Évaluation de l'intelligibilité de la parole par apprentissage profond : vers plus d'interprétabilité en phonétique clinique

Sondes Abderrazek

AVIGNON UNIVERSITÉ

THÈSE DE DOCTORAT D'AVIGNON UNIVERSITÉ

École Doctorale n°536
Agrosciences et Sciences

Mention de doctorat :
Informatique

Laboratoire Informatique d'Avignon

Présentée par
Sondes ABDERRAZEK

Assessment of Speech Intelligibility using Deep Learning: Towards Enhanced Interpretability in Clinical Phonetics

Application to speech disorders due to Head and Neck Cancers

Soutenue publiquement le 02/05/2023 devant le jury composé de :

Jean Hennebert	Professeur	HES-SO, Fribourg	Rapporteur
Damien Lolive	Professeur	ENSSAT, Université de Rennes	Rapporteur
Isabel Trancoso	Professeure	IST, Université de Lisbonne	Examinatrice
Virginie Woisard	Praticienne Hospitalière Professeure Associée	CHU de Toulouse Univ. de Toulouse Jean-Jaurès	Examinatrice
Anthony Larcher	Professeur	Le Mans Université	Examinateur
Jean-François Bonastre	Professeur	Avignon Université, INRIA	Examinateur
Corinne Fredouille	Professeure	Avignon Université	Directrice de thèse

LABORATOIRE INFORMATIQUE D'AVIGNON

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

الحمد لله الذي بِنِعْمَتِهِ تتم الصالحات
اللهم انفعني بما عَلمتني، وعَلمني ما يَنفعُني، وزِدني عِلمًا وانفع بي

إلى أبي و أمي بلسم حياتي ونور قلبي ♥♥

Acknowledgment

I want to thank my defense committee, Pr. Jean Hennebert, Pr. Damien Lolive, Pr. Isabel Trancoso, PH. Virginie Woisard, Pr. Anthony Larcher, and Pr. Jean-François Bonastre. Their valuable feedback and constructive criticism greatly enriched this dissertation and enhanced its rigor and relevance.

"On ne rencontre pas les gens par hasard, ils sont destinés à traverser notre chemin pour une bonne raison." [1]

To my dear Corinne. Thank you for being more than just a thesis director, you have been a true inspiration. You believed in me and that was all I needed. Thank you for the countless hours you dedicated to our meetings, and for all the "where were we?" phrases that can be easily qualified as our official meeting catchphrase and became a testament to the depth and richness of our discussions. These discussions were so engaging that, if left unchecked by the security agent, we might have revolutionized the lab culture with a groundbreaking concept – "Nocturnal Neuron Networks". Nonetheless, I acknowledge the security agent's intervention reminding us of the existence of doors and giving us a gentle nudge back into the space-time continuum. You always reminded me that 'my thesis is my baby.' Your dedication to helping me nurture and develop this academic 'baby' was a cornerstone of my success. I am truly grateful for your unwavering guidance and belief in me throughout this challenging yet rewarding endeavor.
BREF! "WE" broke the cycle... ;)

To the RUGBI project team in which I have been involved. I want to extend my deepest gratitude for the invaluable lessons and experiences I have gained from each and every one of you. Being a part of this multidisciplinary project has been an enriching experience beyond measure. The diversity of expertise and perspectives has opened my eyes to a world of knowledge and understanding. Through long discussions, brainstorming sessions, and shared victories, I have come to appreciate the true essence of teamwork, the importance of resilience, and the power of collective passion. To Virginie (tu occupes une place spéciale dans mon cœur!), Alain & Muriel (je vous adore!), Christine, Jérôme, Gille, Marie, Anna, Robin, Sebastião, Vincent, Thimothé, Mathieu (j'aurais aimé mettre un gif près de ton prénom :p), Corine, Julie et Julien. Thank you all for being a crucial part of this enriching adventure. In the same context, I acknowledge the financial funding provided by ANR [2] which allowed me to do my research and contribute to the academic community.

[1] Sandrine FILLASSIER

[2] RUGBI project: Grant n°ANR-18-CE45-0008-04

To my dear laboratory friends and colleagues, to all the LIA permanents with whom I've had the privilege to interact, Yannick, Jean-françois, Teva (plus de jupyter sur les serveurs!), Driss (saalit), Tania, Rosa, Vincent, Bassam, Stéphane, Michèle, Laure, Philippe and Juan. To all the wonderful people who crossed my path in LIA; Salima, Imen, Afaf, Sarkis, Chaima, Samira, Aran, Mohammed, Mathias, Antoine, Sylvain, Ahmed, Tesnim, Randa, Dina, Jawid & Mursal, Céline, Cédric, Adrien & Anaïs, Natasha, Titouan, Virgile, Lucas, PG, Thibault, Luis, Timothée, Arthur, Tuan, Lucas, Julio and Sahand. Thank you for being a part of this remarkable journey!

♥ ♥ ♥ ♥ ♥ ♥

To the dynamic duo who brought me into this world and survived my teenage years without developing too many gray hairs: Mom and Dad, thanks for giving me life, even though you had no choice in the matter! Your love has been the guiding light of my life, illuminating my path. With every achievement and triumph, I carry your dreams in my heart. I am who I am because of the love and care you bestowed upon me. I dedicate my life's journey to honoring you, making your sacrifices meaningful, and making you proud.

لأمي : ما دمت انت باقيةً فليذهب كل شيء

To my beloved Mom, from the first beat of my heart, you have been my protector and my source of boundless love. I carry your strength within me...

لأبي : قسماً بمن أحل القسم أن لذة الحياة أنت ولا رجل يستحق الحب غيرك أنت

To the man of my heart, Dad, you are my forever happy place and my constant pillar of support. Your presence has been my rock, providing shelter during life's storms. I walk in your footsteps...

To my brothers, you have shaped my world. To my beloved Sofiene, you are destined for greatness, with the qualities of a true leader already shining through. My dear Hamza, I have watched you grow, learn, and face challenges with a spirit that makes me immensely proud. I believe in the amazing future that awaits both of you.

In loving memory of a great soul, whose departure coincided with the week of my defense. Your spirit lives on in the hearts of those you have touched, including mine. Though our paths never crossed, I am dedicated to honoring your memory. May your spirit find eternal peace.

To all my friends and beloved, as I pen down these words, my heart swells with an overwhelming sense of gratitude and love for each one of you.

سَـلامٌ عَلى الدُّنْيـا إِذَا لَمْ يَكُـنْ بِـهَا صَـدِيقٌ صَدُوقٌ صَادِقُ الوَعْدِ مُنْصِفَـا[3]

لا شَيْءَ فِي الدُّنْيـا أَحَـبُّ لِنَاظِـرِي مِـنْ مَنْظَـرِ الخِـلَّانِ والأَصْحَـابِ

وأَلَـذُ مُوسِيقَـى تَسُـرُ مَسَامِعِـي صَوْتُ البَشِـيرِ بِعَـوْدَةِ الأَحْبَـابِ[4]

[3] الإمام الشافعي

[4] الشاعر القروي

To my biggest supporter, my beloved Chaouki, you have walked beside me through the highs and lows. Your belief in my abilities and dreams has spurred me on to reach for the stars. Your unwavering support, boundless love, and constant presence have been my greatest blessings. You have made my journey brighter, and I am grateful beyond words for your presence in my life and for everything you have done for me.

To my dearest childhood friend, Sirine. Your presence in my life stands out as a vibrant thread of joy, nostalgia, and great memories. From the innocence of childhood to the complexities of adulthood, you have been an unwavering companion on this journey. Here's to the incredible moments we have shared, and to the many more ahead.

الصداقة هي أن تبقى القلوب على العهد حتى و إن طالت المسافات.. باختصارٍ هي أنت

To my soulmate and partner-in-crime Marwa. In this journey of life, I have come to realize that destiny couldn't have bestowed upon me a better sister than you. Thank you for illuminating even the darkest of times. You complete my world, and I am endlessly grateful for your presence in my life.

أنت الاخت التي نسيت الحياة أن تعطيها لي

My dear Salima and Imen, you have been more than just workmates, you are true friends. You stood by my side in the most difficult moments of this thesis. I couldn't have asked for better companions on this journey. Imen, you're up next, and I have no doubt you'll shine brilliantly! To my dear Sarkis and Afaf, together, we have faced the storms and celebrated the victories. For every sacrifice made and for every goal pursued, I am proud to have walked this path with you. Here's to the journey, the library weekends, the sleepless nights, and the promise of a future shaped by our persistent dedication. To my dear Fatma, the purest heart, may you find a husband who cherishes and adores you just as you deserve, a man who would finally set me free from your 24/7 persistent desire and nagging about finding a husband :p. To my confidant, Abderrahim, you have become more than a friend, you are a brother. To my cherished Skon, miles apart, your dedicated attendance at my special defense day filled me with a deep sense of honor and pride. Your friendship holds a special place in my heart! To my dear Fedi & Rihab, Congratulations on your marriage! I wish you a lifetime of love and happiness as you embark on this wonderful journey together. To my dear Chaima, I look forward to the day when we can meet again. To Aran, haval, I will miss our deep discussions during coffee breaks! To my old university friends Khalil and Chaima, thank you for honoring me with your presence on my defense day! To all my friends with whom life has created distance Rihab, Tayeb, and Basma. I look forward to the day when we can bridge the gap and reunite. Until then, know that you are missed, loved, and forever appreciated. To a friend from the digital world, Atef, thank you for your kind words and prayers they meant a lot to me and lifted my spirit during challenging times. Despite never having the chance to meet in person, I find immense strength and inspiration in knowing you. My thoughts and prayers are with you as you continue to fight your battle.

To the extraordinary professor who enriched my mind and soul, Mr. Nidhal Jelassi. I am grateful for the incredible impact you have had on my academic journey and, even more profoundly, on my personal growth and understanding of the world.

A heartfelt thank you to the one person who has been there through it all — me. From the inception of this academic escapade to the last full stop in this dissertation, I have been my own

constant companion, cheerleader, and source of motivation. Thank you, dear self, for summoning the determination to embark on this scholarly voyage and weathering the storms of doubt and uncertainty. Thank you for the late nights spent hunched over a keyboard, the countless cups of coffee that kept the neurons firing, and the endless revisions and edits to refine this document. A special shoutout to my caffeine companions and to the delivery drivers, whose consistent service ensured my survival during the long late-night writing escapades. In conclusion, thank you all for making this Ph.D. adventure an unforgettable experience. Also, a light-hearted thanks to the cosmos for not collapsing before I completed my thesis...

Résumé

L'intelligibilité de la parole est une composante essentielle d'une communication efficace. Elle peut être définie comme le degré avec lequel le message d'un locuteur peut être compris par un auditeur. Cette capacité peut être entravée par des troubles de la parole, entraînant potentiellement une diminution de la qualité de vie pour les individus. Dans le cas du cancer de la tête et du cou, la parole peut être affectée par la présence de tumeurs dans l'appareil de production de la parole. Néanmoins, la cause principale est généralement le traitement de la tumeur, impliquant notamment la chirurgie, la radiothérapie, la chimiothérapie ou une combinaison de ces traitements. Dans de tels cas, l'évaluation de la qualité de la parole est cruciale pour évaluer le déficit de communication des patients et élaborer des plans de traitement ciblés. En pratique clinique, les mesures perceptives sont considérées comme un standard pour l'évaluation des troubles de la parole. Bien que ces mesures soient largement utilisées, elles présentent plusieurs limites, la plus importante étant leur subjectivité. Par conséquent, l'évaluation automatique des troubles de la parole s'est révélée être une alternative prometteuse aux mesures perceptives dés les années '90.

Dans cette thèse, nous explorons le potentiel des techniques d'apprentissage profond pour évaluer les troubles de la parole tout en abordant les limites des outils d'évaluation existants. Dans ce contexte clinique sensible où les enjeux sont élevés et la confiance primordiale, nous considérons l'explicabilité et l'interprétabilité de ces outils comme une caractéristique obligatoire plutôt qu'optionnelle. Nous proposons une méthodologie en trois étapes basée sur l'apprentissage profond et dédiée à l'évaluation interprétable de l'intelligibilité dans le contexte des troubles de la parole.

Dans la première étape, nous abordons un problème majeur dans les outils automatiques actuels dédiés à l'évaluation de la parole altérée, à savoir une connaissance limitée sur la relation entre les troubles de la parole et le score d'évaluation qui en découle. À cette fin, nous mettons en place un modèle basé sur l'apprentissage profond, entraîné sur de la parole saine et dédié à une tâche intermédiaire de classification des phonèmes du français. Ce choix méthodologique a deux vocations. La première est de tirer bénéfice des connaissances au niveau phonème apportées par la tâche de classification pour répondre au problème majeur évoqué précedemment. La seconde est en lien avec l'utilisation de la parole saine (normale). Elle permet de pallier la quantité très limitée de données pathologiques à disposition, tout en répondant aux exigences élevées en matière de quantité de données de l'apprentissage profond.

Dans la deuxième étape, l'objectif majeur est de garantir le développement d'une solution interprétable, en vue de son acceptation en pratique clinique. Dans cet optique, nous étudions la capacité du modèle de classification des phonèmes à produire des connaissances pertinentes liées aux caractéristiques des troubles de la parole ciblés. Nous proposons ainsi un cadre analytique

général et original, nommé *Neuro-based Concept Detector - NCD*, spécialement conçu pour interpréter les représentations profondes d'un modèle. Ce cadre permet de mettre en évidence au sein du modèle de classification issu de la première étape une représentation des caractéristiques acoustiques et articulatoires de la parole saine en terme de traits phonétiques, facilement interprétables en matière d'altérations en cas de troubles de la parole.

Enfin, la troisième étape est consacrée à la prédiction d'un score final évaluant l'intelligibilité de la parole d'un individu. Cette étape repose sur les différents niveaux de représentation apportés par les deux étapes précédentes, permettant de mettre en relation le score d'intelligibilité prédit avec le degré d'altération de la parole au niveau phonème et traits phonétiques. Cette méthodologie globale apporte ainsi une interpretation du score d'évaluation dans le domaine de la phonétique à destination des clinicians. Les résultats prometteurs obtenus sur une population de patients atteints de cancer de la tête et du cou laissent envisager le potentiel d'une telle méthodologie pour suivre les progrès d'une thérapie ou développer des protocoles de rééducation sur mesure qui amélioreraient la capacité du patient à communiquer efficacement et, par conséquent, sa qualité de vie. La validation de cette méthodologie en pratique clinique est l'une des nombreuses perspectives de ce travail de thèse.

Abstract

Speech intelligibility is an essential component of effective communication. It refers to the degree to which a speaker's intended message can be understood by a listener. This capacity can be hampered as a consequence of speech disorders, which results in a reduced quality of life for individuals. In the case of Head and Neck Cancer (HNC), speech may be affected due to the presence of tumors in the speech production system, but the main cause of speech impairment is typically the tumor treatment including surgery, radiotherapy, chemotherapy, or a combination of these treatments. In such cases, the evaluation of speech quality is crucial to assess the communication deficit of patients and develop targeted treatment plans. In clinical practice, perceptual measures are considered the gold standard for assessing speech disorders. Although these measures are widely used, they suffer from several limitations, the most important of which is their subjectivity. Consequently, the automatic assessment of speech disorders has emerged as a promising alternative to perceptual measures since the 90s.

In this thesis, we explore the potential of deep learning (DL) techniques to evaluate speech disorders while addressing the shortcomings of existing tools. In this sensitive clinical context where the stakes are high and trust is paramount, we consider the explainability and interpretability of DL tools as requirements rather than optional features. Therefore, we propose a three-step methodology based on deep learning and dedicated to an interpretable assessment of speech intelligibility in the context of speech disorders.

In the first step, we tackle a major issue in the current automatic tools dedicated to disordered speech assessment which is the limited insight into the relationship between speech disorders and the resulting assessment. To this end, we implement a DL-based model, trained on healthy speech and dedicated to an intermediate task which is French phoneme classification. This methodological choice serves two purposes. The first is to take advantage of the phoneme-level knowledge obtained from the classification task to answer the major problem mentioned above. That is, it will enable the provision of insightful information about the final assessment score at the phoneme level in a subsequent stage. The second is related to the use of healthy (normal) speech. Indeed, this allows overcoming the very limited amount of pathological data available while meeting the high data quantity requirements of deep learning.

In the second step, the primary objective is to guarantee the interpretability of the developed solution, thereby ensuring its acceptance within the clinical practice setting. Thus, we investigate the capacity of the implemented phoneme classifier in yielding relevant knowledge related to the characteristics of speech pathology. We then propose *Neuro-based Concept Detector (NCD)*, our general analytic framework for the explainability of the deep representations of a DL-based model. This framework highlights, within the classification model resulting from the first step, a representation of the acoustic and articulatory characteristics of healthy speech in terms of

phonetic features, easily interpretable in terms of alterations in the case of speech disorders. We, therefore, hit two targets with one shot through this methodological choice. Indeed, not only do we actively take steps to mitigate the impact of the black-box nature of DL models, but also we ensure an additional level of granularity that clinicians can use to link and interpret the final intelligibility assessment.

Finally, the third step is dedicated to the prediction of a final score assessing the speech intelligibility of a person. This step is based on the different levels of representation provided by the two previous steps, allowing to relate the predicted intelligibility score to the degree of speech alteration at the phoneme and phonetic feature levels. The overall proposed methodology thus provides an interpretation of the speech assessment score in the field of phonetics for clinicians. The promising results obtained on a population of HNC patients suggest the potential of such a methodology to monitor the progress of therapy or to develop tailored rehabilitation protocols that would improve the patient's ability to communicate effectively, leading consequently to improved quality of life. The validation of this methodology in clinical practice is one of the many perspectives of this thesis.

ملخص

وضوح الخطاب لبنة أساسية في كل عملية تواصل ناجحة و مثمرة. يمكن تعريفه بأنه درجة فهم المخاطب للمنطوق لفظا من قبل المتكلم. قد تتأثر هذه القدرة بشكل سلبي نتيجة اضطرابات النطق، ممّا يعقد حياة بعض الأفراد. ففي حالات مرضى سرطان الرأس والعنق قد يتأذى الجهاز المسؤول عن إنتاج الكلام بعدة أورام إلى مستوى يمنع المصاب من النطق بشكل صحيح، ولكن السبب الحقيقي وراء صعوبة النطق قد يكون ناتجًا عن العلاج وليس بسبب المرض نفسه. لهذا يعد تقيم جودة النطق أمرا بالغ الأهمية لوضع خطط علاجية موجهة بدقة و عناية. إن التقييمات لاضطرابات النطق، عند الممارسة السريرية، تستند كليا إلى المعايير الحسية الطبيعية للمتلقي. على الرغْم من أنّ هذه المعايير هي الأكثر استخداما إلا أنها مكبلة بعديد القيود و غارقة في الذاتية. ونتيجة لذلك، برز التقييم التلقائي لاضطرابات الكلام كبديل واعد للتدابير الإدراكية منذ التسعينات.

في هذه الأطروحة، نستكشف قدرة تقنيات التعلم الآلي العميق على تجاوز أوجه القصور في الأدوات التقليدية المعتمدة في تقييم اضطرابات النطق. بما أن السياق الصحي حساس، حيث المخاطر عالية و الثقة أمر في منتهى الأهمية، لا نكتفي بعرض نتائج ما تقدمه هذه التقنيات الحديثة بل نعتبر قابلية فهم و تفسير أدوات التعلم العميق واجبا لا مناص منه. لذلك، نقترح منهجية من خطوات ثلاث تمكننا من الاستفادة من آليات التعلم العميق في تقييم درجة وضوح الكلام، عند المصابين باضطرابات النطق، بطريقة أقرب للموضوعية و أكثر قابلية للتفسير.

في أولى الخطوات، نعالج مشكلة مفصلية في الأدوات الآلية الحالية لتقييم النطق، وهي الفهم المحدود للعلاقة بين حدة اضطراب النطق والتقييم الناتج. تحقيقا لهذا المأرب، قمنا بتطوير نموذج قائم على أسس التعلم العميق، تم تدريبه على النطق السليم لغرض تمييز و تصنيف المقاطع الصوتية البسيطة المكونة للغة الفرنسية (الفونيمات). تخدم هذه الخطوة المنهجية غرضين إثنين: أولاً، الاستفادة من تصنيف المقاطع الصوتية البسيطة لحل المشكلة المذكورة سابقًا، وثانيًا، استعمال أمثلة للنطق السليم لتدريب النموذج لوفرتها بدلا من أمثلة للنطق الخطأ أو المضطرب وذلك لتفادي

ندرتها وتلبية لحاجة التعلم العميق إلى كميات هائلة من البيانات.

أمّا في ثاني خطواتنا المنهجية، فالمبتغى هو ضمان قابلية التفسير للحلول المطورة كي يتم اعتمادها مستقبلا في المعاينات السريرية لاضطربات النطق. لذا نقوم بدراسة قدرة النموذج الآلي المطور لغرض تصنيف المقاطع الصوتية على إكتشاف وتحديد خصائص النطق المضطرب. تمت هذه الدراسة ضمن إطار تحليلي عام من ابتكارنا أسميناه " كاشف المفاهيم القائم على الأعصاب " . صنع هذا الكاشف من أجل تحليل الأعصاب العميقة في جوف نموذج التعلم الآلي المطور سابقا في الخطوة الأولى. من خلال هذه الدارسة يمكن تمييز الأنماط الصوتية و مخارج الحروف الأكثر تواترا في النطق السليم و التي تسطعصي على معظم المصابين باضطربات النطق. بهذا الاختيار المنهجي نكون قد حققنا هدفين في خطوة واحدة. فنحن لا نقوم فقط باتخاذ خطوات نشطة للتخفيف من تأثير طبيعة الصندوق الأسود لنماذج التعلم العميق، ولكننا نضمن أيضًا مستوى إضافيًا من التفصيل الدقيق يمكن للأطباء استخدامه لربط وتفسير التقييم النهائي لاضطربات النطق.

في ثالث و آخر خطواتنا، نسعى إلى إسناد نتائج نهائية عامة و مرقمة لدرجات وضوح الكلام. يعتمد هذا التقييم على النتائج المستنتجة من الخطوات السابقة، والتي تعتمد على تحليل مفصل للمقاطع الصوتية القصيرة. وهذا سيمكننا من فهم العلاقة بين النتيجة النهائية العامة المسندة لكل فرد ودرجة الاضطرابات في النطق لديه على مستوى الأنماط والسمات الصوتية المتواترة. النتائج الواعدة التي تم الحصول عليها من خلال الاختبارات المنفذة على أصوات مجموعة من مرضى سرطان الرأس والرقبة تشير إلى القدرة العالية لنموذج التعلم العميق على متابعة تطورات حالات المرضى واقتراح بروتوكولات لإعادة التأهيل مخصصة لكل فرد. هذه المنهجية الجديدة لتقييم النطق قد تكون أكثر فاعلية في تحسيين القدرة التواصلية لمصابي السرطان خاصة و الناس عامة. ما زالت هذه المنهجية تحتاج إلى المصادقة من قبل الأطباء و المختصين عند المعاينات السريرة لمزيد تحسينها و تطويرها.

Contents

List of Figures

List of Tables

Introduction

The ability to communicate through complex language is a unique achievement of human evolution, setting us apart from all other species. It relies on a combination of linguistic and cognitive skills, enabling individuals to build complex social structures that promote their mental and physical well-being. Speech, in particular, is the primary mode of communication. The ability to use it effectively is fundamental for various aspects of life such as social interaction, education, and career opportunities. It is, therefore, not surprising that any impairment of this "vital" capacity can have far-reaching negative consequences including social isolation, reduced job prospects, anxiety, depression, and other outcomes that can significantly lower an individual's quality of life. Considering the seriousness of these outcomes, it is essential that speech disorders are taken seriously and treated with appropriate interventions.

Speech disorders can stem from a range of factors, such as neurological diseases (e.g., Parkinson's disease, Amyotrophic Lateral Sclerosis, stroke), structural abnormalities (e.g. cleft palate), or sensory/perceptual disorders (e.g. hearing loss). These pathologies can hamper the anatomical and functional capabilities of the speech articulators, leading to challenges in speech production [Kent, 1992]. **Head and Neck Cancer** (HNC) is one of those conditions that can have significant functional consequences on the speech production system due to the treatment procedures involved, such as radiotherapy, chemotherapy, and/or surgery [Meyer et al., 2004].

Given the potential impacts in the case of disorders [Woisard et al., 2022], the evaluation of speech quality is crucial. It allows clinicians to assess the functional communication deficit of patients and develop targeted treatment plans to improve their speech production. Furthermore, it is essential to ensure that patients receive appropriate and effective treatment for their speech disorders to improve functional outcomes and quality of life. For HNC, a functional deficit in communication is usually examined within a 2-step clinical assessment [Ghio et al., 2021]. The first step is to estimate the impairment, which refers to *"the loss or abnormality of anatomical structures"*, as defined by the World Health Organization (WHO) for cancer cases [Badley, 1993]. This assessment involves examining various components of speech production, including respiration, phonation, velopharyngeal function, and oral articulatory structures (jaw, tongue, and lips). The second step aims to identify functional limitations, which refers to *"the lack of ability to perform an action in a manner considered normal due to impairment"* [Badley, 1993]. The goal of this second level of assessment is to measure how effectively HNC patients can use their preserved articulators to produce the intended acoustic output. As outlined by Yorkston et al. [Yorkston et al., 1996], speech intelligibility can be seen as an important indicator of functional limitations at this assessment level.

In clinical practice, **perceptual evaluation** is the most commonly used method to assess

speech and voice disorders; it is often considered as a gold standard. This assessment method involves listening to the patient's speech and observing various aspects of their speech production (e.g. articulation, resonance, overall quality). It is usually conducted by clinicians, or speech-language pathologists (SLP) to diagnose the speech disorder, monitor the progress of therapy, or evaluate the effectiveness of different treatments. **Speech intelligibility** is usually involved in speech disorder assessment protocols. It is defined in speech communication as a measure of how comprehensible speech is in given conditions. Another definition of this concept in speech disorders is proposed by Kent [Kent, 1992] as *"the degree to which the speaker's intended message is recovered by the listener."* Despite their widespread usage, these definitions are prone to causing significant ambiguity as we will see later in the document.

Although its popularity in clinical practice, the perceptual assessment of speech and voice disorders has been heavily criticized, mainly for its ***subjectivity*** [Revis, 2004]. Indeed, the reliability of this assessment method can be affected by various factors. Among them, we can cite factors related to expert listeners such as their professional experience in the clinical domain, familiarity with the patients and knowledge about their pathology [Ghio et al., 2013], familiarity with both the assessment task and the linguistic material used [Lalain et al., 2020], etc. In addition, a variation in perceptual measures of speech disorders can occur when the same rater assesses the same speech sample multiple times. This problem is referred to as ***intra-rater variability*** and can arise due to multiple reasons such as differences in the rater's subjective judgments, attention, fatigue, and so on. On the other hand, different listeners may have different interpretations of the same speech sample and lack agreement. This issue is referred to as ***inter-rater variability***. Moreover, the perceptual assessment varies with the nature of the spoken material (e.g., linguistic structure and length of utterance), the context of communication (e.g., the quality of the acoustic transmission of the speech signal), and other factors that are not necessarily related to the raters. For instance, this assessment may not be sensitive enough to detect small changes in speech production over time, making it difficult to measure progress in therapy. Consequently, the combination of these factors, along with the time and cost associated with perceptual measures, raise the need for more reliable, objective, and automatic tools.

Automatic assessment of speech and voice disorders has emerged as a promising alternative to perceptual measures, improving the accuracy, efficiency, and objectivity of speech and voice assessment and diagnosis [Middag et al., 2009]. In this work, we are particularly interested in **DL-based assessment tools** which have attracted considerably the attention of researchers in the last years. Indeed, these tools have shown to be more objective and reliable than perceptual measures as they eliminate the subjectivity associated with human listeners. Additionally, their ability to learn complex patterns and relationships in speech data has made them highly accurate, and capable of capturing different speech disorders and variations in individual speech production. However, it is well known that DL-based tools require large amounts of high-quality speech data to be trained effectively. This can be a challenge in some clinical settings, where limited data may be available. In addition, DL-based tools are often considered as ***black-box*** models, meaning that it can be difficult to understand how they reach their assessments. This can be a limitation in a medical context where transparency and ***interpretability*** are essential requirements rather than optional features.

This thesis is conducted in the context of **RUGBI** project which stands for *"looking for*

***R**elevant linguistic **U**nits to improve the intelli**GiBI**lity measurement of speech production disorder"*. It is a collaborative effort that involves multiple disciplines with the goal of creating an objective assessment tool for speech intelligibility in the context of speech disorders. This multidisciplinary approach brings together experts from fields including the clinical domain (Ear, Nose et Throat department, and speech therapy), computer science, covered by automatic speech processing, linguistics, and clinical phonetics. Within this project, we are concerned with exploring the contribution of deep learning tools to an objective assessment of disordered speech. The central research question addressed in this thesis is whether it is possible to develop an objective assessment tool for speech disorders that incorporates the advantages of deep learning methods while addressing the limitations of current assessment tools. These limitations include providing only a single score related to one aspect of speech, as well as limited visibility and understanding. Additionally, the study aims to ensure that the developed solution is interpretable and reliable, to be accepted within clinical practice.

To this end, we propose a three-step methodology as illustrated in figure 1. We can see the solution as a whole in the form of a DL-based approach assessing the intelligibility of speech disorders while addressing the aforementioned limitations and concerns. First, we tackle the issue of the *limited insight into the relationship between speech disorders and the resulting assessment*, present in the vast majority of current automatic assessment tools. To this end, and as a **first step**, we implement a DL-based model (Convolutional Neural Network - CNN) dedicated to an intermediate task which is a French phoneme classification. The model is trained exclusively on healthy speech in order to address the issue of *limited data availability in speech pathology* while also meeting the *relatively high data requirement of deep learning* applications. This methodological choice allows us to have deep representations of French phonemes (hidden layers of the CNN) in addition to the phoneme dimension (output layer of the CNN). Looking at the methodology as a whole, this step provides more detailed and insightful information for the final assessment. By requiring the transition of speech signal through the intermediate and understandable dimension of phonemes, it allows later linking of the intelligibility assessment to the specific linguistic units affecting it. In the **second step**, our main goal is to investigate the capacity of the DL-based model in yielding relevant knowledge related to the characteristics of speech pathology. Our contributions in this step are noteworthy since we have proposed methods that are tailored to our particular context, while also having the capability to handle a range of other applications. Particularly, we design and propose the framework we named **Neuro-based Concept Detector (NCD)**, a general analytic framework for the explainability of hidden neurons/layers of a DL-based model performing a classification task. By applying NCD for the proposed CNN explainability, we bring to light an extra-interpretable dimension of great relevance in the clinical phonetics context which is phonetic features. Subsequently, we propose a scoring approach we named **Artificial Neuron-based Phonological Similarity (ANPS)** to retrieve fine-grained interpretations of the speech impairment based on the emergent dimension of phonetic features. This scoring approach is associated with heatmaps to facilitate the visualization and understanding of interpretable information by clinical experts. In an overall view of the proposed methodology, we hit two targets with one shot through this step. Indeed, not only do we actively take steps to mitigate the impact of the black-box nature of DL models and alleviate the mistrust among experts in a clinical context, but also we ensure an additional granularity level (i.e. phonetic features) with which we can link and interpret the final intel-

ligibility assessment. More interestingly, the interpretation of this extra-dimension is of great practical relevance in the clinical phonetics context since it allows the establishment of a clearer connection between the final intelligibility assessment and the physiologic characteristics of impaired speech. Finally, the **third step** is dedicated to the prediction of a final score assessing the speech production of a speaker in the context of speech disorders. Basically, the aim of this step is to transform the speech signal represented at the phoneme level (issued from step 1), to provide a final assessment, resulting in an intelligibility score, and related interpretable information.

Overall, this study sheds light on a relatively unexplored area which is deep learning interpretability for speech disorder assessment and characterization. To the best of our knowledge, no prior work has explored and explained the hidden representation inside a DL speech model to provide a deeper understanding and interpretation of the final assessment of the disordered speech. By examining this speech in terms of production at the phonemes and phonetic feature levels, clinicians can gather more useful information for speech therapy. This approach can help identify the specific linguistic units that affect intelligibility from an acoustic point of view and enable the development of tailored rehabilitation protocols to improve the patient's ability to communicate effectively, and thus, his/her quality of life.

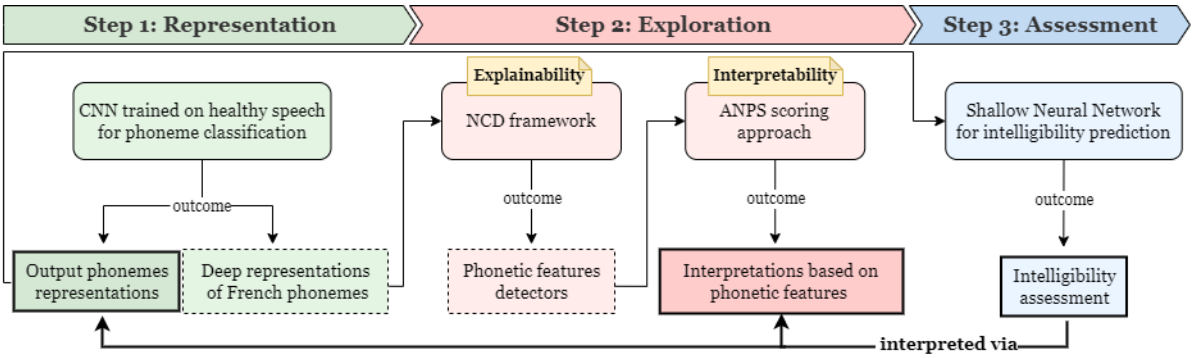


Figure 1: *Proposed methodology for an interpretable objective intelligibility assessment of disordered speech*

This dissertation is structured in two main parts. In the first part, the chapters introduce the different fundamental concepts in relation to this thesis. More specifically, in chapter 1, we focus on the terms related to the speech production system and speech disorders with which the reader is likely to be unfamiliar. We present several perceptual methods widely used in clinical practice to assess speech disorders and shed light on their subjectivity and limitations. Subsequently, chapter 2 is more related to deep learning concepts and reports various research works implementing these tools in a speech pathology context. In this chapter, we particularly emphasize studies designed for the assessment of speech intelligibility. We, therefore, draw upon insights from the shortcomings of these related works and introduce the foundational choices that we take into account to tackle these limitations. In the following chapter 3, we approach deep learning from a different perspective by focusing on its interpretability and explainability. We define relevant terminology and present a taxonomy of techniques, highlighting their im-

portance in medical applications where the stakes are high. Having established the theoretical framework in the previous section, we now turn to the empirical investigation of our research question, which we illustrate in part 2 of this dissertation. In chapter 4, we introduce the general context within which this thesis is conducted and present the different corpora we have in our possession. Then, we give an overview of the three-step methodology that we propose, without getting into details. Building on that, we organize each step of the proposed methodology as a chapter in the rest of this document. Chapter 5 introduces the first step of our proposed methodology, with the different methodological and technical choices and results. We finish this chapter with a discussion containing the related challenges and outcomes. This step was published in a conference paper [Abderrazek et al., 2020] and a JPC (Journées de Phonétique Clinique) summary [Abderrazek et al., 2021]. Moving on, chapter 6 outlines the main objectives and research questions addressed in the second step. The different outcomes and contributions of this step were published in the following conference and journal papers: [Abderrazek et al., 2022b], [Abderrazek et al., 2022c], [Abderrazek et al., 2023a] and [Abderrazek et al., 2022a]. Turning to the third step, chapter 7 illustrates the details of this step implementation and how it builds upon the two previous steps. We then summarize the key points and emphasize the importance of this final step in achieving our final objective. A JPC summary is published for this step [Abderrazek et al., 2023b]. Finally, we conclude this thesis in chapter 7.6 in which we provide a summary of the key findings and their significance. We also highlight the contributions of the research to the field of clinical phonetics and offer some final thoughts and reflections on the potential implications and opportunities presented by the research work carried out in this thesis.

Part I

Fundamental concepts and Literature review

Chapter 1

Speech production & Speech pathology

Contents

Context

This chapter provides an overview of the complex process involved in producing speech sounds. We start by exploring the anatomy and physiology of the speech production system, including the various organs and muscles involved in the production of speech. We then present the process of French sound production, in which we introduce the phonetics-related terms that we use throughout this thesis. Next, we introduce the different types of speech disorders that we cover in this work, including primarily the speech pathology resulting from head and neck cancer, then dysarthria as a motor speech disorder, and dysphonia as a voice disorder. Then, we cover the classical perceptual measures used to assess speech disorders, including speech intelligibility, comprehensibility, and severity. We end up shedding light on the subjectivity of these perceptual

measures widely used in clinical practice. Overall, this chapter aims to provide a comprehensive understanding of the terms related to the speech production system, speech disorders, and the methods used for assessing them, with which the reader is likely to be unfamiliar.

1.1 Speech Production System

Speech production is a complicated process that has been investigated from a variety of theoretical and methodological approaches including linguistic, psycholinguistic, neuropsychology, and cognitive neuroscience. It is a fundamental aspect of human communication, allowing us to express ourselves through language, convey emotions, and interact with others.

Basically, the speech production process is made up of three functions. First, *"motor control"* is a function of the human brain that develops an idea of what to say and then sends control signals to the speech-production organs via motor nerves. Then, the second function known as *"articulatory motion"* takes place. It involves the movement and shaping of the speech-production organs in response to the control signals from the motor control unit, depending on the words to be spoken or the sound to be created. Third, *speech creation* involves expulsing air from the mouth and nasal cavities, creating an acoustic wave sent out into the communication environment.

Speech production is a highly complex motor act involving the coordinated cooperation of the respiratory, phonatory, and articulatory systems. For instance, in order to pronounce the word "gap", the back of the tongue must be raised to the soft palate for a limited period of time. The sudden release of this airflow forces the vocal cords to vibrate in order to produce phonation. To generate the correct vowel, the tongue and jaw should be down, and the air should flow unobstructed. The cords relax when the lips close. Everything must be orchestrated perfectly in time and sequence in order to produce the straightforward word "gap". This section is intended for the definition of key terms related to each of these systems. The highlighting of certain terms is motivated by their later usage in this manuscript when addressing the context of speech pathology.
Fig. 1.1 depicts the sagittal view of the human speech production system.

1.1.1 Respiration

Primarily responsible for breathing, this system is also the fuel behind speech production. As shown in fig. 1.1, the respiratory system includes the lungs, rib cage, trachea, and diaphragm. All begins with taking a breath (inhale) and then starts speaking while exhaling slowly, in synchrony with the speech flow. The air expelled from the lungs moves up through the trachea to the larynx, where it passes over the vocal cords which keep vibrating until we stop talking or run out of breath.

1.1.2 Phonation

As important in the speech production process, the phonatory system is responsible for producing sound using the air that is pumped through the throat by the respiratory system. This passage of the airflow results in vocal fold vibration also called phonation, which is the sound source. The phonation function primarily involves the larynx which contains the vocal cords.

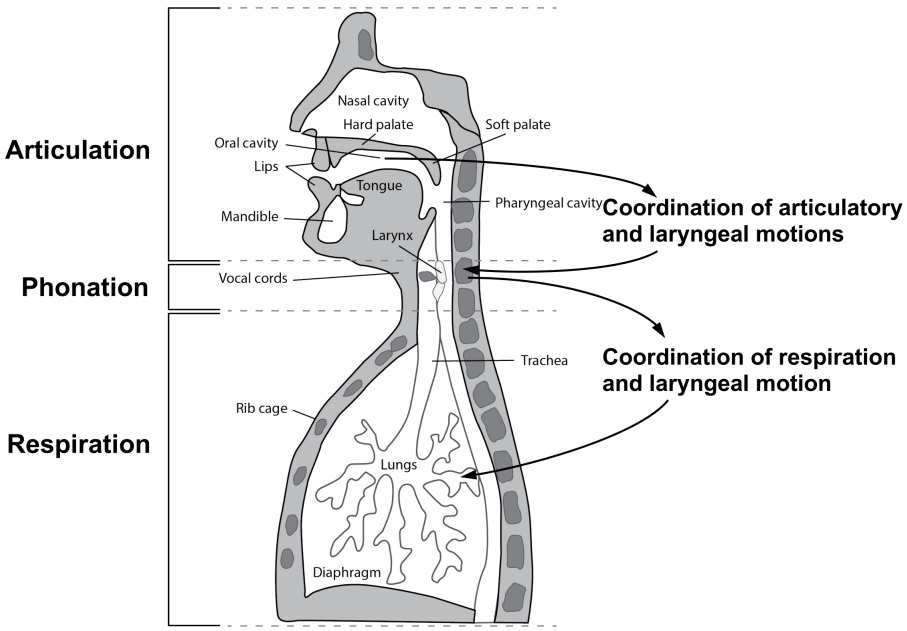


Figure 1.1: *Sagittal view of the human speech production subsystems of articulation, phonation, and respiration and their relationships. (Source: [Talkar et al., 2020])*

- **The larynx:** commonly called the voice box, is a short passageway formed by cartilage just below the pharynx in the neck. The voice box contains the vocal cords. It also has a small piece of tissue, called the epiglottis, which moves to cover the voice box to prevent food from entering the air passages.

