *Proceedings of Interspeech, Graz, Austria*, pages 3038–3042.

[Jensen and Taal, 2016] Jensen, J. and Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE Transactions on Audio, Speech and Language Processing*, 24(11):2009–2022.

[Jiao et al., 2018] Jiao, Y., Tu, M., Berisha, V., and Liss, J. (2018). Simulating dysarthric speech for training data augmentation in clinical speech applications. *Proceedings of ICASSP, Calgary, AB, Canada*, pages 6009–6013.

[Kang, 2013] Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402–406.

[Karita et al., 2019] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., and Yoshimura, T. (2019). A comparative study on transformer vs RNN in speech applications. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

[Kent, 1992] Kent, R. D. (1992). Intelligibility in speech disorders: Theory, measurement and management. *John Benjamins*.

[Kent et al., 1994] Kent, R. D., Miolo, G., and Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology*, 3(2):81–95.

[Kent et al., 1989] Kent, R. D., Weismer, G., Kent, J. F., and Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54(4):482–499.

[Ketkar, 2017] Ketkar, N. (2017). Stochastic gradient descent. in: Deep learning with python. *https://doi.org/10.1007/978-1-4842-2766-4_8*.

[Kewley-Portand et al., 2007] Kewley-Portand, D., Burkle, T. Z., and Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 4:2365–2375.

[Khan et al., 2020] Khan, U., Pericás, H., and Javier, F. (2020). Unsupervised training of siamese networks for speaker verification. *Proceedings of Interspeech, Shanghai, China*, pages 3002–3006.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *ICLR*.

[Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. *Proceedings of Interspeech, Dresden, Germany*, pages 3586–3589.

[Kuyk et al., 2018] Kuyk, S. V., Kleijn, W. B., and Hendriks, R. C. (2018). An instrumental intelligibility metric based on information theory. *IEEE Signal Processing Letters*, 25(1):115–119.

[Laaridh et al., 2018] Laaridh, I., Fredouille, C., Ghio, A., Lalain, M., and Woisard, V. (2018). Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers. *Proceedings of Interspeech, Hyderabad, India*, pages 2943–2947.

[Laaridh et al., 2017] Laaridh, I., Kheder, W., Fredouille, C., and Meunier, C. (2017). Automatic prediction of speech evaluation metrics for dysarthric speech. *Proceedings of Interspeech, Stockholm, Sweeden*, pages 1834–1838.

[Lalain et al., 2020] Lalain, M., Ghio, A., Giusti, L., Robert, D., Fredouille, C., and Woisard, V. (2020). Design and development of a speech intelligibility test based on pseudowords in french: Why and how? *Journal of Speech, Language and Hearing Research*, 63(7):2070–2083.

[LeCun and Bengio, 1995] LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images and time-series. *MIT Press*.

[Leea et al., 2018] Leea, S. J., Pyob, H. Y., and Choi, H.-S. (2018). Normative data of cepstral and spectral measures in korean adults using vowel phonation and passage reading tasks. *Communication Sciences & Disorders*, 23:208–216.

[Liao et al., 2002] Liao, Y., Moody, J., and Wu, L. (2002). Chapter 9. *Handbook of Neural Network Signal Processing*, pages 37–45.

[Lieskovská et al., 2021] Lieskovská, E., Jakubec, M., Jarina, R., and Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Human Computer Interaction for Intelligent Systems*, 10:1163.

[Lin et al., 2021] Lin, S. Y., Cheng, S. C., and Si, D. (2021). Dementia detection using transformer-based deep learning and natural language processing models. *IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 509–510.

[Lin and Tseng, 2021] Lin, Y.-S. and Tseng, S.-C. (2021). Classifying speech intelligibility levels of children in two continuous speech styles. *Proceedings of ICASSP, Toronto, ON, Canada*, pages 7763–7767.

[Liu et al., 2021] Liu, S., Geng, M., Hu, S., Xie, X., and et al (2021). Recent progress in the CUHK dysarthric speech recognition system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2267–2281.

[Lodefoged, 1990] Lodefoged, P. (1990). The revised international phonetic alphabet. *Linguistic Society of America*, 6(3):550–552.

[Logemann et al., 1993] Logemann, J. A., Pauloski, B. R., Rademaker, A. W., and Johnson, J. (1993). Speech and swallow function after tonsil/base of tongue resection with primary closure. *Journal of speech and hearing research*, 36:918–926.

[Luz et al., 2020] Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *Proceedings of Interspeech, Shanghai, China.*

[Marczyk et al., 2020] Marczyk, A., Ghio, A., Lalain, M., Rebourg, M., Fredouille, C., and Woisard, V. (2020). Have a cake and eat it too: Assessing discrimination performance of an intelligibility index obtained from a reduced sample size. *12th Conference on Language Resources and Evaluation*, pages 1784–1788.

[McAuliffe et al., 2017] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of Interspeech, Stockholm, Sweden*, pages 498–502.

[Meyer et al., 2004] Meyer, T. K., Kuhn, J. C., Campbell, B. H., Marbella, A. M., Myers, K. B., and Layde, P. M. (2004). Speech intelligibility and quality of life in head and neck cancer survivors. *The Laryngoscope*, 114(11):1977–1981.

[Middag, 2012] Middag, C. (2012). *Automatic analysis of pathological speech.* Ghent University, Department of Electronics and information systems, Ghent, Belgium, Doctoral Dissertation.

[Middag et al., 2009a] Middag, C., Martens, J.-P., Nuffelen, G. V., and Bodt, M. D. (2009a). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing.*

[Middag et al., 2009b] Middag, C., Martens, J.-P., Nuffelen, G. V., and Bodt, M. D. (2009b). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing.*

[Miller, 2017] Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *arXiv:1706.07269.*

[Mirsamadi et al., 2018] Mirsamadi, S., Barsoum, E., and Zhang, C. (2018). Automatic speech emotion recognition using recurrent neural networks with local attention. *Proceedings of ICASSP.*

[M.Kates and H.Arehart, 2014] M.Kates, J. and H.Arehart, K. (2014). The hearing-aid speech perception index (haspi). *Speech Communication*, 65:75–93.

[Morales-Álvarez et al., 2019] Morales-Álvarez, P., Santos-Rodríguez, P. R. R., Molina, R., and Katsaggelos, A. K. (2019). Scalable and efficient learning from crowds with gaussian processes. *Information Fusion, volume 52.*

[Mueller and Thyagarajan, 2016] Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. *Proceedings of Thirtieth AAAI Conference on Artificial Intelligence*, 30(1):2786–2792.

[Mueller et al., 2018] Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40:917—-939.

[Mukaka, 2012] Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the International Conference on Machine learning (ICML), Haifa, Israel*, pages 807–814.

[Naqa and Murphy, 2015] Naqa, I. E. and Murphy, M. J. (2015). What is machine learning? *Machine Learning in Radiation Oncology.*

[Narasimhan and Vishal, 2017] Narasimhan, S. V. and Vishal, K. (2017). Spectral measures of hoarseness in persons with hyperfunctional voice disorder. *Journal of voice: official journal of the Voice Foundation*, 31(1):57–61.