- **The vocal folds:** also called vocal cords, are membranes stretched across the larynx. When air is pushed through the glottis, it causes pressure to drop in the larynx. This in turn makes the vocal folds vibrate, and this vibration is what produces *"voicing"*. Voicing characteristics are associated with different vibratory patterns of the glottis[1]. Depending on the aperture of the arytenoid cartilages[2], different phonation types are realized [Ladefoged, 1971, Gordon and Ladefoged, 2001], as sketched in figure 1.2. Since we refer later to some specific voice characteristics in the context of voice and speech disorders, we briefly introduce them below:

  - **Glottal closure:** it refers to the extent of vocal fold closure during the closed phase of phonation. Glottal closure is generally described as complete, incomplete, or inconsistent. A complete glottal closure refers to the situation when the vocal folds are brought completely together, resulting in a complete interruption of the flow of air through the glottis and allowing pressure to build up. This pressure build-up is necessary for producing sounds that require a sudden release of air, such as stops.
  - **Creaky phonation:** When the vocal folds are lax but tightly approximated, this can lead to cycles which are closed for a longer proportion of the cycle and which are

[1]The glottis is the opening between the vocal folds.
[2]The arytenoid cartilages are a part of the larynx and control the movement of the vocal folds.

irregular in duration.

- **Whispering phonation:** When the vocal folds are tensed and rigid but held slightly apart, the rigidity prevents the folds from vibrating. The partially opened glottis forms a narrow opening which causes turbulence.
- **Modal phonation:** The most common phonation type which yields maximum vibration of the vocal cords through an optimal combination of airflow and glottal tension.
- **Breathy phonation:** The vocal folds are tensed appropriately for vibration but not fully approximated so that complete closures do not occur.
- **Voiceless phonation:** The vocal folds are held apart, allowing air to pass through the glottis without causing the vocal folds to vibrate.

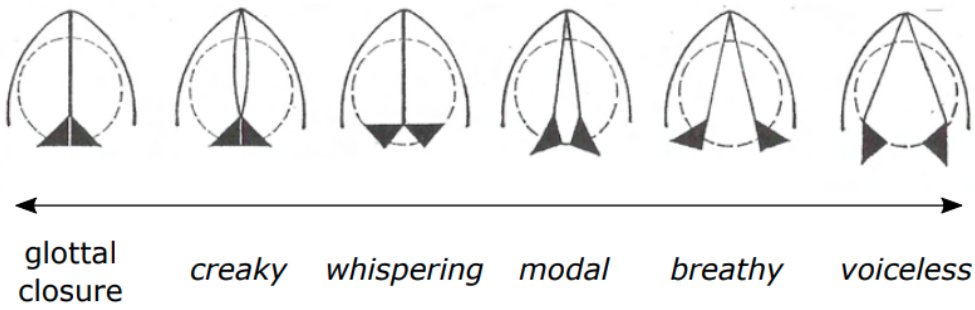


Figure 1.2: *Phonation types from closed to open glottis. The triangles represent the arytenoid cartilages, the lines connected to these triangles are the vocal cords. (source:[Jany-Luig, 2017])*

1.1.3 Articulation

The function of the articulatory system is of paramount importance in speech production. Articulation refers to shaping sounds into recognizable words, which involves forming precise and accurate vowels and consonants. In this part, we briefly describe the various components of the articulatory system and relate them to the terms used later to refer to the place of articulation of French sounds (see figure 1.3). The movement of these components can lead to the realization of about 150 different sounds, which are the basis of all the languages of the world [Teston, 2007].

- **Teeth:** Used to produce dental sounds.
- **Lips:** Used to produce labial sounds.
- **Tongue:** Considered as the primary articulator, used to create a wide variety of sounds. The tongue can be divided into several parts, including the tip, blade, front, back, and root.
- **Alveolar ridge:** The bony ridge behind the upper front teeth, is used to create alveolar consonants.

- **Palate:** The roof of the mouth, is divided into two parts: the hard palate used to create palatal consonants, and the soft palate or velum used to create velar consonants.
- **Uvula:** The small fleshy projection at the back of the mouth, is used to produce the French uvular r-sound.

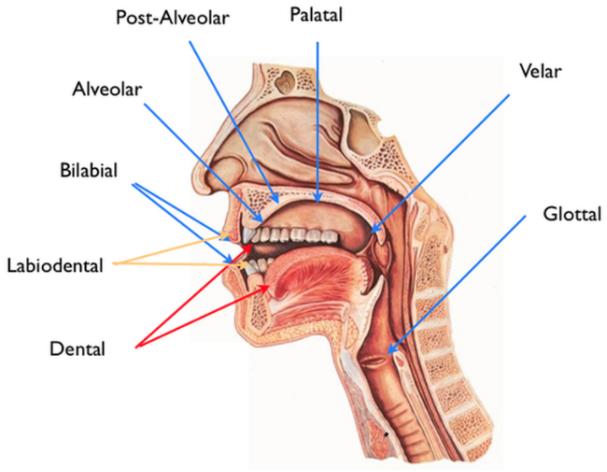


Figure 1.3: *Places of articulation (source: www.studysmarter.us/explanations/english/phonetics/place-of-articulation/)*

1.2 French Phonemes and Phonetic Features

In this section, we introduce the phonetics-related terms that are necessary for the understanding of this work. We start by defining two important notions on which this work is mainly based: phonemes and phonetic features. Afterward, to introduce the French sounds and their corresponding phonetic features, we divide this section into two parts. The first part focuses on vowels, while the second focuses on consonants.

–Phoneme: In phonetics and linguistics, a phoneme is the smallest unit of speech distinguishing one word from another in a particular language. For instance, the p-sound in the word *"tap"* separates this word from the word *"tag"*. These two words form what we call a **minimal pair** since they differ in meaning through the contrast of a single phoneme. On the other hand, a **phone** is the realization of any distinct speech sound or gesture, regardless of whether it distinguishes one word from another. For the purposes of this study, we are not making this distinction and will be referring to all speech sounds as phonemes. In transcription, linguists often use the International Phonetic Alphabet (IPA[3]) to note particular phonemes and conventionally place symbols for phonemes between slash marks (e.g. /p/). In this study, we focus on the French Phonetic Alphabet which consists of 37 phonemes. Related to the notion of the phoneme, we define here the notion of archiphoneme that we use later. An **archiphoneme** is a phonological unit that expresses the common features of two or more phonemes that are involved in neutralization. **Neutralization** in phonology refers to the elimination of certain distinctive features

[3]IPA is a standardized alphabet for phonetic notation composed of a comprehensive set of symbols and diacritical marks used to transcribe the speech sounds of all languages in a uniform fashion.

between two phonemes in specific conditions.

–Phonetic Feature: In linguistics, the term "phonetic features" refers to the way each phoneme is coded according to its acoustic and/or articulatory characteristics. Different categorizations may exist to define the set of phonetic features composing and distinguishing these phonemes. In this work, we rely on the categorization proposed by Ghio et al. [Ghio et al., 2020] characterizing the French phonemes in terms of phonetic features, based firstly on a separation between vowels and consonants. This distinction, as proposed earlier by Ghio in [Ghio, 1997], provides a definition of the set of phonetic features that is more phonetically and acoustically pertinent.

Phonemes are generally categorized into two groups: vowels and consonants. The primary distinction between these two groups is that the airflow from the lungs is either partially or completely blocked during the production of consonants, while it flows freely during the articulation of vowels. In the following sections, we provide a more detailed description of the various sounds associated with each of these categories.

1.2.1 Vowels

Among the 37 French phonemes, there is a total of 16 vowels including 4 nasalized vowels. In IPA, the nasalization is indicated with a small tilde above the vowel in question. In table 1.1, we illustrate the IPA notation of this set of vowels as well as the notation used by the LIA speech processing system implied in this work, with an example of a French word including each phoneme. It is worth mentioning that some vowels are really hard to distinguish for an untrained ear. It can already be seen that the LIA speech processing system does not make any distinction between /ɑ/ and /a/, or between /ø/ and /ə/. This reduces the number of distinct phonemes to 14. In addition, we further reduce this set of vowels to 10 by considering the following archiphonemes; Ê={e,ɛ}, Û={œ,ø}, Ô={o,ɔ} and µ={œ̃,ɛ̃}. This reduction can be done since there is no practical benefit in distinguishing between these phonemes in the clinical field.

To characterize the set of vowels, different criteria can be taken into account. Figure 1.4 illustrates the IPA representation of the different oral vowels of spoken languages where French vowels are circled in red. This diagram is called a vowel chart and is a visual representation of where the tongue is while articulating a vowel. The vertical axis of the chart represents the tongue height for each vowel where sounds higher on this axis have the tongue in a higher position, and those lower have a lower position. The horizontal axis shows the relative front-to-back position of the tongue. The trapezoidal shape of this chart represents the side view of the human mouth and reflects the fact that as the tongue moves lower, it tends to move further back. An additional criterion related to lip rounding is shown in pairs of vowels separated with a point where the second vowel is a rounded vowel. Simply put, four dimensions are often emphasized to characterize the vowels:

a) **Vertical position of the tongue:** The movement of the tongue higher or lower changes the shape of the opening through which air flows and, in turn, changes the type of vowel produced. This movement results in four levels of tongue height; *high*, *mid-high*, *mid-low*,

IPA	LIA	Example
a	aa	f**e**mme
ɑ	aa	p**â**te
e	ei	n**e**z
ɛ	ai	ch**ai**se
œ	ee	fl**eu**r
ø	eu	d**eu**x
ə	eu	ch**e**mise
o	au	gr**o**s
ɔ	oo	p**o**mme
u	ou	f**ou**
y	uu	r**u**e
i	ii	v**i**e
ã	an	v**en**t
œ̃	un	br**un**
ɛ̃	in	p**ain**
ɔ̃	on	citr**on**

Table 1.1: *The IPA and LIA notation of French vowels with examples*

and *low*. For instance, when moving from /u/ to /a/, the tongue is being *LOWERED* for a more *OPEN* vowel. Inversely, when moving from /a/ to /u/, the tongue is being *RAISED* for a more *CLOSED* vowel. In phonetics, this dimension is called *vowel height*.

b) Horizontal position of the tongue: The movement of the tongue from the front to the back of the mouth gives rise to three horizontal positions: *front*, *central* and *back*. For instance, when moving from /i/ to /u/, the tongue is retracted backward for a more *BACK* sound. Inversely, when moving /u/ to /i/, the tongue is extended forward for a more *FRONT* sound. In phonetics, this dimension is called *vowel backness*.

c) Lip rounding: When producing French vowels, the lips are either rounded or unrounded.

d) Nasality: When a nasal vowel is pronounced, the velum is slightly lowered so that air passes through both the oral and nasal cavities.

Phonetic features of vowels:

The main criteria for describing vowels previously detailed can also be seen as a set of binary phonetic features. In this work, we rely on the set of phonetic features for French vowels proposed by Alain Ghio et al. [Ghio et al., 2020], illustrated in table 1.2. This table is readable in two main ways. The first one involves examining each phoneme column by column. In this way, we are able to characterize each phoneme based on the set of binary values of phonetic features. To refer to the phonetic features describing a specific phoneme, we write the feature in brackets with a + or − to indicate its value, where + stands for the presence of the phonetic feature (i.e. value equals to 1), and − stands for the absence of the phonetic feature (i.e. value equals to 0). For example, /i/ is [−nasal], [−back], [−round], [+high], and [−low]. It has to be noted that in some cases no value is specified. This implies that neither the presence nor the absence of the phonetic feature is significant for the phoneme in question. This status occurs for two vowels:

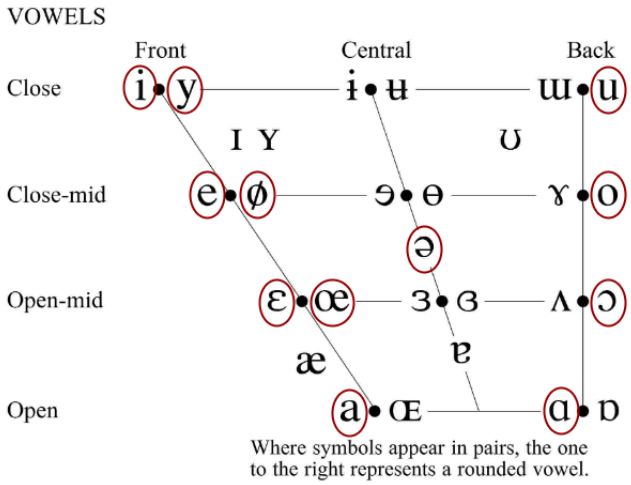


Figure 1.4: *IPA vowel chart with French vowels circled in red. (source: www.internationalphoneticalphabet.org)*

- /a/ is [−nasal], [−round], [−high], and [+low]. Neither the presence nor the absence of the back feature characterizes the vowel /a/.
- /µ/ is [+nasal], [−back], [−high], and [−low]. The round feature is neutralized in the archiphoneme µ={/œ̃/,/ɛ̃/} and thus not specified.

The second way to read the table is by considering rows. In this way, we can specify the set of phonemes forming a specific phonetic feature.
For instance, [+round]={/Û/, /Ô/, /u/, /y/, /ɔ̃/}, [−round]={/a/, /Ê/, /i/, /ã/}, etc.

Table 1.2: *Phonetic features of French vowels (source: [Ghio et al., 2020])*

	a	Ê	Û	Ô	u	y	i	ã	µ	ɔ̃
nasal	0	0	0	0	0	0	0	1	1	1
back		0	0	1	1	0	0	1	0	1
round	0	0	1	1	1	1	0	0		1
high	0	0	0	0	1	1	1	0	0	0
low	1	0	0	0	0	0	0	1	0	0

1.2.2 Consonants

French has a total of 20 consonant sounds with, in addition, the Americanized phoneme /ŋ/ (e.g. in the word park**ing**). Among these consonants, three are considered semi-consonants (also called semi-vowels) which are the /j/, /w/, and /ɥ/. We summarize in table 1.5 the list of these consonants, all with an example in a French word. In order to characterize the articulatory mechanism involved in consonant production, basically three dimensions can be considered: the manner of articulation, the place of articulation, and the voicing. Table 1.5 illustrates the IPA classification of the consonants based on the manner of articulation in the rows and the place of

articulation in the columns. The French consonants are circled in red. In cells where there are two consonants, the one to the right is voiced.

IPA	LIA	Example
p	pp	**p**ont
b	bb	**b**eau
t	tt	**t**oit
d	dd	**d**ans
k	kk	**q**uand
g	gg	**g**are
f	ff	**f**ois
v	vv	**v**ase
s	ss	**s**eul
z	zz	**z**èbre
ʃ	ch	**ch**ien
ʒ	jj	**j**our
m	mm	**m**oi
n	nn	**n**om
ɲ	gn	campa**gn**e
l	ll	**l**ac
R	rr	**r**oi
j	yy	h**ie**r
w	ww	s**oi**r
ɥ	uy	j**ui**n

Table 1.3: *The IPA and LIA notation of French consonants with an example*

a) Manner of articulation: The manner of articulation is the configuration and interaction of the articulators when making a speech sound. Represented by the rows of the table 1.5, the manner of articulation is roughly organized according to the degree of obstruction in the vocal tract during the production of a consonant. The top row comprises the sounds that involve the most significant degree of obstruction, while the bottom row represents sounds that are nearly unobstructed but still retain some consonantal quality rather than being classified as vowels.

- **Plosive:** Also called stops, they are sounds produced with complete closure in the vocal tract. French has six plosive sounds /p/, /b/, /t/, /d/, /k/, and /g/.
- **Nasal:** Nasal consonants are produced with complete obstruction in the mouth only, however, the air is allowed to flow out through the nose. The three French nasals are /m/, /n/, and /ɲ/.
- **Fricative:** Fricative sounds involve only a partial blockage of the vocal tract so that air has to be forced through a narrow channel. That is, they are made with high-speed turbulent airflow that results in a “hissy” or “noisy” sound. In French, the fricative sounds are /f/, /v/, /s/, /z/, /ʃ/, /ʒ/ and /R/.

- **Approximant:** Also called glides, semi-vowels, or semi-consonants, they are pronounced like a vowel but with the tongue closer to the roof of the mouth, so that there is slight turbulence. French has three glides where only /j/ appears in table 1.5. The two other glides /w/ and /ɥ/ are on a different chart.
- **Lateral approximant:** Usually shortened to lateral, it is a type of approximant produced with the air flowing around the tongue. French /l/ is a lateral.

b) Place of articulation: The second way to classify consonants is according to the place in the mouth where the air is blocked or restricted. IPA identifies 7 different places of articulation for French consonants, which are represented in the columns of Table 1.5. These consonants can be broadly classified into three categories: labials, dentals, and velopalatals. In the following, we briefly introduce each of these places of articulation.

- **Bilabial:** Bilabial consonants occur when the two lips are brought together, which either partially or completely obstructs the airflow from the lungs.
- **Labiodental:** To produce labiodental consonants, the lower lip is placed against the upper teeth, which either partially or completely obstructs the airflow from the lungs.
- **Alveolar:** Alveolar consonants are articulated with the tip or blade of the tongue against the alveolar ridge. To produce alveolar consonants, the tongue is brought into contact with the alveolar ridge, which either partially or completely obstructs the airflow from the lungs.
- **Post-alveolar:** Post-alveolar consonants are articulated with the tongue near or just behind the alveolar ridge (known as the postalveolar region). That is, the tongue is raised towards the back of the alveolar ridge and the roof of the mouth, creating a constriction that either partially or completely obstructs the airflow from the lungs.
- **Palatal:** To produce palatal consonants, the tongue is brought into contact with the hard palate, which either partially or completely obstructs the airflow from the lungs.
- **Velar:** To produce velar consonants, the back of the tongue is brought into contact with the velum, which either partially or completely obstructs the airflow from the lungs.
- **Uvular:** Uvular consonants are articulated with the back of the tongue against or near the uvula.

c) Voicing: In a third way, consonants are classified as either voiced or voiceless, depending on whether or not the vocal folds are vibrating during their production. Voiced sounds are produced when the vocal folds vibrate, while voiceless sounds are produced when the vocal folds do not vibrate. In table 1.5, this feature is reflected by the position of the phoneme in a cell, where all the phonemes to the right in a cell are voiced, while those to the left are voiceless.

These different dimensions are combined to describe unique phonemes. Two examples are given in figure 1.6 where, obstruent, plosive, unvoiced, and labial features are combined to describe the French phoneme /p/. Changing the plosive feature to fricative and the bilabial feature to labiodental describes the phoneme /f/.

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Figure 1.5: *IPA classification of the consonants based on the manner of articulation (row) and the place of articulation (column). The French consonants are circled in red. (source: www.internationalphoneticalphabet.org)*

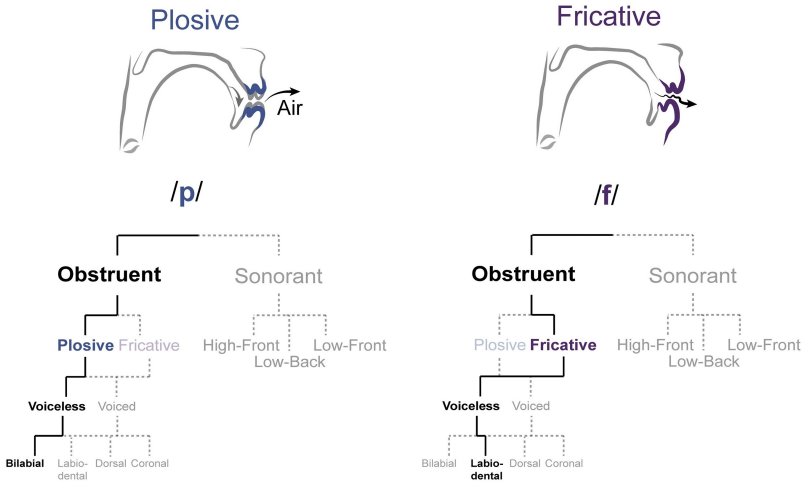


Figure 1.6: *Exemples of the two phonemes /p/ and /f/ described as the combination of acoustic-phonetic features. (source: adapted from [Yi et al., 2019])*

Phonetic features of consonants:

Following the same logic as in vowels, French consonants can be characterized by a set of binary phonetic features related to the above-cited dimensions. Similarly, we rely on the set of phonetic features for French consonants proposed by Alain Ghio et al. [Ghio et al., 2020], illustrated in table 1.4.

- The sonorant feature: distinguishes the nasal consonants, /l/ and semi-vowels [+sonorant] from the obstruents (occlusives and fricatives) [-sonorant].
- The continuant feature: distinguishes fricatives, /R/ and semi-vowels [+continuant] from occlusives and nasal consonants [−continuant] (based on [Chomsky and Halle, 1968] p.317).
- The nasal feature: distinguishes the nasal consonants [+nasal] from the oral consonants [−nasal].

- The voice feature: distinguishes voiced [+voiced] from voiceless consonants [−voiced].
- The compact feature: denotes *"the consonant articulated against the hard or soft palate"* (based on [Jak, 1951] p.27) [+compact]. We say [−compact] or diffuse for the other consonants.
- The acute feature: is defined as *"gravity characterizes labial consonants as against dentals, plus velars vs. palatals"* (based on [Jak, 1951] p. 30). This defines the dentals, velars, and palatals consonants as [+acute], as opposed to the labial consonants which are [−acute] (also grave).

Table 1.4: *Phonetic features of French consonants (source: [Ghio et al., 2020])*

	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ɲ	l	R	j	w	ɥ
sonorant	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1		1	1	1
continuant	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0		1	1	1	1
nasal	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
voiced	0	0	0	1	1	1	0	0	0	1	1	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	1	0	0
acute	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	0

1.3 Speech and Voice Disorders

A communication disorder is defined by the American Speech-Language-Hearing Association (ASHA) as an impairment in the ability to receive, send, process, and comprehend concepts or verbal, nonverbal, and graphic symbol systems. This definition refers to four major types of communication disorders, which are voice disorders, speech disorders, language disorders, and hearing disorders. Our focus in this work primarily centers on speech disorders, but we also touch upon voice disorders. Speech disorders are any condition that affects a person's ability to produce correct sounds, due to a range of underlying pathologies. These difficulties may arise due to functional issues, neurological causes like dysarthria or apraxia, or structural deficits, which include malformations of the oral/pharyngeal apparatus. Moreover, speech disorders can occur as after-effects of radiotherapeutic and/or surgical treatment of head and neck cancers. In

Table 1.5: *Definition and categorization of the types of distortions according to the place of articulation, manner of articulation, and voicing criteria.*

Distortion	Description
Manner of articulation	
Incomplete closure	The presence of noise or formants during the closure portion of the stops (i.e. spirantization[4] and gliding[5]).
Stopping	The presence of closure and/or burst in fricatives (e.g. f→t).
Hyponasality	Not enough nasal resonance on nasal sounds due to a blockage in the nasopharynx or nasal cavity, i.e. denasalization of nasal sounds (e.g. m→b).
Hypernasality	The presence of sound energy in the nasal cavity during the production of voiced, oral sounds, i.e. nasalization of oral sounds (e.g. d→n).
Place of articulation	
Backing	Backing of the place of articulation (e.g. t→k).
Fronting	Fronting of the place of articulation (e.g. g→d)
Labialization	Replacing tongue tip consonants with labial consonants (e.g. t→b)
Voicing (laryngeal articulation)	
Voicing	Partial or total voicing of voiceless consonants (e.g. f→v).
Devoicing	Partial or total devoicing of voiced consonants (e.g. v→f).
More general	
Imprecise consonants	The clarity and accuracy of consonants are affected. Consonants show slurring, inadequate sharpness, distortions, and lack of crispness.
Vowels distorted	The production of vowels is disrupted, resulting in altered or unusual-sounding vowels.

[4] Spirantization is the change where oral stops turn into fricatives.
[5] Gliding is a phonological process in which a continuant consonant is replaced with a glide consonant.

the following, we address both speech disorders resulting from dysarthria and head and neck cancers. On the other hand, a voice disorder occurs when voice quality, pitch, resonance, and loudness differ or are inappropriate for an individual's age, gender, cultural background, or geographic location [Aronson and Bless, 2009]. We address next a specific type of this disorder which is dysphonia.

In order to later simplify linking the distortions to the phonetic features and phoneme realization, we introduce and categorize different types of distortions in the table 1.5 according to the place of articulation, manner of articulation, and voicing criteria.

1.3.1 Head and Neck Cancer

Head and Neck Cancer (HNC) refers to a group of cancers that start in the tissues and organs located in the head and neck area. The most common type of head and neck cancer is squamous cell carcinoma, which develops in the squamous cells that line the mucosal surfaces of the head and neck. While squamous cell carcinomas are still the most prevalent type of head and neck cancers [Chow, 2020], there are other less common types of HNC that can originate from the salivary glands, sinuses, or muscles and nerves. The risk factors for developing head and neck cancer include tobacco and alcohol consumption [Hashibe et al., 2007, Gandini et al., 2008], exposure to certain chemicals and toxins, and infection with certain types of human papillomavirus (HPV) [Chaturvedi et al., 2011].

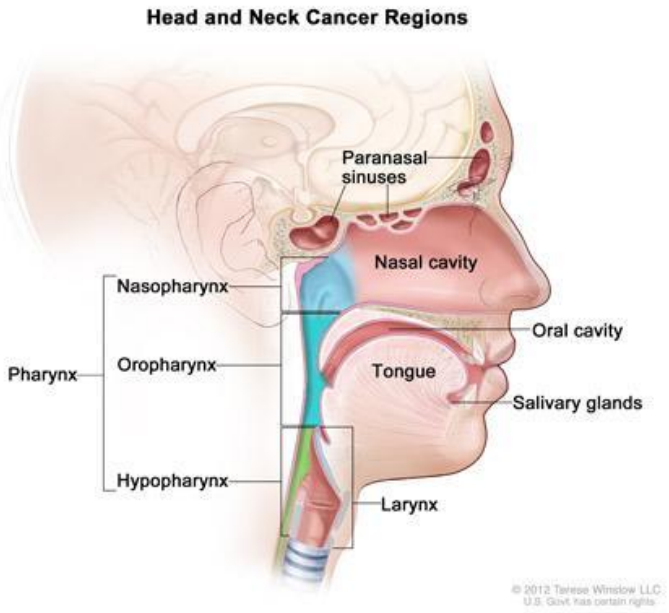


Figure 1.7: *Head and neck cancer regions (source:* `www.cancer.gov/types/head-and-neck/head-neck-fact-sheet`

As depicted in figure 1.7, cancers of the head and neck can form in different regions that we

detail in the following:

- **Oral cavity:** In oral cavity cancer, the malignant cells can be formed in the lips, the front two-thirds of the tongue, the gums, the lining inside the cheeks and lips, the floor (bottom) of the mouth under the tongue, the hard palate, and the small area of the gum behind the wisdom teeth.
- **Pharynx:** Pharyngeal cancer includes cancer of the nasopharynx (the upper part of the throat behind the nose), the oropharynx (the middle part of the pharynx, including the soft palate [the back of the mouth], the base of the tongue, and the tonsils), and the hypopharynx (the bottom part of the pharynx).
- **Larynx:** In laryngeal cancer, malignant cells form in the tissues of the larynx.
- **Paranasal sinuses and nasal cavity:** A paranasal sinus tumor is a cancer that has grown inside the paranasal sinuses (the small hollow spaces in the bones of the head surrounding the nose). This tumor can begin in the cells of the membranes, bones, or nerves that line the nasal cavity area.
- **Salivary glands:** Salivary gland cancer is a rare disease in which malignant cells form in the tissues of the salivary glands. The major salivary glands produce saliva and are located in the floor of the mouth and near the jawbone. Minor salivary glands are located throughout the mucous membranes of the mouth and throat.

Speech can be impacted in cases of head and neck cancers, with the presence of the tumor itself in the mouth or throat. However, the primary factor affecting speech is often the treatment of the tumor. Depending on the tumor size, location, and stage, patients suffering from those cancers are often treated by radical surgery (such as glossectomy[6], or laryngectomy[7]), radiology, chemotherapy or a combination of these treatments. These treatments can alter the anatomical structures and properties of speech organs such as the tongue and vocal folds. As a result, the patient's ability to speak and perform critical speech-related functions may be negatively impacted, as well as other functions like swallowing and thus, more generally, the patient's quality of life.

Speech distortions resulting from head and neck cancers can vary depending on many factors. For instance, a surgery conducted on the tumor mass combined with radiotherapy is more harmful than radiotherapy alone [Barrett et al., 2004]. In addition, treatment for tumors in the oral cavity adversely affects speech more than those in the oropharynx [Dwivedi et al., 2009], since the anatomical structures involved in speech production are concerned in a greater way in the oral cavity. These examples, but also the tumor size, the age of the patient and other factors reported in [Balaguer et al., 2019] highly impact the types of speech distortions resulting from HNC.

1.3.2 Dysarthria

As defined by Darley et al. [Darley et al., 1969a], dysarthria refers to a group of motor speech disorders resulting from a disturbance in muscular control over the speech mechanism due to

[6] A glossectomy is the surgical removal of all or a part of the tongue

[7] A laryngectomy is the surgical removal of all or a part of the larynx.

damage of the central or peripheral nervous system. This damage can lead to *"weakness, slowing, incoordination, altered muscle tone and inaccuracy of oral and vocal movements"* [Palmer and Enderby, 2007]. In this work, we focus on three specific types of dysarthria, namely, Cerebellar Ataxia (CA), Amyotrophic Lateral Sclerosis (ALS) and Parkinson's Disease (PD). A brief description of the main speech alterations resulting from these pathologies is presented in the following.

Cerebellar Ataxia

Cerebellar Ataxia (CA) is a type of neurological disorder that results from cerebellar damage which disrupts the coordination of muscular activity and causes ataxic dysarthria. Ataxic dysarthria is a speech disorder in which patients have difficulty coordinating the movements of the lips, tongue, and throat that are necessary for producing speech. These disturbances in the force, speed, timing, and direction of the muscles negatively impact speech performance and result in many distortions. According to Darley et al. [Darley et al., 1969b], imprecise consonants are the predominant distortion in patients with ataxic dysarthria. More precisely, Kent et al. [Kent et al., 1975] highlighted the fact that consonant distortions contribute to the slurred aspect of ataxic dysarthria, with stops often described as fricated and unreleased. In addition, the rate is slowed and the timing of phonemes is abnormal. Distortions include equal and excessive stress on syllables, irregular articulatory breakdowns, distorted vowels, and deviant loudness and pitch. Distortions better describing the ataxic dysarthria are summarized in table 1.6, as mentioned by [Darley et al., 1969b].

Table 1.6: *Distortions marking the CA (source: [Darley et al., 1969b])*

Imprecise Consonants
Excess and Equal Stress
Irregular Articulatory Breakdown
Vowels Distorted
Harsh Voice
Phonemes Prolonged
Intervals Prolonged
Monopitch
Monoloudness
Slow Rate

Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is the result of upper and lower motor neuron damage, most typically manifests with muscle weakness and atrophy, and leads to a mixed type of dysarthria. Distortions of manner and voice were reported previously as two of the most commonly observed speech characteristics in ALS [Kent et al., 1990]. The hallmark features of ALS include imprecise consonants, which demonstrate the greatest decline with disease progression. Hypernasality is also a distinctive criterion marking ALS patients due to velopharyngeal impairment. In addition, the weakness of speech musculature can lead to a harsh or breathy voice, a decrease in respiratory support, a slow speaking rate, and difficulty with the physical production

of speech. Table 1.7 summarizes the main distortions in mixed dysarthria resulting from ALS, based on the work of Darley et al. [Darley et al., 1969b].

Table 1.7: *Distortions marking the ALS (source: [Darley et al., 1969b])*

Imprecise Consonants
Hypernasality
Harsh Voice
Slow Rate
Monopitch
Phrases Short
Vowels Distorted
Low Pitch
Monoloudness
Excess and Equal Stress
Intervals Prolonged
Reduced Stress
Phonemes Prolonged
Strained-Strangled Voice
Breathy Voice (Continuous)
Audible Inspiration
Inappropriate Silences
Nasal Emission

Parkinson's Disease

Parkinson's disease (PD) is a neurodegenerative disorder caused by basal ganglia[8] damage which possible reasons are encephalitis[9], degeneration of nerve cells due to aging or arteriosclerotic[10] changes, repeated head injuries, congenital diseases, exposure to certain toxins, and certain tranquilizing drugs. Parkinson's disease causes slowness in movement and movement initiation, rigidity, unintended or uncontrollable movements, such as shaking, stiffness, difficulty with balance and coordination, and speech pathology. These symptoms usually begin gradually and worsen over time. Speech pathology due to Parkinson's disease is characterized by a pattern of motor speech disorders known as hypokinetic dysarthria. The key characteristics of this speech pathology are illustrated in table 1.8 issued from the study of Darley et al. [Darley et al., 1969b].

[8]The basal ganglia is a group of brain structures linked together, handling complex processes and best known for their role in controlling the body's ability to move.

[9]Encephalitis is inflammation of the brain.

[10]Arteriosclerosis occurs when the blood vessels that carry oxygen and nutrients from the heart to the rest of the body become thick and stiff which sometimes restrict blood flow to the organs and tissues.

Table 1.8: *Distortions marking the PD (source: [Darley et al., 1969b])*

Monopitch
Reduced stress
Monoloudness
Imprecise Consonants
Inappropriate Silences
Short Rushes
Harsh Voice
Breathy Voice (Continuous)
Low Pitch
Variable Rate

1.3.3 Dysphonia

According to ASHA, the term dysphonia encompasses the auditory-perceptual symptoms of voice disorders. A voice disorder occurs when voice quality, pitch, and loudness differ or are inappropriate for an individual's age, gender, cultural background, or geographic location [Aronson and Bless, 2009, Boone et al., 2005]. It can be categorized as either physiological voice disorders that result from alterations in respiratory, laryngeal, or vocal tract mechanisms, or functional voice disorders that result from inefficient use of the vocal mechanism when the physical structure is normal. Table 1.9 includes more details about the voice alteration due to dysphonia, as reported by ASHA.

Table 1.9: *Perceptual signs and symptoms of dysphonia (source: www.asha.org/practice-portal/clinical-topics/voice-disorders)*

Rough vocal quality
Breathy vocal quality
Strained vocal quality
Strangled vocal quality
Abnormal pitch
Abnormal loudness/volume
Abnormal resonance
Aphonia (loss of voice)
Phonation breaks
Asthenia (weak voice)
Gurgly/wet-sounding voice
Pulsed voice
Shrill voice
Tremorous voice (shaky voice)

1.4 Perceptual evaluation of speech and voice disorders

In clinical practice, perceptual evaluation is the most commonly used method to assess speech and voice disorders and is often considered a gold standard. Usually conducted by a speech and language pathologist (SLP), this assessment method involves listening to the patient's speech and observing various aspects of their speech production. In this section, we give an overview of the definition of some classical perceptual measures in the literature and provide a disambiguation of the terminology. Next, we introduce some perceptual protocols to evaluate the quality of speech and voice used in the clinical context.

1.4.1 Classical perceptual measures and disambiguation of the terminology

Speech pathology sits at the interface of several disciplines, i.e. *linguistics*, *psychology*, *medicine*, and *sociology* [Tanner, 2006]. It has been reported in the literature [Walsh, 2005, Denman et al., 2019] that, because of the complex evolution and diverse parentage of speech pathology, terminology in the field is sometimes ambiguous, improperly defined, and used inconsistently. The *"terminology problem"* was expressed since 1971 by Kenneth Scott Wood [Wood, 1971], which still seems disconcertingly relevant: *"This growth of speech pathology and audiology [...], has generated hundreds of terms, some of which are interchangeable, some of which have different meanings to different people, some of which are now rare or obsolete"*. In particular, there seems to be a lack of consensus regarding the terminology of the perceptual concepts related to speech, as well as how to assess them. Indeed, in a recent clinician survey in French-speaking countries [Pommée et al., 2021], Pommée and al. revealed a lack of standardization of the speech assessment, regarding its overall structure, but also the assessment tasks and stimuli used. Furthermore, the terms used by the speech-and-language pathologists in this study indicated a lack of clarity, specifically regarding intelligibility and comprehensibility definitions. In the following, we introduce the three most commonly used measures in the perceptual evaluation of speech pathology including intelligibility, comprehensibility, and severity. While defining these concepts as reported in the literature, we focus as well on works proposing terminological disambiguation when a lack of consensus is involved.

Different definitions were proposed for **speech intelligibility** in the literature. As defined by Hustad [Hustad, 2008], *"intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener"*. Hodge and Whitehill [Hodge and Whitehill, 2010] report: *"Intelligibility, or how understandable one's speech is to another, is a functional indicator of oral communication competence. It reflects a talker's ability to convert language to a physical signal (speech) and a listener's ability to perceive and decode this signal to recover the meaning of the talker's message"*. If these definitions agree that intelligibility-related information is carried by the acoustic signal, they use different terminology to define it which could lead to ambiguity. That is, varying interpretations of intelligibility by different individuals can arise due to variations in the terminology. Additionally, the definitions do not explicitly outline the technical aspects of how intelligibility should be assessed. Now, if we consider the intelligibility definition proposed by Kent [Kent, 1992] *"the degree to which the speaker's intended message is recovered by the listener"*, this definition is even more ambiguous since it does not even reveal that intelligibility-related information is carried by the acoustic signal. The confusion is particularly around the term *"intended message"* which may be influenced by factors that are not

always captured by the speech signal solely, such as context and nonverbal cues. To this end, this definition could be considered as a definition of comprehensibility as reported by Lalain et al. [Lalain et al., 2020]. The term comprehensibility will be further defined as we progress in this section. In their turn, the intelligibility definition of Ghio et al. [Ghio et al., 2019] (translated from French[11]) is : *"The perception of speech is a complex process that integrates both an ascending flow of information from the speech signal and a descending flow based on high-level information held by the listener. The bottom-up flow is mainly an acoustic-phonetic decoding operation that consists in identifying phonemes from the speech signal. Phonemes, which can be considered as the smallest units for opposing meanings, are the basic elements of speech intelligibility. [...] Acoustic-phonetic decoding is therefore the fundamental process for perceptually measuring a speaker's intelligibility"*. Obviously, this definition provides a more detailed and technical understanding of speech perception and how it relates to speech intelligibility, compared to the other definitions discussed earlier. Most importantly for us, it explicitly highlights the role of acoustic-phonetic decoding, which involves identifying phonemes from the speech signal, as the fundamental process for measuring intelligibility. **In this study, we adopt this definition as our reference definition of intelligibility for its clarity.**

Moving on, we find that terminological disambiguation is important at this stage. As regards **intelligibility and comprehensibility**, while both concepts are linked and both contribute to functional human communication, they relate to two different aspects of speech. Yorkston et al. [Yorkston et al., 1996] explained: *"The term intelligibility refers to the degree to which the acoustic signal (the utterance produced by the dysarthric speaker) is understood by a listener. [...] The concepts of comprehensibility and intelligibility may be distinguished by the fact that comprehensibility incorporates signal-independent information such as syntax, semantics, and physical context"*. In their turn, Ghio et al. [Ghio et al., 2021] outlined this difference based on the Lindblom model [Lindblom, 1990] which proposes that two types of information are essential for comprehension during spoken communication. The first is *"signal-dependent information"*, which is extracted from the speech signal through a *bottom-up* process known as *"acoustic-phonetic decoding"*. This process involves identifying phonemes in the speech signal, which are considered the fundamental units of speech intelligibility. Acoustic-phonetic decoding is the primary process used in perceptual measures of speech intelligibility. The second type of information is *"signal-independent"* and is the result of *top-down* processes where the listener constructs the message using all available information at different levels, including lexicon, communicative context, shared knowledge, and psychosocial context. In line with this, comprehensibility was defined by Fontan et al. [Fontan et al., 2015] as *"the integration of both acoustic-phonetic information and all relevant information independent of the signal in order to understand a spoken message in a particular communicative situation"*. An elaboration of a comprehensive definition of intelligibility and comprehensibility and their assessment in both the clinical and scientific fields was also proposed by Pommee et al. [Pommée et al., 2022] in their

[11]Original definition: *"La perception de la parole est un processus complexe qui intègre à la fois un flux ascendant d'informations provenant du signal vocal mais aussi un flux descendant fondé sur les informations de haut niveau détenues par l'auditeur. Le flux ascendant (« bottom-up ») est principalement une opération de décodage acoustico-phonétique qui consiste à identifier les phonèmes à partir du signal de parole. Les phonèmes, pouvant être considérés comme les plus petites unités permettant d'opposer du sens, sont les éléments de base de l'intelligibilité du discours. [...] Le décodage acoustico-phonétique est donc le processus fondamental pour mesurer perceptivement l'intelligibilité d'un locuteur."*

consensus study. Indeed, the authors revealed that intelligibility refers to the acoustic-phonetic decoding of the utterance, while comprehensibility relates to the reconstruction of the meaning of the message. To summarize, if comprehensibility is considered to be based on both signal-dependent and signal-independent sources of information, intelligibility, on the other hand, is defined as the amount of speech understood solely from signal-dependent information.

Furthermore, **intelligibility and severity** are another couple of measures to consider in terminological disambiguation. Indeed, speech intelligibility is related to the accuracy of speech perception and the ability to extract meaning from spoken language. It is typically evaluated as the amount of speech understood from the acoustic signal, such as word or sentence recognition accuracy [Keintz et al., 2007, Hustad, 2008]. **Speech disorder severity**, meanwhile, can be seen as a more global measure, referring to the degree of alteration of the speech signal. In this case, various elements of the vocal signal are taken into account, such as the quality of the speech rate, acoustic-phonetic decoding, the consonant and vowel precision, and other prosodic parameters relating to the perceived speech impairment [Kent et al., 1989, Yorkston et al., 1996, Auzou, 2007].

1.4.2 Overview of extra perceptual assessment protocols and scales

Perceptual evaluation typically involves the use of a protocol which is a standard procedure for systematically describing and quantifying an impairment. Using auditory perception, it is made by speech therapists who affect different scores to assess the speech quality of a speaker. Numerous perceptual protocols, measurements, or scales are available to evaluate the quality of speech and voice in clinical practices.

- **GRBAS:** Among the most popular protocols, we can cite GRBAS Scale. Developed in 1981 [Hirano, 1981], this scheme is designed for the evaluation of dysphonic voice quality. It assesses 5 components: (1) Grade (the overall grade of hoarseness); (2) Roughness; (3) Breathiness; (4) Asthenia (voice weakness); and (5) Strain. Each component is rated on a 4-point scale, where 0 is normal, 1= slight, 2= moderate, and 3 = severe.

- **CAPE-V:** Developed by the ASHA [Kempster et al., 2009], the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) is a tool for clinical auditory-perceptual assessment of voice. The CAPE-V indicates the six salient perceptual vocal attributes which are: (a) Overall Severity; (b) Roughness; (c) Breathiness; (d) Strain; (e) Pitch; and (f) Loudness. The CAPE-V displays each attribute accompanied by a 100-millimeter visual analog scale where the clinician indicates the degree of perceived deviance from normal. Additional features may also be used by clinicians to rate additional prominent attributes required to describe a given voice as commenting about resonance.

- **ASSIDS:** Assessment of Intelligibility of Dysarthric Speech (ASSIDS) is a tool for perceptual dysarthric speech assessment [Yorkston and Beukelman, 1981]. It provides different measures including a percentage of intelligibility at both word and sentence levels, a total speaking rate, a rate of intelligible speech expressed as intelligible words per minute, and a communication efficiency ratio.

- **FDA:** We can cite the Frenchay Dysarthria Assessment (FDA) in its original and second edition (FDA-2) [Enderby, 1980, Enderby, 1983, Enderby and Palmer, 2008] and its French version proposed by [Ghio et al., 2019] for the assessment of motor speech disorders and associated orofacial impairments. With FDA, the subject's overall clinical intelligibility level and articulator motor functionality are assessed based on the 28 criteria including coughing and swallowing reflexes, laryngeal and respiratory functioning, the position of lips, jaws, palate, vocal cords, and tongue etc. Each criterion was rated on a 9-point alphabetical scale (i.e., a = normal function to e = no function).

- **BECD:** The BECD (*Batterie d'Evaluation Clinique de la Dysarthrie* in French) [Auzou and Rolland-Monnoury, 2006] is the most commonly used test by clinicians for French speech. The perceptual assessment of the BECD is based on the scoring of features of voice, articulation, prosody, respiration, and intelligibility in order to characterize dysarthria. Each of these items is rated on a 5-point scale (i.e., 0 = normal to 4 = severely impaired).

A famous clinical scale including a specific item dedicated to speech disorders can also be cited:

- **UPDRS:** The Unified Parkinson's Disease Rating Scale (UPDRS) was developed in 1987 by neurologists as a gold standard to measure the severity and progression of Parkinson's disease. It enables monitoring the response to medications used to decrease the signs and symptoms of PD through the assessment of multiple items on a 0-4 rating scale. Including a specific item dedicated to speech disorders, it is rated as if 0 = normal speech, 1 = slight decrease in intonation and volume, 2 = monotonous, garbled but understandable speech, clearly disturbed, 3 = marked speech disturbance, difficult to understand, and 4 = unintelligible speech.

1.5 Reliability and validity of perceptual measures

The reliability of the perceptual assessment of pathological speech is mainly reflected by both inter-rater and intra-rater reliability. *Inter-rater reliability* measures the agreement between subjective ratings of the same phenomenon by multiple judges while *intra-rater reliability* refers to the consistency of the judgments by one rater over several trials and is best determined when multiple trials are administered over a short period of time. Studies such as [Pommée et al., 2021] have shown that speech and language pathologists lack of reliability in the currently available assessment tools in clinical practices. Below, we report some of the variables affecting the perceptual assessment of speech disorders, particularly, the intelligibility measurement:

– **Ambiguity of perceptual measure definitions:** As already mentioned, the various definitions used for different perceptual measures may create confusion and lead to a lack of agreement among judges. This, in turn, makes the assessment of speech disorders challenging to reproduce, but also very subjective and variable.

– **The speaker's task:** It is different to score intelligibility when the speaker's production is a word or a sentence. Indeed, a sentence provides more context for the listener to evaluate the speaker's production than a single isolated word.

- **The listener's task:** Forced-choice word selection and sentence completion will yield higher scores than orthographic transcription for instance. As well, evaluation based on a word transcription is more objective than a global measurement of intelligibility based on a rating scale.
- **The transmission system:** Live voice will often yield higher intelligibility scores than transcriptions of tape-recorded utterances.
- **Predictability of the test items:** Clinicians' judgment is influenced by their familiarity with both the assessment task and the linguistic material used. Indeed, as reported by Lalain et al. [Lalain et al., 2020], this can be explained by the compensatory mechanisms, which integrate the effects of lexicality[12], the effects of frequency of the words (i.e. the most frequent words are the most easily recognized), the phonotactic rules of the language (e.g. a sequence [rsit] is not very likely in French), the shared knowledge about the context of the interaction, etc. That is, the experts use this top-down information to restore the production of impaired speech and mask the real difficulties of the patient. As a result, we obtain an overestimation of intelligibility, but more generally unreliable assessments of the speech disorder.
- **Knowledge about the pathology of the patient:** Clinicians' judgment is influenced by their familiarity with the patients and their care pathway. It has even been shown that only knowledge about the details of speech pathology has a significant impact on speech quality assessment [Ghio et al., 2013].

Although perceptual measures remain the most commonly used method for assessing speech and/or voice disorders in clinical practice, the numerous shortcomings outlined and others shed light on the fact that they are not only subjective but also non-reproducible and time-consuming. Controlling these factors will certainly enhance the reliability level of these measures. Indeed, one of the suggested ways to manage the listeners' variability is the use of a large number of test items combined with a random selection including pseudo-words. This also can be discussed since employing unnatural speech material (e.g., nonsense words) cannot fully exclude errors due to listener bias. To conclude, the limits reported above raise not only the need for reliable, optimized, and accurate tools for disordered speech assessment but also the need for standardization of these tools for French-speaking adults as evoked in [Pommée et al., 2021].

1.6 Conclusion

In this chapter, we started by providing an overview of the speech production system, including the respiratory, phonatory, and articulatory subsystems. We described the various speech sounds present in the French language based on these subsystems. Moving on, we introduced the different types of speech/voice disorders we are going to address in this work, namely dysarthria, dysphonia, and speech pathology resulting from head and neck cancers. For each disorder, we provided a brief description of the related speech and voice distortions. Subsequently, we outlined the perceptual measures which are the most commonly used method for assessing these

[12] a phonetically ambiguous sound [t/d] will be preferentially perceived /t/ in front of a sequence [aʃ], with reference to the French word "tâche" (task), but it will be perceived /d/ if it is placed in front of a sequence [isk], with reference to the French word "disque" (disk).

disorders in clinical practice. We discussed their limitations, and how they can be highly subjective and inconsistent with an emphasis on the need for proper and accurate assessment tools. The automation of these tools was largely proposed as an alternative since 1992 [Ferrier et al., 1992]. In the next chapter, we introduce an overview of these methods, basically based on deep learning, that we consider relevant in our context.

Chapter 2

Deep Learning and Speech Pathology

Contents

Context

Although perceptual measures are considered as the gold standard for assessing speech disorders in the clinical context, they have some drawbacks. They are time-consuming, expensive, difficult to reproduce, and can be influenced by factors such as the listener's familiarity with the patient's speech disorder and the linguistic context of the speech tasks being evaluated. To better support clinicians in their assessments, there is a need for automatic measures that can offer frequent, reliable, and objective intelligibility assessments that are also cost-effective. To cope with the limitations of perceptual evaluation listed in the previous chapter, automatic approaches have emerged very early as potential solutions to provide objective assessment tools. We can see these methods basically as those based on speech signal processing, on machine learning methods, and more recently those based on deep learning approaches.

In this chapter, we start by presenting fundamental concepts related to deep learning, that we use in this manuscript. Following that, we discuss various applications of deep learning in speech pathology, with a particular emphasis on the assessment of speech intelligibility. Finally, we conclude the chapter by outlining the foundational choices upon which our work is based, drawing upon insights from related work shortcomings.

2.1 Deep Learning key concepts

To start, it is relevant to briefly introduce deep learning within the broader context of **artificial intelligence (AI)**. AI is the broadest term used to classify machines that mimic human intelli-

gence. These machines are programmed to perform tasks that would normally require human intelligence, such as perception, reasoning, learning, and problem-solving. **Machine learning (ML)** is a subfield of AI that focuses on developing algorithms and statistical models that enable computers to learn from data, without being explicitly programmed. By going deeper, we find **deep learning (DL)**, a more advanced subfield of ML that uses artificial neural networks with multiple layers to learn from complex data [Goodfellow et al., 2016]. Based on the structure and function of the biological brain, DL has revolutionized many fields such as computer vision, natural language processing, and robotics. This field is the main focus of the present work. In this section, we introduce fundamental concepts related to the deep learning field to provide the reader with the proper scientific basis to better understand this study. Certain notations are emphasized due to their importance in later chapters.

2.1.1 Artificial Neuron

The idea behind creating artificial neurons is to mimic the biological neuron ability to receive, process, and transmit information through electrical and chemical signals. The basic structure of an artificial neuron is therefore strongly inspired by biological neurons (see figure 2.1), yet certainly simplified and does not fully capture the complexity of the biological system. The **inputs** of an artificial neuron are analogous to the dendrites of a biological neuron, which receive signals from other neurons or from sensory receptors. The **weights** in an artificial neuron are numerical values that represent the strength of the connection between the input and the neuron. They are similar to the synapses between biological neurons, which determine the strength of the connection between neurons. Based on this, a **weighted sum** of the inputs, referred to as z, is calculated with generally the addition of a bias b.

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{2.1}$$

This value is then fed to the neuron **activation function**, which will define if the neuron is activated, and at which intensity.

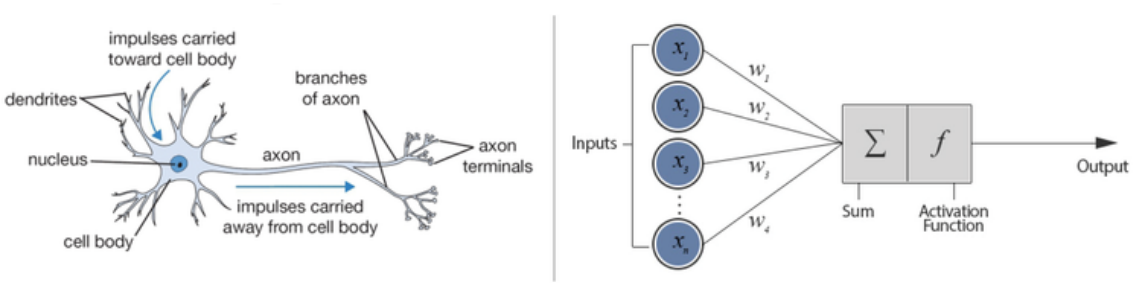


Figure 2.1: *Artificial neuron vs. biological neuron*

Activation Function

The activation function in an artificial neuron is also inspired by biological neurons. It is similar to the cell body of a biological neuron, which processes the inputs and generates an output signal.

The activation function is a mathematical function that takes the weighted sum of the inputs and produces an output based on the function mathematical properties. The most important feature of an activation function is its ability to add **non-linearity** into a neural network. This non-linearity plays a critical role in allowing the network to learn complex, non-linear relationships between inputs and outputs. Among the most common activation functions we present below the Rectified Linear Unit (ReLU), the sigmoid, the softmax, and the hyperbolic tangent (Tanh) [Karlik and Olgac, 2011].

- **ReLU:** ReLU is currently the most widely used activation function in DL models due to its computational efficiency and effectiveness. It outputs the input directly if it is positive, and outputs zero if it is negative.

$$f(z) = \max(0, z) \tag{2.2}$$

- **Sigmoid:** Regardless of the input, sigmoid always outputs a value between 0 and 1 allowing very large values to be mapped to 1 and very small values to be mapped to 0. It is ideally used in binary classification problems.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$

- **Softmax:** The softmax is a more generalized form of the sigmoid used in multi-class classification problems. Similar to sigmoid, it produces values in the range between 0 and 1. Applied in the final output layer of a classifier with N classes, softmax takes a vector of value $z_1, z_2, ..., z_N$ (also called **logits**) and outputs a vector of probabilities assigned to each class $\sigma_1, \sigma_2, ..., \sigma_N$ such that each σ_i is a non-negative number between 0 and 1, and the sum of all σ_i equals 1. The formula for the softmax function is:

$$\sigma_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{2.4}$$

- **Tanh:** Unlike the sigmoid function, the Tanh function is zero-centered where the output is a real number between -1 and 1.

$$f(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{2.5}$$

2.1.2 Artificial Neural Network

An **Artificial Neural Network (ANN)** is a type of machine learning algorithm that is inspired by the structure and function of the biological brain. It is composed of an interconnected group of artificial neurons typically aggregated in layers, that work together to learn patterns in data and perform a specific task. The connections between neurons are weighted and this weight is adjusted as learning proceeds. The larger the weight, the stronger the signal at the connection is. Figure 2.2 shows the general structure of an ANN. During training, an ANN is presented with a set of input data along with the desired outputs or labels. This is what we call supervised learning. **Supervised learning** involves comparing the output of a neural network to the label or *ground truth* associated with the input, using a **loss function** to measure the difference between

the two. **Backpropagation** [Amari, 1993] is then used to update the weights of the network to minimize the loss function and improve the performance of the model. **Gradient descent** is one of the commonly-used optimization algorithms for minimization of the cost function of a model. It uses the **gradient** (or derivative) of the cost function to update the model parameters according to a **learning rate**. Since deep learning usually makes use of large amounts of data, **batch learning** is introduced to randomly sample inputs and perform **stochastic gradient descent (SGD)** based on the mean loss of each batch. When a DL model becomes too complex, it begins to fit the noise in the training data instead of the underlying pattern. This problem is referred to as **overfitting** and can lead to poor performance on new unseen data, as the model has essentially memorized the training set instead of learning the general pattern. **Regularization techniques** are usually used in these cases to prevent overfitting. Batch normalization and dropout are two regularization techniques commonly used in DL to prevent overfitting and improve generalization performance. **Batch normalization** (also known as batch norm) [Ioffe and Szegedy, 2015] is a technique used to make training of artificial neural networks faster and more stable through normalization of the layer inputs by re-centering and re-scaling. **Dropout**, on the other hand, is a technique that randomly drops out (i.e., sets to zero) a proportion of neurons in the network during training. This helps to prevent overfitting by forcing the network to learn redundant representations of the input data.

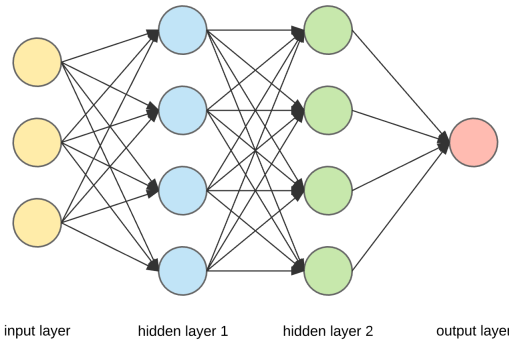


Figure 2.2: *Artificial Neural Network*

To ensure clarity throughout this thesis, we use the term "Artificial Neural Networks" to refer to all types of neural networks presented. In addition, we specifically use the term **"Deep Neural Networks" (DNNs)** to describe artificial neural networks that have more than two hidden layers, and **"Shallow Neural Networks" (SNNs)** to refer to networks that have two or fewer hidden layers.

2.1.3 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are very popular in DL since they play a major role in very fast-growing and emerging areas such as computer vision tasks (e.g. localization and segmentation, video analysis, obstacle recognition in self-driving cars) [Voulodimos et al., 2018, LeCun et al., 2010], natural language processing [LeCun and Bengio, 1998], speech recognition [Abdel-Hamid et al., 2014], etc. CNNs are motivated by the imitation of biological vision systems and have been widely adopted for computer vision and image-related tasks, such as

reading pathology slides [Acs et al., 2020] or brain images [Bernal et al., 2019]. The aim is to simulate the hierarchical nature of neurons in the vision cortex. Overall, the architecture of a CNN typically consists of a series of convolutional and pooling layers followed by one or more fully connected layers and an output layer. We detail these components in the following.

Convolution layer

The convolution layer is the core building block of a CNN. It is responsible for extracting features from the input by applying a series of **filters** (also known as **kernels**). Each filter slides over the input, performing element-wise multiplications and sums to produce a **feature map** (also called an **activation map**). **Stride** determines the amount by which the filter moves, for example, a stride of 1 as shown in figure 2.3 will cause the filter to shift by one row or column. Filters are capable of learning to recognize meaningful features from the input that are directly related to the final task. For example, to identify a table, representations such as sharp edges or a flat surface might be captured automatically.

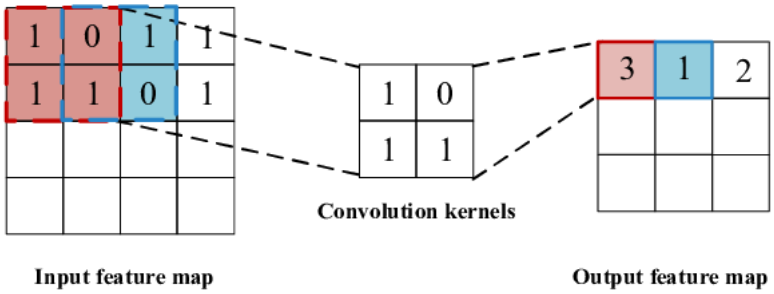


Figure 2.3: *Convolution layer*

Pooling layer

The purpose of the pooling layers is mainly to down-sample the feature maps and thus learn larger-scale features characterized by a spatial invariance to small local transformations (e.g. translation, scaling, and rotation). As a consequence of this resolution downsampling, the pooling layer reduces the computational complexity and the memory requirements while preserving important features that are needed for processing by the subsequent layers. The necessity of pooling layers stems out of the need to learn complex features, from different image resolutions while keeping the number of parameters and the computational cost as low as possible. In addition, pooling layers act as a very effective mechanism to control overfitting and increase the invariance of the learned model parameters. There are mainly two types of pooling, as shown in figure 2.4. **Max pooling** returns the maximum value from the portion of the image covered by the Kernel. On the other hand, **average pooling** returns the average of all the values from the portion of the image covered by the Kernel.

Classification - Fully Connected Layers

After extracting relevant features from the input by the convolutional and pooling layers, a flattening of the output into a one-dimensional vector is performed. This vector is generally fed

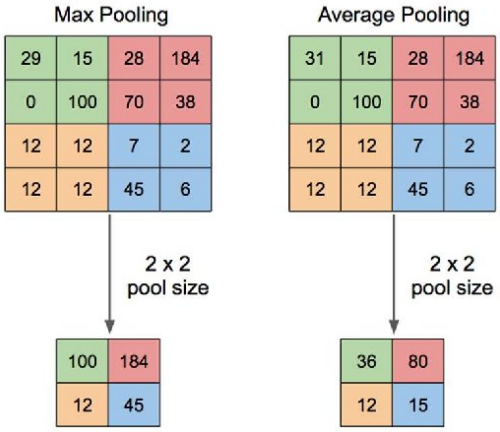


Figure 2.4: *Pooling layer: Illustration of max pooling and average pooling*

to one or more fully connected layers. Adding these fully-connected layers is usually a way to learn non-linear combinations of the extracted features and perform the final task of classification. Typically, the output layer of a CNN is a softmax layer that outputs a probability distribution over the different classes in the dataset. The class with the highest probability is then selected as the final prediction.

It is worth mentioning that CNNs are characterized by some properties over the rest of DNNs. These properties primarily include *"weight sharing"* and *"locality"*, which we elaborate on below:

Weight sharing:

Weight sharing (also called ***parameter sharing***) is the fact that all neurons in a particular feature map share the same weights (kernel parameters). In traditional neural networks, each weight matrix element is only used once during a layer output computation. It is multiplied by a single input element and not used again. Contrastingly, in CNNs, each kernel parameter is utilized at almost every input position. Figure 2.5 is a graphical depiction of how parameter sharing works. Black arrows indicate the connections that use a particular parameter. At the top, we have the case of a convolutional layer, where the black arrows indicate the connection that uses the central parameter of a 3-element kernel. Because of parameter sharing in this model, this single parameter is used at all input locations. Just below, we have an example of a fully connected model where the black arrow indicates the use of the central element of the weight matrix. Since no notion of parameter sharing exists in such a model, each parameter is used only once. Therefore, weight sharing is a key feature in the convolutional layer since it reduces the number of trainable parameters which results in significantly greater memory efficiency and ultimately helps the model to improve generalization and prevent overfitting.

Locality and sparse connections:

In CNNs, locality (also local connectivity) means that each neuron in a convolutional layer is only connected to a small and restricted region of the input (called the neuron ***receptive field***).

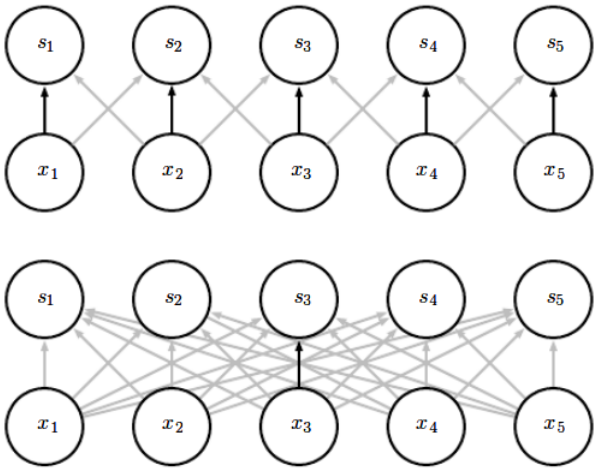


Figure 2.5: *Weight sharing in a convolutional layer (Top) vs. weights in a fully connected layer (bottom) (source:[Goodfellow et al., 2016])*

Typically, this is also referred to as ***sparse connectivity***. This, indeed, is better explained in figure 2.6, where we highlight the sparse connections in a convolutional layer (in the top), while comparing it to the connections in a traditional neural network (in the bottom) where all the neurons are fully connected (i.e. every output unit interacts with every input unit). This allows the network to learn local features such as edges, corners, and other patterns. For example, when processing an image, the input image might have thousands or millions of pixels, but we can detect small, meaningful features such as edges with kernels that occupy only tens or hundreds of pixels. In addition, in a deep CNN, units in the deeper layers may indirectly interact with a larger portion of the input, as shown in figure 2.7. This allows the network to efficiently describe complicated interactions between many variables by constructing such interactions from simple building blocks that each describe only sparse interactions. That is, even though direct connections in a convolutional net are very sparse, units in the deeper layers can be indirectly connected to all or most of the input units which allows the model to capture higher-order information (e.g. spatial or structural information).

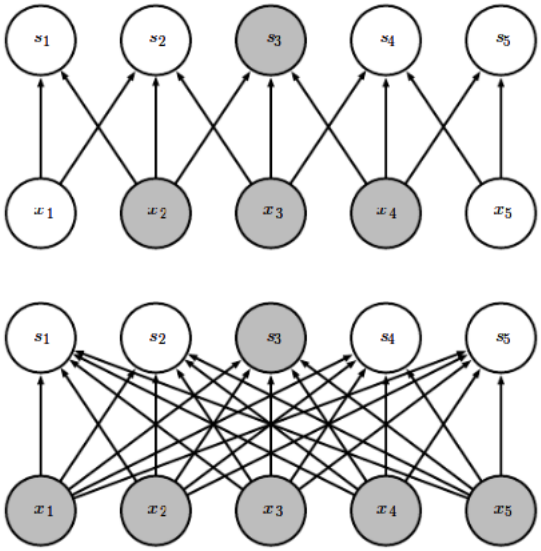


Figure 2.6: *Locality and sparse connections: a particular output unit is highlighted (s_3) with the corresponding input units in x that affect it. (Top) When s is the feature map resulting from a convolution of x with a kernel of width 3, only three inputs affect s_3. (Bottom) When s is formed by matrix multiplication, no longer sparse connectivity exists, and all the inputs affect s_3. (source:[Goodfellow et al., 2016])*

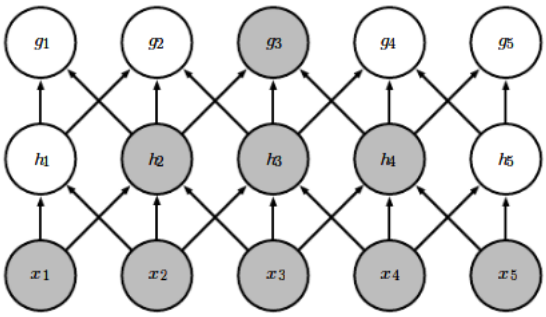


Figure 2.7: *Sparse connectivity in the deeper layers does not restrict the units from being linked to all or most of the input units. (source:[Goodfellow et al., 2016])*

2.2 Applications of DL in speech and voice pathology

Over the last few years, significant progress has been made in understanding pathological speech, thanks to advancements in scientific tools and approaches. As a result, research in this area has become inherently cross-disciplinary, with various fields collaborating. Speech signal processing is one of these fields widely used in the speech pathology context, for instance, to improve the accuracy of pathological voice detection. This was achieved via many methods such as time-frequency approaches [Umapathy et al., 2005], Mel frequency cepstral coefficients (MFCCs) [Fraile et al., 2009], MFCCs with Gaussian mixture models (GMM) [Godino-Llorente et al., 2006], MFCCs with hidden Markov model (HMM) [Costa et al., 2008], and wavelet coefficients

[Fonseca et al., 2005]. Traditional machine-learning-based algorithms have been also investigated to diagnose voice disorders or classify them according to rating scales, using, for instance, GMM [Pouchoulin et al., 2007] SVM [Chen et al., 2007, Markaki and Stylianou, 2011, Arjmandi and Pooyan, 2012], Naive Bayes [Dahmani and Guerti, 2017], KNN [Dahmani and Guerti, 2018, Chen et al., 2021] and have achieved good performance.

More recently, deep learning has shown great strides in several speech pathology-related tasks. In particular, multiple studies have been focusing on the automatic detection and classification of various types of neurodegenerative disorders such as Parkinson's Disease [Chaki and Woźniak, 2023]. These disorders typically worsen over time with no known cure, making early detection and treatment crucial in relieving symptoms. DL can offer an efficient alternative to the manual detection of these disorders. For example, researchers have utilized various deep learning architectures, such as CNNs [Vásquez-Correa et al., 2017, Trinh and O'Brien, 2019, Vavrek et al., 2021], and LSTM [Rizvi et al., 2020, Quan et al., 2021] to classify patients with Parkinson's disease compared to healthy control subjects. In their turn, Aal and al. [Aal et al., 2021] proposed a method for early detection of PD patients using speech features with RNN and LSTM. Deep Autoencoder was used by Hoq and al. [Hoq et al., 2021] for the same objective. Transfer learning techniques were also explored and shown to achieve competitive performance in voice and speech pathology detection and classification [Alhussein and Muhammad, 2018] even when the task includes various languages [Vásquez-Correa et al., 2021]. Furthermore, researchers have also investigated the application of DL to speech pathology analysis and recognition. We can cite the works [España-Bonet and Fonollosa, 2016] and [Zaidi et al., 2021] both dedicated to dysarthric speech analysis and recognition with different DL models including hybrid DNN-HMM, CNN and LSTM. Speech enhancement is another important DL application to consider in the speech pathology context. It is generally applied to improve the accuracy and reliability of speech recognition systems by improving the quality and intelligibility of pathological speech [Sidi Yakoub et al., 2020]. For example, Bhat et al. [Bhat et al., 2018] have proposed a denoising autoencoder based on a time-delay neural network (TDNN) to enhance dysarthric speech features before performing DNN-HMM–based recognition. It is worth mentioning that all of these approaches and tasks have the potential to improve the diagnosis and treatment of speech disorders. In this work, we focus on a particular application, which is the assessment of speech pathology.

Pathological speech assessment

Assessing pathological speech is a crucial diagnostic measure for understanding the effects of a particular treatment on a patient (through a longitudinal comparison), monitoring the progress of pathology, and evaluating the effectiveness of speech therapy. Many studies are available in the literature for the automatic assessment of speech disorders, and particularly intelligibility which is the focus of this work. Among them, we can distinguish different research orientations and applications. From a technological perspective, these approaches can be categorized considering whether they are based on typical acoustic or prosodic features issued from speech signal processing associated with a classifier or regression system, or whether they imply machine learning and deep learning approaches for speech disorder modeling. Briefly, in the first category, extraction methods involve specific features issued from speech signals such as spectral features, articulatory features, prosody features, or voice quality features. This feature extrac-

tion is then coupled with classical classification or prediction approaches (GMM, SVM, DNN, etc.) to achieve speech intelligibility or severity assessment depending on the corpora used. For instance, a feature selection considering three speech dimensions, namely phonetic quality, prosody, and voice quality was proposed by authors in [Kim and Kim, 2012, Kim et al., 2015]. In their turn, [Hahm et al., 2015] and [Orozco-Arroyave et al., 2016] were based on the estimation of articulatory features. In [An et al., 2015], the authors combine the use of classical low-level spectral and voice quality features with speech rate features (syllable and silence duration, syllable amounts, etc.) and phonotactic features (phone duration and monophone-biphone-triphone distributions). In [Narendra and Alku, 2021], a parameterization of glottal flow waveforms is used to extract glottal features, combined with classical spectral and temporal features to improve speech intelligibility assessment rates.

Due to the latest advances in machine learning, and more specifically in deep learning, the majority of the recent approaches tend to touch on these two topics. They belong to the second category of approaches dedicated to speech disorder assessment, which is of great interest in this thesis.

Intelligible speech from healthy speaker signals is exploited in different manners to measure the deviation of impaired speech from a clean reference signal. One approach is directly based on automatic speech transcription using an ASR pre-trained on healthy speech. Speech intelligibility can therefore be estimated as the error rate given by this system while considering pathological speech [Doyle et al., 1997, Schuster et al., 2005, Schuster et al., 2006, Christensen et al., 2012]. For instance, authors in [Maier et al., 2007] and [Riedhammer et al., 2007] have demonstrated high correlations between the outputs of an ASR (e.g. word accuracy and word error rate) trained on healthy speech and the intelligibility scores of patients with speech pathology including cancer of oral cavity. Other kinds of scores were computed by comparing the original word (that the patient has to pronounce from a list of known pseudo-words/words) and the recognized string provided by an ASR system or other automatic speech processing, as done by clinicians in some perceptual assessment tests, or by analyzing phonological features derived from the ASR outputs [Sharma et al., 2009, Maier et al., 2010, Middag et al., 2009, Tripathi et al., 2020, Fredouille et al., 2019]. Although these approaches may be effective for some speech tasks, they may not be appropriate for every type of task. For example, ASR systems may not perform at all for speech tasks involving pseudo-words, pseudo-sentences, or semantically unpredictable sentences that are commonly used in clinical contexts. This is because these systems lack customized language models that are specifically designed for such data. Even the greatest improvements observed with recent DL-based ASR systems, studies conducted in [Green et al., 2021] have shown that only personalized ASR models (trained on specific patient's speech productions) performed very well on short speech utterances of individual dysarthric patients compared to classical ASR systems trained on "healthy" speech (which perform very poorly in disordered speech context).
Finally, in a different way, [Janbakhshi et al., 2019] proposed to align pathological speech signals with reference signals estimated on multiple healthy speakers to bring out deviance. Dynamic time warping and the divergence between the two signals are quantified using the short-time or spectral correlation.

In parallel, other kinds of approaches are proposed to deal with speech intelligibility as-

sessment. Inspired by the speaker recognition field, i-vector-based systems, and more recently x-vector paradigms are explored for instance to model acoustic features associated with disordered speech [Martínez et al., 2015, Laaridh et al., 2018, Quintas et al., 2020]. In addition, the detection of abnormalities in disordered speech at the phone level, based on forced and semi-constrained speech alignment, is investigated in [Laaridh et al., 2015] to assess speech alteration locally and bring information about patients' speech impairment severity. Finally, [Fernández Díaz and Gallardo-Antolín, 2020] proposed the use of an LSTM-based system enhanced by the incorporation of an attention mechanism that is able to determine the more relevant frames for speech intelligibility prediction.

Although the advances in the study of pathological speech using DL architectures, only a few studies addressed this subject from the interpretability/ explainability point of view. Indeed, interpretable/explainable DL-based models are crucial in the medical domain since a lack of transparency can lead to a lack of confidence among healthcare practitioners and patients. This need is further detailed in the next chapter. In the specific context of DL-based models interpretability/ explainability for disordered speech assessment, we can cite mainly two works. Both [Tu et al., 2017] and [Xu et al., 2023] implemented an interpretable solution based on DNNs, and dedicated to the speech severity assessment of dysarthric patients. The main idea of the authors was to incorporate an interpretable layer in the DNN with clinically interpretable labels (e.g. vocal quality, articulatory precision). In this way, clinicians would have a final predicted severity score and can interpret it via this intermediate layer. More details about these approaches are given in the section dedicated to the related works in chapter 7.

2.3 Conclusion

In the majority of presented approaches, the final solution for assessing speech disorders lacks a crucial factor, which is a detailed evaluation of speech intelligibility. This indeed leaves us with limited visibility and understanding of how the scores are derived. Moreover, we can see that an interpretability-performance dilemma exists, wherein DL-based approaches lack interpretability, while handcrafted feature-based approaches are less performant. We try to find a compromise in our proposed approach by formulating an assessment solution that integrates technical and methodological choices driven by these limitations and resources at our disposal. In the following, we briefly introduce the main axes and choices we were based on.

1. **Deep Learning based approach:** Deep learning models are capable of *automatically learning and extracting relevant features* from raw input data, such as speech signals. This eliminates the need for manual feature engineering, which can be time-consuming and may require domain expertise. Besides, DL-based approaches generally outperform handcrafted feature-based approaches in pathological speech classification tasks due to the ability of deep learning models to capture complex patterns and *non-linear* relationships in the data.

2. **Reference-based Assessment:** Deep learning-based approaches require large amounts of training data to achieve high performance, while handcrafted feature-based approaches can often achieve good results with smaller datasets. Obtaining a substantial quantity of data, particularly in case of pathological speech can be a challenging and costly endeavor,

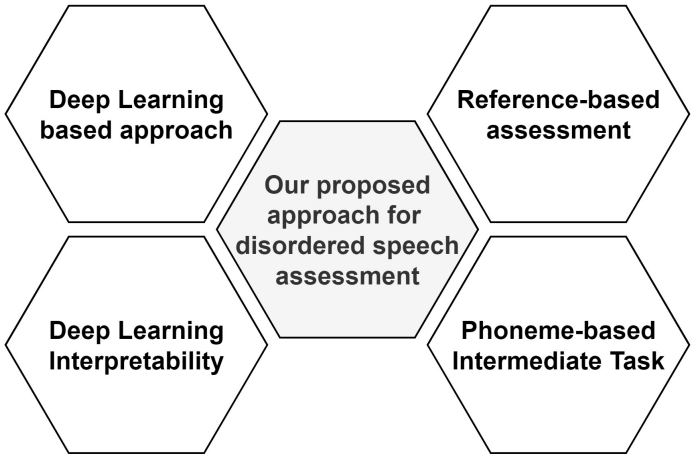


Figure 2.8: *The main choices around our proposed approach for disordered speech assessment*

making this factor a significant point to consider. By considering *healthy speech* in our proposal, notably for the training of certain models, and, therefore reference-based assessment, we partially tackle this issue.

3. **Deep Learning Interpretability**: Handcrafted feature-based approaches are generally more interpretable than deep learning-based approaches, as the features used in the classification model can be directly related to the underlying speech disorder. They also require less computational resources. In contrast, deep learning models are often considered as black boxes, making it more difficult to understand the specific features that are driving the classification. Since we adopt a DL-based approach in this work, we lay a special focus on the interpretability of these tools for transparency and reliability reasons.

4. **Phoneme-Level Intermediate Task:** All the studies cited above share the same objective of evaluating speech intelligibility in the clinical context, by providing a single score, considering one or more dimensions of speech. If these approaches take into account, often effectively, speech disorders and their impact on the speech signal in terms of alterations, the vast majority of them do so implicitly. Indeed, these approaches are ultimately unable to precisely define the link between speech disorders and the intelligibility score they obtain and make their decision in a way that we consider as *blind*. However, this analysis is important to guide the clinician in his/her clinical evaluation, whether upstream of therapeutic management or to assess the benefit of rehabilitation.

Chapter 3

Literature review on the interpretability/explainability

Contents

The explainability and interpretability concepts are utilized in several fields, spanning from mathematics, physics, computer science to engineering, psychology, medicine, and social sciences [Abdul et al., 2018]. In this chapter, we introduce these concepts, more oriented toward the medical context. This chapter is then organized into three distinct parts. In the first part, we highlight the need for interpretability from different perspectives focusing more on the medical field due to the high stakes of medicine-concerned applications. In the second, we introduce the terminology to clarify its subsequent use, followed by the taxonomy. Finally, we briefly report some of the challenges that can be encountered when trying to systematically bring interpretability.

3.1 The need and application

In this section, we shed light on the relevance of DL interpretability in clinical practice not only from a medical point of view, which is the context of this work, but also from a legal and technological perspective [Amann et al., 2020].

3.1.1 Legal perspective

From the legal perspective, we will discuss to what extent explainability in AI is legally required. Under the European Union's General Data Protection Regulation (GDPR) [GDPR, 2016], a description of the decision-making process of a system performing automated processing of personal data should be possible whenever the user asks for it. Indeed, article 15(h), reported in appendix B.1.1, sets out the right for individuals to obtain an explanation of the inference(s) automatically produced by a model. In addition, article 22, reported in appendix B.1.2, grants individuals the "right of human intervention" under which they may ask for a human to review the AI's decision to determine whether or not the system made a mistake, particularly when it might have a negative legal, financial, mental or physical effect on the individual. More recently, a proposal for a regulation on artificial intelligence was announced by the European Commission in April 2021, the so-called "Artificial Intelligence Act" (AI act) [AIA, 2021]. The requirements for transparency of high-risk AI systems laid down in the AI act are certainly a step in the right direction. Indeed, Art. 13, reported in appendix B.2.1, states that: *"High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately."* While this article refers to the interpretability, we found that Recital 38 of the AI act calls for "explainable" AI systems, (reported in appendix B.2.2), which brings to the fore the issue of terminology ambiguity that we are going to address in the next sections. In addition, the AI act has been criticized for the non-specification of the technical controls that need to be taken to ensure that AI systems are sufficiently transparent and not left to the discretion of the AI system providers, as revealed in [Ebers et al., 2021]. The increasing interest given by the legislation in this context is certainly a step in the right direction to create a safe and reliable regulatory environment for AI, however, certainly many improvements and clarifications need to be made.

3.1.2 Technological perspective

From the technological point of view, explainability has to be considered both in terms of how it can be achieved and to what it is beneficial from a development perspective. Explainability is of a great interest for getting insights into what is called black-boxes to understand the what and the why of a decision-making process. Indeed, it is very important to demonstrate that these tools learned valid and generalizable properties, ruling out the possibility that their performance is based on meta-data and spurious correlations rather than the data itself. This phenomenon is referred to as the *"Clever Hans"* phenomenon [Lapuschkin et al., 2019]. A famous non-medical example of this phenomenon is the "Husky vs Wolf" classifier, which predictions were proven to be solely driven by the identification of a snowy background rather than real differences between huskies and wolves [Ribeiro et al., 2016], see fig. 3.1.