[Navarro, 2001] Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys.*

[Ng and Lee, 2020] Ng, S.-I. and Lee, T. (2020). Automatic detection of phonological errors in child speech using siamese recurrent autoencoder. *Proceedings of Interspeech, Shanghai, China*, pages 4476–4480.

[Nuffelen et al., 2008] Nuffelen, G. V., Middag, C., Bodt, M. D., and Martens, J.-P. (2008). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders*, 48(6):716–730.

[Owren and Cardillo, 2006] Owren, M. J. and Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of The Acoustical Society of America*, 119(3):1727–1739.

[Pappagari et al., 2020] Pappagari, R., Wang, T., Villalba, J., Chen, N., and Dehak, N. (2020). X-vectors meet emotions: A study on dependencies between emotion and speaker recognition. *Proceedings of ICASSP, Barcelona, Spain.*

[Peddinti et al., 2015] Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. *Proceedings of Interspeech, Dresden, Germany*, pages 3214–3218.

[Pedersen et al., 2020] Pedersen, M., Kolbaek, M., Andersen, A. H., Jensen, S. H., and Jensen, J. (2020). End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks. *Proceedings of Interspeech, Shanghai, China*, pages 1151–1155.

[Pham-Gia and Hung, 2001] Pham-Gia, T. and Hung, T. (2001). The mean and median absolute deviations. *Mathematical and Computer Modelling*, 34:921–936.

[Plisson et al., 2017] Plisson, L., Pillot-Loiseau, C., and Crevier-Buchman, L. (2017). Intelligibilité de la parole après le traitement d'uncancer de l'oropharynx: étude descriptive chez sept patients en prétraitement et en post-traitement précoce. *7èmes Journées de phonétique clinique (JPC), Laboratoire de Phonétiqueet Phonologie, hôpital Européen G. Pompidou, Paris, France.*

[Pommée et al., 2022] Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., and Woisard, V. (2022). Intelligibility and comprehensibility: A delphi consensus study. *International Journal of Language & Communication Disorders*, 57(1):21–41.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

[Quintas et al., 2022a] Quintas, S., Abad, A., Mauclair, J., Woisard, V., and Pinquier, J. (2022a). Utilisation de modèles transformers pour la prédictionde l'intelligibilité de la parole de patients atteintsde cancers des voies aérodigestives supérieures. *XXXIVe Journées d'Études sur la Parole, Île de Noirmoutier, France*.

[Quintas et al., 2020] Quintas, S., Mauclair, J., Woisard, V., and Pinquier, J. (2020). Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer. *Proceedings of Interspeech, Shanghai, China*, pages 4976–4980.

[Quintas et al., 2022b] Quintas, S., Mauclair, J., Woisard, V., and Pinquier, J. (2022b). Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer. *Proceedings of Interspeech, Incheon, South Korea*, pages 3608–3612.

[Ravanelli et al., 2020] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. *Proceedings of ICASSP, Barcelona, Spain*, pages 6989–6993.

[Ravì et al., 2017] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21.

[Raykar et al., 2010] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

[Reddi et al., 2018] Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of Adam and beyond. *ICLR*.

[Rizos and Schuller, 2020] Rizos, G. and Schuller, B. W. (2020). Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. *IPMU - International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 42–55.

[Rodrigues and Pereira, 2019] Rodrigues, F. and Pereira, F. (2019). Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1611–1618.

[Roger et al., 2022] Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2022). Création d'une mesure entropique de la parole pour évaluer l'intelligibilité de patients atteints de cancers des voies aérodigestives supérieures. *XXXIVe Journées d'Études sur la Parole, Île de Noirmoutier, France*, pages 117–125.

[Saravanan et al., 2016] Saravanan, G., Ranganathan, V., Gandhi, A., and Jaya, V. (2016). Speech outcome in oral cancer patients - pre- and post-operative evaluation: A cross-sectional study. *Indian J Palliat Care*, 22(4):499–503.

[Schneider et al., 2019] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). Wav2Vec: Unsupervised pre-training for speech recognition. *Proceedings of Interspeech, Graz, Austria*, pages 3465–3469.

[Schutte et al., 2020] Schutte, K., Moindrot, O., Hérent, P., Schiratti, J.-B., and Jégou, S. (2020). Using stylegan for visual interpretability of deep learning models on medical images. *arXiv:2101.07563*.

[Serrano and Smith, 2019] Serrano, S. and Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

[Serrà et al., 2021] Serrà, J., Pons, J., and Pascual, S. (2021). SESQA: Semi-supervised learning for speech quality assessment. *Proceedings of ICASSP, Toronto, ON, Canada*, pages 381–385.

[Shun-ichiAmari, 1993] Shun-ichiAmari (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*.

[Snyder et al., 2015] Snyder, D., Chen, G., and Povey, D. (2015). MUSAN: A music, speech, and noise corpus. *arXiv:1510.08484v1*.

[Snyder et al., 2018a] Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. (2018a). Spoken language recognition using x-vectors. *Proceedings of Interspeech, Hyderabad, India*, pages 105–111.

[Snyder et al., 2017] Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. *Proceedings of Interspeech, Stockholm, Sweeden*, pages 999–1003.

[Snyder et al., 2018b] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018b). X-vectors: Robust DNN embedings for speaker recognition. *Proceedings of ICASSP, Calgary, AB, Canada*, pages 5329–5333.

[Taal et al., 2011] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2125–2136.

[Tian and Zhu, 2012] Tian, Y. and Zhu, J. (2012). Learning from crowds in the presence of schools of thought. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[Tjaden and Wilding, 2004] Tjaden, K. and Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria. *Journal of Speech, Language and Hearing Research*, 47(4):766–783.

[Tschiatschek et al., 2018] Tschiatschek, S., Singla, A., Rodriguez, M. G., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. *WWW '18: Companion Proceedings of the The Web Conference, Geneva, Switzerland*, pages 517–524.

[Vachhani et al., 2018] Vachhani, B., Bhat, C., and Kopparapu, S. K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. *Proceedings of Interspeech, Stockholm, Sweeden*, pages 471–475.

[van den Oordan et al., 2016] van den Oordan, A., Ze, S. D. H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[Vasquez-Correa et al., 2019] Vasquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., and Noth, E. (2019). Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech. *Proceedings of Interspeech, Graz, Austria*, pages 549–553.

[Vaswani et al., 2017] Vaswani, A., Parmar, N. S. N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing System, Long Beach, CA, USA*.

[Verma and Das, 2015] Verma, P. and Das, P. (2015). i-vectors in speech processing applications: A survey. *International Journal of Speech Technology*, 18:529–546.

[Waibel, 1989] Waibel, A. (1989). Modular construction of time-delay neural networks for speech recognition. *Neural Computation*, 1(1):39–46.

[Wan et al., 2018] Wan, L., Wang, Q., Papir, A., and Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. *Proceedings of ICASSP, Calgary, AB, Canada*.