(a) Husky classified as wolf (b) Explanation

Figure 3.1: *Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task [Ribeiro et al., 2016]. (a) The image of the husky being misclassified as a wolf, (b) Explanation showing that the misclassification was driven by the identification of the snowy background.*

The *"Clever Hans"* phenomenon has also been reported by [Zech et al., 2018] in the medical field. Indeed, researchers from Mount Sinai Hospital developed a model that performs very well in distinguishing high-risk patients from non-high-risk patients based on X-ray imaging. It turned out that the model was not based on clinically relevant information from the images related to the risk of patients, but rather on a simple distinction of the machine used for imaging. In analogy to the snowy background in the example introduced above, the prediction was based on hardware-related meta-data tied to the specific x-ray machine that was used to image the high-risk patients exclusively at Mount Sinai. In summary, explainability methods allow AI experts to have insight into their models and identify errors and biases (e.g. the Clever Hans predictors) before AI tools go into clinical validation, which saves time and development costs. More generally, it is important ***to justify*** the decision made by a model, ***to control*** its functioning allowing its debugging and the identification of potential flows, ***to improve*** the accuracy and efficiency of a model and finally ***to discover*** the knowledge acquired by the model and the hidden patterns.

3.1.3 Medical perspective

Looking at the issue of explainability from a medical perspective emphasizes the importance of considering the interaction between human actors and medical AI [Kundu, 2021]. In clinical practice, AI often comes in the form of clinical decision support systems (CDSS), assisting clinicians in the diagnosis of disease and treatment decisions [Sutton et al., 2020]. With the significant advancements of AI in healthcare [Rajpurkar et al., 2022], AI-based CDSSs emerged to support decision-making in many aspects. Notably, we can cite CDSS for breast cancer diagnosis and classification on ultrasound images [Ragab et al., 2022], the prediction of the quality of life of ALS patients [Antoniadi et al., 2021], the identification of prescriptions with a high-risk of medication error [Corny et al., 2020] and so many others. Yet, trust in AI-driven CDSS is not yet established and explainability may be a pivotal driver to uptake these systems in clinical practice [Cutillo et al., 2020, Tonekaboni et al., 2019]. The World Health Organization AI Guidelines for Health have highlighted the need to guarantee explainability as a guiding factor

for the effective application of AI in healthcare [World Health Organization, 2021]. In the medical field, the aim of explainability is to demonstrate to clinicians how various factors contribute to the final recommendation. Based on their experience and clinical judgment, clinicians can therefore make an informed decision about whether or not to rely on the system's recommendations. Particularly in cases where the CDSS produces recommendations that are strongly out of line with a clinician's expectations, explainability allows verification of whether the parameters taken into account by the system make sense from a clinical point of view. An example of CDSS proving that this explainability characteristic is paramount for healthcare professionals is the model of [Caruana et al., 2015] which demonstrated high performance in pneumonia[1] risk detection. When analyzing the predictions of this model, it has been reported that cases of pneumonia with concurring asthma were assigned a lower risk of death than those without, despite the fact that the presence of this underlying condition has been always known to worsen the severity of the cases. A correct prediction therefore would have been the opposite diagnosis. The misleading correlation (i.e. presence of asthma and thus low risk of death from pneumonia) was rather a consequence of the effective care given to these patients by healthcare specialists. Given these considerations, explainability assisted the clinicians to capture this misleading feature and identify false negatives in which they must not rely on the system's recommendations, consequently, strengthening their trust in the system. That is an appropriate reliance on a CDSS, where healthcare practitioners follow the correct outputs and reject incorrect ones, as illustrated in Fig. 3.2. The misuse of the CDSS, on the other hand, can be due to over-reliance (i.e. healthcare practitioners putting too much trust on the CDSS, even following the incorrect outputs) or self-reliance (i.e. healthcare practitioners neglect the correct outputs).

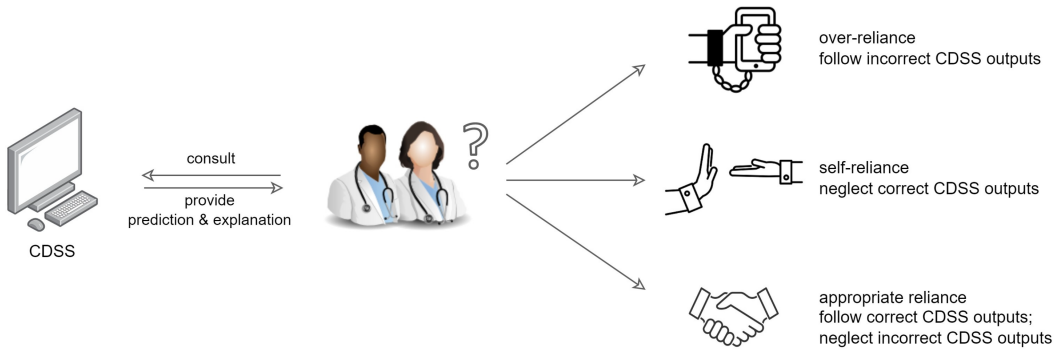


Figure 3.2: *Interaction between healthcare practitioners and a CDSS, classified by [Du et al., 2022] into over-reliance, self-reliance, and appropriate reliance*

3.1.4 The patient perspective

When considering explainability from the patient's point of view, the question that arises is whether using AI-powered decision aids is consistent with the core principles of patient-centered care. Patient-centered care considers patients as active partners in the care process, emphasizing their right to understand risks and outcomes, to explore the available options, and to determine which course of action best fits their goals and priorities [Politi et al., 2013]. It is clear that

[1] Pneumonia is an infection that inflames the air sacs in one or both lungs.

the so-called *'black-box'* in medicine conflicts with core ideals of patient-centered care. Indeed, it makes the clinicians no longer able to make sense of the inner functioning of these tools and, therefore, not able to provide explanations of the decision-making process to the patients [Bjerring and Busch, 2021]. In particular, black-box medicine is not conducive to supporting informed decision-making based on shared information between patient and clinician. The capacity of explainability to address this issue is evident in this case since it allows providing clinician and patient with a personalized conversation. The concept of contestable AI decision-making in a clinical context was presented in [Ploug and Holm, 2020]. Taking a patient-centric approach the authors argue that patients should be able to contest the diagnoses of AI diagnostic systems. To effectively contest AI diagnoses on patient-relevant aspects, it is necessary to have access to diverse information concerning the AI system. In other words, contestability means that the decision algorithm has to provide information about the data used, any system biases, system performance in terms of algorithmic metrics, and the decision responsibility carried by humans or algorithms.

3.2 Terminology

In this section, we provide the background of the key concepts of interpretability and explainability. Then, we point out the meaningful differences between them based on the literature. An overview of the concepts connected with these terms in the machine learning field is also presented. This section is a real challenge in a field where terminology is very ambiguous and its foundation stays in the intersection of several fields such as psychology [Malle, 2011], social science [Miller, 2019], cognitive science [Cimpian and Salomon, 2014] and philosophy [Thagard, 1978]. We are thus aware that the definition of these concepts goes beyond what we present in the following, but a selection has to be made focusing on those from the AI community.

3.2.1 Interpretability

Regarding interpretability, the definitions of this term generally lack mathematical formality and rigorousness. In his book, Molnar notes [Molnar, 2022] that *"interpretable machine learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans"*. In agreement with Biran and Cotton [Biran and Cotton, 2017]'s definition of interpretability, Miller [Miller, 2019] reported that: *"Interpretability is the degree to which a human can understand the cause of a decision"*. Kim et al. [Kim et al., 2016] describe interpretability as *"the degree to which a human can consistently predict the model's result"*. Doshi-Velez and Kim define interpretability in [Doshi-Velez and Kim, 2017] as the *"ability to explain or to present in understandable terms to a human"*. Given the lack of a "formal technical meaning", an interesting point of view was given by Lipton [Lipton, 2018] revealing that the concept of interpretability is not a monolithic one, but in fact, reflects several distinct ideas such as trust or transparency.

3.2.2 Explainability

According to [Guidotti et al., 2018], explainability is associated with the notion of *explanation* seen as an interface between humans and a decision maker that is, at the same time, both an

accurate proxy of the decision maker and comprehensible to humans. More recently, authors in [Barredo Arrieta et al., 2020] proposed a novel definition of explainability that places the audience and the purpose as the key aspects to be considered when explaining an ML model, and thus proposed the following definition: *"Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand"*. Their reasoning is summarized in figure 3.3

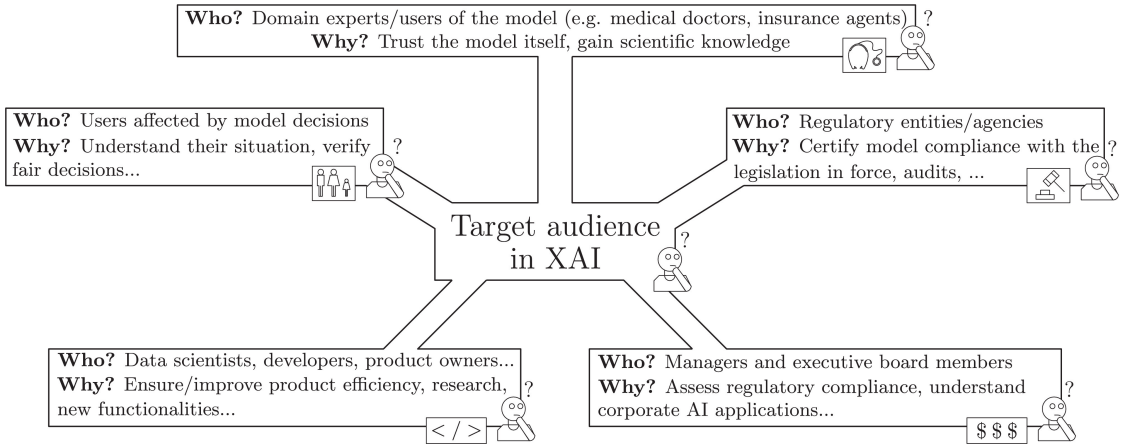


Figure 3.3: *Diagram showing the different purposes of explainability in ML models sought by different audience profiles. [Barredo Arrieta et al., 2020]*

3.2.3 Interpretability vs. explainability

Interpretability and explainability are both continuums, sometimes with blurred edges of where interpretability ends and explainability begins. These terms have recently been the focus of many researchers from different fields, and given the complexity of the subject, there is still no agreement on a single definition or taxonomy.

Even though *"explainable"* is a keyword in the XAI appellation, Adadi and Berrada reported in their survey [Adadi and Berrada, 2018] that the term *"interpretable"* is more used in the ML community, being confirmed by the Google trends comparison between the use of both terms in the scientific context until 2018. Practically speaking, the terms interpretability and explainability are frequently used interchangeably by researchers as [Miller, 2019] and [Molnar, 2022]. While these authors equated the terms interpretability/explainability (i.e. interpretable/explainable), they distinguished them from the term explanation which they defined as an answer to a **why-question**.

On the other hand, while these terms are very closely related, several studies attempt to highlight the distinction between them [Vilone and Longo, 2021, Gilpin et al., 2018]. To help make the distinction clearer, we illustrate in this section some of the works in which the definition is based on contrasting both terms.

A distinction between the concepts of interpretation and explanation was proposed in [Montavon et al., 2018]. On one hand, authors defined **an interpretation** as *"the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of"*, for instance, images or texts as they can be inspected by human unlike vector spaces (e.g. word embeddings, wav2vec output). On the other hand, they defined **an explanation** as *"the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)"*. An example of explanation can therefore be seen as a heatmap highlighting which pixels of the input image most strongly support the classification decision.

In their turn, authors in [Barredo Arrieta et al., 2020] highlighted that the interchangeable misuse of both terms interpretability and explainability is a clear issue that hinders the establishment of common grounds in the field. To clarify the difference, they defined **interpretability** as referring to *"a passive characteristic of a model referring to the level at which a given model makes sense for a human observer"*. This feature is also expressed as transparency. By contrast, they defined **explainability** as *"an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions"*.

More recently, [Namatēvs et al., 2022] attempted to define the boundaries between interpretability and explainability in DL based on an extensive literature review, and revealed the following: *Interpretability means the ability of a human to understand and trust the decision of the DL model's results.* Explainability is the ability by which a human can justify the cause of the explanatory rule of the DL model's results.

3.2.4 Key related concepts

In view of this diversity, a wide range of aspects that are significantly related to interpretability/explainability appeared. A great number of systematic literature surveys were therefore proposed in the last years aiming to define these key concepts in machine learning but focusing mostly on deep learning context. We sum up some of these key concepts as well as their definitions in table 3.1.

Notion	Definition and references
Causality	The capacity of a method for explainability to clarify the relationship between input and output [Lipton, 2018]
Effectiveness	The capacity of a method for explainability to support good user decision-making [de Fine Licht and de Fine Licht, 2020]
Explicitness	The capacity of a method to provide immediate and understandable explanations [Alvarez Melis and Jaakkola, 2018].
Informativeness	The capacity of a method for explainability to provide useful information to end-users [Lipton, 2018]
Justifiability	The capacity of an expert to assess if a model is in line with the domain knowledge [de Fine Licht and de Fine Licht, 2020]
Transparency	The capacity of a method to explain how the system works even when it behaves unexpectedly [Lipton, 2018, de Fine Licht and de Fine Licht, 2020]

Table 3.1: *Definition of the notions related to the concept of interpretability/explainability*

3.3 Taxonomy

Methods for ML interpretability/explainability can be classified according to various criteria. Many works tended to define this taxonomy, among which we found Molnar's book [Molnar, 2022], [Barredo Arrieta et al., 2020] and [Carvalho et al., 2019]. Figure 3.4 is a summary of this taxonomy performed in [Linardatos et al., 2021].

3.3.1 Model-specific vs. model-agnostic

Broadly speaking, ML interpretability can be categorized into model-specific or model-agnostic approaches. This duality is described as the difference between interpretation tools that are limited to specific model classes as opposed to those that can be applied to any ML model disregarding its inner processing or internal representations (model-agnostic). Some examples of model-specific methods designed for DNNs include guided backpropagation [Springenberg et al., 2015], integrated gradients [Sundararajan et al., 2017], and Gradient-weighted Class Activation Mapping (Grad-CAM) [Selvaraju et al., 2017]. Among the model-agnostic explanation methods, we can cite the SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] and LIME (Local Interpretable Model-agnostic Explanations) [Ribeiro et al., 2016].

3.3.2 Local vs. global

Seen from another angle, the interpretability methods can be categorized into local methods focusing on the explainability of an individual prediction (i.e. understanding the reasons for a specific decision), as against global methods targeting the explainability of the entire model behavior (i.e. the whole logic of a model can be understood, and following the entire reasoning leads to all the different possible outcomes). In addition to the above-mentioned methods,

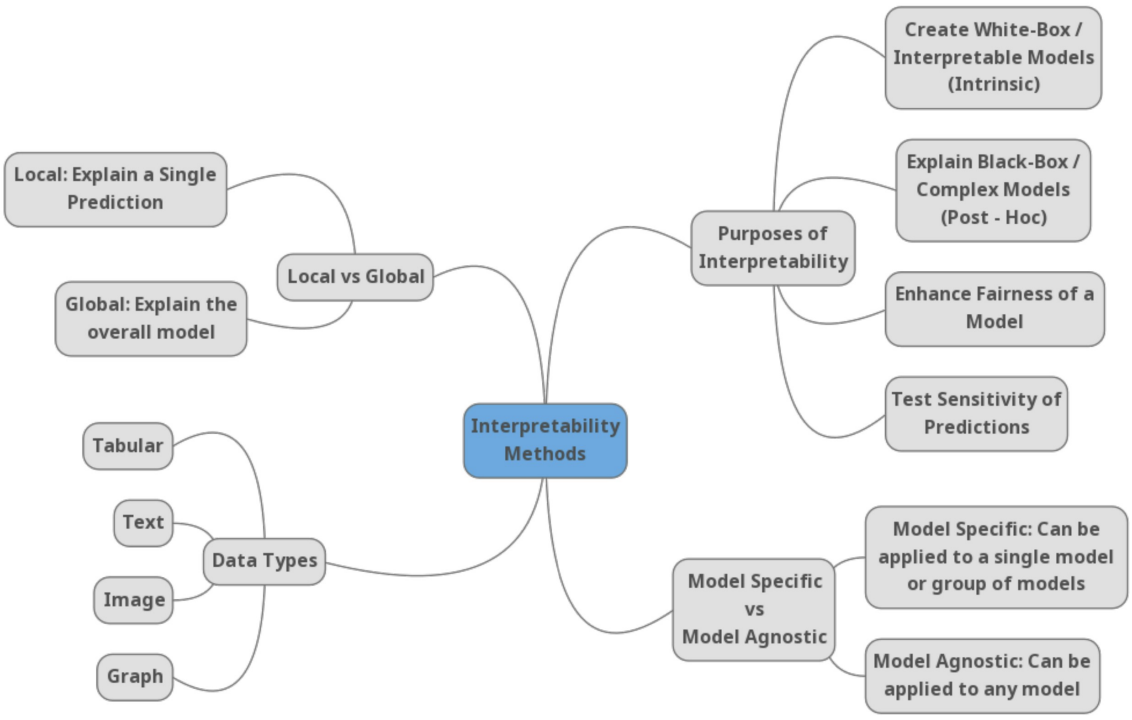


Figure 3.4: *Taxonomy mind-map of ML interpretability techniques [Linardatos et al., 2021]*

SHAP and LIME are considered as local interpretability methods to explain prediction yielded by a single instance. Other examples are the Layer-wise Relevance Propagation (LRP) [Bach et al., 2015], and the deep Taylor decomposition [Montavon et al., 2017], which, for a given input instance, decompose the output of a neural network into contributions of this instance by backpropagating the explanations from the output layer to the input.

3.3.3 Intrinsic vs. post-hoc

From another point of view, these methods can be classified into intrinsic or post-hoc methods. This criterion distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post-hoc), see figure 3.5.

In the literature, this duality was also regarded as the difference between two kinds of models : (1) models that are interpretable by design but with relatively low performance (e.g. linear models, rule-based models, decision trees). These models are also known as *transparent* or *white-box models*. (2) models that are more complex and achieve better performance (e.g. CNN, RNN, ensemble models) while having a lower explainability. They are therefore considered as black-boxes that need to be explained by means of post-hoc XAI strategies. For instance, commonly used strategies are reported in figure 3.6 including *explanations by example*, *explanations by simplification*, *feature relevance explanations*, and *visual explanations*.

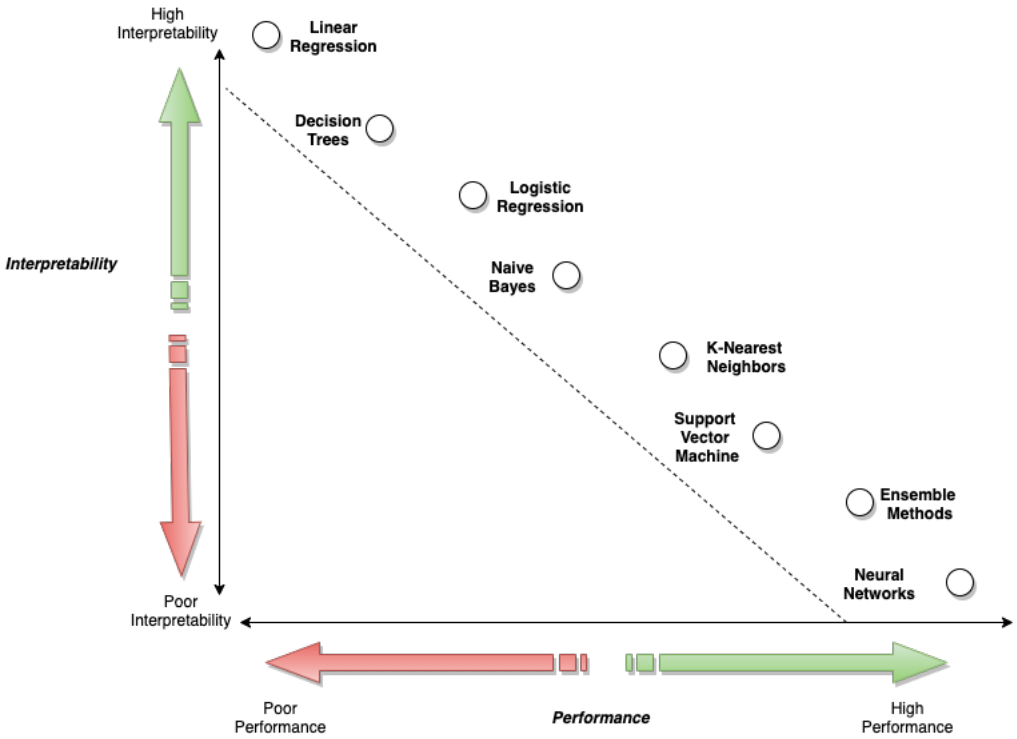


Figure 3.5: *Interpretability versus performance trade-off given common ML algorithms (source: https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html)*

3.4 Challenges

All along the previous sections, we highlighted the reasons making interpretability/explainability a valuable and even an indispensable property in some cases. Clearly, for all these reasons, the awareness and demand for it are growing in various domains. Nonetheless, it is worth raising the question "Why it is not evidence and thus everyone uses it?"
Upon identifying the challenges to systematically bring interpretability for every model, many surveys addressed this point considering multiple aspects [Adadi and Berrada, 2018, Rudin, 2019, Fan et al., 2021]. In this section, we highlight some of these challenges.

Algorithmic Complexity

Despite the fact that non-linearity may not necessarily result in opacity (e.g. a decision tree), it becomes more complicated to understand the inner working of a model once this non-linearity is spread over many hidden layers as in the case of DNNs. In addition to non-linear activations, other specificities of deep architectures make even more difficult the task of interpretability such as convolution, pooling, shortcuts, the high number of trainable parameters, etc. There can be no doubt that all of these complex factors are behind the extraordinary performance achieved by these tools since they allow DNNs to intrinsically consider high-degree interactions between

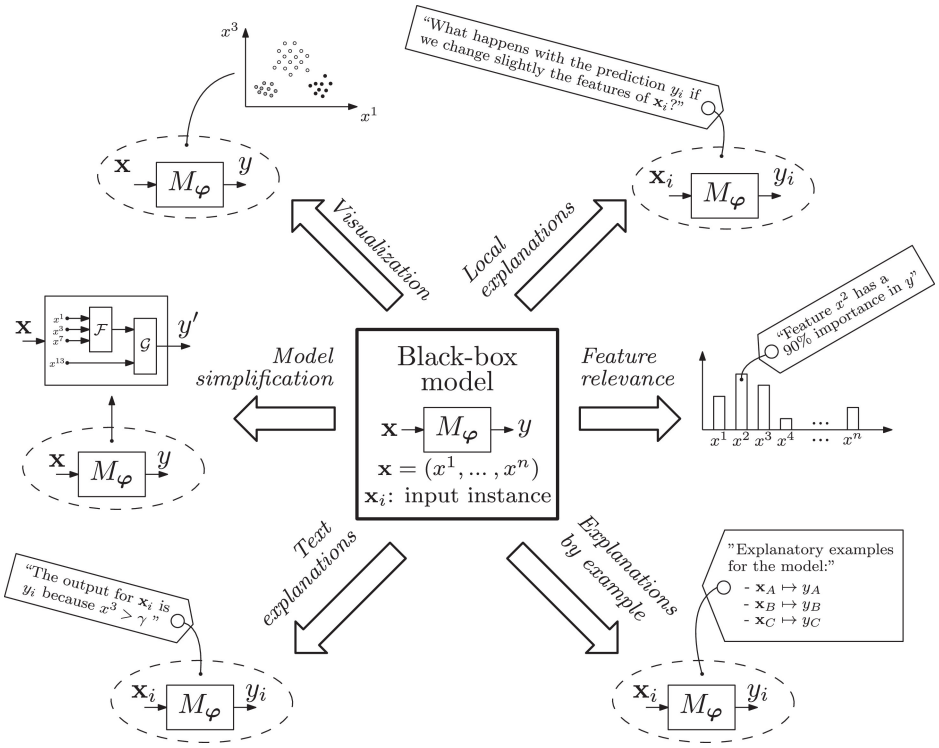


Figure 3.6: *Conceptual diagram showing the different post-hoc explainability approaches available for an ML model [Barredo Arrieta et al., 2020]*

input features. However, traducing such interactions into human understandable form is a super difficult task and potentially even a questionable one. Interestingly, [van der Maas et al., 1990] have shown that even simple neural networks can reveal a chaotic behavior (i.e. tiny changes of initial inputs may lead to huge outcome differences in these models), which once more confirms the complexity of interpretation of such tools.

Lack of objective evaluation metrics

One of the great challenges of the XAI is to establish an objective measure of what constitutes a good explanation. Actually, this is strongly dependent on the audience to which this explanation is addressed as pointed out in section 3.2.2, since providing an explanation will clearly not be the same for an expert in the field, a policy-maker or a user without ML knowledge [Langer et al., 2021]. Until the adoption of such an objective metric, it appears necessary to make an effort to rigorously formalize evaluation methods.

Commercial barrier

In the commercial world, companies' motivation is basically the high performance of a system regardless of its level of transparency. Indeed, prototyping an interpretable model may cost too much in terms of financial, computational, and other resources. In fact, existing open-source good models can be used to quickly build a well-performing algorithm for a certain task. How-

ever, it takes significantly more effort to produce an accurate and consistent knowledge of the behavior of the resulting model. Interestingly though, the black-box property is a considerable advantage for companies as long as customers are satisfied since it will prevent their competitors from stealing their intellectual properties easily [Rudin, 2019].

Data wildness

Real-world data are increasingly more accessible in many domains. Obviously, such data are characterized by heterogeneity, inconsistency, and high dimensionality and are subject to different types of measurement errors and biases [Liu and Demosthenes, 2022]. These characteristics hamper not only the accuracy of ML models but also the construction of interpretability.

3.5 Conclusions

In this chapter, we introduced the background related to interpretability and explainability in a deep learning context. This introduction is of great interest since we would like to be aligned with both the terminology and taxonomy of the literature. We will, therefore, be based on these introduced concepts to justify our later choices in the upcoming chapters. It is worth mentioning that due to significant confusion in the literature review, we have used the terms "explainability" and "interpretability" interchangeably throughout this chapter. However, we are convinced and we consider later that there is a difference between these two concepts. Therefore, it becomes essential to specify the definitions that we adopt in the rest of this document.

Mostly aligned with the **interpretability** definition of [Montavon et al., 2018], this study considers interpretability as *"the mapping of an abstract concept (i.e. a prediction) into a domain that the human can make sense of"*. On the other hand, we emphasize alignment with the **explainability** definition proposed in [Gilpin et al., 2018]: *"The explanation of deep network representations aims to understand the role and structure of the data flowing through these bottlenecks"*. This choice of these definitions is driven by our application domain. It will be further detailed and justified in the next chapter.

Part II

Contribution

Chapter 4

General Context

Contents

This chapter is an introduction of the general context within which this thesis is conducted. We will start with a brief description of the RUGBI project and the composition of its consortium in order to better explain our contribution in this project. We, therefore, outline the corpora we used to achieve our objective. Finally, we present an overview of the proposed methodology, without getting into details. We organize each step of the proposed methodology as a chapter in the rest of this manuscript.

4.1 RUGBI Project

The RUGBI acronym of the project stands for *looking for **R**elevant linguistic **U**nits to improve the intelli**GiBI**lity measurement of speech production disorder*. It aims to develop an objective evaluation tool for speech intelligibility in the context of speech disorders. It is a multidisciplinary project involving a coordinated effort that brings together several disciplines, including the pathology branch of medical sciences dedicated to ear, nose, and throat (ENT) clinical specialty, speech therapy, linguistics, and computer sciences involving automatic speech processing, and more specifically clinical phonetics. This multidisciplinarity is certainly a key asset offering multiple perspectives and a broad range of expertise for generating unique and creative solutions.

RUGBI consortium is composed of 4 academic partners and one university hospital. In the following, we highlight the role of each of these partners in order to clarify our contribution to this project:

- **CHU:** Toulouse Hospital is the provider of the dataset for disordered speech and represents the clinical expertise to which the outcome of this project is destinated.
- **LIA:** The role of the Laboratoire Informatique d'Avignon (LIA), that we represent in RUGBI project, is to investigate the **acoustic-phonetic units** responsible for speech intelligibility, and how both units and intelligibility are impacted in the context of speech disorders. All the contributions brought in this thesis work respond to these objectives.
- **LPL:** The Laboratoire Parole et Langage (Speech and Language Lab or LPL) is involved in the identification of relevant units and tasks for the evaluation of intelligibility in speech disorders in the **acoustic-phonetic** domain.
- **IRIT:** Institut de Recherche en Informatique de Toulouse (IRIT) is involved in the automatic evaluation of speech disorders based on DL approaches and also in the automatic identification of **prosodic linguistic units** responsible for speech intelligibility in speech disorders.
- **LNPL:** The Laboratoire de NeuroPsychoLinguistique (LNPL) is the main collaborator of the analysis in the **prosodic domain** involved in the modeling and interpretation of prosodic linguistic units in disordered speech.

4.2 Data corpora

In this section, we introduce the main corpora used in this thesis, namely **BREF** [Lamel et al., 1991] dataset as a reference for healthy read speech and **C2SI** [Woisard et al., 2021] for speech disorders due to head and neck cancer, both in the French language.

4.2.1 BREF: Reference dataset for healthy speech

Developed in 90's at LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur), BREF-120 corpus [Lamel et al., 1991] is composed of French read-speech recordings produced by 120 speakers (65 women and 55 men) recruited from Paris. BREF-120 contains 100 hours of speech with approximately 650 sentences per speaker. The speakers' ages range from 20 to 65 years. This corpus was designed to provide continuous speech for the development and evaluation of ASR and dictation systems as well as the study of phonological variations. The recordings were made in stereo in a sound-isolated room. The textual content is sentences selected from the French newspaper *"Le Monde"* in order to maximize the number of phonemic contexts and the number of different words. Some distributional properties of BREF corpus are given in table 4.1.

The choice of **BREF** as a healthy speech reference was driven by different reasons. First of all, various comparative experiments (not presented in this document) were conducted, comparing the use of **BREF** with French Broadcast news corpora like ESTER, and ETAPE (focusing

only on the specific conditions of prompted and prepared speech) for the tasks targeted by the thesis work [Abderrazek, 2019]. *BREF* corpus was found to be best suited to the targeted objectives. Our assumption is that *BREF* recording conditions (clean conditions, read speech) are closer to those of classical impaired speech recording protocols. More recently, similar French corpora have been made available for research. We can cite *LibriVox*, which is composed of 140 hours of recordings selected from French books read by native and non-native speakers. The main issue of this corpus is the lack of information regarding the speakers and their origins to serve as reference speech. Still, considerations regarding the recording conditions lead us to dismiss another more recent French corpus, *French CommonVoice*.

unit	**value**
#sentences	167.359
#words	4.244.810
#phones	16.416.738
#distinct phones	35

Table 4.1: *Distributional properties of BREF corpus*

4.2.2 C2SI: Dataset for disordered speech due to Head & Neck Cancers

The main data for disordered speech are issued from the C2SI project (Carcinologic Speech Severity Index), granted by the INCa (*"Institut National du Cancer"*). The C2SI study objective is to assess how treatment for upper aerodigestive tract cancers (i.e. pharynx and oral cavity) affects speech production using both perceptual and automated speech processing techniques.

Population

C2SI corpus includes 87 patients and 40 healthy controls, where 7 patients were recorded twice. The mean age is 56.9 years old (range 35-79) for healthy controls and 65.8 years old (range 36-87) for patients. The patients were visiting the IUCT Oncopole *"Institut Universitaire du Cancer Toulouse Oncopole"* for a follow-up appointment after treatment for oral or oropharyngeal cancer between 2015 and 2016. They had to have completed the treatment plan six months prior to enrollment and be in clinical remission in order for their speech impairment to be as stable as possible. These aspects enabled patients to be in a context of chronicity. It is worth mentioning that patients with speech disorders that might be related to another pathology, such as those following a cerebrovascular accident or disorders of fluency like stuttering, were excluded from the study. As well, C2SI corpus includes some clinical information about the patients such as the treatment type (surgery, radiotherapy, chemotherapy), cancer region, values of T and N criteria from UICC Tumor/Node/Metastasis (TNM) classification [Sobin et al., 2011], etc. Table 4.2 shows the distribution of patients according to the anatomical region affected by the cancerous lesion and the tumor size T (T1 reflects a tumor size equal or less than 2cm; T2 is for a tumor size from 2 to 4cm; T3 tumor size larger than 4cm; and T4 is for a tumor invading surrounding structures), two criteria of interest for us for later analysis.

Criteria	#Patients
Tumor region	
Oral cavity	35
Oropharynx	52
Tumor size	
T1	11
T2	33
T3	12
T4	31

Table 4.2: *Distribution of patients according to the tumor size and region.*

Different recorded tasks

In order to acquire the best quality and prevent biases from expert evaluations, the patients were sat in an anechoic room, located in the onco-rehabilitation unit of the IUCT Oncopole, in front of a microphone with a pop shield filter. Audio files sampled at 44 or 48 kHz were recorded with a digital recorder. Several tasks were recorded, each conceived for a specific type of analysis. In the following, we briefly introduce these tasks (the reader may refer to [Woisard et al., 2021] for more details):

- **Sustained Vowels (AAA):** These recordings are made up of three sustained /a/ sounds. A sustained vowel provides details regarding voice quality, phonation time, stability of production, harmonic content, noisy speech due to speech disorders, etc.
- **Pseudo-words (DAP):** Each speaker had to pronounce 52 pseudo-words. Each pseudo-word has a specific phonotactic structure: C(C)1V1C(C)2V2, where C(C)i is an isolated consonant or a consonant cluster as detailed in [Ghio et al., 2016].
- **Passage Reading (LEC):** This task consists of reading the first paragraph of the tale *"La chèvre de M. Seguin"* by Alphonse Daudet. The length and coverage of all French phonemes were behind the choice of this passage. This text is widely used in clinical phonetics in France and can be found in the appendix A.
- **Picture Description (DES):** 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 on the basis of the oral explanations. This task was designed in order to reduce speech predictability.
- **Spontaneous speech (SPO):** The patient was required to express his/her thoughts on a questionnaire that must be completed prior to the recording session. He/she had to talk for at least three minutes. With no restrictions on the sentences, this activity enables the collection of spontaneous speech recordings.
- **Prosodic tasks:** Three prosodic tasks were designed to evaluate which structural functions of prosody are most affected by the types of cancer in C2SI corpus:
    1. **Modality Function (MOD):** prosodic marking of assertion, question, and injunction, by intonation contour shapes and directions.

2. **Pragmatic Focus (FOC):** This task required speakers to mark the pragmatic focus by highlighting the important information of an utterance by sole prosodic cues.
3. **Syntactic Disambiguation (SYN):** speakers had to solve syntactic ambiguity by prosodic means in syntagms composed of two nouns and an adjective, where the adjective either applied to both nouns (high syntactic attachment) or to the last noun (low syntactic attachment).

- **True/False Sentences (SVT):** 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.

Perceptual evaluation

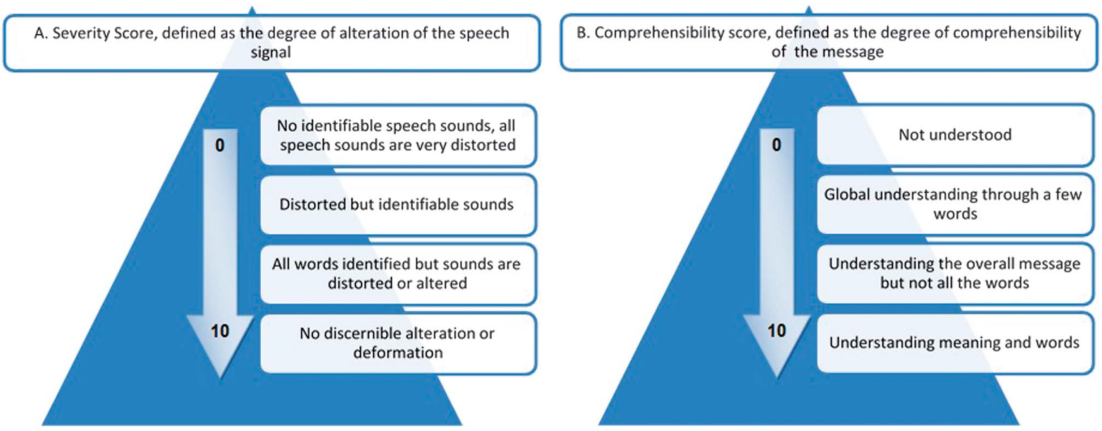
The recordings were therefore analyzed by a jury composed of six clinicians whose expertise area is speech disorder evaluation. It is worth noting that an Interclass Correlation Coefficient (ICC), $r > 0.69$ was reached, which is considered as a good degree of concordance between the jury ratings for the set of tasks as reported in [Woisard et al., 2021].
Among the different perceptual measures that were conducted, we outline the most significant ones in relation to the current study in the following. We use an abbreviation to note these measures with a specification of the task on which they were obtained.

- **Degree of Alteration - Voice quality, Resonance, Prosody, and Phonemic Alteration (Phnm-DES):** Experts were asked to evaluate the degree of alteration following four perceptual dimensions including voice quality analysis, resonance, prosody, and phonemic alteration (Phnm-DES), on the picture description task-based recordings (DES). No definition of these concepts was given to the experts. The alteration index is between 0 (normal) and 3 (severe impairment).

- **Intelligibility and Severity on the picture description task (Intel-DES and Sev-DES):** An index of severity (Sev-DES) and intelligibility (Intel-DES) were provided by each expert on the recordings of the picture description task. The instructions given to the experts included the following definitions [Balaguer et al., 2019]. Intelligibility is defined as *"the comprehensibility of the message sent by the signal"*, while severity is defined as *"the degree of the overall deterioration of the audible signal"*. The severity and intelligibility are assessed on a scale from 0 to 10, where 0 corresponds to the strongest alteration/unintelligible speech, and 10 corresponds to the absence of alteration/perfectly intelligible speech. Figures 4.1a and 4.1b show in more details the scales used in C2SI to assess these measures as reported by [Lalain et al., 2020].

- **Intelligibility and Severity on the passage reading task (Intel-LEC and Sev-LEC):** Similarly to the previous evaluations conducted on picture description, the same set of judges evaluated the intelligibility (Intel-LEC) and severity (Sev-LEC) on the recordings of the passage reading task.

- **Perceived Phonological Deviation (PPD-DAP):** A Perceived Phonological Deviation (PPD-DAP) score was proposed by Lalain et al. [Lalain et al., 2020]. This score is obtained with an acoustic-phonetic decoding of pseudo-words produced in the DAP task. It reflects the average number of features altered per phoneme for each speaker. It is considered as an objective perceptual measure of speech intelligibility in the C2SI project.



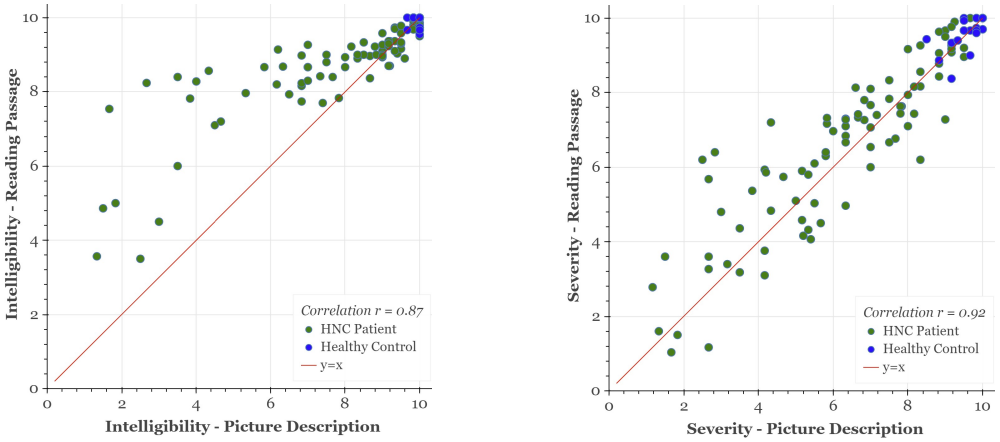
(a) Severity **(b)** Intelligibility in C2SI (Comprehensibility)

Figure 4.1: *Clinical subjective assessment (source: [Lalain et al., 2020])*

In order to briefly explore the perceptual assessments at our disposal, we present in the following an analysis based on the Pearson correlation between pairs of the perceptual measures aforementioned. Table 4.3 sets out the obtained results. These correlations cannot only reflect the degrees of similarity between the measures, but 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. Indeed, when analyzing table 4.3, we can see that Intel-LEC is the perceptual measure with the lowest correlation with the majority of the other measures. The main reason is that Intel-LEC is more considered as a comprehensibility measure. As reported in [Lalain et al., 2020], it integrates contextual information in addition to the acoustic-phonetic information in the speech decoding process. In fact, the text used in the protocol of the passage reading task is relatively short, classically used in the community, and can be easily memorized during evaluation by experts. This leads to an overestimation of the speech intelligibility of patients since experts can deduce the heard text despite speech production errors. Still, the scatter plots provided in figure 4.2 confirm the previous observations and conclusions. Here, each dot on the graphs represents a speaker, where green is for patients and blue is for healthy controls. We can see that there is a ceiling effect when it comes to the intelligibility assessed on the reading task, whereas this is clearly not the case for the same measure assessed on the picture description task. Conversely, this effect is not visible regarding the severity measure, which confirms once more the overestimation of the Intel-LEC measures. In addition, we assume that the task of image description leads to less predictable linguistic content and, therefore, to more valuable perceptual assessment by the experts.

	Intel-LEC	Sev-LEC	Intel-DES	Sev-DES	Phnm-DES	PPD-DAP
Intel-LEC	1.00	0.89	0.87	0.82	-0.77	-0.78
Sev-LEC	—	1.00	0.83	0.92	-0.86	-0.82
Intel-DES	—	—	1.00	0.93	-0.87	-0.84
Sev-DES	—	—	—	1.00	-0.93	-0.85
Phnm-DES	—	—	—	—	1.00	0.84
PPD-DAP	—	—	—	—	—	1.00

Table 4.3: *Correlations between the different C2SI perceptual measures of interest in this study.*



(a) Scatter plot for intelligibility

(b) Scatter plot for severity

Figure 4.2: *Scatter plots showing trends of the two pairs of perceptual measures (Intel-LEC,Intel-DES) and (Sev-LEC,Sev-DES)*

4.2.3 SpeeCOmco: An additional dataset for disordered speech due to Head & Neck Cancers

Proposed by Balaguer [Balaguer, 2021], SpeeCOmco is a corpus including 25 patients treated for cancer of the oral cavity or oropharynx. The innovative aspect of this corpus lies in the inclusion of spontaneous speech recordings by the patients during a semi-directed interview. Similarly to C2SI, the patients in SpeeCOmco corpus were subject to several perceptual assessments. To carry out this task, a jury of six experts was recruited. In this study, we are particularly interested in the intelligibility and severity measures of these patients, which were assessed on the recordings of the semi-directed interview. These assessments were conducted using exactly the same instructions and rating scales (i.e. from 0- very severe/unintelligible to 10 - absence of alteration/completely intelligible), as those used in C2SI corpus. Figure 4.3 depicts the perceptual measures of intelligibility (in purple) and severity (in orange) of patients in the SpeeCOmco corpus. We can observe that the degree of speech impairment among the patients is rather balanced between the low and high scores. We can also observe that the intelligibility scores are systematically higher than the severity scores for all patients, with varying difference values (up to two point difference on scales depending on patients).

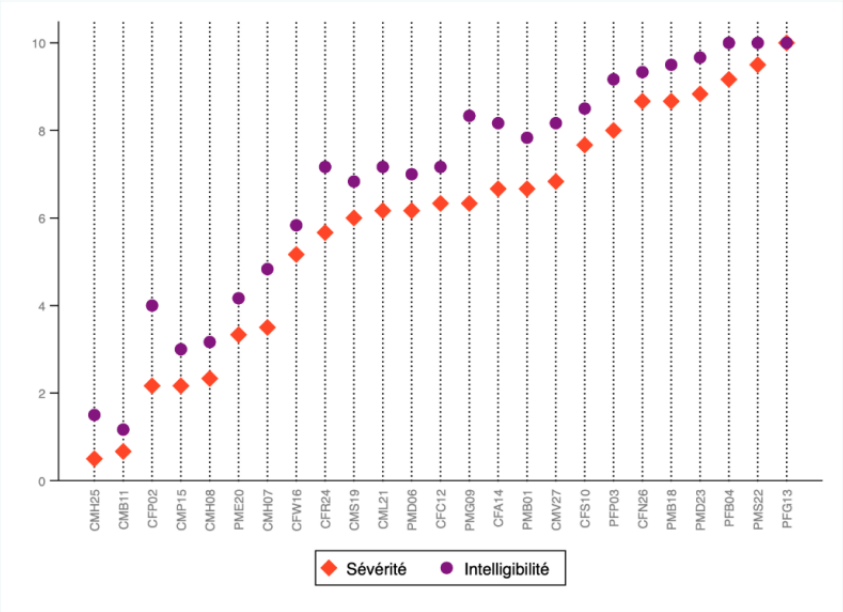


Figure 4.3: *Intelligibility and severity scores of patients in the SpeeCOmco corpus (source: [Balaguer, 2021])*

4.2.4 Automatic speech alignment

All audio recordings were segmented automatically at the phoneme level thanks to a forced-alignment system developed at LIA. The automatic forced alignment consists in providing the temporal segmentation of the known phoneme sequence present in the speech signal. By taking as input the target segment, its sequence of phonemes, and the speech signal produced by the speaker, the automatic processing is based on a decoding of the speech signal, involving the Viterbi algorithm and statistical HMM. The HMM-based models are built thanks to the Maximum Likelihood Estimate paradigm from about 200 hours of French radiophonic speech recordings [Galliano et al., 2005].

4.3 Proposed methodology

While perceptual measures still remain the gold standard in clinical settings, as we pointed out in chapter 1, this procedure is mainly characterized by its subjectivity since it depends on a lot of intrinsic variables (e.g. familiarity of the judge with the test items, speaker, and/or speech pathology, ...). In this section, we introduce our solution to tackle these problems. We start by presenting an overview of the proposed methodology with a brief description of the reasoning behind the steps. Beyond that, we take a position in our particular context in relation to the terminology and taxonomy of the interpretability/ explainability that we invoked in chapter 3.

4.3.1 An overview of the proposed approach

A novel method for approaching speech pathology assessment is presented in this work. Taking advantage of the advances in neuroscience as well as speech and deep learning technologies, the aim of this work is to develop an objective assessment tool for speech intelligibility providing reliable analysis of speech production from a quantitative perspective (i.e. a score assessing the speech quality) and qualitative perspective (i.e. a focus on acoustic and articulatory degradation). The characterization of the disordered speech in the phoneme and phonetic feature dimensions will offer a better basis for providing the clinician with information that is directly related to speech therapy. Consequently, this characterization will enable the identification of the linguistic units that best contribute to the maintenance or loss of intelligibility from an acoustic point of view. This would provide the basis for the design of appropriate protocols for rehabilitation purposes to enhance the patient's intelligibility. Based on this, the derivation of an objective intelligibility score is a multi-stage process involving an acoustic preprocessing of the speech data followed by the three main steps of our methodology, a phoneme-based analysis, a phonetic feature-based analysis, and an intelligibility prediction. In the following, we provide a brief description of each step.

Step 1: Phoneme level representation

In this first step, our aim is to encode the French phoneme characteristics of *healthy speech*. To do so, we put in place a DL model taking as an input the acoustic features of the speech signal and performing phoneme classification (hereafter, **the base task**). The particular choice of the phoneme dimension as well as the other details related to this first step are discussed in the next chapter 5. Although the task is relatively simple, we assume that it is relevant and suitable for the main target task of speech intelligibility prediction.

Step 2: Phonetic feature level exploration

This step is addressed in detail in chapter 6. In this step, we peer into the trained DL model of the previous step to find meaningful representations that were automatically learned by the model. We investigate the inner representations performed by individual units and layers and reveal their capacity to locate emergent concepts relevant to our specific context of clinical phonetics, which is the phonetic feature. By considering this extra dimension of phonetic features, we increase both the experts' trust and the interpretability of the global model performing the target task (i.e. intelligibility prediction) that we present in the next step hereafter.

Step 3: Speech quality assessment and interpretation

This step is entirely based on the two previous steps. At this stage, we have a skillful model for the base task of phoneme classification, the outcome of the first step. We do not stop at simply considering this task as a base for intelligibility prediction (hereinafter referred to as the **target task**). We choose, instead, to get into the details and ensure that the acquired knowledge is also suitable for this target task and can serve as an extra dimension for the interpretability of the final intelligibility score. This is the outcome of the second step. The aim of this third step is to predict an intelligibility score for a given speaker and to interpret this result in terms of phonetic feature alteration. That is, with means of a shallow neural network, we transform

the output of the CNN, reflecting the production of a speaker within the phoneme dimension into an intelligibility score. We can thereupon investigate the two dimensions of phonemes and phonetic features and their capacity in yielding a reasonable interpretation of the phonemic unit contribution to speech intelligibility and its variation (improvement or alteration).

4.3.2 Take position in the interpretability/ explainability dilemma

As seen in chapter 3, one of the recurring themes in the ML interpretability literature is the constant efforts toward a universally accepted terminology. Despite efforts, thus far, we highlighted the fact that there has not been an established consensus on how this terminology should be best defined in the context of ML.

In our view, we find it useful to clarify the meaning we assign to these words throughout this dissertation, as well as the type of techniques that are adopted. We will make sure, therefore, to avoid using both terms interchangeably and invoke commonly used synonyms in order to minimize confusion.

Our methodology, as presented in the previous section, can be approached from two different perspectives in order to answer the question of whether we are in the context of interpretability or explainability:

(1) The perspective of the *whole model* predicting an assessment of the speech quality that can be interpreted by the production of low-level units. This perspective arises from the clinical need to find the linguistic units responsible for speech intelligibility. In this case, we consider that we are more in an interpretability task.

(2) The perspective in which we seek to get insights into the internal representations of the *local model* performing a phoneme classification task, and uncover relevant meaningful representations that will enhance the interpretability of the *whole model*. Here, we consider that we are more in an explainability context.

We begin our discussion with the former. Mostly aligned with the interpretability definition of [Montavon et al., 2018], this study considers interpretability as the mapping of an abstract concept (i.e. a prediction) into a domain that the human can make sense of - that is to say in our case, mapping the score assessing the speech quality into the domains of phonemes and phonetic features which are understandable by humans and relevant in this disordered speech context. In light of the taxonomy presenting the different types of interpretation, see section 3.3, we can suggest that this work takes place in the *post-hoc* category (i.e. a trained model is given and our goal is to understand what the model predicts in terms of what is readily interpretable). Generally speaking, the input domain in most cases is the one considered as readily interpretable (e.g. images considered as arrays of pixels, or texts considered as sequences of words) since a human can look at them and read them respectively. However, it can also be not interpretable as the example of abstract vector spaces (e.g. word embeddings, or speaker embeddings like x-vectors). In our case, the input domain is the acoustic parameters of speech data. That is, we have insight into the time dimension (frames extracted from the speech waveform), but not into the filterbank feature dimension. Thus, we are not in the case of a fully interpretable domain, which makes the task more challenging. Consequently, simple post-hoc interpretability would not be enough since the input domain is not fully interpretable. Hence, we are more concerned

with incorporating interpretability directly into the structure of the model.

This discussion leads us to the following perspective which is the preparation of the intermediate domain to interpret the speech quality assessment. In this perspective, we will bring to light two intermediate interpretable domains. The first one is the phoneme domain which derives from the dedicated French phoneme classification as an intermediate task before the speech quality assessment. As regards the second dimension, we choose to peer into the black box and shed light on its internal representations in order to find meaningful and explainable ones that were automatically learned by the DL model. By revealing the capacity of individual deep units to locate emergent concepts of relevance in our clinical context (e.g. phonetic features), we enhance the degree of interpretability of the final model without the need to explicitly train on an extra intermediate task. From another point of view, it can also increase the confidence of clinical experts. Indeed, our contributions allow them to see that meaningful representations can emerge systematically, satisfy their need to see that there is a logic in the inner workings of the model, similar to their own way of assessing speech quality.

In our point of view, this step has to be considered as explainability in its own right. Even, we emphasize alignment with a specific type of explainability detailed in [Gilpin et al., 2018] focused on the explanation of deep representations for understanding the role and structure of the data flowing through the neural network. Illustrated in fig. 4.4, this type of explainability involves different levels of analysis : layer level, unit level, and representation vector level considering the granularity examined.
What we are concerned with in this work is the understanding of the role of individual units in our CNN based architecture. That being said, a relevant interpretation related to a particular assessment of speech quality does not necessarily involve a fully detailed understanding of all the mechanics of the black-box based model. Instead, we assume that it can be simply based on both the dimensions of phonemes and phonetic features we target.

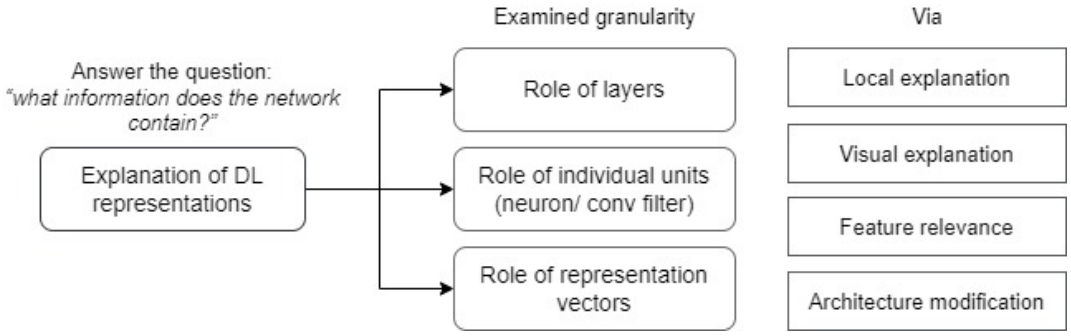


Figure 4.4: *Explainability of deep learning representations drawn from [Barredo Arrieta et al., 2020] and extended from the categorization of [Gilpin et al., 2018]*

4.4 Conclusion

In this chapter, we briefly introduced the general context of our work, including the project and the data used. An overview of the proposed methodology was also provided, with a brief

outline of the reasoning behind its various steps. In particular, we have set out our position on the interpretability/explainability dilemma in this context, by referring to the literature review we proposed in chapter 3. In the rest of this document, we detail the proposed methodology, presenting each step in a dedicated chapter.

Chapter 5

Step 1: Phoneme-Level Representation

Contents

5.1 Specific context

In this chapter, we introduce the first step in our methodology which is the characterization of French phonemes through the training of a DL model, on healthy speech, for the basic task of phoneme classification. We start by explaining the different choices that we made spanning from the base task as such, to the choices related to the type of features in the input data and the type of DL architecture. Afterward, we describe the experimental setup in which we unveil the details related to the data preprocessing, the model architecture, and the training phase. Thereafter, we explore the behavior of the model when exposed to a disordered speech dataset and report the different analyses we made within this step. We finish this chapter with a discussion which brings us to the next chapter detailing the second step of our methodology.

5.1.1 Why phoneme classification task?

At this stage, it is of great importance to justify the choice of phoneme classification task that we made since it is considered as the **base task** on which the rest of this work is founded. To answer the question *why a phoneme classification task*, we address the subject from two points of view. Indeed, as already presented in the background chapter on speech production 1, phonemes are the basic sound units in any given language forming higher-level linguistic representations such as syllables, words, and sentences. Phonemes provide the smallest inventory units, and given the relatively small number of phonemes in any language, a phoneme-based analysis would be relatively efficient. This efficiency becomes even more clear if we consider the definition of speech intelligibility. Intelligibility, as defined in section 1.4.1, is the accuracy with which the acoustic signal produced by the speaker is decoded by a listener. It is an analytical acoustic-phonetic decoding concept [Ghio et al., 2021], which refers to the pronunciation quality of *"low-level"* units (i.e. phonemes, phonemic groups, syllables). Let us not forget that this work aims to address the lack of interpretation of intelligibility assessment with regard to local alterations in speech units. Based on that need and all of the above, we find that the phoneme level is a suitable choice. This was also outlined by [Pommée et al., 2021]. Nevertheless, it is worth mentioning that such a choice is also challenging because of the coarticulation effect that makes the physical realization of the phoneme highly variable [Liberman et al., 1967] and consequently not easily segmented and identified.

5.1.2 Why CNN architecture?

Deep learning is becoming unavoidable in new trends, particularly dealing with speech, because of its huge success in more or less complex tasks including phoneme classification [Malakar and Keskar, 2021]. It is important to note that DNNs are inherently flexible, and there is no presumption of a "definitive" way to fit a particular type of data. In many cases, data can fit well with more than one method or a mixture of methods. The choice of the appropriate architecture, therefore, depends on several reasons as the case may be. In this work, we consider a standard and generally successful class of architectures, the convolution neural network (CNN). This architecture has demonstrated its performance in the phoneme classification task because of certain exciting features [Palaz et al., 2013, Abdel-Hamid et al., 2014, Glackin et al., 2018]. It was even shown that it works successfully on small training sets [Poliyev and Korsun, 2020]. In the following, we explain the main reasons behind this choice.

To start, the speech signal exhibits local similarities in both the spectral and temporal dimensions. CNNs have three key properties that make them ideal for such speech-related tasks: *locality*, *weight sharing*, and *pooling*, refer to section 2.1.3 for more details. It is worth recalling that CNNs run small filters over the input to extract relevant features. Weight sharing refers to the decision to use the same weights at every position of the filter. On the other hand, locality refers to the fact that individual units computed at a particular positioning of the filter depend upon features of the local region of the input that the window is currently observing. Locality allows more robustness against non-white noise as good features can be computed locally from cleaner parts of the spectrum and only a smaller number of features are affected by the noise. This gives a better chance to higher layers of the network to handle this noise, which is better than simply handling all input features in the lower layers as in standard fully connected neural networks. Moreover, weight sharing improves model robustness and reduces overfitting as each

weight is learned from multiple locations in the input instead of just one single location. As regards pooling, this technique pools together feature values computed at different locations and represents them by one value. This technique is actually shown to be useful in handling small frequency shifts that often occur in speech signals [Abdel-Hamid et al., 2014]. However, when considering other models such as fully connected DNNs, these pattern shifts become more difficult to handle due to the need for many hidden units. On the other hand, a reason behind the choice of CNN architecture could be the trade-off explainability/performance. Indeed, the road to explainability for CNNs is easier than for other types of models, as human cognitive skills favor the understanding of visual data. Once we would choose to explain the convolutional layers of the CNN, feature maps are easier to inspect as they preserve the 2D structure of the input.

5.2 Experimental setup

In this section, we describe the experimental context that we set up to obtain the adequate phoneme classifier. We begin with a description of the data preprocessing, followed by the architecture and training details, and then a highlight of the principle factors taken into account in the CNN for later explainability.

5.2.1 Data preprocessing

At this stage, we describe the data preparation performed on raw data for later processing with DL model. Figure 5.1 is a schematic representation of this phase. It is worth mentioning that we are involved in a supervised learning task. This explains that we have the speech signal sampled at a 16KHz frequency, representing the input for the CNN. In addition to the speech signal, we have the corresponding time-aligned phoneme transcription generated automatically which represents the target. At first, we are interested in detailing the input preprocessing which starts with splitting the speech signal into a sequence of frames by a fixed-size window of 20 ms; with the assumption that within this frame size, the signal is stationary. An overlap of 10ms between two adjacent frames is considered to preserve information. Subsequently, feature extraction is performed on each frame. Given the multitude of features that can be used, we opt for the Mel Filterbank features. Subsequently, each frame is a 40-dimensional log Mel Filterbank features concatenated with first- and second-order derivatives. The features are normalized to zero mean and unit variance per speaker utterance level.

As shown in figure 5.1 (b), a sliding context window of 11 frames is used as input for the CNN which results in an input matrix of dimension [11x120]. That is, one input sample for classification is composed of a central frame to which have been added the five previous and five following neighboring frames. The target label of this input sample is the phoneme corresponding to the central frame. **For the sake of simplicity, we will refer to an input sample as a *frame* in the rest of this work, even though an input sample is actually a contextual window of 11 frames as explained.**

5.2.2 Architecture

The CNN architecture adopted in this study is drawn from the work of Pellegrini and Mouysset [Pellegrini and Mouysset, 2016] and trained on a supervised task of French phoneme classification. Briefly, the architecture consists of two pairs of convolution and pooling layers, followed

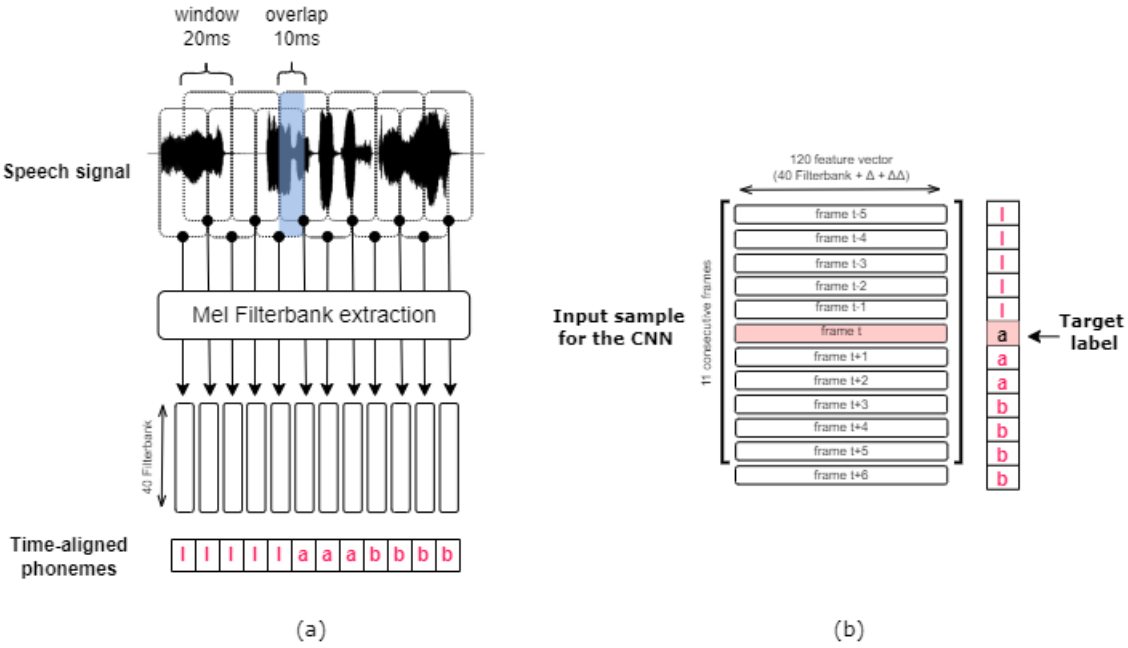


Figure 5.1: *Data preprocessing: (a) The extraction of Mel Filterbank features from the speech signal (b) The preparation of the input samples and target labels for the CNN training*

by three fully connected layers and a final output layer. In more detail, the convolution layers apply a set of [3x5] filters to extract the local characteristics, then produce 32 and 64 activation maps respectively. The max-pooling layers apply [1x3] and [1x2] filters respectively providing lower frequency resolution features that contain more useful information to be processed by higher layers of the CNN. The classification task is then performed by three fully connected layers of 1024 neurons (namely FC1, FC2, and FC3). A ReLU activation function as well as a dropout of 0.4 are applied to the output of each of the fully connected layers. Finally, a *softmax* layer corresponds to the posterior probability of each of the 32 final classes associated with the 31 French phonemes and silence. It is worth mentioning that the 31 French phonemes are:

- **11 vowels:** {a, ɛ, e, Û, Ô, u, y, i, ã, ɔ̃, µ}. The set of vowels includes the following archiphonemes[1]: Û={œ,ø}, Ô={o,ɔ} and µ={œ̃,ɛ̃}.
- **3 semi-vowels:** {w, ɥ, j}
- **17 consonants:** {l, ʁ, n, m, ɲ, p, t, k, b, d, g, f, s, ʃ, v, z, ʒ}

5.2.3 Factors taken into account in the CNN architecture for later explainability

As we have already mentioned, our major concern is not to have the best performance ever, which explains the fact that we did not make a comparison with the accuracy of state-of-

[1] An archiphoneme is a phonological unit that expresses the common features of two or more phonemes that are involved in neutralization.

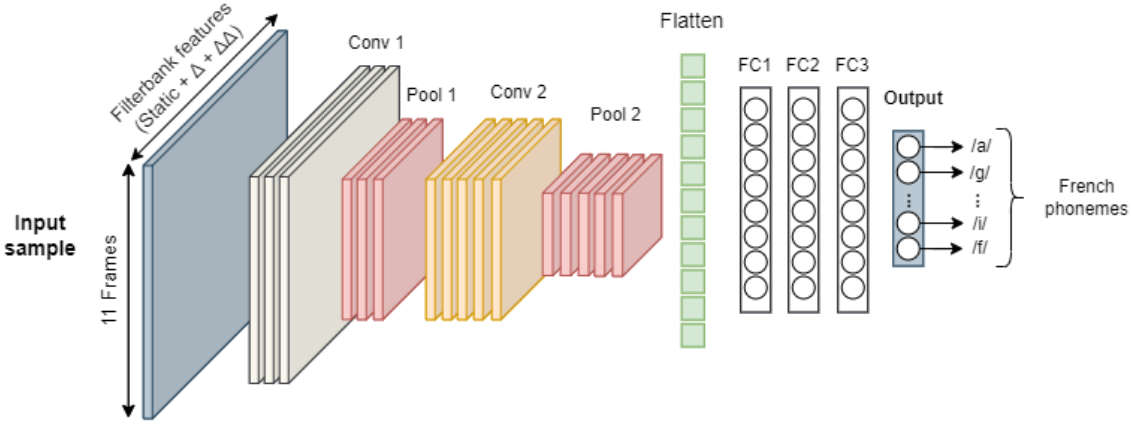


Figure 5.2: *The CNN architecture*

the-art models on the same task. Indeed, what is important for us is the trade-off performance/complexity since we are later involved in an explainability task. In this section, we cite some factors that we took into account for maximum later explainability of the model.

ReLU activation function:
The choice of the ReLU activation function is important. ReLU for short is a piecewise linear function [Montufar et al., 2014], that outputs the input directly if it is positive, otherwise, it outputs zero. Representational sparsity is one of the benefits of ReLU activation function, since, unlike the tanh and sigmoid activation functions that learn to approximate a zero output, this function is capable of outputting a true zero value. This means that negative inputs can output true zero values allowing the activation of hidden layers in neural networks to contain one or more true zero values. This is called a sparse representation and is a desirable property in representational learning as it can accelerate learning and simplify the model. This resulting simplicity is what we are looking for in the next step of explainability.

Avoiding batch normalization:
In the first experiments, we used batch normalization layers between fully-connected layers as a regularization technique. Nevertheless, it turns out that such a technique is not adequate in our case despite the fact that it improves the CNN classification performance and accelerates training. Indeed, authors in [Amorim et al., 2020] have shown that interpretability appears to be higher with lower regularization values. Related to batch normalization, our empirical observations as well as other works, [Bau et al., 2020], [Morcos et al., 2018] confirm that this regularization technique seems to decrease interpretability significantly. Consequently, despite the fact that we already published work including batch normalization layers [Abderrazek et al., 2020], we later adopted a model without batch normalization given the objective of this work which priority is the trade-off performance/complexity.

5.2.4 Frame Selection: Training, validation and testing details

Regarding CNN training, an initial learning rate of 0.001 following an exponential decay schedule and early stopping strategies are used. The main goal is to minimize the categorical cross-entropy loss function using the stochastic gradient descent computed on a mini-batch of 64 samples. It is worth mentioning that the model is trained, validated, and tested on healthy speech issued from BREF corpus described in section 4.2.1 after being subject to the data preprocessing described in section 5.2.1. Therefore, we have mainly three different datasets extracted from BREF corpus used for these purposes. Since the phoneme distribution within BREF corpus is highly imbalanced, a random undersampling technique is adopted to handle the disproportional distribution of classes. This prevents the classifier from being biased toward the majority class. As shown in figure 5.3, this technique reduces the count of samples falling under the majority class and results in a balanced dataset. Almost 3M samples were obtained in our case and were split into 90%-10% partitioning corresponding to training and validation sets. Regarding the

Figure 5.3: *Undersampling strategy for an imbalanced dataset with two classes*

testing dataset, it is important to underline that this dataset is specially conceived to be representative of healthy speech for later explainability and interpretability. Referred to as *BREF-Int*, the usefulness of this dataset is actually highlighted in the next chapters, however, we introduce it at this level since we find it is important to report the different analyses related to the classification performance of this dataset.

BREF-Int:

BREF-Int, is a subset of BREF corpus dedicated to the interpretability phase. More specifically, it is devoted to the fine-grained neuron-level analysis of the representations learned by the trained model. Obviously, this dataset should never be seen in the training/validation phases of the model. It is important to point out that a special selection of the frames forming *BREF-Int* was performed, and that the choice is not random. Indeed, the included frames are associated with speech segments (yielded by the automatic forced alignment) related to a complete phoneme production. In addition, these frames reflect different phoneme contexts and speakers available in BREF corpus. To this diversity must be added the fact that the dataset is balanced in order to achieve a roughly equal distribution of frames over the 31 phonemes. All in all, this leads to a subset including almost 82K samples, referred to as *BREF-Int*, which we consider as representative of healthy speech.

5.3 Results

This section summarizes the different results related to this first step, reflecting that the proposed architecture is suitable for the next steps of the proposed methodology. We first report results on the classification performance of the proposed model. We want to show that, despite the restrictions we took into account to avoid the architectural complexity, our choices are still acceptable and do not significantly impact the performance. To this end, the test dataset *BREF-Int* is used. Furthermore, we use other datasets, basically issued from C2SI, to explore the capacity of the trained model to generalize well to data collected under different conditions and protocols. Later on, the model behavior when exposed to different levels of speech degradation was explored, with the aim of demonstrating the capacity of the model to characterize potential deviation with impaired speech, even though it had been trained exclusively on healthy speech and had never seen any pathological speech. **It is worth recalling that the phoneme classification is made at the frame level (i.e. mapping a speech frame to a final phoneme label). Consequently, all the model classification accuracies reported in the following sections are calculated based on the model decision taken at the frame level.**

5.3.1 Classification Performance

This section presents an analysis of the classification performance of the trained model when exposed to a dataset variation. The goal is to confirm that the model is not overfitting on BREF, the dataset on which it was trained and validated. Consequently, in this case, we can assume that the model can generalize on new datasets that have neither the same recording conditions nor the same speakers.

Dataset	#Samples	Balanced Accuracy
BREF-Int	82K	81.4%
C2SI-LEC HC speakers	43K	72.2%
C2SI-DAP HC speakers	78K	69.2%

Table 5.1: *Number of samples and balanced accuracy for the studied datasets.*

To this end, the classification performance of the model is calculated for the different datasets *BREF-Int* as well as *C2SI-LEC* and *C2SI-DAP* both for healthy control (HC) speakers. Table 5.1 summarizes the different results. We use balanced accuracy as an evaluation metric for the classifier. Especially useful when the classes are imbalanced, this accuracy is calculated as the average of the correct classification rates obtained on each individual class. Even though *BREF-Int* is already balanced, this metric is mainly used to reflect the performance of the model on other datasets that are highly imbalanced. We do not consider the silence class in the accuracy calculation since we want it to reflect the performance of the model exclusively on the French phonemes. It is worth mentioning also that /e/ and /ɛ/ were considered as two separate phonemes for the CNN training, but are considered in the rest as one class referred to as the archiphoneme Ê={e,ɛ}, reducing the number of classes involved in all the calculation of balanced accuracies to 30 French phonemes.

Obviously, the performance calculated on different datasets takes into account only HC

speakers since the first objective is to study the cross-corpus generalization capacity of the trained model. Therefore, the differences between corpora have to be solely related to the conditions of recordings and not related to speech pathology.

As illustrated in table 5.1, we can see that a drop in the performance of the classifier is seen when comparing *BREF-Int* with *C2SI-LEC* (i.e. from 81.4% of balanced accuracy on *BREF-Int* to 72.6% of balanced accuracy on *C2SI-LEC*). This drop can be explained, indeed, by the difference between the two datasets, for instance, in the recording equipment (e.g. microphone), the environment, the recruitment region of the speakers (implying accent), and other factors that are either related to the recording conditions or to the speakers' characteristics. On the other hand, one can observe that this drop is slightly more important when considering *BREF-Int* and *C2SI-DAP* datasets. This can be explained by the fact not only do the two datasets differ in terms of factors related to the recording conditions and speakers' characteristics, but they also differ in terms of the task being performed. Unlike *BREF* and *C2SI-LEC* consisting of a reading task, *C2SI-DAP*, as described in section 4.2.2, consists of a pseudo-word pronunciation task. This task is a little bit specific and difficult when compared to the reading task. Indeed, the pseudo-words are not natural (i.e. words that do not exist in the dictionary), and are also generated using specific patterns (i.e. combination of vowels and consonants each in a specific place in the word). Therefore, the pronunciation of certain pseudo-words can be very difficult when they include a combination of consonants that are not easily articulated even for an HC speaker.

A closer look at the classification performance

In this part, we are going to look closer at the phoneme classification performance of the trained CNN-based model. To do so, we carry out an analysis of a complete speech segment produced by the HC speaker *TIO-000020* while reading the passage "La chèvre de M. Seguin" (i.e. one of the HC speakers' recordings belonging to C2SI-LEC). Figure 5.4 illustrates the speech signal with its phoneme alignment underneath. In the phoneme alignment sub-figure, each point denotes one frame which coordinates are, the true phoneme label in the X-axis and its classification probability (i.e. the output of the softmax layer of the CNN). The point color denotes whether the frame in question is well-classified (black color) or miss-classified (red color). The blue vertical bars represent the phonemes boundaries, the outcome of the forced-alignment system. It is worth clarifying that the probability of a miss-classified frame (a red point) corresponds to the probability of the phoneme predicted by the CNN and not the probability of the true phoneme (i.e. the maximum probability given by the softmax layer). For the sake of clarity, a focus is done in the same figure, on the first part of the speech segment "Monsieur Seugin n'avait jamais eu de bonheur avec ses chèvres".

This analysis reveals that the classification performance depends on the frame position in the phonetic segment. To start with, we can clearly see that the majority of well-classified frames (i.e black points) are concentrated in the upper part of the plot corresponding to the higher values of classification probabilities. On the other hand, the classification probabilities corresponding to the majority of miss-classified frames (i.e. red points) are noticeably lower.

In addition, it is worth mentioning that the model tends to well classify the middle frames belonging to a phoneme segment (i.e. one segment delimited by two blue bars) with higher confidence than those situated on the phoneme boundaries. Notice that within a phoneme segment,

the probabilities of frames classification form an arc shape. It is even remarkable that a significant number of the frames located on the phoneme boundaries are miss-classified, and that, in general, most of the miss-classification performed by the model occurs in the phonemes boundaries. As a matter of fact, this reflects that the classifier is not confident about its prediction since the frames are close to the decision boundary the model has learned from the training data.
This, indeed, can be explained by two main reasons. The first reason comes from the errors that can be generated by the forced-alignment system since the latter is used to locate the phonemes boundaries and no manual correction was carried out after this alignment. The second reason is related to the composition of frames considered in this work. Let us recall that a frame in our case is actually a misnomer to refer to an input sample and is not one isolated frame, but rather a frame with the five previous and five following neighboring frames (i.e. 11 consecutive frames concatenated together), and which target label is the phoneme label of the central frame. Assume that we are in the last frame of the phoneme /g/ (noted 'gg') in the word 'Seguin'. Thus the frame is actually five frames of /g/, one central frame always /g/, and five frames from the following neighboring phoneme /μ/ (noted 'un'). That is, even if no error was made by the forced-alignment system, almost 45% (the five succeeding frames) of the input sample that was labeled /g/ belongs to the next produced phoneme /μ/.

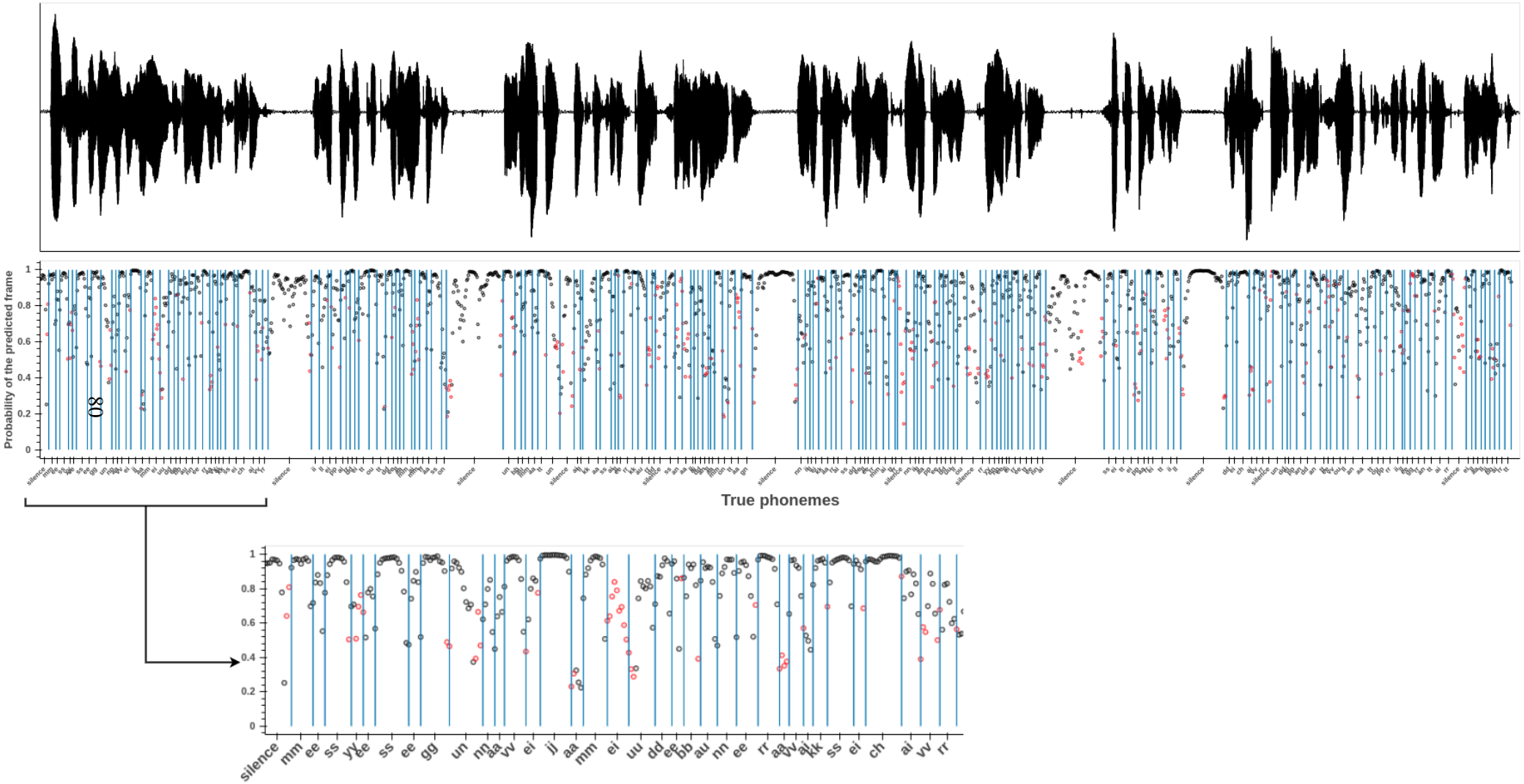
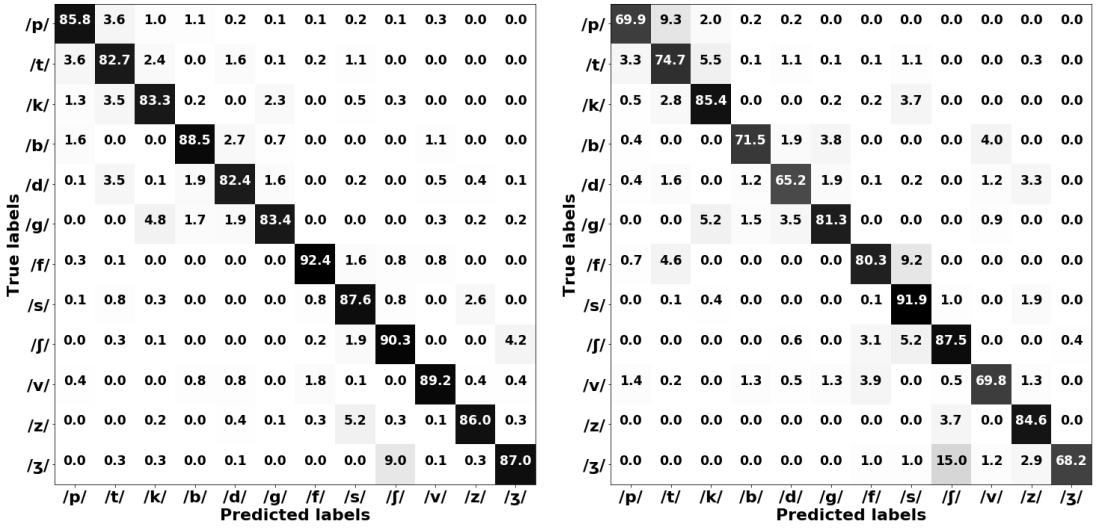


Figure 5.4: *Speech segment of the control speaker TIO-000020 on the reading task: the waveform and the aligned phoneme segmentation with the classification probabilities at the frame level.*

5.3.2 Analysis of Confusion Matrices

In this section, we carry out analyses based on observations issued from confusion matrices (CMs) on both datasets *BREF-Int* and *C2SI-LEC* HC speakers. This will allow us to look closely at the types of confusion made between different phonemes, and reveal a pattern of the most cooccurring ones. A split of the confusion matrices into two sub-confusion matrices (sub-CMs) is performed for readability purposes, as well as to highlight the most relevant confusions. To observe in more detail the discussed effect of dataset variation, figures 5.5 and 5.6 are organized in a way to put in parallel the two sub-CMs grouping the same phonemes, with *BREF-Int* dataset on the left and *C2SI-LEC* HC speakers on the right side.