[Wang and Liu, 2021] Wang, F. and Liu, H. (2021). Understanding the behaviour of contrastive loss. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wang et al., 2019] Wang, J., Qin, Y., Peng, Z., and Lee, T. (2019). Child speech disorder detection with siamese recurrent network using speech attribute features. *Proceedings of Interspeech, Graz, Austria*, pages 3885–3889.

[Woisard et al., 2020] Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D., and Roger, V. (2020). C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55:173–190.

[Xiong et al., 2019] Xiong, F., Barker, J., and Christensen, H. (2019). Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. *Proceedings of ICASSP, Brighton, united Kingdom*, pages 5836–5840.

[Xue et al., 2019] Xue, W., Cucchiarini, C., van Hout, R., and Strik, H. (2019). Acoustic correlates of speech intelligibility. the usability of the eGeMAPS feature set for atypical speech. *Proceedings of SLaTE: ISCA Workshop on Speech and Language Technology in Education*, pages 48–52.

[Yan et al., 2010] Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., and Dy, J. (2010). Modeling annotator expertise: Learning when everybody knows a bit of something. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy*, pages 932–939.

[Yorkston et al., 1996] Yorkston, K. M., Strand, E. A., and Kennedy, M. R. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, 5(1):55–66.

[Yu1 et al., 2018] Yu1, J., Xie, X., Liu, S., Hu1, S., Max.W.Y.LAM, Wu, X., Wong, K. H., Liu, X., and Meng, H. (2018). Development of the cuhk dysarthric speech recognition system for the UASpeech corpus. *Proceedings of Interspeech, Hyderabad, India*, pages 2938–2942.

[Zargarbashi and Babaali, 2019] Zargarbashi, S. and Babaali, B. (2019). A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language. *arXiv:1910.00330*.

[Zhang and Yang, 2018] Zhang, Y. and Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, 5(1):30–43.

[Zhao et al., 2018] Zhao, H., Zarar, S., Tashev, I., and Lee, C.-H. (2018). Convolutional-recurrent neural networks for speech enhancement. *Proceedings of ICASSP, New Orleans, LA, USA*, pages 2227–2231.

[Álvaro Barbero Jiménez et al., 2007] Álvaro Barbero Jiménez, Lázaro, J. L., and Dorronsoro, J. R. (2007). Finding optimal model parameters by discrete grid search. *AINSC - Advances in soft Computing*, 44:120–127.

Résumé long

La prédiction automatique de l'intelligibilité de la parole peut être considérée comme une alternative pertinente aux évaluations perceptives réalisées en clinique. Ces évaluations sont connues pour être subjectives, biaisées et variables, puisque l'évaluation peut être conditionnée par une variété d'aspects, tels que la connaissance préalable du patient ou la tâche d'évaluation de la parole elle-même. Dans ces conditions, le développement d'une approche automatique peut être considérée comme une alternative plus robuste, reproductible et objective.

Dans ce document, nous présentons différentes approches pour la prédiction de l'intelligibilité de la parole basées sur les différents niveaux de granularité d'analyse, que sont la phrase, le mot et le phonème. Étant donné la complexité du processus qui sous-tend les évaluations cliniques dans lesquelles l'intelligibilité est évaluée, il devient pertinent d'obtenir une mesure automatique qui soit non seulement capable d'être hautement corrélée aux évaluations perceptives, mais aussi facilement explicable et reproductible. L'approche granulaire, décrite dans l'illustration 1, peut résoudre ce problème en évaluant différentes parties de la communication parlée, et fournir un score qui peut évaluer l'intelligibilité d'un point de vue général (par exemple, la phrase) jusqu'aux unités vocales les plus irréductibles (par exemple, les phonèmes). Tous ces aspects sont connus pour être fortement liés à l'intelligibilité de la parole.

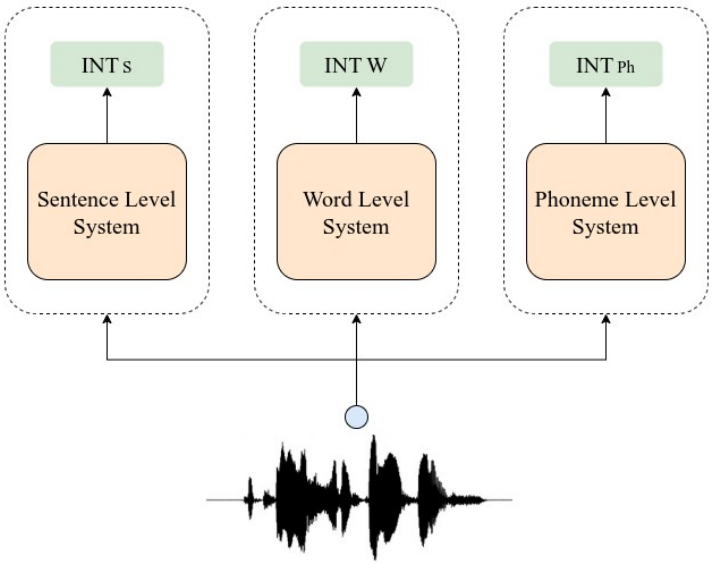


Figure 1 : Aperçu de la prédiction de l'intelligibilité aux trois niveaux granulaires de la phrase, du mot et du phonème.

Dans la littérature, on peut trouver différentes approches pour prédire automatiquement l'intelligibilité de la parole. Ces approches peuvent promouvoir des scores plus fiables, objectifs et reproductibles [1, 2]. Afin de prédire automatiquement l'intelligibilité de la parole, on peut utiliser des approches qui régressent un score à partir du taux d'erreurs de mots obtenu via un ASR (reconnaissance automatique de la parole), cependant ce type d'approche a tendance à être moins performant sur les locuteurs ayant des déficiences sévères de la parole [3]. D'autres approches font appel à des technologies de traitement de la parole basées sur l'extraction de caractéristiques pertinentes, telles que les MFCC, les banques de filtres, etc. [4, 5]. Néanmoins, il est connu que ces types de systèmes basés sur les données ont tendance à nécessiter des quantités importantes de données afin d'être efficaces [6]. L'évaluation perceptive de l'intelligibilité de la parole diffère également de celle appliquée au paradigme de la parole dans le bruit [7]. Si la définition de l'intelligibilité peut être similaire dans les deux cas, le décodage perceptif diffère entre les deux. Les prédicteurs d'intelligibilité traditionnels utilisés dans l'amélioration de la

parole, tels que STOI ou E-STOI, nécessitent l'utilisation de signaux alignés dans le temps, même dans les approches end-to-end [8], ce qui est irréalisable pour la parole pathologique. Des travaux récents, tels que [9] et [10], ont utilisé des systèmes d'estimation de similarité, tels que les réseaux siamois, dans un contexte de parole pathologique. Dans les deux cas, les systèmes ont été développés pour la détection des troubles de la parole chez les enfants, en se concentrant sur la tâche binaire de détection de mauvaises prononciations spécifiques. Ces systèmes deviennent pertinents dans l'avènement d'un système d'intelligibilité qui opère au niveau du phonème, le dernier niveau de granularité mentionné. Cette analyse automatique basée sur la phonétique est importante car les phonèmes individuels, en particulier les consonnes, sont très importants pour l'intelligibilité perçue [11,12,13].