When analyzing the confusion matrices grouping obstruents, illustrated in Fig. 5.5, we can

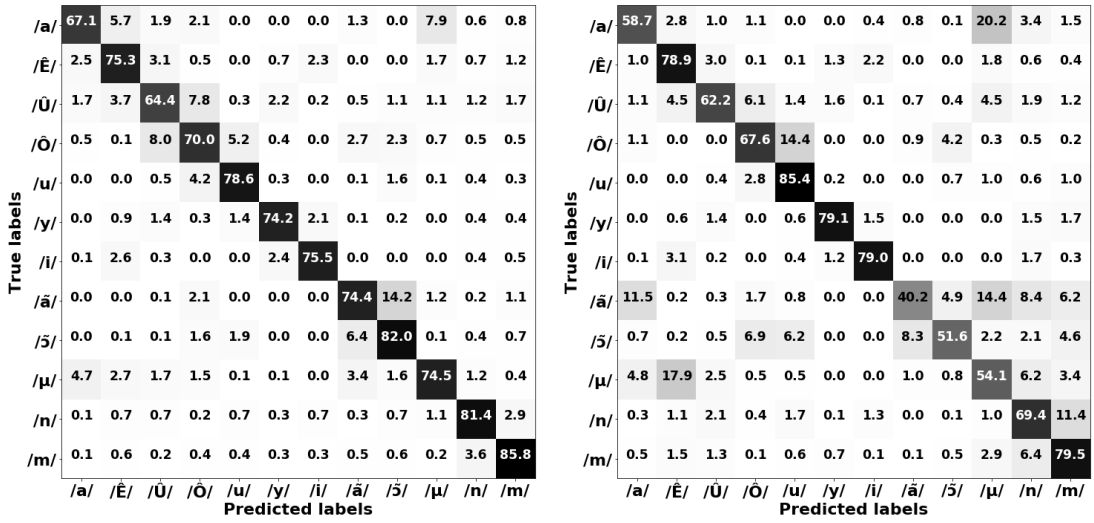


(a) BREF-Int

(b) C2SI-LEC HC speakers

Figure 5.5: *Confusion matrices grouping obstruents*

clearly observe that classification errors are somehow explainable and make sense since confusions are generally made between phonemes sharing most of their phonemic features. For instance, a confusion of 9% is observed between the phones /ʒ/ and /ʃ/ on the sub-CM of *BREF-Int* in figure 5.5a. When examining the phonological representation of these two phonemes, based on table 1.4, it can be clearly seen that they share all their phonetic features except for the voicing feature (i.e. /ʒ/ is voiced and /ʃ/ is voiceless). In special cases, this observation can be interpreted as the archiphoneme phenomenon where for two separate phones differing in one distinctive feature, the contrast is neutralized in certain positions. Taking the example of the voiced obstruents (/b/, /d/, /g/, /v/, /z/, /ʒ/) opposed respectively to voiceless obstruents (/p/, /t/, /k/, /f/, /s/, /ʃ/), in French, when a voiced consonant is followed by a voiceless phone, it may lose its distinctive feature of voicing and thus be pronounced like its corresponding voiceless consonant (eg. the p-sound of /b/ in the word "obtus"). Another type of confusion is illustrated by the phone /p/ being confused at 9.3% with /t/ on *C2SI-LEC* HC speakers (see figure 5.5b) which is



(a) BREF-Int

(b) C2SI-LEC HC speakers

Figure 5.6: *Confusion matrices grouping oral/nasal vowels and nasal consonants*

related to the loss of the distinctive feature of acuteness. To generalize the above-mentioned observations, it can be seen that the highest confusions between consonants take place generally as a result of losing either the distinctive feature related to the place of articulation (i.e. acuteness/ compactness) or the loss of voicing distinctive feature.

Regarding vowels, illustrated in fig. 5.6 grouping oral/nasal vowels and nasal consonants, a particular classification behavior is observed in the right sub-CM corresponding to *C2SI-LEC* HC speakers (see figure 5.6b). Indeed, nasal vowels produced by HC speakers of *C2SI-LEC* are subject to strong confusion with both oral vowels and nasal consonants. For instance, /ɑ̃/ is confused with 11.5% with the oral vowel /a/, as well as with nasal consonants, 8.4% and 6.2% for /n/ and /m/ respectively. In contrast, nasal vowels of *BREF-Int* corpus, are not subject to such confusions but rather to more evident ones within the class of nasal vowels such as 14.2% of confusion between /ɑ̃/ and /ɔ̃/ (see figure 5.6a). This difference in confusion pattern between BREF and C2SI more generally, can be explained by the recruitment region of speakers for both corpora, exhibiting a major Parisian accent for the BREF corpus (the closest to standardized French), and a major southwestern accent for the *C2SI* corpus. Indeed, it is reported in the literature that nasal vowels can be produced with a less complete nasalization in speech exhibiting a southwestern accent, no more dealing with nasal vowels but rather with a combination of an oral vowel followed by a nasal consonant, totally coherent with our observations regarding nasal vowel confusions on *C2SI-LEC* corpus.

5.3.3 Correlation Analysis

So far, we have demonstrated the performance of the proposed model when exposed to a dataset variation. Therefore, we can assume that any significant degradation in the model performance,

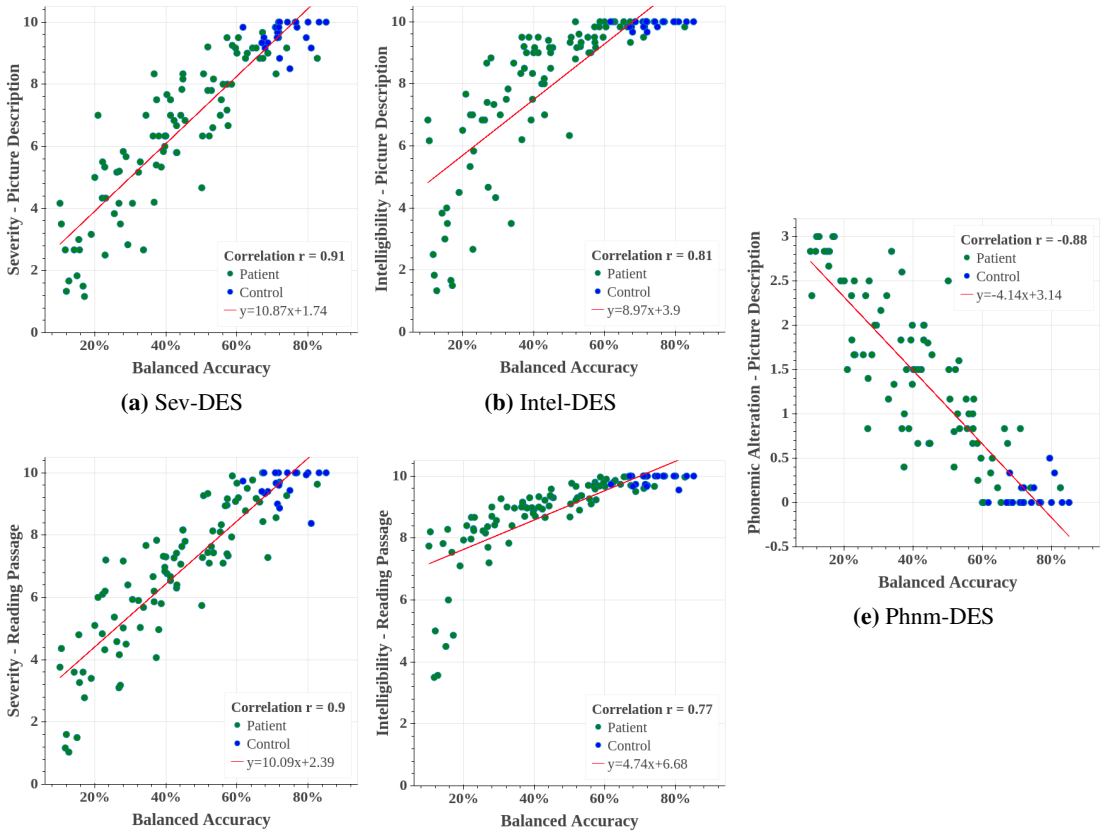
while typically testing on patients' utterances issued from *C2SI-LEC* corpus, is consistent with the degree of speech quality degradation, and thus, with the perceptual ratings of that patient. To highlight this idea, we evaluate the performance of the classifier on individual speakers (i.e. HC speakers and patients), issued from the two corpora of C2SI at our disposal C2SI-LEC and C2SI-DAP, again using the balanced accuracy metric for the same abovementioned reason. We further do the same process on speakers belonging to SpeeCOmco corpus, to ensure we obtain consistent results since this dataset will be used in later stages for the validation of the interpretability approach. This corpus does not include HC speakers, that is why we could not analyze the generalization capacity of the model on this dataset following the same logic as in section 5.3.1.

For each dataset, the Pearson correlation coefficient (noted r in the following) is calculated between the resulting balanced accuracies for the overall set of speakers and each of the perceptual measures. The measures taken into account include intelligibility, severity, phonemic alteration, and perceived phonological deviation described in section 4.2.2. Points are graphed on a scatterplot, with blue and green points being HC speakers and patients respectively. The coordinates are the balanced accuracy given by the model for the speaker recording on the X-axis and the perceptual measure in question on the Y-axis. A best-fit line, as well as the correlation coefficient r, are also given in the same graph. Figures 5.7, 5.8 and 5.9 are the result of this analysis on *C2SI-LEC*, *C2SI-DAP*, and SpeeCOmco datasets, respectively. This analysis also aims to discover which perceptual measure is the closest to the model objective. To this end, we did not correlate the balanced accuracies calculated for the recordings of a dataset with the perceptual measures assessed exclusively on the same dataset, but rather with all the perceptual measures at our disposal. That is, for instance, we correlate the balanced accuracies issued from C2SI-LEC recordings with Intel-LEC and Sev-LEC but also with Intel-DES, Sev-DES, and Phnm-DES. In the following, we analyze these different correlation results.

C2SI-LEC:

Figure 5.7 illustrates the balanced accuracy calculated for each speaker recording from the passage reading task against the perceptual measures Sev-LEC, Intel-LEC, Sev-DES, Intel-DES, and Phnm-DES respectively, and provides the corresponding r-values. These figures, whatever the perceptual measure observed, show a coherent behavior by comparing the HC speakers and patients in terms of balanced accuracy but also in terms of perceptual ratings. Indeed, the blue dots are concentrated on the upper right (resp. down right for the Phnm-DES) where we have the highest severity and intelligibility scores (resp. the lowest phonemic alteration scores) as well as the highest balanced accuracies, reflecting a high quality of speech. Moreover, we can clearly see that both severity ratings Sev-DES and Sev-LEC, resp. in sub-figures 5.7a and 5.7c, have the strongest correlation with the CNN classification accuracy ($r \geq 0.9$). As to the intelligibility ratings, we were expecting a low correlation with Intel-LEC and were not at all surprised when obtaining an *r-value* equal to 0.77 because of all the biases [2] related to this perceptual measure assessment that we previously mentioned in section 1.4.1. Further, we were expecting a stronger correlation with Intel-DES roughly in line with the strong correlations obtained for

[2]These biases are still confirmed by the ceiling effects and the lower positive slope of the best line fit (i.e. a flatter upward tilt) in figures 5.7d and 5.8e related to Intel-LEC.



(a) Sev-DES **(b)** Intel-DES

(e) Phnm-DES

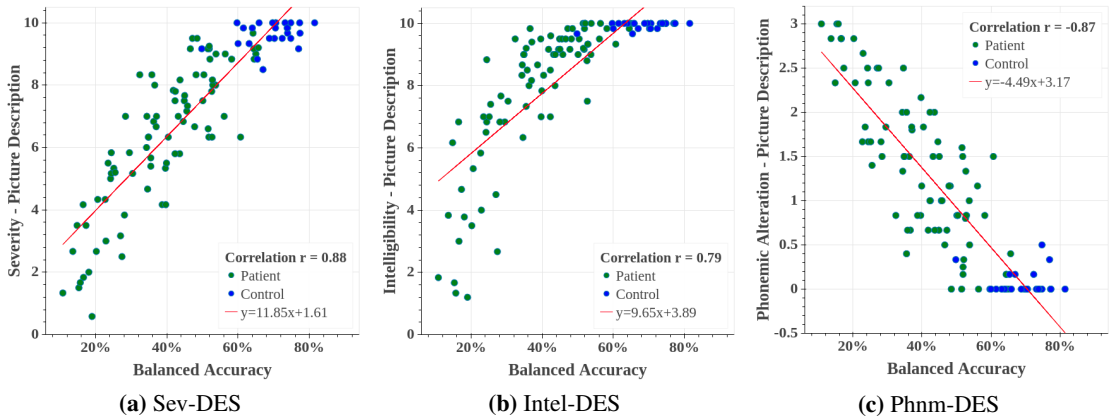
(c) Sev-LEC **(d)** Intel-LEC

Figure 5.7: *Scatter plots of different perceptual measures vs. model balanced accuracy on the C2SI-LEC HC and patient speakers*

severity. This is actually not the case since the correlation for Intel-DES deteriorates to 0.81. This *r-value* decrease could be explained by the fact that the severity rating is focusing more on speech sounds - which is closer to our CNN model objective - rather than on the spoken message as is the case for the intelligibility rating. Finally, the correlation between the model accuracy and the phonemic alteration rating (Phnm-DES), shown in figure 5.7e, is slightly lower than for severity score but still strong reaching up to -0.88. Regarding the initial goal of modeling the characteristics of phonemic units, this high *r-value* still confirms the phonetic modeling capabilities of the adopted CNN-based architecture.

C2SI-DAP:

Similarly to the analysis done in the previous section on *C2SI-LEC* dataset, we report in this section the results obtained on *C2SI-DAP* recordings. As regards this dataset, we can explore the correlation of the model accuracy with an additional perceptual measure which is the phonological perceived deviation (PPD). The corresponding scatterplots and correlation values are given in figure 5.8. In general, we can see that the correlation pattern obtained on *C2SI-LEC* still holds true on *C2SI-DAP*, with a slight downward trend on overall perceptual measures. It is notewor-



(a) Sev-DES **(b)** Intel-DES **(c)** Phnm-DES

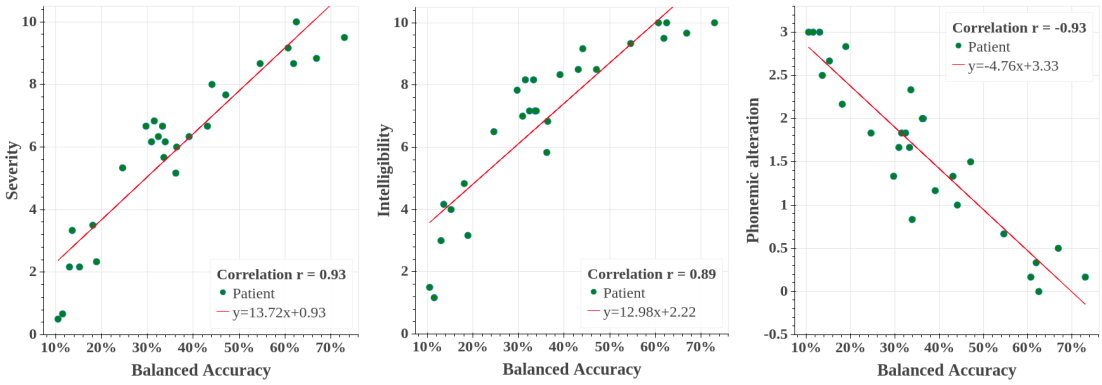
(d) Sev-LEC **(e)** Intel-LEC **(f)** PPD-DAP

Figure 5.8: *Scatter plots of different perceptual measures vs. model balanced accuracy on the C2SI-DAP HC and patient speakers*

thy, moreover, that the strongest correlation of the balanced accuracies obtained on *C2SI-DAP* recordings is with the PPD perceptual rating with an *r-value* equal to −0.89. Indeed, as previously defined in section 4.2.2, PPD provides a way to metrically measure the difference between the distorted phonetic realization of linguistic units from the expected forms that occur in normal speech. Closely related to the way PPD is assessed, our model is trained exclusively on healthy phonemes and tested on distorted ones. Correspondingly, the performance of the model on a patient recording reflects the misclassified phonemes due to the speech disorder, which is quite similar to the PPD score calculated as the average number of phonological features misidentified by the listeners due to the articulatory disorders of the speakers. It is noticeable that the major difference is that the accuracy of the model is calculated considering the phoneme labels while the PPD is calculated based on the phonetic features, and the strong correlation between both can reflect, once more, the model capacity to encode finer phonetic characteristics so far.

SpeeCOmco:

As already mentioned, the same analysis is carried out on SpeeCOmco dataset to ensure consistency with the previously observed trends. That is, the main objective is to make sure no



(a) Severity **(b)** Intelligibility **(c)** Phonemic alteration

Figure 5.9: *Scatter plots of different perceptual measures vs. model balanced accuracy on the patients of SpeeCOmco dataset*

odd behavior is brought forward by the model on this dataset since it will be used later in the validation of the methodology. The results are shown in figure 5.9. Obviously, this dataset includes fewer patients and no HC speakers, yet, it is still representative since their corresponding perceptual ratings are distributed over the full range of values. We can notice that actually, the correlation values are consistent with expectations and previous observations, with a noticeable overall increase. Indeed, for both severity and phonemic alteration, an absolute *r-value* equal to 0.93 is obtained when correlating these perceptual ratings with the balanced accuracies obtained on SpeeCOmco recordings. Not only this but also the strong correlation of these balanced accuracies with the intelligibility measure achieving 0.89 confirms our choice for this dataset for the validation of later stages.

5.4 Discussion

In this chapter, we proposed a CNN-based model for the classification of French phonemes from speech acoustic features. We have shown that, while it was exclusively trained on healthy speech and has never been exposed to pathological data, the model is able to reflect the degree of speech alteration. It is important to recall that, despite the constraint of the trade-off complexity/performance that we have and that largely limited our choices, the phoneme classification task itself remains challenging. In the following, we highlight some of the important challenges in designing such a task, part of which is already pointed out in the results. Indeed, various sources of variability such as style of speech, speaker characteristics, noisy environment, and co-articulation effect can create difficulty in phoneme classification. The co-articulation phenomenon is one of the major challenges. As the articulators (e.g. tongue, lips, glottis, etc.) change their position smoothly from one phoneme to the next, the effect of neighboring phonemes influences the current phoneme, thereby, its acoustic property. As well, characteristics related to speakers such as speaking rate affecting both temporal and spectral properties of speech signal [Siegler and Stern, 1995], age, gender, and so on have an adverse effect on phoneme classification. Moreover, accent induces a large variability in pronunciation and affects acoustic features. Thereby, the classification boundaries drawn during training may

change if the accent changes. Environmental noise is also a factor to consider. Many sorts of noises could be induced in the signal and negatively affect the task in question (e.g. echo, microphone quality, a background interfering speaker, etc.). Finally, accurate segmentation and optimal feature extraction are difficult to achieve since they may retain unwanted information or inadvertently discard important information for phoneme classification [Chibelushi et al., 2002].

Still, the results of this first step aiming to represent the French phonemes and highlight their relevance in our clinical context look very promising. We are therefore ready to move on to the next step which is the exploration of the resulting model (see figure 5.10). This step will be detailed in the next chapter. Nevertheless, it is worth noting that the choices we made so far can be criticized. Indeed, CNN mostly reduced the need for handcrafted features due to its ability to learn the problem-specific features from the raw input data. In speech-related tasks, it is quite common to apply CNN directly on raw speech [Palaz et al., 2015, Ghahremani et al., 2016, Passricha and Aggarwal, 2018]. In this case, the first convolutional layer acts automatically as a filterbank learned in a data-driven manner, and consequently, the data preprocessing stage is no more needed. It could be interesting in our case to try such an alternative instead of starting from a spectral-based representation of the speech. Moreover, even the CNN architecture can be discussed when there are newer and more advanced architectures available for speech representation such as wav2vec and pase+. These newer architectures may have better performance and accuracy. However, the challenge would mainly be related to their explainability.

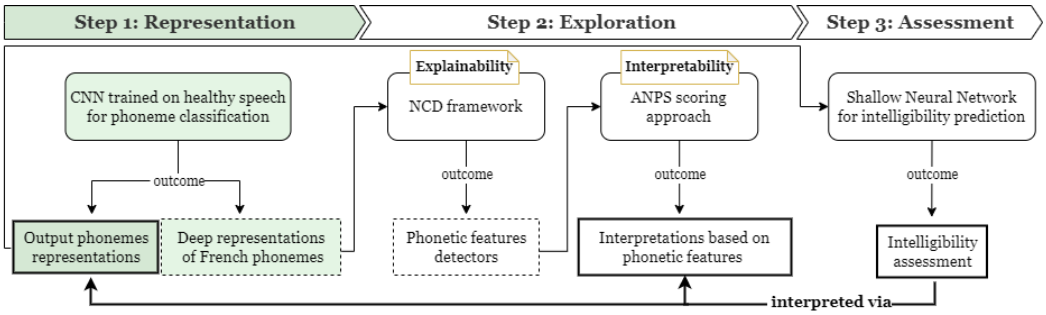


Figure 5.10: *A step forward in the achievement of the proposed methodology: the accomplishment of step 1*

Chapter 6

Step 2: Exploring the Phonetic Feature level

Contents

6.1 Specific Context

In this chapter, we address the second step of our proposed methodology. This step consists in peering into the CNN-based phoneme classifier and exploring its internal representations in order to find meaningful concepts of relevance in our clinical context. This chapter is composed as follows. We start with a brief state-of-the-art of the existing works most relevant to our current study. Thereafter, we raise the research questions to which we look for an answer in this second step. Subsequently, we introduce our proposed approach to answer these questions one by one. It is important to note that this approach itself is subdivided into two main parts, including an explainability phase and an interpretability phase. We later validate this approach on different types of speech pathology to show its capacity to reflect the characteristics of each of them. We end this chapter with a discussion summarizing the key strengths and limits of our proposed approach.

6.1.1 Related Work

With the increasing use of DL in several fields, researchers are showing a growing interest in the interpretability and explainability of these tools, particularly in sensitive domains. In the medical field, interpretable and explainable models are even more crucial for the reasons outlined in Chapter 3, as a lack of transparency can cause distrust among healthcare professionals and patients, hindering widespread adoption. As such, researchers are exploring various techniques to make DL models more transparent and understandable. We briefly introduced these techniques in section 3.3 dedicated to the taxonomy of interpretability and explainability. In this introductory section of the second step, we introduce some works related to the investigation of hidden representation, which are most relevant to our current work.

Recently, researchers have investigated DL models for auxiliary knowledge they encode in their learned representations. These works took place mainly in the computer vision field [Zeiler and Fergus, 2014, Selvaraju et al., 2017]. Particular interest was paid to the investigation of the internal representations of CNNs. Notably, authors in [Gonzalez-Garcia et al., 2018] have shown that semantic part detectors emerge in object classifiers (e.g. bird, car, and cat). In addition, [Zhou et al., 2015, Bau et al., 2020] were able to discover automatically meaningful object detectors in their CNN trained for scene classification. Zhou et al. [Zhou et al., 2018] introduced a more general method for interpretability, called Network Dissection, measuring the alignment between convolutional units and visual interpretable concepts. As work is in full growth in the computer vision domain when it comes to interpretability and explainability, a growing interest in these techniques is being shown in the speech domain. For instance, Dalvi et al. [Dalvi et al., 2019] aim to the analysis of individual neurons in order to identifying linguistically meaningful in deep NLP end-to-end models. In their turn, [Krug et al., 2018, Krug et al., 2021] analyzed neuron activation profiles (NAPs) to explain a CNN-based speech recognizer, and revealed via clustering techniques that neurons and layers implicitly learned intermediate representation related to phonemes and graphemes. In works [Nagamine et al., 2015] and [Pellegrini and Mouysset, 2016], authors have shown that clustering techniques applied to the neurons of a DNN-based phoneme classifier and the filters of a CNN-based phoneme classifier respectively, uncovered the presence of phonetic features in the hidden representations of these models. A neuron-level analysis was also carried out by authors in [Qian et al., 2016] and

[Durrani et al., 2020] to investigate the hidden representations of a sequential neural model of sentence and pre-trained language models, respectively. The analyses reveal that neural vector representations often contain a substantial amount of linguistic information. Similar studies reveal interesting findings such as how different linguistic properties (e.g. morphology and syntax) are captured within an end-to-end dialect identification model [Chowdhury et al., 2020], or pre-trained speech models for speaker, language, and channel properties [Chowdhury et al., 2021].

It is worth noting that we focused more on studies related to DL models' interpretability and explainability in the speech domain since it is our work context. Yet, generally speaking, these techniques are mostly applied to the computer vision domain. This special focus can be explained by the ease of handling of image features since they are visually interpretable, compared to speech data where features are more complex, variable, and obviously less visual. Speech signals are long, have variable lengths, and are of complex hierarchical structure.

While the field of interpretability and explainability has recently made inroads in speech applications, it is even less developed in the context of speech pathology and can be considered close to being absent. Proposed very recently, the work [Klumpp et al., 2022] is the closest we found to our proposal. In their work, Philipp Klumpp et al. proposed a phonetic recognizer trained only on healthy speech data to examine the impact of Parkinson's Disease on the phonetic footprint of patients. They defined the phonetic footprint as resembling the distribution of production probabilities among different phonemes or phonological classes for an individual speaker. They have shown that their model discovered patterns that have been previously reported in the literature and enabled the phonetic profiling of Parkinson's Disease patients. Compared to our work, we both support the idea that it is not necessary to train systems with pathological speech data so that they could reflect characteristics related to the pathology. The main difference is that the approach of the phonetic footprint proposed by authors in [Klumpp et al., 2022] is mainly based on the output probabilities of the phoneme classifier. They have also shown that an intermediate feature vector can encode PD-related information since they were able to observe noticeable differences in hidden states between phoneme productions of HC and PD. However, they did not get inside the model to explain the nature of encoded representations and enhance its interpretability to gain the trust of the experts. Nor did they provide a formal means to translate the observed differences in the hidden states of the model into a measure reflecting the production deviation of the phoneme or phonetic class in question for an individual speaker. They also did not report results on different types of pathology other than Parkinson's Disease.

In our ongoing efforts, we believe that we are making a significant contribution to the characterization of speech pathology through the use of deep learning. Furthermore, we are actively taking steps to mitigate the impact of the black-box nature of these models and alleviate the mistrust among experts in this field to the greatest extent possible. In the following section, we outline the research questions that will be discussed in this chapter.

6.1.2 Research Questions

We carry this second step with the following research questions (RQs):

- **RQ1:** Can we find a concept that is relevant in the clinical phonetics context (e.g. phonetic features) captured in the network learned representations?

- **RQ2:** Where in the network is it preserved and how localized or distributed is it? Is there a relationship between the complexity of the property and the number of neurons required to encode it?

- **RQ3:** How can we retrieve information related to speech pathology based on the outcome of the explainability process laid down?

- **RQ4:** Does the investigation of the pre-identified neurons capturing relevant concepts bring out a fine-grained analysis of the speech quality? Is it able to reflect the characteristics of each speech pathology when exposed to different ones?

To the best of our knowledge, this is the first work that carries a layer-wise and fine-grained neuron-level analysis of a DL model trained on healthy speech, to later take advantage of the interpretable resulting neurons to bring knowledge about speech pathology.

6.2 Neuro-based Concept Detector: Our proposed Framework for Neurons Explainability

In this section, we address the first and second research questions (RQ1 & RQ2). Overall, our aim is to explore the hidden representations of the previously trained CNN to see if a relevant concept, in regard to the clinical phonetics context, could be automatically captured in the layer and neuron levels. For this purpose, we first explored an explainability method that was applied in the computer vision domain by [Rafegas et al., 2020]. Based on the calculation of a Class Selectivity Index (CSI) for each neuron, the aim of this method is the identification of neurons selective for one specific final class. We detail this approach, how we adjusted to fit our application domain, as well as the obtained results in section C.1 of the appendix C. We conclude that CSI is not adequate in our case since it does not really reflect the encoding properties of a given neuron. To address these shortcomings, we then propose *Neuro-based Concept Detector (NCD)*, our general analytic framework for the explainability of hidden neurons/layers of a DNN performing a classification task. Of a wide application, this framework involves a representation vector of neuronal activity characterizing hidden neurons, coupled with a score measuring the capacity of these neurons to detect a specific concept related to the final task. If we shall ensure overall consistency with what we already argued in section 4.3.2, our designed framework serves as the explainability tool described in the second perspective. More precisely, it is worth mentioning that this framework is dedicated to the explainability of neurons within fully-connected layers. Even though filters in the convolutional layers hide certainly valuable knowledge that could be considered further to take advantage of the internal representations of phonemes in our clinical context (see section C.3 of the appendix C), we chose not to include them in our explainability study. NCD is composed of three stages organized in the following subsections.

6.2.1 Representation Vectors of Neurons

This first stage is more dedicated to the visualization of the organization of neurons in different fully-connected layers. Although optional, this stage is still important. Indeed, visualization is a central human cognitive ability, seen as a powerful tool for exploratory data analysis and one that enables inductive reasoning in a natural, seamless manner [Vellido, 2020]. In our specific case, visualization can help generate knowledge about the internal organization of the neurons within a layer and how this organization evolves through layers, leading to potential hypotheses about the role and structure of the data flow over the network.

To this end, we start by defining key notions necessary for the understanding of the proposed approach. Let $h_{n,i}$ be the activation value of the neuron n, given the i^{th} input frame of a stimulus set, (refer to section 5.2.1 for a reminder of the frame notion). A normalized activation $a_{n,i}$ is calculated for each neuron by dividing the initial activation values of the neuron for different input frames of the dataset by the maximum value reached over all these values. That is, $a_{n,i} = \frac{h_{n,i}}{hmax_n}$ where $hmax_n = \max h_{n,j} \forall j$.
Based on that, a process to generate representation vectors reflecting the neuronal activity is set up as illustrated in figure 6.1 and detailed below:

- For a neuron n, a histogram is generated for each phoneme k in order to approximate the distribution of the neuron activations as a response to all the frames having the phoneme k as a true label.
  The histogram displays the number of frames falling into each interval of normalized activation, also called bins, which have equal widths and divide the entire range of normalized activation [0, 1]. Here, the number of bins is fixed at 20.

- Subsequently, a vector $V_{n,k} \in N^{20}$ containing the number of frames appearing in each bin is derived from each histogram. Thus, $V_{n,k}$ is a representational vector characterizing the response of the neuron n to the phoneme k.

- Finally, a concatenation of these vectors generated for each of the 30 phonemes for a given neuron results in a 600-dimensional representation vector, and is considered later as characterizing the neuron n for visualization purposes.

The aforementioned process is therefore applied on *BREF-Int*, the reference dataset of healthy speech. For each layer of the fully-connected layers, the set of 600-dimensional vectors representing the hidden neurons is prepared. Subsequently, a projection of these representation vectors into a 2-dimensional space is performed by using a t-Distributed Stochastic Neighbour Embedding (t-SNE) [van der Maaten and Hinton, 2008]. Since t-SNE applies a non-linear dimensionality reduction technique where the focus is on keeping the very similar data points close together in a lower dimensional space, we expect to observe neuron clusters sharing similar encoding properties. Figure 6.2 is the visualization of this projection. As expected, interesting insights about the organization of neurons per layer can be observed. Indeed, we can mention the presence of clusters of neurons in the different examined layers. This is very promising to get on with the next step in which we try to uncover the encoded information carried out by these neurons and check if the outcome of this upcoming step is coherent with the cluster organization obtained so far.

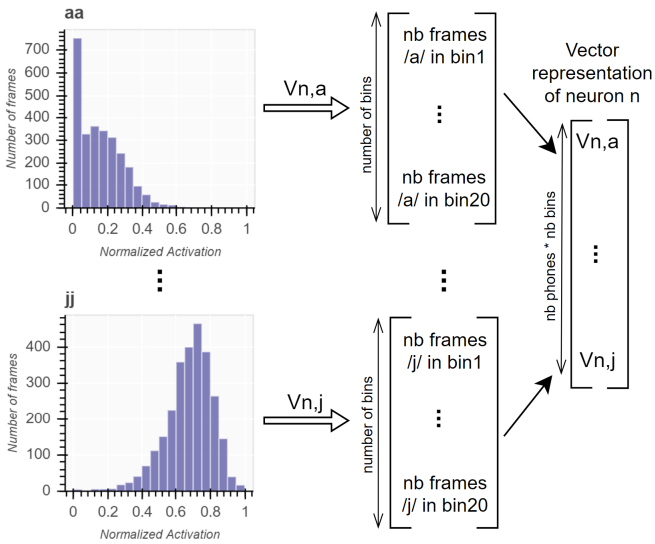
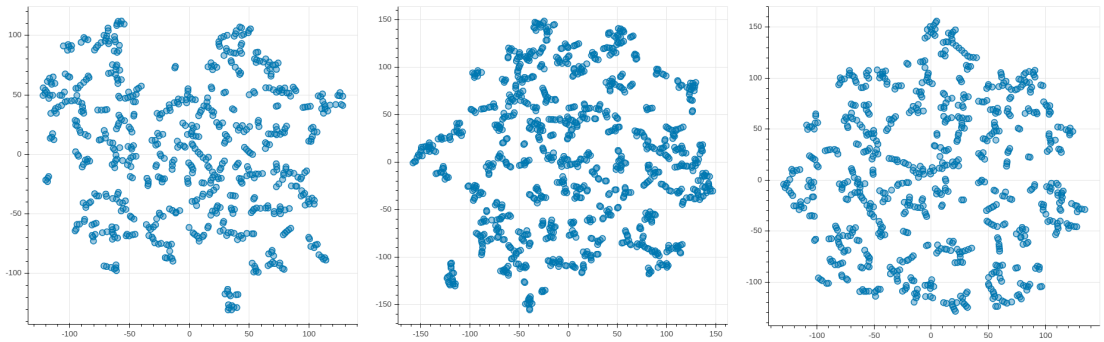


Figure 6.1: *Process of representation vector generation for a given neuron*

6.2.2 Fixing the concept to explore: Why phonetic features?

At this stage, we find it useful to recall the later usage of the concept in our global methodology, already highlighted in section 4.3.1. Let's not forget that the final goal of this work is to set up a solution based on DL assessing the speech intelligibility of disordered speech while providing an interpretation of this assessment at a more fine-grained level. To do so, we already incorporated one intermediate interpretable domain (i.e. phonemes) via the phoneme classifier already presented in step 1. Now, the idea is to come up with an extra-interpretable domain by locating an emergent meaningful concept that was automatically learned by the phoneme classifier. This concept will indeed enhance the interpretability of the global model performing the target assessment task. But what are the criteria for choosing this concept? The response to this question lies in the first research question itself (RQ1): *"Can we find a concept that is relevant in the clinical phonetics context (e.g. phonetic features) captured in the network's learned representations?"* In fact, the concept that we want to explore has to be of great relevance in the clinical



(a) FC1 **(b)** FC2 **(c)** FC3

Figure 6.2: *Visualization of the representation vectors of neurons by fully-connected layer*

phonetics context in which we are involved. For instance, through this concept, we must have the possibility to link the physiologic characteristics of impaired speech with the final speech intelligibility loss.

Phonetic features can therefore be a great choice since they can be directly related to the physiology of the speech production system. To clarify this idea, let's take the example of nasal phonetic feature for consonants. An alteration in the production of the nasalized consonants (i.e. /n/,/m/,/ɲ/) is directly linked to a decreased nasal airflow, in other words, a hyponasal resonance[1]. Hyponasal resonance may be associated with any cause of nasal obstruction, such as adenoid hypertrophy or congested nasal turbinates. Signs of hyponasal resonance include alteration of /m/ to sound like /b/, /n/ to sound like /d/, and /ɲ/ to a hard /g/ within the context of speech. The velopharynx continues to be active during the production of words and phrases with nasal consonant sounds, but it is maintained between a relaxed and a closed state. As well, there are systematic patterns related to phonetic features that have been observed among speech disorders and which provide clues to the basis of the deficit. For instance, aphasic patients typically make phoneme substitution errors involving one phonetic feature. These errors are driven by a hierarchy, with substitution of the place of articulation being more common, then voicing, and fewest, manner of articulation [Blumstein, 1998].

To conclude, finding that phonemes are adequately characterized in a space of phonetic features within the trained CNN will be a great exploration since such a characterization will offer a better basis for providing the clinician with information that is directly related to speech therapy. This exploration will be carried out through a special scoring approach that we detail in the next section.

6.2.3 Characterizing the Neuron Ability to detect the concept of phonetic features

At this stage, we aim to measure the alignment of each hidden neuron belonging to the fully-connected layers, with the concept of phonetic features that we have fixed in the previous section. Therefore, we will be based on the definition of French phonetic features proposed by Ghio et al. [Ghio et al., 2020] detailed in section 1.2. Consequently, we propose a score that aims to quantify the degree to which a neuron detects the presence of a phonetic feature. This degree reflects the contrast between the neuron activations for phonemes that present this phonetic feature and the neuron activations for phonemes that do not present it.

To this end, the normalized activation values of the individual neurons as a response to *BREF-Int* dataset were gathered and visualized for later analysis. An example of visualization is given for neuron 214 of the layer FC2 in the form of a Jitter plot, shown in fig. 6.3a. Each point of the plot corresponds to the normalized activation of the neuron in question in response to a single frame from the dataset stimuli *BREF-Int*. An organization of these points along the Y-axis is performed based on the true labels of the frames (i.e. the phonemes obtained from the forced alignment). As it can be visually observed, this neuron has a distinctive response for the three nasal consonants (i.e. /n/,/m/,/ɲ/).

Based on these observations, we design the following score. For each neuron n, let $A_{n,k}^{BREF}$ be

[1] Hyponasal resonance occurs when there is not enough nasal resonance on nasal sounds due to a blockage in the nasopharynx or nasal cavity [ASHA].

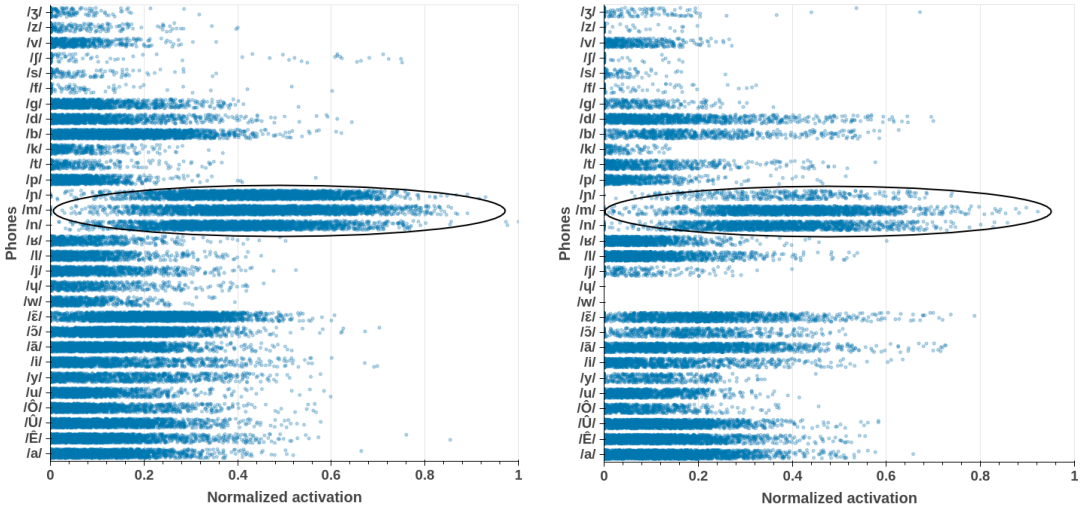
the set of normalized activations of the neuron n for all the frames having the phone k as a true label and belonging to *BREF-Int*. We note the median activation value of the neuron n for the phone k as $m_{A^{BREF}_{n,k}}$. The choice of the median is discussed later in the final section of this chapter. The score S_{n,T_x}, quantifying the degree to which a neuron detects the presence/absence of a phonetic feature, is therefore calculated for each neuron n and phonetic feature T_x as follows:

$$S_{n,T_x} = \frac{1}{|[+T_x]|} \sum_{k \in [+T_x]} m_{A^{BREF}_{n,k}} - \frac{1}{|[-T_x]|} \sum_{k \in [-T_x]} m_{A^{BREF}_{n,k}} \tag{6.1}$$

where $|.|$ is the cardinality of a set. Since phonetic features are binary concepts characterizing vowels and consonants separately, this distinction is incorporated to the score thanks to $x \in [v, c]$, that denotes the macro-class of either vowels or consonants, v and c respectively. Consequently, T_v and T_c denote respectively a vowel and a consonant phonetic feature where:

- $T_v \in PF_v$ *where* PF_v = *{nasal, back, round, high, low}*
- $T_c \in PF_c$ *where* $PF_c = \{sonorant, continuant, nasal, voiced, compact, acute\}$

It is important to mention that the score $S_{n,T_x} \in [-1; 1]$. That is to say, a strong value close to 1 reflects that the neuron is a strong detector for the presence of the phonetic feature in question since it distinguishes the phonemes presenting the feature by a high activation level. At the same time, a very low activation level distinguishes the complementary set of phonemes not presenting this feature. At the other extreme, when a neuron has a very low score close to -1, it means that it is a strong detector for the absence of the phonetic feature, which is also relevant. With a score reflecting how well a neuron encodes a given phonetic feature, we consider that the neuron n is a detector of the presence of phonetic feature T_x, noted $[+T_x]$, if S_{n,T_x} exceeds a given threshold (e.g. > 0.25). Conversely, if S_{n,T_x} is below the opposite threshold (e.g. < -0.25) then the neuron



(a) *BREF-Int* **(b)** *C2SI-LEC* control speakers

Figure 6.3: *Jitter plot visualizing the normalized activations for unit 214 of FC2 layer according to phone frames (distinctive response for nasal consonants is circled)*

n is considered as a detector of the absence of the phonetic feature T_x, noted $[-T_x]$. Clearly, different thresholds could lead to a different number of neurons selected as phonetic feature detectors across layers. However, we observe that it does not result in a significant change in terms of the distribution of this set of neurons over the different phonetic features. Thus, we have empirically fixed the threshold to be ± 0.25[2]. Additionally, given that a neuron can be a detector for several phonetic features (associated with relevant scores respecting the threshold), the top phonetic feature is chosen in this case. Suppose the neuron is identified as a detector for multiple phonetic features belonging to both vowel and consonant macro-classes. In that case, it will be considered as a detector for the top phonetic feature for both vowels and consonants.