Dans le présent travail, nous visons à explorer trois questions de recherche distinctes dérivées du sujet en question :

1 - ***L'apprentissage profond peut-il être utilisé de manière fiable pour prédire l'intelligibilité de la parole ?***

2 - ***Comment pouvons-nous établir la confiance dans ces systèmes pour qu'ils soient appliqués cliniquement ?***

3 - ***Une approche granulaire peut-elle fonctionner pour prédire l'intelligibilité de la parole ? Si oui, quels sont les avantages supplémentaires ?***

Les sections suivantes décrivent les trois systèmes mis en œuvre aux niveaux de granularité susmentionnés. Tous les systèmes sont accompagnés d'un diagramme schématique pour faciliter la compréhension, suivi d'un tableau de résultats et de références supplémentaires. Le développement du présent travail s'est appuyé sur le Corpus français du cancer de la tête et du cou (C2SI) [14]. Ce corpus contient une variété de tâches vocales enregistrées et d'évaluations perceptives, qui seront également présentées avec chacun des systèmes.

Prédiction automatique de l'intelligibilité de la parole au niveau de la phrase

En suivant les niveaux de granularité proposés, nous commençons par présenter le système qui opère au niveau de la phrase : le système Sentence-Xvec. Il est clair que ce niveau de granularité joue un rôle important dans l'intelligibilité de la parole, puisque c'est celui qui imite le mieux les évaluations perceptives cliniques habituellement réalisées, ainsi que la communication quotidienne.

Lors du développement de ce système, les enregistrements effectués lors de la tâche de lecture de passages du Corpus français du cancer de la tête et du cou (C2SI) sont utilisés comme signal d'entrée. Ce choix permet d'uniformiser le contenu phonétique émis pour les différents locuteurs. Comme mesure d'intelligibilité de référence, l'évaluation perceptive obtenue par 6 juges sur la tâche de description d'images du même corpus a été privilégiée. Ce choix de score de référence permet d'atténuer le biais lié à la connaissance préalable de la tâche de lecture du passage par les juges, rendant ce score plus robuste.

En ce qui concerne la méthodologie utilisée, nous avons utilisé le paradigme des *x-vectors* speaker embeddings [15]. Ces représentations de longueur fixe, d'abord adaptées à la reconnaissance automatique de la parole, visent à représenter les caractéristiques du locuteur par un vecteur de faible dimension, et peuvent avoir de nombreuses applications pour le traitement automatique de la parole pathologique. Les embeddings sont extraits au moyen d'un réseau neuronal profond pré-entraîné, composé de 5 réseaux neuronaux temporisés (TDNN), d'une couche de mise en commun des statistiques, de 2 couches entièrement connectées et d'une noramlisation de type softmax. La méthodologie globale du système Sentence-Xvec, utilisé pour la prédiction de l'intelligibilité, est présentée dans l'illustration 2. Les *x-vectors* sont extraits de la tâche de lecture de passages segmentés mentionnée précédemment. Ils sont en entrée d’un réseau neuronal peu profond pour prédire automatiquement l'intelligibilité de la parole. Ce dernier réseau contient une couche d'entrée et deux couches entièrement connectées, qui utilisent des fonctions ReLU comme non-linéarités.

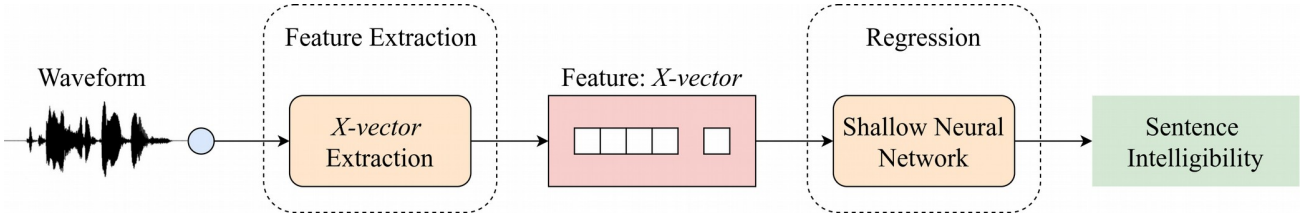


Figure 2 : Vue d'ensemble du système de prédiction de l'intelligibilité au niveau des phrases.

Le système Sentence-Xvec a été entraîné en utilisant la tâche de lecture de passages segmentés du corpus C2SI. Au total, 105 locuteurs, 21 témoins et 84 patients présentant différents degrés d'intelligibilité de la parole, ont été utilisés. Afin d'augmenter artificiellement le volume de données durant la phase d'apprentissage et accentuer le facteur de variabilité, les passages ont été segmentés 8 segments individuels. Le système a été entraîné en utilisant un schéma de validation croisée à 5 niveaux : à chaque étape, 21 locuteurs ont été laissés de côté pour les tests, choisis de manière aléatoire. Le système a été évalué à l'aide du coefficient de corrélation de Spearman (p) et de l'erreur quadratique moyenne (RMSE), calculé entre les scores prédits et et les mesures d'intelligibilité perceptive obtenues à partir de la tâche de description d'images du corpus C2SI. Deux types de score ont été expérimentés, les résultats sont illustrées dans le tableau 1. Le premier correspond à la moyenne de la prédiction individuelle sur chacun des 8 segments individuels du locuteur : un gain de corrélation élevé et une erreur réduite sont observés par rapport à une approche précédente classique basée sur le paradigme *i-vector*. L'examen des scores sur chaque segment individuel a démontré que le score d'intelligibilité automatique basé uniquement sur le segment 2 favorise une corrélation légèrement meilleure et une erreur plus faible que l'approche moyenne.

Table 1: Scores obtenus par le système de prédiction de l'intelligibilité au niveau des phrases.

Type de Système		p	RMSE
I-Vectors	Scores Moyens	0.72	2.121
X-Vectors: Système Sentence-Xvec	Scores Moyens	0.81	1.716
X-Vectors: Système Sentence-Xvec	Segment 2	0.82	1.434

Les résultats présentés par le système Sentence-Xvec suggèrent que nous pouvons prédire de manière fiable l'intelligibilité de la parole au niveau de la phrase en utilisant le paradigme des *x-vector* speaker embeddings. De plus, le choix du segment utilisé peut influencer le score final. Ici, le segment 2 a montré le meilleur gain de corrélation. Le choix du segment peut avoir un impact sur la prédiction automatique de l'intelligibilité de la parole, montrant que pour chaque locuteur, il y a des phrases qui sont capables de transmettre une estimation plus précise de l'intelligibilité.

Les résultats présentés ont été acceptés à Interspeech 2020 sous le nom de "Automatic Prediction of Speech Intelligibility based on X-vectors in the context of Head and Neck Cancer" [16].

Prédiction automatique de l'intelligibilité de la parole au niveau du mot

Le niveau de granularité correspondant au mot est considéré comme un niveau intermédiaire entre la phrase et le phonème. La principale pertinence d'une évaluation à ce niveau réside dans l'analyse des co-articulations contextuelles spécifiques entre phonèmes consécutifs, qui sont liées à l'intelligibilité de la parole. Cette analyse peut également être considérée comme plus objective, ce qui est très pertinent lors du développement d'outils destinés à une application clinique.

Lors du développement de ce système, la tâche des pseudo-mots du corpus C2SI a été utilisée. Dans cette tâche, on a demandé aux locuteurs d'enregistrer 52 pseudo-mots qui, bien qu'inexistants, respectent les règles phonotactiques et orthographiques du français. Un score d'intelligibilité basé sur la déviation phonologique perçue (PPD) a été attribué à chaque pseudo-mot. Ce score a été obtenu à partir de la distance entre le pseudo-mot transcrit (par 3 juges naïfs) et la vérité terrain originale, en prenant en compte un coût matriciel des voyelles et des consonnes. Chaque pseudo-mot suit une structure de C(C)1V1C(C)2V2, où C(C)i est soit une consonne unique soit un groupe de consonnes et Vi une voyelle unique. Chaque ensemble de 52 pseudo-mots comporte un sous-ensemble de 16 mots avec une occurrence de double consonne (d.c.) au début, 16 mots avec une double consonne au milieu, et 5 avec les deux [17].

La méthodologie utilisée a fait appel à un réseau de neurones récurrents (Recurrent Neural Network) couplé à un mécanisme d'auto-attention. Ce système appelé Word-RNN qui se trouve dans l'illustration 3, reçoit des ensembles de prononciations de pseudo-mots comme entrées de fichiers audio, où 40 bancs de filtres sont extraits comme caractéristiques. Le système utilise un bloc récurrent bidirectionnel avec 3 unités récurrentes gated (GRU), chacune ayant une dimension de 40 et une dimension cachée de 100. Un mécanisme d'auto-attention est ajouté à la sortie du bloc récurrent, qui permet au système de se concentrer davantage sur des parties particulières de l'entrée, tout en ignorant les parties moins pertinentes. Après le mécanisme d'attention, le vecteur de longueur fixe obtenu est transmis à un ensemble neuronal formé de trois couches entièrement connectées, avec une dimension de 100 unités chacune et des fonctions de type ReLU comme non-linéarités. Enfin, une couche de mutualisation Global Max est utilisée pour obtenir le score individuel pour chaque pseudo-mot.

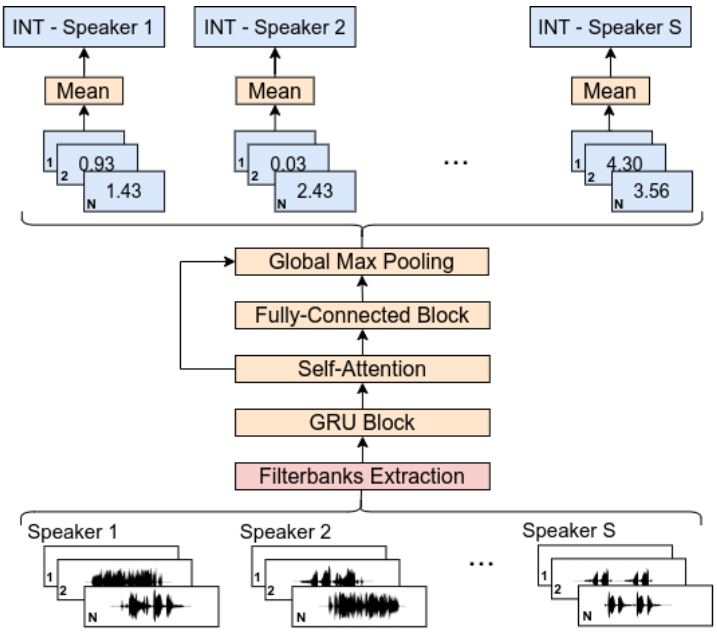


Figure 3 : Vue d'ensemble du système Word-RNN, utilisé pour prédire l'intelligibilité au niveau des mots. INT représente le score d'intelligibilité. N correspond au nombre de pseudo-mots utilisés pour le score.

Le système a été entraîné avec un schéma de validation croisée 10 fois. Un total de 126 locuteurs a été utilisé, correspondant aux locuteurs qui ont enregistré la tâche de pseudo-mots. À chaque étape de validation, 113 locuteurs, patients et témoins, ont été utilisés pour l'entraînement, tandis que les 13 autres ont été utilisés pour le test. Le système a été entraîné avec 200 itérations. Un taux d'apprentissage de 0,0001 a été utilisé avec une décroissance polynomiale jusqu'à 0,0001 pendant les 50 premières époques. Une taille de lot de 16 et l'optimiseur Adam ont été utilisés pendant l'apprentissage. Comme pour le système précédent, le *p* de Spearman et le RMSE ont été utilisés comme métriques d'évaluation. Les résultats sont présentés dans le tableau 2, suivis des résultats en limitant l'évaluation du score sur les sous-ensembles de pseudo-mots comportant des occurrences de consonnes doubles, mentionnés précédemment. En outre, pour chaque locuteur, différents ensembles réduits de pseudo-mots ont été expérimentés au lieu de l'ensemble complet de 52 pseudo-mots. Ces ensembles correspondent directement aux différentes structures de pseudo-mots trouvées dans le corpus, basées sur les occurrences de consonnes doubles.

Table 2: Les scores obtenus avec le modèle proposé, suivis des scores obtenus par la réduction de la quantité de pseudo-mots utilisés lors de l'inférence.

Modèle	**Quantité des pseudo-mots utilisés**	***p***	**RMSE**
Reconaissance automatique de la parole (Wagner-Fischer)	52 (total)	0.72	0.792
X-vector Speaker Embeddings		0.80	0.447
	52 (total)	0.87	0.370

Word-RNN System	16 avec d.c. au début	0.85	0.370
	16 avec d.c. au milieu	0.85	0.375
	26 sans d.c.	0.84	0.398
	5 avec d.c. au début et milieu	0.79	0.413

Les résultats généraux suggèrent une corrélation élevée par rapport à deux approches précédentes, ainsi qu'une erreur plus faible. La première approche (Wagner-Fischer) correspond à la ligne de base précédente basée sur la distance entre la transcription automatique des pseudo-mots et la vérité terrain. La deuxième approche correspond au système précédent, basé sur les encastrements de locuteurs du vecteur x. De plus, les ensembles réduits de pseudo-mots suggèrent que le système peut atteindre des performances similaires en utilisant de plus petites quantités de données (jusqu'à moins de 10% de la quantité originale de pseudo-mots), tout en restant comparable au système général et aux deux lignes de base. Cet aspect devient très pertinent sur le plan clinique. D'après les valeurs de corrélation et d'erreur obtenues à partir des ensembles réduits, nous pouvons conclure sans risque que les mots comportant des occurrences de double-consonnes sont très pertinents pour l'intelligibilité de la parole, et devraient être privilégiés lorsqu'ils sont utilisés dans des évaluations similaires.