6.2.4 Results of the application of NCD framework: Emergence of phonetic feature detectors

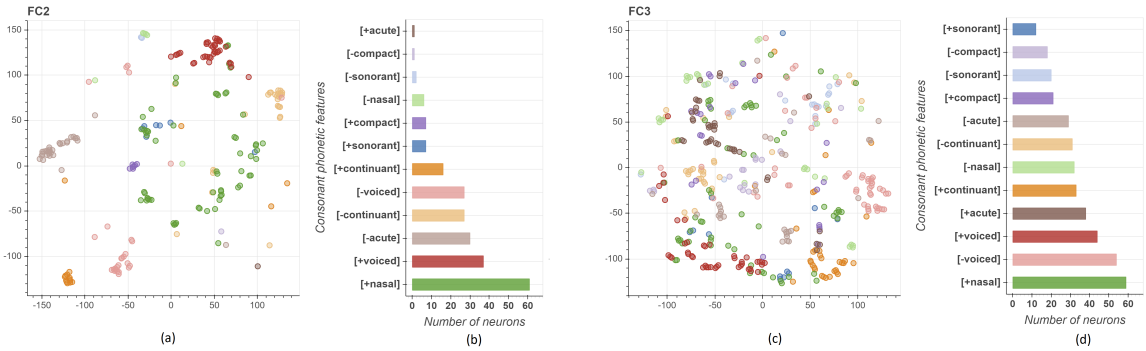


Figure 6.4: *t-SNE visualization highlighting neurons with phonetic feature encoding properties for consonants in: (a) FC2 (c) FC3*
(b) & (d) Sorted counts of neurons detecting each of the consonant phonetic features in FC2 & FC3 resp.

In this section, we report the results of the application of the NCD framework to the fully-connected layers of the trained phoneme classifier. Our focus at this stage is basically to analyze if phonetic features, the concept that we have fixed in section 6.2.2, emerge in these visualized layers (i.e. response to the RQ1). And if it does, we want to locate in which layers it happens and, more specifically, identify the neurons that detect this concept (i.e. partial response to RQ2). To do so, we apply the scoring approach proposed in section 6.2.3 on the different neurons across layers in order to measure their alignment with phonetic features. Thus, the outcome can be seen as the set of neurons identified as phonetic feature detectors based on the different constraints of the scoring approach, all with the corresponding phonetic features they detect. To illustrate the result of this application, we choose once more the visualization technique. We will be based on the neuron visualization already presented in figure 6.2, after the projection of the neurons representation vectors (see section 6.2.1). Only neurons identified as detectors are taken into consideration in the following visualization plots. Moreover, an extra variable is added to this visualization which is the color of phonetic features detected by the neurons. For the sake

[2]The interval of scores $[-1; 1]$ mentioned above is theoretical. Empirical experiments rather show a value interval between $[-0.5; 0.5]$, hence this fixed threshold value.

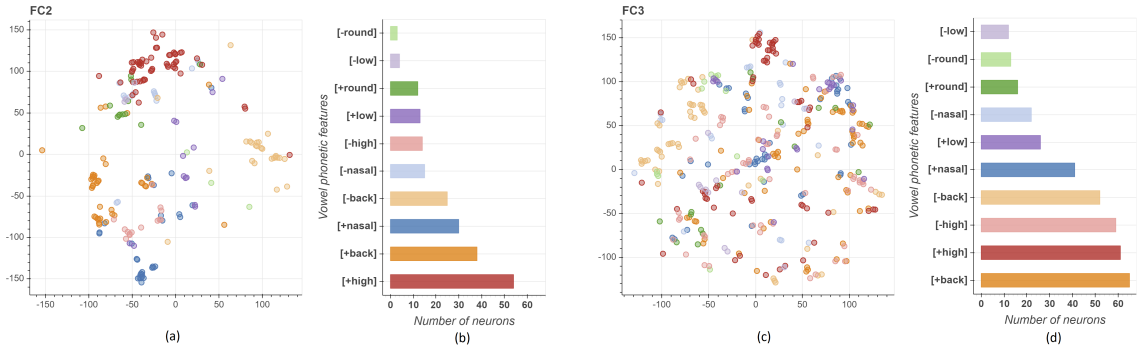


Figure 6.5: *t-SNE visualization highlighting neurons with phonetic feature encoding properties for vowels:*
(a) & (c) Plots of the embedded neurons of FC2 & FC3 resp. (b) & (d) Sorted counts of neurons detecting each of the vowel phonetic features in FC2 & FC3 resp.

of clarity, this visualization is presented in the form of two figures 6.4 and 6.5, considering separately the neurons detectors of consonant phonetic features and those detectors of vowel phonetic features.

This visualization reveals fascinating insights. Firstly, we have to mention that the absence of FC1 in plots is due to the fact that none of its neurons are identified as phonetic feature detectors, neither for vowels nor for consonants (according to the threshold used). Secondly, it is worth mentioning the presence of dense neuron clusters with homogeneous colors, automatically identified as encoding the same phonetic feature. This outcome is in line with our expectations that these neurons are more than just randomly distributed; they seem to be grouped according to their encoding properties. On top of that, when analyzing the number of phonetic feature detectors, we can note that it increases by a factor of 1.75 when we go deeper in layers toward the final layer performing phoneme classification. Indeed, although FC2 and FC3 have the same number of hidden neurons, the total number of neurons detecting vowel phonetic features has increased from 208 in FC2 to 367 in FC3. The same trend is observed for neurons detecting consonant phonetic features, slightly more numerous than those detecting vowel ones, with 222 detectors in FC2 and 391 in FC3. This emergence of phonetic feature detectors when going deeper in layers suggests that these features allow discrimination among phoneme classes, in the classification task.

To give a more focused analysis, figures 6.4.b and 6.4.d show the sorted count of neurons detecting consonant phonetic features in FC2 and FC3 respectively. Our analysis reveals that neurons specifically detecting the phonetic feature of nasality *[+nasal]* exhibit the greatest presence in both layers of the model. A similar study is performed on the vowel phonetic features, summarized in figure 6.5, showing roughly the same patterns. Still, in both layers, we find that the phonetic features *[+high]* and *[+back]* appear in the top two as the phonetic features having the greatest number of neurons assigned to their detection. Furthermore, the illustration in the presented figures highlights the substantial discrepancy in the number of detectors between various phonetic features.

Consequently, at this stage, the question that can be raised is as follows: *Is there a relationship between the complexity of the property and the number of neurons required to encode it?* (the second part of RQ2). To answer this question, we need to define the notion of complexity related to phonetic feature production. The complexity of phonetic feature production refers to the level of difficulty in producing specific sounds or phonetic features. This can involve the coordination of various physiological processes, such as the movement of the lips, tongue, and vocal cords, as well as the precise control of airflow and pressure. For example, sounds produced at the back of the mouth, such as the velar sounds /k/ and /g/, can be more challenging to produce than sounds produced at the front of the mouth, such as the bilabial sounds /p/ and /b/. Additionally, sounds involving complex tongue movements, such as the lateral sound /l/ can be more difficult to produce than other sounds. However, it would be an over-simplification if we generally assume that the phonetic feature *[+compact]* is more complex to produce than *[−acute]* since complexity can vary depending on many factors including the individual's native language, age, and any speech disorders they may have. Ultimately, the complexity of phonetic feature production is determined by a combination of multiple factors which make difficult a direct answer to the raised question.

Overall, the results of this analysis suggest that the concept of phonetic features is an important part of the representation built by the CNN-based model to obtain discriminative information for the final task of phoneme classification. We are able to examine and identify neurons detecting these phonetic features more thoroughly. We further evaluate the effectiveness of our NCD in a secondary study (refer to appendix C.2) by ablating the identified neurons and reevaluating the model classification performance.

6.3 Tapping into Phonetic Feature Detectors to interpret Speech Alteration: Artificial Neuron-based Phonological Similarity

As seen in the previous section, we design an original framework, *NCD*, to determine the set of interpretable neurons per layer considered as detectors of specific phonetic features in healthy speech. In this section, we set up a scoring approach through which we retrieve fine-grained information related to speech pathology based on the resulting set of phonetic feature detectors (i.e. response to RQ3).

To start, each selected interpretable neuron, n, is being labelled with the specific phonetic feature t it detects : $[+T_x]$ (i.e. the presence of T_x) or $[-T_x]$ (i.e. the absence of T_x). Let t be the finer notation of T_x including the information of its presence/absence (i.e. t is equal to either $[+T_x]$ or $[-T_x]$). Let $N_{L,t}$ denotes the set of interpretable neurons belonging to layer L and selected as detectors for the phonetic feature t. Considering a new speaker and his/her associated speech production issued from the speech pathology dataset, a characterization of the overall response of the layer for a given phonetic feature can be seen as the evoked response of all neurons belonging to $N_{L,t}$ for only the phonemes presenting the feature t. Subsequently, the similarity of this speech production compared with a standard can be expressed as a ratio including this latter and a reference (i.e. BREF corpus). Similarly to $A^{BREF}_{n,k}$ introduced in section 6.2.3, we note $A^{s}_{n,k}$ the set of normalized activations of the neuron n for all the frames belonging to the phoneme k and produced by the speaker s belonging to the speech pathology dataset in

question. In the same way, we note $m_{A^s_{n,k}}$ the median value of this set of normalized activations.

6.3.1 Local ANPS

For a speaker s and phonetic feature t, we define the score, named *Artificial Neuron-based Phonological Similarity (ANPS)*, as the following ratio:

$$ANPS_{s,t} = \frac{\sum_{n \in N_t} \sum_{k \in t} m_{A^s_{n,k}}}{\sum_{n \in N_t} \sum_{k \in t} m_{A^{BREF}_{n,k}}} \tag{6.2}$$

It should be noted that this score can be declined as a similarity score depending on layers, in this case, $N_t = N_{L,t}$. As well, it can be seen as a global score over all the examined layers.

Consequently, $$N_t = \bigcup_{L \in \{FC2,FC3\}} N_{L,t}$$

It is important to underline that even though we are no more dealing with balanced data in terms of number of samples per phoneme (compared to *BREF-Int*), we assume that the proposed score, as well as, the overall framework, is relatively robust when exposed to highly unbalanced data (recordings from speech pathology dataset). Indeed, it is based on the median computation per phoneme distribution which is less sensitive to outliers than the mean. To support this argument, fig. 6.3b shows the same neuron 214 of FC2 as the one illustrated in fig. 6.3a, but rather on the data *C2SI-LEC* issued from the set of control speakers. Despite differences between data - *C2SI-LEC shares neither the same conditions (e.g. recording equipment and location) nor the same speakers of BREF on which the model is trained* - the neuron preserved its overall behavior and thus, its capacity to detect the phonetic feature *[+nasal]* on consonants. The slight differences between both figures lie in the density of the set of points, pointing back that *C2SI-LEC* dataset is unbalanced, and the total absence of the two semi-vowels /w/ and /ɥ/ in fig. 6.3b since they are not present in *C2SI-LEC* dataset.

Back to the similarity score $ANPS_{s,t}$, the one allows to assess how well acoustic/articulatory characteristics related to phonetic feature t are produced by speaker s, based on the corresponding set of detectors. It can range from zero to an unbounded maximum value. However, a value greater than 1 does not provide more information than a perfect production of the phonetic feature by the speaker in question. Hence, we constrained this score to a maximum value of 1. While a low score close to 0 implies that almost none of the detectors for the phonetic feature in question has explicitly provided a selective response for phonemes presenting the phonetic feature and produced by the speaker s. Consequently, we can assume that the speech production of that speaker does not exhibit typical acoustic characteristics, but rather severely impaired speech.

6.3.2 Global ANPS

To assess a speaker's overall production for vowels or consonants, the corresponding *local ANPS* scores of the speaker in question are averaged, taking into account all phonetic features belonging to the relevant macro-class. Let's recall that $x \in [v, c]$ denotes the macro-class of either

vowels or consonants, v and c respectively, and correspondingly, PF_x is the set of phonetic features of either vowels or consonants.
For a speaker s and a macro-class x, the corresponding *global ANPS* score is defined as:

$$ANPS\,{}^{x}_{s} = \frac{1}{|PF_x| \times 2} \sum ANPS_{s,t}\ ,\ \forall t \in \bigcup_{T_x \in PF_x} \{[+T_x], [-T_x]\} \tag{6.3}$$

6.4 Application in Disordered Speech Context: A Comparative Study

At this point, we consider that we have covered all the necessary steps to finally respond to the most important question raised in RQ4, which is: *Does the investigation of the pre-identified neurons capturing relevant concepts bring out a fine-grained analysis of the disordered speech quality? Is it able to reflect the characteristics of each speech pathology when exposed to different ones?*
Therefore, our aim in this section is to examine to what extent we can rely on this set of neurons detecting phonetic features to extract relevant interpretations of speech intelligibility considering different types of disordered speech. We conduct our study on the C2SI-LEC database to analyze our approach on head and neck cancers, then we extend the analysis to other disorders including dysphonia and different types of dysarthria.

6.4.1 Head & Neck Cancers

Analysis based on global ANPS score

Analyses based on Pearson correlation coefficients are conducted between the *global ANPS* scores per macro-class and each of the perceptual measures for the overall set of C2SI speakers. Table 6.1 is a sum-up of the obtained correlations, illustrating in more detail the *global ANPS* scores per-macro class calculated taking into account the phonetic feature detectors in each layer (i.e FC2 and FC3 separately), or considering all the phonetic feature detectors across layers (i.e. FC2&FC3 simultaneously).

The first point to mention is that the results are coherent with our previous findings illustrating the correlation of the balanced accuracies of the CNN obtained on C2SI-LEC recordings with the different perceptual measures (see section 5.3.3). Indeed, while the strongest correlations are observed between severity measure and *global ANPS* scores, regardless of the layer or macro-class, we find that the correlation with intelligibility is less important. Once more, this comes as a reaffirmation that the CNN model objective is closer to the severity measure, which assessment focuses more on speech sounds, rather than on the spoken message as with the intelligibility measure. Secondly, it can be noted that there is no particular trend in the correlation between different perceptual measures and *global ANPS* scores issued from either FC2 and FC3 separately or considering both of them. This reflects that the proposed score is not sensitive to the number of phonetic feature detectors. More globally, this also tends to denote that the scoring approach but also the entire proposed framework could be properly generalized to any emergent concept, with no constraint on the number of interpretable neurons selected as detectors. In what follows and for the sake of clarity, we report all the following results of the *global ANPS* score considering all the detectors across layers (i.e. FC2&FC3). In the last row

of table 6.1, we add extra information related to the correlation between the *global ANPS* score calculated as the mean of all the phonetic features regardless of the macro-class separation and different perceptual measures. This *global ANPS* score can be seen as a single measure characterizing the production of all phonetic features by a speaker. Obviously, these correlations reflect the same general trend that we previously mentioned.

Table 6.1: *Correlation between global ANPS scores and perceptual measures for HNC*

		Sev-DES	**Intel-DES**	**Phnm-DES**
FC2	Vowels	0.84*	0.74*	-0.80*
	Consonants	0.90	0.82	-0.86
FC3	Vowels	0.83*	0.72*	-0.77*
	Consonants	0.89*	0.81*	-0.85*
FC2&FC3	Vowels	0.84*	0.74*	-0.79*
	Consonants	0.90*	0.82*	-0.86*
	Total (VC)	0.90*	0.81*	-0.85*

(*) The correlation coefficient is statistically significant (P< 0.05).

Analysis based on local ANPS scores

In this section, attention is paid to each individual phonetic feature. Indeed, with regard to our initial long-term objectives, this individual focus is of great interest for linking the impaired speech of the patients to deeply learned knowledge about each phonetic feature in order to generate a meaningful interpretation of the intelligibility loss. Therefore, this analysis is carried out based on *local ANPS* scores computed for each individual phonetic feature per speaker.

For visualization, heatmaps are used to plot *local ANPS* scores, where the X-axis represents the C2SI speakers sorted from the least severely affected (on the right) to the most severely affected (on the left), according to the severity measure Sev-DES. The Y-axis represents the phonetic features of the macro-class in question. A sequential color scale shows the progression from the most to the least opaque shades of red color, representing low to high score values.

As a first global observation, we can clearly mention that cells with high opacity are concentrated on the left side, which is consistent with the increasing severity level of the corresponding patients. Regarding fig. 6.6 dedicated to vowels, we can firstly mention that the feature *[+nasal]* has deteriorated even for HC speakers. Although it can be surprising, we are convinced that the set of selected neurons associated with this feature has been satisfactory in fulfilling their task. Indeed, this can simply reflect the confusion observed on nasal vowels produced by the C2SI speakers, even the HC speakers, when compared to BREF speakers. This confusion is due to the difference in the recruitment regions of speakers (i.e. different regional accents), already reported in the first step (see section 5.3.2) while analyzing the CNN classification performance via confusion matrices. Furthermore, since the heatmaps clearly show that speech degradation does not have the same effect on the different phonetic features, these initial results are supported by a correlation analysis between the different *local ANPS* scores and the perceptual measures. The details of this correlation analysis are presented in table 6.5. Same as global scores, local scores

Table 6.2: *Correlation between local ANPS scores and perceptual measures of C2SI speakers*

	Phonetic features	Sev-DES	Intel-DES	Phnm-DES
Vowels	**[+nasal]**	0.40*	0.31*	-0.41*
	[−nasal]	0.72*	0.62*	-0.68*
	[+back]	**0.77***	0.73*	**-0.77***
	[−back]	0.55*	0.48*	-0.47*
	[+round]	0.73*	0.67*	-0.65*
	[−round]	0.26*	0.19	-0.26*
	[+high]	**0.81***	**0.76***	**-0.76***
	[−high]	0.17	0.07	-0.1
	[+low]	0.22*	0.12	-0.15
	[−low]	0.48*	0.49*	-0.42*
Consonants	**[+sonorant]**	0.41*	0.37*	-0.34*
	[−sonorant]	**0.79***	0.74*	**-0.76***
	[+continuant]	0.72*	0.73*	-0.73*
	[−continuant]	**0.80***	0.68*	**-0.76***
	[+nasal]	0.60*	0.58*	-0.50*
	[−nasal]	**0.76***	0.71*	**-0.75***
	[+voiced]	0.54*	0.50*	-0.38*
	[−voiced]	0.69*	0.67*	-0.62*
	[+compact]	**0.84***	0.74*	**-0.86***
	[−compact]	0.57*	0.48*	-0.58*
	[+acute]	**0.77***	0.72*	**-0.76***
	[−acute]	0.70*	0.60*	-0.69*

(*) The correlation coefficient is statistically significant (P< 0.05).
For clarity, correlation values above 0.75 (resp. below -0.75) are in **bold**.

follow similar correlation trends with perceptual measures, all with a marked preference for the severity ratings. Besides, to follow up the analysis on vowel phonetic features, we can observe that *[+high]* has the greatest correlation with Sev-DES with r equals to 0.81, closely followed by *[+back]* with r equals to 0.77. In lay terms, the *[+high]* feature reflects the characteristic of vowels articulated while the tongue is positioned high in the mouth. The *[+back]* feature reflects the characteristic of vowels articulated while the tongue is positioned relatively back in the mouth. These observations demonstrate a strong relationship between impaired speech and tongue movements in the speech production considered here, which is consistent with the clinical data. This also has been confirmed through a perceptual phonetic analysis conducted on the pseudo-word production task performed by all the speakers involved in the *C2SI corpus* [Rebourg, 2022].

Regarding the consonant phonetic features, we can first observe from Fig. 6.7 that voicing feature is almost not impacted even for patients with very severe degradation. This indicates that the vocal cords of patients can normally vibrate to produce voiced phonemes. This observation is consistent given that patients in *C2SI corpus* do not suffer from laryngeal cancer, thus no direct impact on their vocal cords is supposed to occur. Secondly, as shown in tab. 6.5,

strong correlations exist between different phonetic features and perceptual measures. In particular, scores for the phonetic feature *[+compact]* reached the best correlation with both severity Sev-DES ($r = 0.84$) and phonemic alteration Phnm-DES ($r = -0.86$). Clearly, *[−continuant]*, *[−sonorant]* and *[+acute]* have also shown noticeable correlations with the different perceptual measures. Simply stated, these correlations, particularly for the *[+compact]* and *[+acute]*, underline once again how the tongue strongly reflects the speech disorder due to HNC. Thanks to those very promising observations, a more detailed analysis of the correlation rates and heatmaps will need to be conducted with clinical experts, taking into account clinical data and individual patient outcomes.

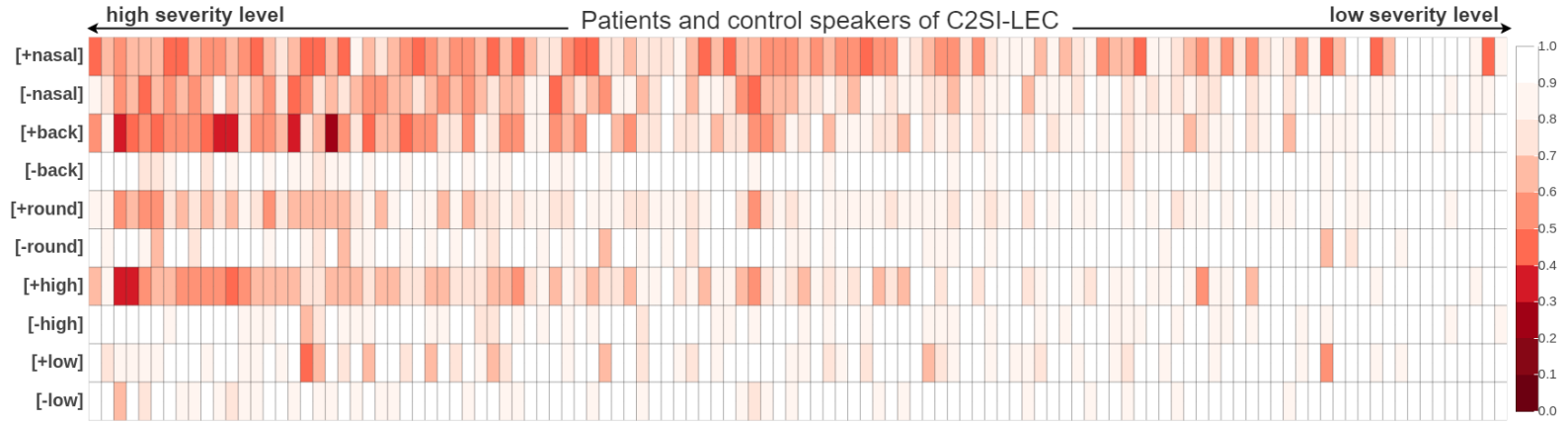


Figure 6.6: *Heatmap showing local ANPS scores per vowel phonetic feature (Y-axis) and C2SI-LEC speakers sorted by Sev-DES (X-axis)*

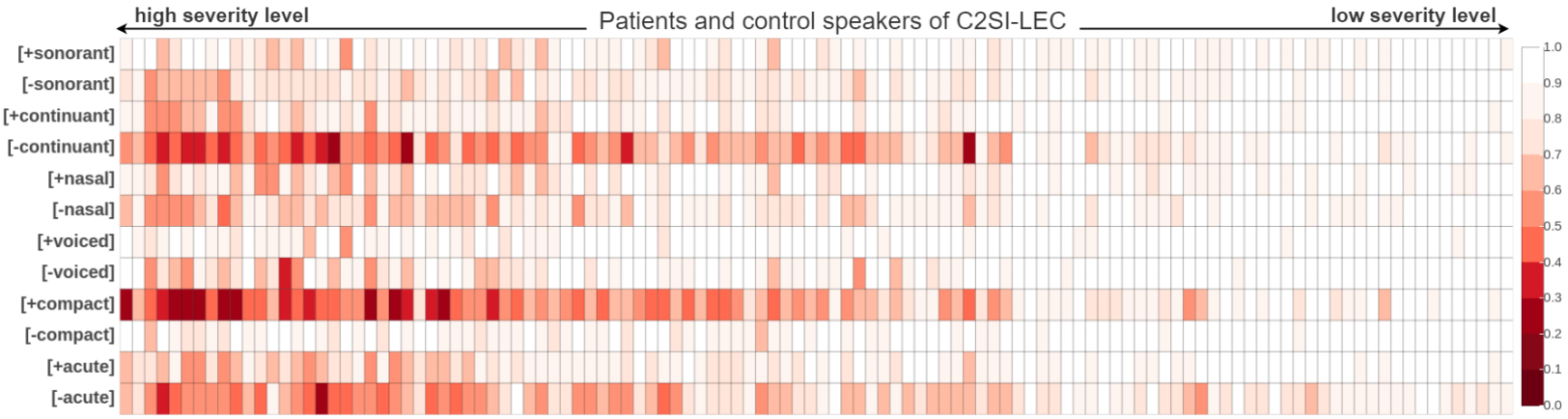


Figure 6.7: *Heatmap showing local ANPS scores per consonant phonetic feature (Y-axis) and C2SI-LEC speakers sorted by Sev-DES (X-axis)*

6.4.2 Dysarthria

Dataset Description

Three types of dysarthria are included in this study. They are all associated with neurodegenerative diseases implying three major neurological systems: the extrapyramidal system with Parkinson's disease, the pyramidal system with Amyotrophic Lateral Sclerosis, and the cerebellar system with Cerebellar ataxia. Patients involved were recruited in different hospitals at

	Parkinson		**CA**		**ALS**	**HC-CCM**
	CCM	**AHN**	**CCM**	**ENT**		
Speakers Metadata						
Region	Paris	Aix	Paris	Marseille	Paris	Paris
#Male/#Female	13/3	10/5	6/5	7/4	14/24	5/5
#Recordings	16	30	11	11	38	10
Age (mean±std)	66.7±8.6		55.7±16.3		65.4±9.4	40±7.3
Age (min.-max.)	48-85		32-86		44-89	32-54
9 items of GEPD scale						
Global Severity	0.82±0.64		1.50±0.56		1.91±0.74	0.16±0.15
Intelligibility	0.56±0.61		0.99±0.48		1.29±0.74	0.05±0.07
Articulation	0.69±0.54		1.36±0.60		1.69±0.72	0.11±0.13
Nasal Resonance	0.31±0.28		0.80±0.42		1.61±0.74	0.18±0.27
Global Regularity	0.79±0.56		1.08±0.44		1.02±0.42	0.13±0.09
Palilalia	0.45±0.43		0.43±0.27		0.17±0.26	0.11±0.14
Reg. of speech rate	0.89±0.64		1.26±0.52		0.82±0.38	0.15±0.16
Melodic fluctuation	-0.89±0.72		-0.50±0.82		-0.88±0.77	-0.21±0.30
Speech rate	0.64±0.92		-0.81±0.94		-1.32±0.93	0.29±0.60
Extra item						
Regional Accent	0.42±0.53	0.73±0.77	0.13±0.16	1.04±0.67	0.07±0.16	0.03±0.09

Items values are reported in terms of mean ± standard deviation

Table 6.3: *Dysarthria and dysphonia data description*

different periods (all patients signed a consent form when required). All of them were recorded on different speech production tasks. In this work, we consider the reading task which includes both the French texts of "Le cordonnier" and "La chèvre de M. Seguin". As a reference, we select an HC group composed of 5 male and 5 female HC speakers issued from the CCM corpus [Fougeron et al., 2010]. A description of metadata of the included patients and the HC speakers is given in the following and summarized in table 6.3.

- **Parkinson's disease (PD):** composed of two sub-groups of patients (min./max. age: 48/85 years; mean/standard deviation: μ=66.7 years/σ=8.6). The first one, composed of 13 male and 3 female patients issued from the CCM corpus described in [Fougeron et al., 2010], was recorded by Dr. Claude Chevry-Muller over 30 years (between 1967 to 1997) in Paris. The second sub-group, composed of 10 male and 5 female patients, referring to the AHN corpus in [Fougeron et al., 2010], was recorded at the Department of Neurology of Aix-en-Provence Hospital (impulsed by Prof. François Viallet). It is worth noting that this second sub-group performed a double task of reading, comprising the same text as the other groups of patients as well as the reading of the French text *"La chèvre de M. Seguin"*.

- **Cerebellar Ataxia (CA):** composed of two sub-groups of patients as well. The first one,

still referring to the CCM corpus, includes 6 male and 5 female patients. The second sub-group was recorded at the Department of Ear, Nose & Throat (ENT) of the Timone Hospital at Marseille (impulsed by Dr. Danielle Robert) and includes 7 male and 4 female patients (min./max. age: 32/86 yrs; μ=55.7 yrs/σ=16.3).

- **Amyotrophic Lateral Sclerosis (ALS):** recorded in the Voice and Speech lab. of the European Hospital Georges Pompidou in Paris by Dr. Lise Crevier Buchman and her colleagues. It includes 14 male and 24 female patients (min./max.: 44/89 yrs; μ=65.4 yrs/σ=9.4).

A perceptual evaluation of speech productions of all patients and HC speakers was performed at the same time by 11 expert judges (10 speech pathologists and 1 neurologist). This evaluation was done according to the 9 items of a French perceptual evaluation scale of dysarthria (GEPD) [Lhoussaine, 2012]. Seven speech dimensions - global dysarthria severity, global speech/voice regularity, speech intelligibility, presence of nasal resonance, palilalia, articulatory accuracy and regularity of the speech rate - were rated on a 4-degree scale (0=normal to 3=severely impaired). The two remaining dimensions: melodic fluctuation and speech rate, were rated on a -3 to 3 scale. That is, these dimensions can also be seen as rated on a 0 to 3 scale with + or – sign to indicate the direction of the abnormal pattern (too fast/slow, hyper/hypo modulated). A final item regarding the presence of a regional accent was evaluated by experts as this may be very significant for some patients.

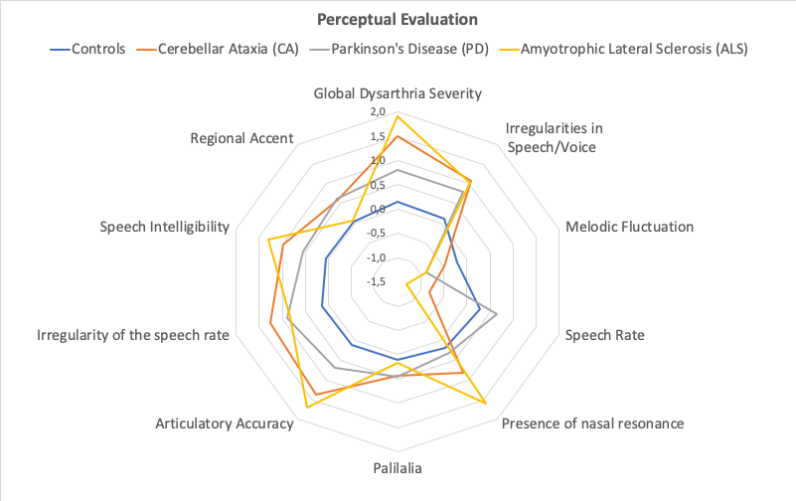


Figure 6.8: *Mean perceptual scores according to 9 GEPD items & the regional accent per dysarthria group and HC speakers.*

The mean scores and standard deviation of the different perceptual items per type of dysarthria, in addition to the HC group, are given in table 6.3. Moreover, a visualization of these mean scores is provided in the spider plot in figure 6.8 in order to give an overview allowing to compare the impact of each dysarthria on the different perceptual items. As reported in [Meunier et al., 2016] on similar patient sub-populations, the ALS group has the most severely rated speech, with the highest scores for global severity, speech intelligibility, and articulation accuracy. It is followed by the CA group, and afterward the PD group, considering the same perceptual items.

The ALS group also presents the largest presence of nasal resonance, which is typically due to the weak or absent closure of velo-pharynx. Although speech rate and speech melody will not be addressed later, abnormal slow speech rate is observed in CA and ALS groups, typical of ataxic or flaccid-spastic dysarthria as opposed to normal or fast observations for the PD group.

Analysis based on global ANPS scores

Similarly to the analysis conducted on patients with HNC in section 6.4.1, in the following we report some analyses to characterize the dysarthric population described above. In table 6.4.1, we summarize the correlation values obtained between *global ANPS* scores per-macro class and different perceptual measures, grouped by type of dysarthria.

Regarding both ALS and PD groups, it is clear that the *global ANPS* scores of these two groups

Table 6.4: *Correlation between global ANPS scores and perceptual measures*

		G. Severity	Intelligibility	Articulation	Nasal Res.
ALS	Vowels	**0.76***	**0.84***	**0.78***	0.73*
	Consonants	**0.81***	**0.88***	**0.83***	0.74*
	Total (VC)	**0.82***	**0.90***	**0.85***	**0.77***
CA	Vowels	0.55*	0.73*	0.56*	0.45*
	Consonants	0.28	0.45*	0.18	0.03
	Total (VC)	0.42*	0.61*	0.36	0.22
PD	Vowels	**0.75***	**0.77***	**0.75***	0.51*
	Consonants	**0.87***	**0.85***	**0.79***	0.61*
	Total (VC)	**0.84***	**0.84***	**0.79***	0.58*

All correlation values are in absolute value (for clarity those ≥ 0.75 are in **bold**).
(*) The correlation coefficient is statistically significant ($P < 0.05$).

strongly correlate with the perceptual measures of global severity, intelligibility, and articulation. The following section will provide a deeper understanding of the observed trends with analysis based on *local ANPS* scores. It is noticeable that the ALS group has a strong correlation with nasal resonance compared to other types of dysarthria. In fact, this is a good indicator reflecting the ability of the global ANPS score to reveal a crucial characteristic of patients with ALS. Indeed, this is consistent with the fact that these patients may have increased nasality in their speech due to the weakness of the muscles that control the airflow through the nose and mouth. Still, this trend has to be confirmed with a more local analysis.

For the CA group, the only noteworthy correlation that can be outlined is the one between the *global ANPS* on vowels and the intelligibility perceptual measure (r=0.73). However, the reasoning behind such an observation cannot be determined at this stage.

Analysis based on local ANPS scores

At this stage, the analysis is carried out for each dysarthria type based on *local ANPS* scores computed for each individual phonetic feature and the concerned patients. As previously reported for HNC, heatmaps are used to visualize *local ANPS* scores of dysarthric patients in

figures 6.10 and 6.11, considering vowel and consonant phonetic features respectively. The X-axis sorts the speakers based on the type of dysarthria, with the least affected appearing on the right and the most affected appearing on the left within each group, using the global perceptual severity measure. The Y-axis represents the specific phonetic feature being analyzed. As a first global observation, we can clearly mention that cells with high opacity are concentrated on the left side of each group, which is consistent with the high global severity level of the corresponding patients. Although ALS, PD and CA are associated with different types of dysarthria, it is worth mentioning that they all show difficulties in consonant production in terms of articulatory alteration or consonant imprecision [Darley et al., 1969b].

- **ALS:** As we revealed before, it is worth noting that the perceptual measure of the nasal resonance is mainly correlated with the *ANPS* scores of patients suffering from ALS when compared to the rest of the pathologies. More specifically, the phonetic features *[-nasal]* for the vowel and consonant macro-classes are among the top two phonetic features with which this measure correlates most strongly, with 0.76 and 0.81 respectively. These correlations, visually clear for the consonant phonetic feature *[−nasal]* in figure 6.11 for ALS, suggest that the ALS patients have a nasal quality voice (i.e. oral phones are badly nasalized). This finding is altogether consistent with the high nasal resonance of the patients, perceived by the experts as reported in table 6.3, and with the well-known hypernasality of mixed dysarthria characterizing ALS patients [Darley et al., 1969b]. Furthermore, regarding the imprecision of consonants, which is one of the characteristics of ALS, it is observed in table 6.5 that *[+compact]* and *[-continuous]* are also strongly correlated with the articulatory measure of ALS, with values equal to 0.85 and 0.75 respectively.

- **PD:** The imprecision of consonants is also an important feature in Parkinson's disease and usually includes distortions due to the reduction of articulatory movements notably. In particular, this imprecision is observable in the table through the strong correlation of the feature *[-continuant]* with related perceptual measures, such as global severity, intelligibility and articulation accuracy. This can be notably explained by the Parkinsonian reduced capacity of completing articulatory occlusion [Ackermann and Ziegler, 1992].

- **CA:** Surprisingly, while the cerebellar patients show rather similar patterns to the other dysarthric groups on the heatmap, no correlation score with the perceptual assessments exceeds 0.7 value. Further patient-by-patient analysis is required here to better understand *ANPS* scores obtained and their consistency with perceptual measures and related dysarthria characteristics.

Table 6.5: *Correlation between local ANPS scores and perceptual measures of dysarthric speakers*

Phonetic features		G. Severity			Intelligibility			Articulation			Nasal Res.		
		ALS	CA	PD	ALS	CA	PD	ALS	CA	PD	ALS	CA	PD
Vowels	**[+nasal]**	0.57*	0.43*	0.49*	0.59*	0.43*	0.48*	0.59*	0.48*	0.45*	0.37*	0.47*	0.21
	[−nasal]	0.50*	0.47*	0.71*	0.60*	0.67*	0.66*	0.51*	0.49*	0.67*	**0.76***	0.41	0.57*
	[+back]	0.74*	0.60*	0.67*	0.74*	0.65*	0.64*	0.73*	0.63*	0.69*	0.64*	0.72*	0.56*
	[−back]	0.07	0.02	0.01	0.19	0.20	0.06	0.12	0.00	0.02	0.13	0.11	0.14
	[+round]	0.59*	0.39	**0.85***	0.67*	0.47*	**0.84***	0.60*	0.29	**0.83***	0.64*	0.40	0.65*
	[−round]	0.09	0.04	0.41*	0.21	0.12	0.48*	0.20	0.04	0.40*	0.17	0.22	0.23
	[+high]	**0.78***	0.27	0.73*	0.67*	0.43*	0.73*	0.68*	0.20	**0.76***	**0.75***	0.00	0.53*
	[−high]	0.15	0.05	0.42*	0.02	0.21	0.43*	0.02	0.13	0.33*	0.25	0.07	0.13
	[+low]	0.30	0.31	0.42*	0.39*	0.45*	0.51*	0.32*	0.41	0.48*	0.20	0.23	0.28
	[−low]	0.31	0.22	0.29	0.32	0.35	0.29	0.24	0.17	0.33*	0.32*	0.17	0.33*
Consonants	**[+sonorant]**	0.62*	0.22	0.49*	0.64*	0.28	0.52*	0.63*	0.15	0.49*	0.52*	0.07	0.35*
	[−sonorant]	0.70*	0.18	0.72*	**0.76***	0.33	0.69*	0.69*	0.06	0.64*	0.63*	0.04	0.50*
	[+continuant]	0.56*	0.18	**0.76***	0.69*	0.16	**0.78***	0.62*	0.41	0.67*	0.49*	0.37	0.54*
	[−continuant]	**0.76***	0.37	**0.82***	**0.78***	0.58*	**0.81***	**0.75***	0.29	**0.76***	**0.75***	0.05	0.63*
	[+nasal]	0.58*	0.05	0.65*	0.64*	0.03	0.64*	0.56*	0.11	0.56*	0.63*	0.22	0.43*
	[−nasal]	0.74*	0.52*	0.72*	0.77*	0.61*	0.70*	0.70*	0.44*	0.73*	**0.81***	0.33	0.63*
	[+voiced]	0.21	0.19	0.23	0.19	0.22	0.25	0.12	0.19	0.23	0.33*	0.11	0.15
	[−voiced]	0.41*	0.06	0.68*	0.54*	0.23	0.59*	0.50*	0.14	0.51*	0.36*	0.10	0.36*
	[+compact]	**0.79***	0.41	0.74*	**0.84***	0.57*	0.74*	**0.85***	0.36	0.71*	0.55*	0.25	0.56*
	[−compact]	0.13	0.06	0.48*	0.19	0.14	0.39*	0.18	0.03	0.38*	0.08	0.10	0.16
	[+acute]	0.54*	0.35	0.67*	0.52*	0.40	0.67*	0.50*	0.20	0.64*	0.45*	0.02	0.43*
	[−acute]	0.47*	0.22	0.58*	0.51*	0.26	0.57*	0.53*	0.30	0.53*	0.42*	0.39	0.50*

All correlation values are in absolute value (For the sake of clarity, those ≥ 0.75 are in **bold**).