Les résultats présentés ici ont été acceptés et présentés aux Journées des Études sur la Parole 2022 (JEP), sous le nom de publication "Utilisation de modèles Transformers pour la prédiction de l'intelligibilité de la parole de patients atteints de cancers des voies aérodigestives supérieures" [18].

Les scripts utilisés pour développer le système proposé peuvent également être trouvés dans le dépôt en ligne suivant : https://github.com/Elquintas/Self-Attention-Speech-intelligibility.

Prédiction automatique de l'intelligibilité de la parole au niveau du phonème

Le troisième et dernier niveau de granularité exploré correspond au phonème. Les phonèmes pouvant être considérés comme les fondements de la communication orale, une analyse de l'intelligibilité à ce niveau peut fournir des indices intéressants. Un score automatique basé sur une analyse phonémique peut également être considéré comme plus objectif et interprétable, du fait qu'il peut être dérivé directement des phonèmes eux-mêmes.

Afin d'obtenir un système de qualité qui effectue une analyse de l'intelligibilité au niveau des phonèmes, il faut d'abord obtenir des phonèmes. Dans ce cas, comme pour le système précédent, nous avons utilisé la tâche des pseudo-mots du corpus C2SI, qui contient plusieurs occurrences de phonèmes français dans différents contextes. Parmi les différents phonèmes, il est supposé que les consonnes jouent un rôle très important pour l'intelligibilité de la parole étant donné qu'elles sont capables de transmettre des informations sur le mot au lieu des informations sur le locuteur plus facilement transmises par les voyelles. Compte tenu de cela et de la structure des pseudo-mots utilisés, le présent système utilise uniquement les consonnes pour l'estimation de l'intelligibilité. Pour comparer les prédictions d'intelligibilité obtenues, nous avons utilisé l'intelligibilité perceptive obtenue à partir de la description d'images comme référence (de la même manière que le système Sentence-XVec), dans le contexte du corpus C2SI.

Le système Phoneme-SN est basé sur un réseau de neurones siamois pour détecter le niveau de similarité entre les phonèmes de même classe (voir illustration 4) et sur la quantité de phonèmes similaires/dissimilaires d'un locuteur donné pour prédire un score d'intelligibilité. Le réseau siamois est composé de deux encodeurs GRU (gated recurrent unit) bidirectionnels avec des poids partagés, qui reçoivent comme caractéristiques 13 coefficients cepstraux selon l’échelle fréquentielle Mel (MFCC) extraits des phonèmes d'entrée. Chaque encodeur est composé de 2 couches GRU bidirectionnelles avec une dimension cachée de 100 chacune, et produit une représentation de longueur fixe. En outre, le réseau siamois calcule la différence absolue entre ces deux représentations et transmet les résultats à un bloc de réseau neuronal profond composé de 3 couches entièrement connectées de taille 200. Enfin, une fonction sigmoïde est ajoutée pour indiquer le niveau de similarité entre les deux phonèmes d'entrée. Le score d'intelligibilité pour chaque locuteur est déduit de la quantité de phonèmes similaires aux représentations de phonèmes canoniques, émis par les locuteurs de contrôle. Plus la quantité de phonèmes similaires est grande, plus le locuteur est intelligible.

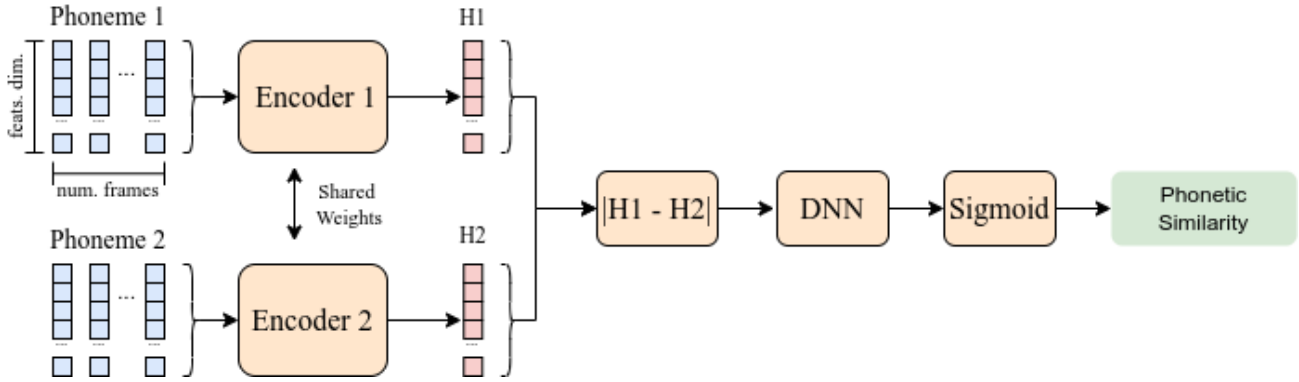


Figure 4: Schéma du système phonème-SN proposé.

Le système Phoneme-SN est entraîné en utilisant les phonèmes prononcés par des locuteurs contrôles (témoins) du corpus C2SI extraits par alignement forcé. 24 témoins sont utilisés pour l'entraînement, 6 patients présentant des valeurs d'intelligibilité élevées (non distinguables des témoins sains) sont utilisés pour la validation et 72 patients pour le test. Le système est entraîné et validé en utilisant des paires de phonèmes identiques, pour imiter des phonèmes similaires, et des paires de phonèmes différents, pour imiter des phonèmes pathologiques. Un modèle individuel est ajusté pour chacune des 16 consonnes françaises. La similarité est évaluée par la fonction sigmoïde à la fin du réseau siamois : un phonème test est considéré comme similaire si la similarité est supérieure à un seuil défini par la similarité médiane des phonèmes de validation (de cette même classe) moins la déviation absolue médiane. Le système n'utilise pas de mesures d'intelligibilité perceptive pendant l'entraînement, et celles-ci ne sont utilisées que pour la comparaison avec les mesures automatiques. Les résultats, illustrés dans le tableau 4, suggèrent une corrélation de base élevée de 0,82, et une erreur plus faible par rapport à deux approches précédentes, l'une basée sur la reconnaissance automatique de la parole et la seconde basée sur le paradigme des *x-vector* speaker embeddings, illustré précédemment. La même tâche enregistrée et la même mesure d'intelligibilité de référence sont utilisées pour le système et les lignes de base précédentes.

Table 3: Résultats de corrélation et RMSE obtenus par le système Phoneme-SN, comparés à deux approches précédentes.

Model	***p***	**RMSE**
Reconnaissance automatique de la parole (Wagner-Fischer)	0.63	2.406
X-vector Speaker Embeddings	0.74	2.488
Phoneme-SN	0.82	2.350

Le présent système peut être considéré comme une approche plus objective pour prédire automatiquement l'intelligibilité de la parole, puisque le score obtenu peut être entièrement relié à la quantité de consonnes similaires ou dissemblables. Les résultats obtenus, malgré l'erreur plus importante par rapport aux deux systèmes granulaires précédemment décrits montrent qu'un système performant peut être réalisé en fournissant une approche interprétable, tout en restant en bonne corrélation avec les mesures perceptives obtenues cliniquement.

Les résultats affichés ont été acceptés et présentés à Interspeech 2022 sous le nom de "Automatic Assessment of Speech Intelligibility using Consonant Similarity for Head and Neck Cancer" [19].

Les scripts utilisés pour le développement du modèle peuvent également être trouvés en ligne, sous le dépôt en ligne suivant: https://github.com/Elquintas/siamese_intel

Conclusions

Le travail développé au cours de cette thèse a exploré la prédiction automatique de l'intelligibilité de la parole en utilisant des méthodologies d'apprentissage profond. Plus précisément, au cours de ce travail, j'ai développé une variété de systèmes qui prédisent l'intelligibilité de la parole à trois niveaux granulaires distincts : la phrase, le mot et le phonème. L'approche

granulaire promue répond à une lacune de la littérature concernant le besoin de prédictions automatiques fiables et explicables de l'intelligibilité de la parole. Les systèmes développés ont montré des gains de performance intéressants et élevés par rapport aux approches et lignes de base issues de la littérature.

Ce travail me permet de répondre aux questions de recherche soulignées au début de ce document. Pour répondre à la première question, *à savoir si nous pouvons réellement utiliser l'apprentissage profond pour prédire l'intelligibilité de la parole*, j'ai proposé un ensemble de modèles tous issus de l'apprentissage profond capables de promouvoir des corrélations élevées avec les mesures perceptives. Ces résultats ont montré qu'en utilisant des méthodes créatives (par exemple, l'augmentation des données, différentes fonctions de notation de l'intelligibilité, etc.), nous pouvons atténuer le besoin de grandes quantités de données généralement requises par les systèmes d'apprentissage profond, tout en obtenant des performances supérieures à la moyenne. Dans la deuxième question de recherche, *centrée sur la manière d'instaurer la confiance dans ces systèmes pour qu'ils soient appliqués en clinique*, j'ai exploré le sujet de l'interprétabilité des solutions proposées. Bien qu'il soit compliqué d'avoir un système entièrement explicable basé sur l'apprentissage profond en raison de la nature de ces approches axées sur les données, une couche d'interprétabilité a toujours été ajoutée aux systèmes développés. En naviguant vers des systèmes interprétables, nous établissons plus facilement la confiance dans ces approches, ce qui était l'un des principaux objectifs fixés au cours de cette thèse. Un système explicable peut également être considéré comme plus digne de confiance, et donc ces méthodologies sont sur la voie de la valeur ajoutée aux pratiques cliniques. Pour la dernière question de recherche, *à savoir si une approche granulaire peut fonctionner et si oui, quels seraient les avantages supplémentaires*, nous avons vu que les différents systèmes étaient capables non seulement d'atteindre une haute performance, mais aussi de compenser les défauts des systèmes individuels, montrant qu'une approche qui prédit l'intelligibilité de la parole en utilisant ces trois niveaux n'est pas seulement valide mais aussi performante et innovante.

Perspectives

Les différentes études réalisées dans le cadre du présent travail permettent de dégager une variété de perspectives. Ces perspectives peuvent englober non seulement les différentes questions de recherche qui ont été abordées, mais aussi certaines propositions sur ce que pourrait être la suite du présent travail. Les **perspectives à court terme** seraient centrées sur les méthodologies nouvelles et à venir qui pourraient avoir une utilisation directe potentielle pour la prédiction automatique de l'intelligibilité de la parole. Ces perspectives concernent les nouveaux aspects techniques et les améliorations possibles des performances des modèles développés. De nouveaux schémas d'augmentation des données et des modèles dits « de bout en bout » sont des options viables pour augmenter la performance des modèles, cependant, pour des raisons d'interprétation clinique, une couche d'interprétabilité devra toujours être conçue.

Les **perspectives à moyen terme** peuvent être orientées vers l'utilisation clinique des modèles développés. Comme l'analyse automatique de la parole pathologique reste encore un domaine très académique, il est naturel que certaines des avancées de la recherche mises en avant prennent un certain temps avant d'être adoptées. C'est pourquoi ces perspectives aborderont la question de savoir comment établir la confiance dans ces approches, et comment elles fonctionneraient dans un scénario clinique. La mise en œuvre d'un « juge automatique » directement dans un hôpital est une perspective intéressante. Dans cet ordre d'idée, une étude sur la fiabilité de la comparaison des performances d'un ensemble de « juges automatiques » à un ensemble de juges perceptifs, ou même à un ensemble mixte, doit être faite. Cet aspect pourrait aider à valider la pertinence des approches automatiques et aussi à augmenter la confiance qu'il peut leur être accordée.

Enfin, en tant que **perspectives à long terme**, nous pouvons aborder certaines questions de recherche de manière plus large et sans contrainte, telles que la définition des différents concepts utilisés dans ce travail, à savoir l'intelligibilité et la sévérité, et la combinaison de ces mesures pour réduire la subjectivité. En raison de la multimodalité de ces concepts qui sont présents non seulement dans le domaine de l'automatique, mais aussi dans les sciences de la parole, l'orthophonie et la linguistique, cette réflexion n'a été qu'initiée au cours de cette thèse. Puisqu'il existe une variété de façons différentes d'évaluer l'intelligibilité (à partir de différentes tâches, en utilisant une combinaison d'autres paramètres, etc.), la création d'une mesure d'intelligibilité universelle ou d'un score de maladie qui prend en considération tous ces paramètres (du perceptuel à l'automatique) pourrait être d'un grand intérêt, malgré cela, la validité et l'utilité de cette même mesure sont sujettes à débat. La création de cette mesure universelle pourrait aussi potentiellement aider au développement de systèmes automatiques en réduisant la subjectivité perceptive, et donc en favorisant un meilleur diagnostic, pronostic et traitement des maladies affectant la parole comme le cancer de la tête et du cou.

Acknowledgements :

Cette thèse a été effectuée dans le cadre du projet TAPAS : Training Network on Automatic Processing of PAthological Speech. (financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne sous la convention de subvention Marie Sklodowska-Curie n° 766287). Dans ce projet, 15 chercheurs en début de carrière ont travaillé au développement de diverses méthodes de détection, de thérapie et d'assistance à la vie quotidienne pour les personnes souffrant de pathologies de la parole. TAPAS adopte une approche interdisciplinaire et multisectorielle. Le consortium comprend des praticiens cliniques, des chercheurs universitaires et des partenaires industriels, dont l'expertise couvre l'ingénierie de la parole, la linguistique et les sciences cliniques. Tous les membres ont une expertise dans un élément de la parole pathologique.

References:

[1] S. Fex, “Perceptual evaluation,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 6, no. 2, pp. 155–158, 1992.

[2] C. Middag, Automatic analysis of pathological speech. Doctoral Dissertation: Ghent University, Department of Electronics and information systems, Ghent, Belgium, 2012.

[3] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, “A comparative study of adaptive, automatic recognition of disordered speech,” Proceedings of Interspeech, 2012.

[4] C. Middag, J.-P. Martens, G. V. Nuffelen, and M. D. Bodt, “Automated intelligibility assessment of patho- logical speech using phonological features,” EURASIP Journal on Advances in Signal Processing, 2009.

[5] Y.-S. Lin and S.-C. Tseng, “Classifying speech intelligibility levels of children in two continuous speech styles,” Proceedings of ICASSP, 2021.

[6] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, “Deep learning for health informatics,” IEEE Journal of Biomedical and Health Informatics, 2017.

[7] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, “Non-intrusive speech intelligibility prediction using convolutional neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018.

[8] M. B. Pedersen, M. Kolbæk1, A. H. Andersen, S. H. Jensen, and J. Jensen, “End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks,” Proceedings of Interspeech, 2020

[9] J. Wang, Y. Qin, Z. Peng, and T. Lee, “Child speech disorder detection with siamese recurrent network using speech attribute features,” Proceedings of Interspeech, 2019.

[10] S.-I. Ng and T. Lee, “Automatic detection of phonological errors in child speech using siamese recurrent autoencoder,” Proceedings of Interspeech, 2020.

[11] G. V. Nuffelen, C. Middag, M. D. Bodt, and J.-P. Martens, “Speech technology-based assessment of phoneme intelligibility in dysarthria,” International Journal of Language & Communication Disorders, Volume48, Issue6, 2008.

[12] G. Saravanan, V. Ranganathan, A. Gandhi, and V. Jaya, “Speech outcome in oral cancer patients - pre- and post operative evaluation: A cross-sectional study,” Indian J Palliat Care, 2016.

[13] Crevier-Buchman, V. J, M. S, and B. D, “Intelligibility of french consonants after partial supra-cricoid laryngectomy,” Revue de Laryngologie - Otologie - Rhinologie, 2002.

[14] V. Woisard, C. Astésano, M. Balaguer, J. Farinas, C. Fredouille, P. Gaillard, A. Ghio, L. Giusti, I. Laaridh, M. Lalain, B. Lepage, J. Mauclair, O. Nocaudie, J. Pinquier, G. Pouchoulin, M. Puech, D. Robert, and V. Roger, “C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers,” Language Resources and Evaluation, 2020

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embedings for speaker recognition,” Proceedings of ICASSP, 2018

[16] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, “Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer,” Proceedings of Interspeech, 2020.

[17] M. Lalain, A. Ghio, L. Giusti, D. Robert, C. Fredouille, and V. Woisard, “Design and development of a speech intelligibility test based on pseudowords in french: Why and how?” Journal of Speech, Language and Hearing Research, 2020.

[18] S. Quintas, A. Abad, J. Mauclair, V. Woisard, J. and Pinquier, J. “Utilisation de modèles transformers pour la prédiction de l’intelligibilité de la parole de patients atteints decancers des voies aérodigestives supérieures”. XXXIVe Journées d’Études sur la Parole, 2022.

[19] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, “Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer,” Proceedings of Interspeech, 2022.

Résumé

La perte d'intelligibilité de la parole est souvent constatée après le traitement de maladies qui affectent les voies aérodigestives, comme les cancers ORL. Les évaluations perceptives restent la méthode la plus utilisée pour évaluer cliniquement l'intelligibilité de la parole. Cependant, ces appréciations sont connues pour être hautement subjectives, biaisées et longues puisque l'évaluation peut être conditionnée par l'expérience du praticien par exemple, ou encore les patients précédemment examinés. Afin de résoudre ces problèmes, une évaluation automatique est une alternative intéressante et viable, qui pourrait fournir des mesures plus objectives, plus rapides et non biaisées. Dans ce travail, nous explorons différentes manières de prédire l'intelligibilité de la parole en nous basant sur différents niveaux de granularité : la phrase, le mot et le phonème. Les résultats des modèles granulaires proposés suggèrent des corrélations avec l'intelligibilité perceptive allant de 0,80 à 0,89 lorsqu'ils sont appliqués sur un corpus français de cancers ORL. Les corrélations atteignent même 0,91, lors de la fusion de tous les systèmes granulaires. Plusieurs conclusions sont tirées de chaque niveau de granularité, notamment en ce qui concerne les types de mots et de phonèmes qui jouent un rôle plus ou moins important dans l'intelligibilité des différents locuteurs. En outre, une étude sur la modélisation individuelle d'un ensemble de juges perceptifs est également présentée. Celle-ci montre que différents profils de juges émergent de l'ensemble des juges perceptifs et automatiques. De plus, les résultats suggèrent qu'une approche automatique peut effectivement être considérée comme plus uniforme et objective qu'un juge humain. Cela laisse la possibilité de mettre en œuvre ces approches dans des environnements cliniques, soit pour servir de second avis, soit pour libérer le praticien afin qu'il puisse effectuer d'autres tâches.

Mots-clés: intelligibilité de la parole, apprentissage profond, parole pathologique, traitement automatique de la parole, cancer ORL

Abstract

Loss of speech intelligibility is commonly found in the post-treatment of conditions that affect the vocal tract, such as head and neck cancer. Due to this, perceptual evaluations are still the most widely used method to clinically assess speech intelligibility. On the other hand, these evaluations are known to be highly subjective, biased and time-consuming since the evaluation can be conditioned by the practitioner, or patients previously assessed. In order to tackle these issues, an automatic assessment has been seen as a growing and viable alternative, that could provide more objective, faster and unbiased measures. In the present work, we explore distinct ways to predict speech intelligibility based on the different granularity levels of sentence, word and phoneme. The results from the proposed granular models suggest correlations with the perceptual intelligibility ranging from 0.80 to as high as 0.89 when applied to the French head and neck cancer speech corpus. The results also suggest a correlation up to 0.91 when merging all granular systems. Several conclusions are drawn from each granularity level, namely concerning specific types of words and phonemes that play different levels of relevance for the intelligibility of distinct speakers. Moreover, a study on the individual modelling of a set of perceptual judges is also presented. The study showcased that different judge profiles emerge from the perceptual and the automatic set of judges. Similarly to the granular systems, the results suggest that an automatic approach can indeed be seen as more uniform and objective. This leaves the possibility of these approaches being implemented in clinical environments to either serve as a second opinion or to free the practitioner to perform other relevant tasks.

Keywords: speech intelligibility, deep learning, pathological speech, automatic speech processing, head and neck cancer