(*) The correlation coefficient is statistically significant (P< 0.05).

6.4.3 Dysphonia

Dataset Description

The corpus of dysphonic voice disorders was recorded in the 2000s at the department of ENT of the Timone Hospital at Marseille. It is composed of 80 records of female voices, including 20 control subjects and 60 dysphonic patients, aged from 17 to 50 years (average 32.2 years) [Pouchoulin et al., 2007]. The set of dysphonic patients underwent a laryngoscopic examination showing dysphonia essentially of functional origin mainly due to nodules, oedemas, polyps, and cysts (gathering 53 patients among 60). All patients were recorded on a reading task of the French text "La chèvre de M. Seguin". The 80 female speakers were selected among a larger corpus in order to be equally distributed into the 4 levels of the Global item of the GRBAS scale [Hirano, 1981]: 20 normal/control voices (i.e. grade G0), 20 voices with mild dysphonia (i.e. grade G1), 20 voices with moderate dysphonia (i.e. grade G2), and 20 voices with severe dysphonia, but still intelligible (i.e. grade G3). The GRBAS-based assessment of the larger corpus was performed by a panel of three clinical experts following a consensus decision.

Analysis based on global ANPS scores

Since dysphonic patients are evaluated on the discrete Global item of the GRBAS scale, we perform a boxplot visualization of the *global ANPS* scores instead of correlation analysis. Figure 6.9 illustrates the *global ANPS* scores (Y-axis) for the vowel and consonant macro-classes grouped by level of the Global item (X-axis). Clearly, the *global ANPS* score does not reflect the different grades of dysphonia. Indeed, we can see that almost all speakers (78/80) have a *global ANPS* score above 0.8. Indeed, these results were not unexpected since dysphonia is a voice disorder, not a speech disorder. In fact, a voice disorder affects the quality of the voice, while a speech disorder affects the ability to produce speech sounds. We remind that the *global ANPS* is calculated as the mean of *local ANPS* scores, thus, it reflects the phonetic features production which majority is related to the place and manner of articulation.

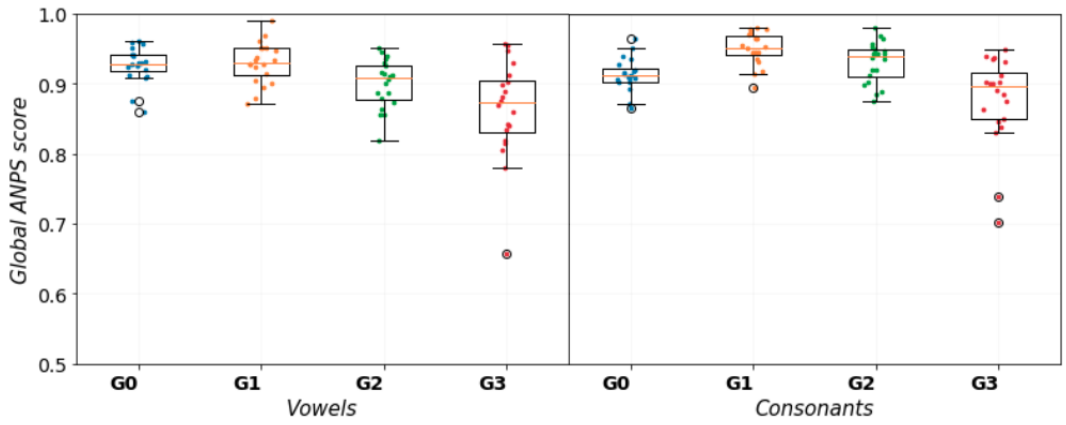


Figure 6.9: *Boxplot of the global ANPS scores per macro-class grouped by dysphonia grade.*

Similarly to what has been done before, a more detailed analysis based on *local ANPS* score is taking place in the next section. We expect to gain a better view with this locality and see the

impact of dysphonia severity on the phonetic features reflecting the voice quality characteristics rather than on those reflecting the articulatory characteristics.

Analysis based on local ANPS scores

As previously mentioned, dysphonia is a voice disorder. We can clearly notice in figure 6.11 that almost none of the phonetic features related to the place or manner of articulation is significantly impaired, as this would be the case of the most affected patients having speech disorders. We would have expected the phonetic feature *[+voiced]* to be affected, almost within G2 and G3 patients. However, three patients only exhibit very low scores (less than 0.5) for that feature. When checking the corresponding recordings, it turns out that two of these patients have the most severe voice disorders (compared with other patients rated G3). Indeed, they are characterized by a weak and whispered speech as well as some large difficulties to produce speech (vocal fatigue).

6.4.4 CCM HC speakers

Since we now have a new set of healthy control (HC) speakers from the CCM dataset (above described in table 6.3), we can compare the results of the ANPS scoring approach on these speakers to that on the HC speakers from the C2SI corpus. It is worth recalling that particular observations were made regarding the *local ANPS* scores related to the vowel nasality for the HC speakers of C2SI corpus and were explained by the southwestern accent they exhibit. If our hypothesis is correct, we would expect the scoring approach to behave differently for the CCM group and not show any degradation on the vowel nasality phonetic feature since the CCM speakers are recruited from the region of Paris. Actually, this is confirmed by figure 6.10 were we can clearly see that HC CCM patients don't show any degradation on the vowel phonetic feature related to the nasality. Thus we can once more confirm our assumption relating the sensitivity of this phonetic feature to the regional accent.

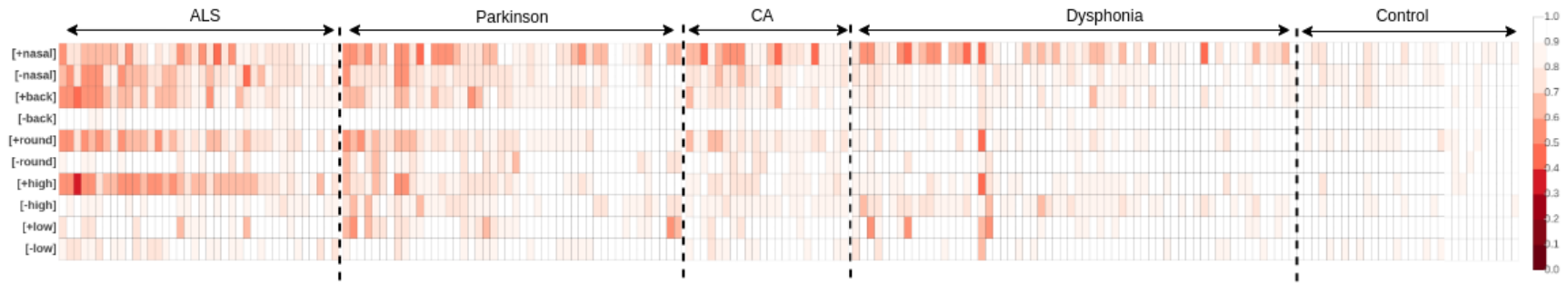


Figure 6.10: *Heatmap showing local ANPS scores per vowel phonetic feature (Y-axis) and patients grouped by pathology and sorted by Global Severity within each group (X-axis), in addition to control speakers*

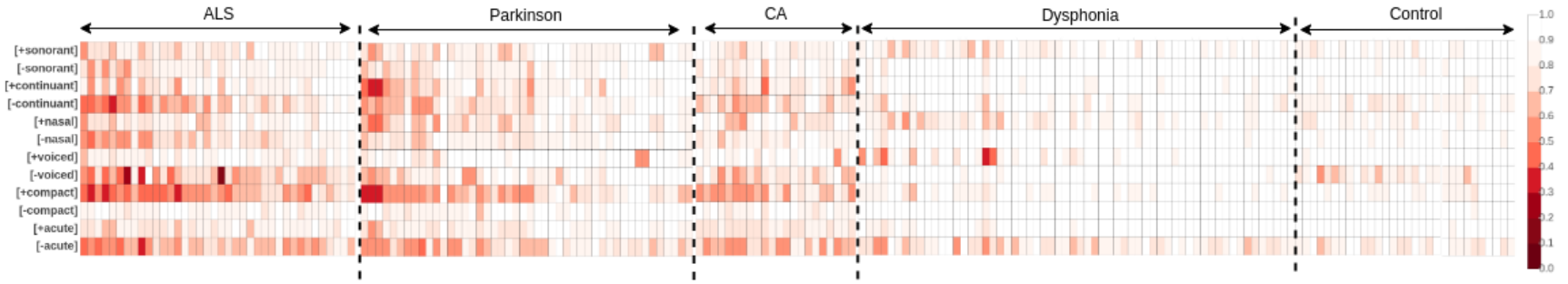


Figure 6.11: *Heatmap showing ANPS scores per consonant phonetic feature (Y-axis) and patients grouped by pathology and sorted by Global Severity within each group (X-axis), in addition to control speakers*

6.5 Impact of diverse factors on ANPS score

In this section, we explore the impact of several factors on the ANPS scoring approach. We start by investigating the effect of linguistic content, a factor that is unrelated to speech pathology. Then we examine if the ANPS score can reflect the tumor size of HNC patients.

6.5.1 Variability of linguistic content

So far, we have shown how our method can effectively reveal some specific characteristics of the pathology, corresponding to the type of speech disorder observed. At this stage, we aim to study the robustness of ANPS scoring approach to the variability of linguistic content. To this end, we observe the *local ANPS* scores of the 15 patients of PD-AHN corpus who were involved in a double reading task on two different texts. Figure 6.12 depicts these various ANPS scores for PD-AHN patients. The heatmap on the left shows the scores for consonant phonetic features, while the one on the right shows the scores for vowel phonetic features. Each column in the heatmap represents the result of one text reading, with the first two successive columns corresponding to the two reading tasks of one patient, and so on.

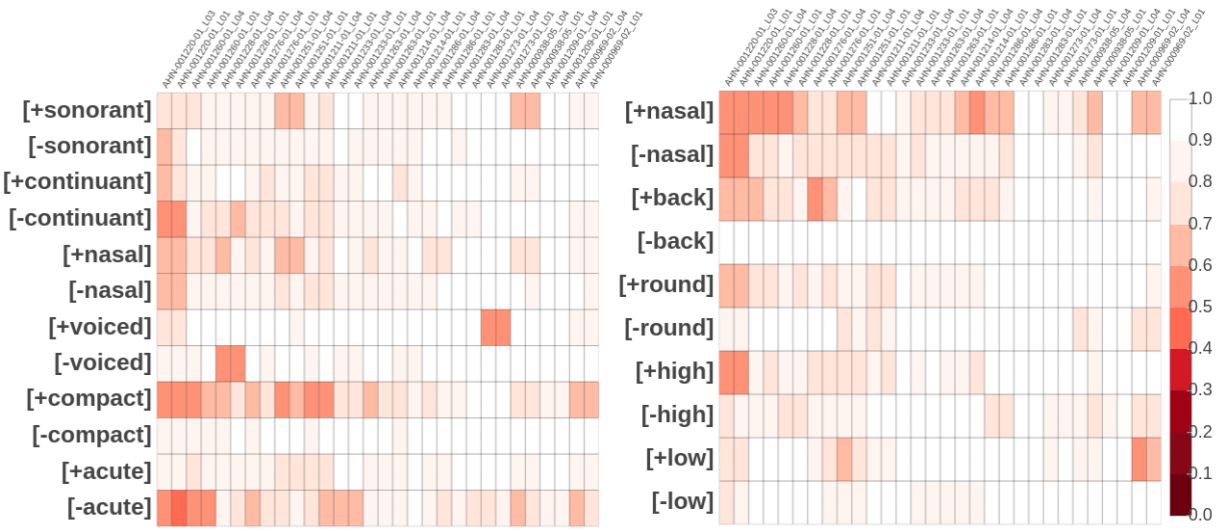


Figure 6.12: *Heatmap showing local ANPS scores of PD patients (Y-axis) involved in a double reading task: two successive columns are the two reading tasks of the same patient.*

.

Let us consider the set of independent paired *local ANPS* scores:

$$(ANPS\,{}^{ch}_{i,t}, ANPS\,{}^{co}_{i,t}),\ \forall t, i = 1, .., 15$$

where $ANPS\,{}^{ch}_{i,t}$ and $ANPS\,{}^{co}_{i,t}$ refer to the local ANPS scores of the phonetic feature t produced by the i^{th} patient on *"La chèvre de M. Seguin"* and *"Le cordonnier"* texts, respectively. In order to analyze the impact of the variability of linguistic content on the *ANPS* scores, we calculate the absolute difference between the two ANPS scores of the same couple. Figure 6.13 depicts the range of the obtained absolute differences (Y-axis) per phonetic feature (X-axis), in the form of

a boxplot. That is to say, each boxplot displays the distribution of fifteen-point data and reflects the impact of linguistic content variability on the local ANPS scores of a particular phonetic feature. Let's recall that the value range of the ANPS score is [0;1], which is also the range for the absolute difference between two ANPS scores. We can see from figure 6.13 that the overall maximum difference is around 0.15, where for most of the phonetic features the maximum does not exceed 0.1. So far, while interpreting the heatmaps of the local ANPS scores of patients

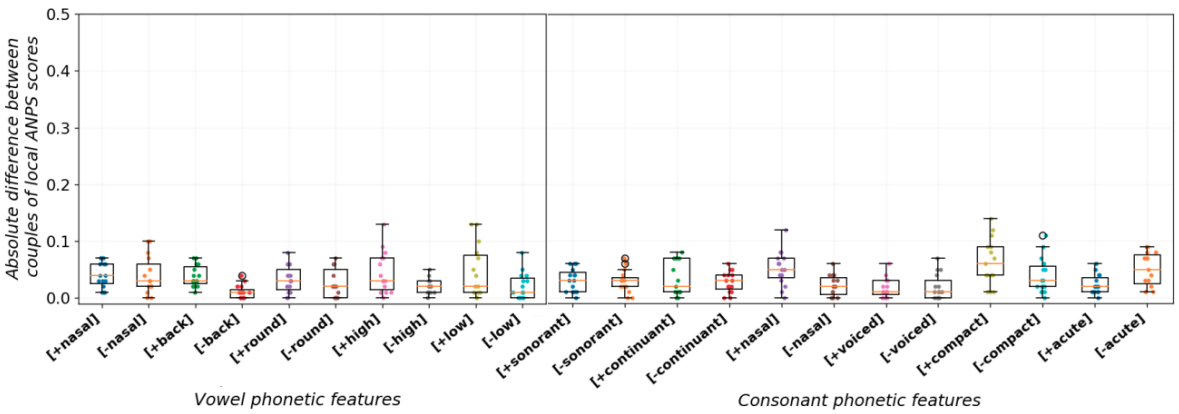


Figure 6.13: *Visualizing the impact of linguistic content variability on local ANPS scores: Boxplot per phonetic feature displaying the absolute difference between a couple of local ANPS scores obtained for each speaker as the result of a double reading task.*

with different pathologies, we never took into account slight changes in the score values. In fact, all the interpretations were based on a strong contrast in color code reflecting the ANPS scores. Now if we consider the value of 0.15 observed above, this value can be translated to a maximum difference of two gradients in the sequential color scale reflecting ANPS values. Therefore, in most cases, the absolute difference results in a zero to one difference of gradient in the color scale. To conclude, when evaluating and comparing the ANPS scoring approach on the two texts *"La chèvre de M. Seguin"* and *"Le Cordonnier"*, the method has maintained its performance and demonstrated an ability to produce consistent results despite variations in linguistic content.

6.5.2 Tumor size factor

In this section, we hypothesize that a tumor size effect exists and can be reflected by the set of interpretable neurons and associated global scores. Under this assumption, both groups of patients, with small tumors (T1+T2) and large tumors (T3+T4) as described in section 4.2.2, are considered and compared with the HC group. Figure 6.14 illustrates the boxplot of the *global ANPS* score ranges for the three groups of speakers. From HC speakers to HNC patients with big tumor sizes passing by patients with small tumor sizes, a significant decrease in the range of the score is clearly visible. Yet, the decreasing trend is more marked in the macro-class of consonants compared to vowels. Similarly, a Student t-test involving both distributions of *global ANPS* scores belonging to control speakers and patients with (T1+T2) tumor size as well as those between patients with (T1+T2) and (T3+T4) tumor sizes, is calculated. It is consequently found that, either for vowels or for consonants, the difference between the two sets of *global ANPS* values for HC speakers and patients with small tumor size (T1+T2) is statistically significant

with ($p < 0.001$). Similarly, a statistically significant difference between the scores of patients with a (T1+T2) tumor and those with a (T3+T4) tumor is demonstrated. Even though there is a significant difference between the distributions of *global ANPS* scores of HNC patients with small and big tumor sizes, we can see that the interquartile and whiskers ranges do not strongly reflect this difference. Indeed, this suggests that the two groups have a similar spread of values and that the difference is primarily driven by the extreme values leading to the difference in means. To conclude, the difference in *global ANPS* scores between HNC patients with small and big tumor sizes may still be statistically significant, however, it's important to consider the practical significance and whether it has any meaningful impact in our use case.

6.6 Discussion

In this chapter, we present the second step of our proposed methodology in which we built on the phoneme classifier proposed in the first step. The principle goal is to investigate the model capacity to yield relevant knowledge related to the characteristics of speech pathology. Figure 6.15 summarizes the steps seen thus far. This second step can be seen as composed of two sub-steps. First, we started with an explainability sub-step in which we explored the hidden representations of the phoneme classifier. To this end, we were based on a general analytic framework, *Neuro-based Concept Detector (NCD)* that we proposed for this explainability stage. The application of NCD revealed the emergence of the concept of phonetic features in the deep representations of the CNN-based phoneme classifier. This sub-step gave rise to an interpretable dimension which is the detectors of phonetic features. Consequently, the second sub-step aims to take advantage of this outcome and to provide interpretations in the context of speech pathology in terms of phonetic feature alteration. For that purpose, we proposed a scoring approach *Artificial Neuron-based Phonological Similarity (ANPS)* dedicated to the assessment of phonetic feature productions performed by a speaker, based on the corresponding detectors. Through this score, we have shown that the perceived decrease in speech quality is very well conserved in the results of the global ANPS scores. Step by step we outlined via local ANPS scores how alterations in phonetic features reflect the characteristics of each speech pathology.

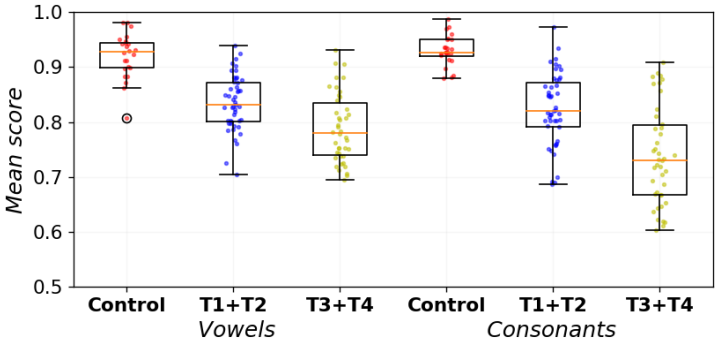


Figure 6.14: *Boxplot of the global ANPS score ranges for control speaker and patients per tumor size*

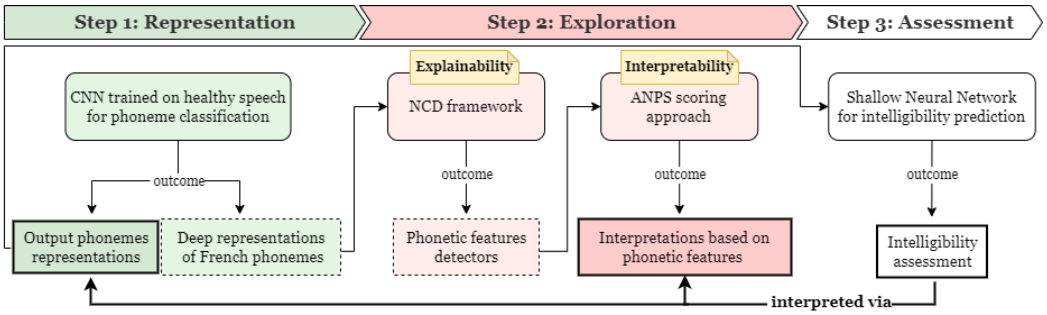


Figure 6.15: *A step forward in the achievement of the proposed methodology: the accomplishment of step 2*

6.6.1 Advantages of the proposed approach

Our proposed methodology holds significant advantages, which we summarize below:

1. **Is not pathology-dependent:** The core idea behind the design of our approach is that it is trained only with speech samples from healthy speakers, and thus, not restricted to a specific type of speech disorder. In this way, we guarantee that we are not acquiring patterns specific to a particular pathology, thereby increasing the generalization ability of the interpretability approach across different types of disordered speech.

2. **Support the varying progression of the disease among patients:** Our proposed approach is not a one-size-fits-all approach. In other words, it is not established to provide interpretations only for a large population (e.g. HNC patients have an issue pronouncing compact phonemes). In such a situation, important details about a particular patient's speech disorder can be easily missed, leading to inaccurate or incomplete assessments. Indeed, each individual experiences different symptoms and the development of those symptoms can differ greatly. The interpretability approach we propose is designed to reflect the varying progression of the disorder among patients. This will allow healthcare providers to make informed decisions adapted to each patient's specific needs (e.g. personalized speech rehabilitation). Since the approach is able to provide interpretations for a particular patient, it would be even very interesting to explore the approach capacity to track the changes in speech over time, within a longitudinal study.

3. **Robustness of the ANPS score to unbalanced data and outliers:** ANPS score is designed to be robust when dealing with highly unbalanced data, such as recordings from speech pathology datasets. This is due to the fact that it uses the median calculation per phoneme distribution, which is less sensitive to outliers than the mean calculation. Indeed, in highly unbalanced data, we can face a very small sample size in some phoneme distributions. In such a case, it may be more appropriate to use the median as a measure of central tendency since outliers may have a significant impact on the mean. A possible source of outliers can simply be related to errors that occur from the automatic phoneme alignment.

4. **Adaptability of the ANPS score to the absence of phonemes:** Even more, the computation of local ANPS scores of phonetic features relies on phoneme-level information. To

handle the fact that a speech production may not contain all phonemes, only the present phonemes are automatically taken into account. The scoring approach is therefore able to calculate a score assessing the production of a phonetic feature even when not all phonemes of this phonetic feature are present in the speaker's production.

6.6.2 Limits and self-criticism

Certainly, our proposed approach provided satisfactory results thus far. Yet, from a critical perspective, it cannot handle many aspects which makes several future work directions possible to ameliorate it. In the following, we emphasize some of these shortcomings:

1. **Data-driven explainability approach:** One drawback of the proposed explainability approach is the fact that it relies on a specific dataset. While we paid attention to the quality and representativeness of the data, we certainly did not cover all possibilities (e.g. all the possible phonetic contexts, articulation, neutralization, archiphoneme, etc.). Consequently, we might be missing some important information since in the final, we are characterizing the model behavior on that limited data distribution. Therefore, the explainability approach may not provide a complete understanding of the inner representations of the model.

2. **Non-generalization on accents:** One weakness in our approach as a whole comes from the training data choice in step 1, which resulted in learning strong accent-dependent patterns. To improve generalization it would be useful to train the model on different accents. Even though we are not concerned with language variability in our specific context, it would be worth training our model with datasets from different languages. As a consequence of such a choice, we assume that final interpretations will reflect the characteristics of speech pathology, regardless of the language since the model did not learn language-dependent patterns.

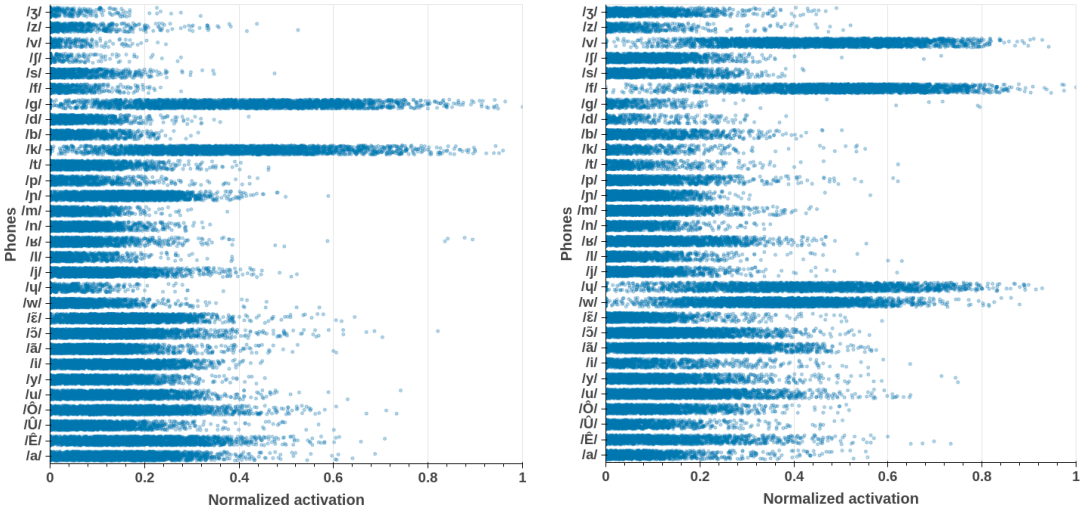
3. **The text-constrained phoneme alignment limits the scalability of the interpretability approach:** The proposed interpretability approach needs a forced text-constrained phoneme alignment. In other words, it needs the true phoneme label that had to be pronounced by a patient based on an orthographic transcription of the reference text, all with its start and end boundaries in the speech recording. Indeed, on the one hand, orthographic transcription is performed by human experts which means that it is time-consuming and expensive, especially if we consider larger pathology datasets. On the other hand, that limits the application of our approach to generating interpretations exclusively on reading tasks.

4. **Not taking advantage of the full interpretable neurons:** Even though the capacity of our explainability approach can go beyond identifying detectors of phonetic features, we did not extend our work to other concepts. Indeed, we are aware that among the non-identified neurons by the *NCD* approach, there are other neurons that are interpretable too and can later provide relevant interpretations in the speech disorders context. For instance, in figure 6.16a we show that neuron 384 of the FC2 has a distinctive response for the phonemic class of stop velars (/k/ and /g/). That is, this neuron encodes a more specific representation of phoneme classes combining both the manner and place of articulation.

Another example is shown in figure 6.16b, where neuron 8 of the FC2 has a distinctive response for the labio-dental fricatives (/f/ and /v/) together with the labio-palatal (/ɥ/) and labio-velar (/w/) approximants.

5. **The color scale in heatmaps:** The color scale is very important when it comes to data visualization. In heatmap visualization, we used a sequential scale of a single hue (red), from the least to the most opaque shades, representing low to high *local ANPS* values. We are aware that, visual distortion can be introduced by the color map choice, and consequently, the conclusions drawn from these visualizations. Indeed, one of the most important criteria of the used color palette is its perceptual uniformity [Crameri et al., 2020]. A perceptually uniform color map weights the same data variation equally all across the dataspace, while a non-uniform color map interprets some small data variations to be more important than others. Such a color-introduced bias can result in misleading interpretations. In our case, the problem does not really arise since we did not interpret the slight variations of intermediate colors, but attention should be paid once we would like to do so in a clinical perspective.

6. **The need for clinical expertise to validate ANPS scores:** While our approach may be effective at identifying certain characteristics of speech disorders, it is important to validate its accuracy and reliability in a clinical context. To properly validate the ANPS scores (implicitly the overall proposed methodology), it is necessary for clinical experts to confront the ANPS scores with the patient's clinical data and his/her speech disorders (by listening to his/her different recordings). By doing this, we can validate the consistency of the ANPS scores and whether they can be used as an effective and interpretable tool for assisting clinicians in diagnosing and treating speech disorders.



(a) Neuron 384 of layer FC2 **(b)** Neuron 8 of layer FC2

Figure 6.16: *Jitter plot visualizing the normalized activations on BREF-Int dataset*

Chapter 7

Step 3: Intelligibility Prediction

Contents

7.1 Specific Context

Throughout the preceding chapters, we have laid the groundwork for an original approach to evaluating speech disorders that not only focuses on the objective of assessment but also provides a deeper understanding of the final assessment. We implemented an intermediate task that operates at the phoneme level, enabling us to achieve a finer granularity of information and insight into the scoring process. Furthermore, we were able to incorporate an additional level of granularity by considering phonetic features, which interpretation is of great practical relevance in the clinical phonetics context. The aim of this third and last step is therefore to implement the **target task**, which is the prediction of the final score assessing the speech intelligibility of an individual and interpreting it based on the outcome of previous steps. The rest of this chapter is organized as follows. For the remainder of this section, we reference a limited number of related

works and emphasize the research questions that we aim to address in this step. The following section 7.2 gives an overview of the process set up to achieve the final score prediction. Section 7.3 is dedicated to the experimental setup presentation, including the dataset, architecture, and training details. At the conclusion of this chapter, we provide a dedicated section 7.5 that illustrates a case study. This example demonstrates how our proposed methodology can be utilized as a comprehensive end-to-end solution for the objective assessment and interpretation of speech disorders in a clinical setting.

7.1.1 Related work

Although the advances in pathological speech assessment using DL architectures, only a few studies addressed this subject from a DL interpretability point of view. In this context, we can find a research work that was conducted with a focus on dysarthric speech by Tu Ming et al. [Tu et al., 2017]. The authors trained a model to predict the severity of dysarthric speech from the input signal. On the other hand, they took steps to make the model interpretable by incorporating a specific bottleneck layer. They used transfer learning to learn both clinically-interpretable labels (perceived by speech-language pathologists such as vocal quality and articulatory precision) and the final severity score. The result is a model that not only improved the accuracy of dysarthria assessment but also provided justifications for its predictions by exhibiting high correlations with the interpretable bottleneck features. An extension of this work was recently proposed by [Xu et al., 2023]. Instead of relying on perceptual labels provided by speech-language pathologists, the authors of this work trained the interpretable layer to learn four acoustic features that characterize different aspects of dysarthria (articulatory precision, consonant-vowel transition precision, hypernasality, and vocal quality). Authors extracted these acoustic features from the speech samples they have in possession. They also applied SHAP [Lundberg and Lee, 2017] as an explanation tool to further analyze the contribution of each acoustic feature in the interpretable layer to the final prediction. Very close to Tu Ming et al., authors in [Korzekwa et al., 2019] proposed a DL model for the detection and reconstruction of dysarthric speech. Their model not only provides interpretable characteristics of dysarthria but also try to reconstruct healthy speech which is their main contribution.

Although these works address one major requirement of DL in a clinical application which is DL interpretability, their methodology based on the incorporation of a bottleneck layer raises the need for a large dataset of speech pathology. Indeed, they need a large amount of data as they train their DL-based models from scratch using a dysarthric dataset. For Tu Ming et al. [Tu et al., 2017], this data requirement is even more important since they need extra labels, in addition to the severity score, for the training of the bottleneck layer. In addition, if we consider that these intermediate labels could be subjective since they are provided by humans (SLPs), this leads to the incorporation of a subjectivity characteristic in the interpretability of the final score.

The design of our proposed methodology sheds light on these issues. Indeed, one of our main contributions is the fact that it addresses the issue of data requirements. Indeed, collecting a significant amount of data, especially for pathological speech, can be a difficult and expensive task, making this factor a crucial aspect to take into account. That is why our starting point is a DL-based model trained on healthy speech that encodes the characteristics of "normal" reference. Moreover, in our case, the interpretable dimension emerges automatically. This dimension serves later to interpret the final assessment of patients. As a result, interpretability can be achieved without the need for additional labels or data, and without introducing any possible

subjective factors.

7.1.2 Research Questions

This third step is carried out with these three main research questions:

- **RQ1:** Can we predict a global score assessing the speech production of patients with speech disorders, based on the outcome of the CNN performing the intermediate task of phoneme classification?
- **RQ2:** Which score to adopt as an intelligibility ground-truth for the regression model, to better reach an objective intelligibility assessment?
- **RQ3:** How to link this final assessment to the outcome of the interpretable dimensions? That is, in a clinical context how the overall proposed methodology can be used?

7.2 An overview of the process of score prediction

In this section, we address the first research question (RQ1). Our aim is to investigate whether the outcome of the CNN-based phoneme classifier can be used to predict a global score that assesses the speech production of patients with speech disorders. To this end, we provide the following overview of the process we propose for predicting this final score, from data preparation to the regression task.

7.2.1 Preparation of input for the task of score prediction

In this section, we present two different approaches to data preparation for the score prediction task, which we illustrate in figures 7.1 and 7.2, respectively. Basically, we take the speech productions from every speaker and apply data preprocessing to make it compatible with the input of the CNN that we previously trained for phoneme classification (see section 5.2.1 of the chapter dedicated to step 1 for further details). The outcome of this data preprocessing stage is a set of acoustic feature matrices at the frame level, that we refer to as CNN input samples. Next, we consider these CNN input samples by blocks of 100 consecutive samples which reflect almost one second of speech produced by a particular speaker. Each of these blocks is then fed to the trained CNN. The choice of one-second segments leading to blocks of 100 consecutive samples has been driven by two main reasons. The first reason relies on the necessity of sufficient data for the intelligibility score prediction process, regarding the speech disorder corpora available in our context, Indeed, we cannot consider the set of overall speech recordings available per patient, but smaller speech segments to augment the processed data. Secondly, we consider that the duration of one second for speech segments can carry sufficient and relevant information related to speech disorder for intelligibility score prediction. The described stages, so far, are shared by the two approaches of data preparation. Now, in the first approach, detailed in figure 7.1, we consider the CNN output vectors. That is to say, for one block of 100 input samples, we obtain 100 output vectors with a dimension of 32. Let us recall that the CNN output is a softmax layer. That is, the output of the CNN for a given input sample is a vector with 32 dimensions, which represents the *probability* that the input belongs to each of the 32 final classes (associated with the 31 French phonemes and silence). If we go more in detail, this softmax layer takes a

vector of real values (the outputs generated by the last linear layer) and normalizes them into a probability distribution over the classes. These raw values before normalization are called logits. In our case, we consider each of these blocks of 100 **logit vectors** as one input sample to the next model responsible for the prediction of the final scores.

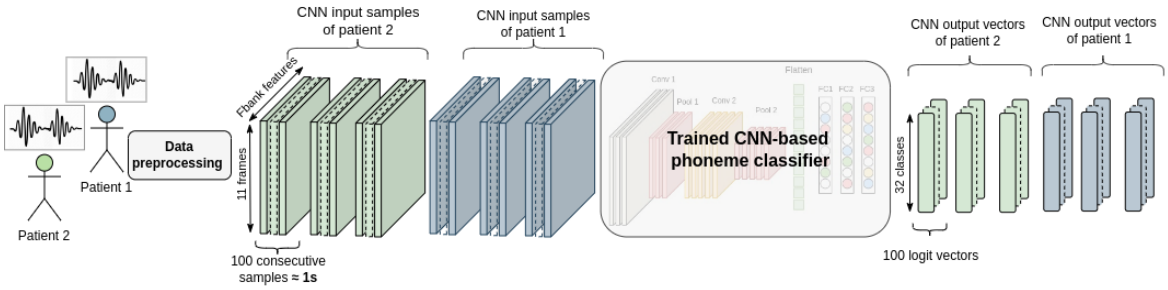


Figure 7.1: *Preparation of the input for score prediction: Logit vectors*

Regarding the second approach to data preparation, illustrated in figure 7.2, we select the set of interpretable neurons across the different fully-connected layers of the CNN. We have already identified these neurons as phonetic feature detectors in step 2 (see section 6.2.4 in chapter 6 for further details). In total, we have 985 interpretable neurons. Now, as aforementioned, we fed the blocks of 100 input samples to the CNN. At this stage, instead of getting the CNN output vectors as done in the first approach, we retrieve the activations of the selected set of interpretable neurons and concatenate them into embedding vectors with a dimension of 985. Here, a single embedding vector matches a single input sample. That is to say, for one block of 100 input samples reflecting almost one second of speech, we obtain 100 embedding vectors. We refer to these resulting embedding vectors as **phonetic feature embeddings** as they represent the input speech signal in terms of phonetic features. A block of 100 phonetic feature embeddings is considered later as one input sample to the next model responsible for the prediction of the final score.

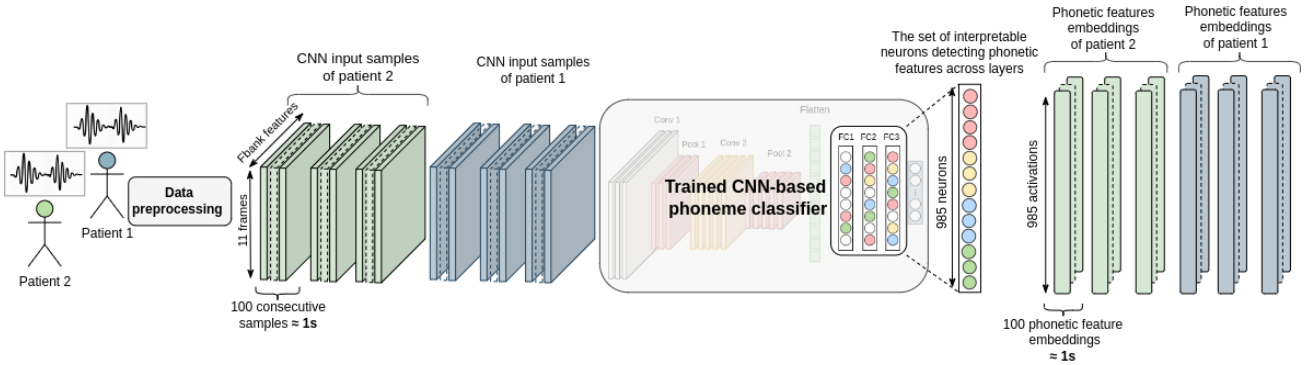


Figure 7.2: *Preparation of the input for score prediction: Phonetic feature embeddings*

7.2.2 The process of score prediction

Building on the previous step, we use the blocks of 100 logit vectors/phonetic feature embeddings, generated for each speaker, as input to a shallow neural network whose aim is score prediction. As detailed in figure 7.3, this shallow neural network generates a score prediction for each block of 100 vectors. In other words, for each speaker, we have an assessment of his/her speech production for almost each second. It is worth noting that it is possible to obtain an overall score for an utterance or a speaker. For instance, to get an utterance-level score, we average the scores generated for each second of the utterance. Similarly, a global score for a given speaker is the result of averaging all the scores generated for each second across all utterances produced by that speaker.

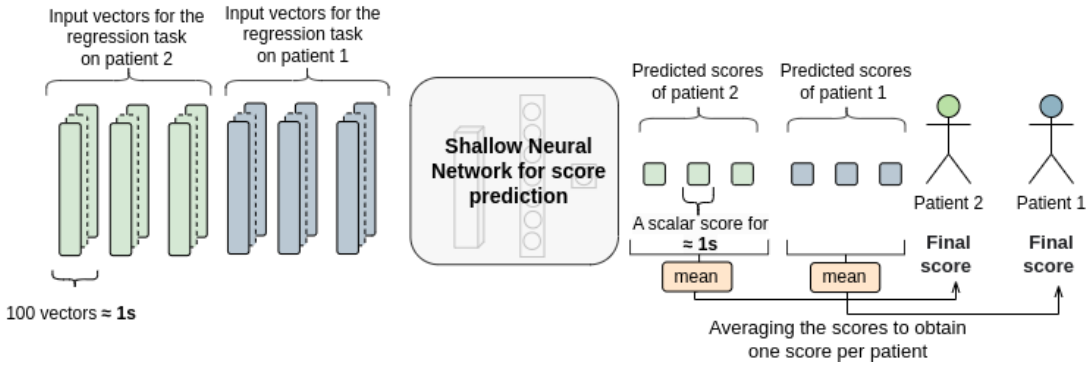


Figure 7.3: *The process of score prediction*

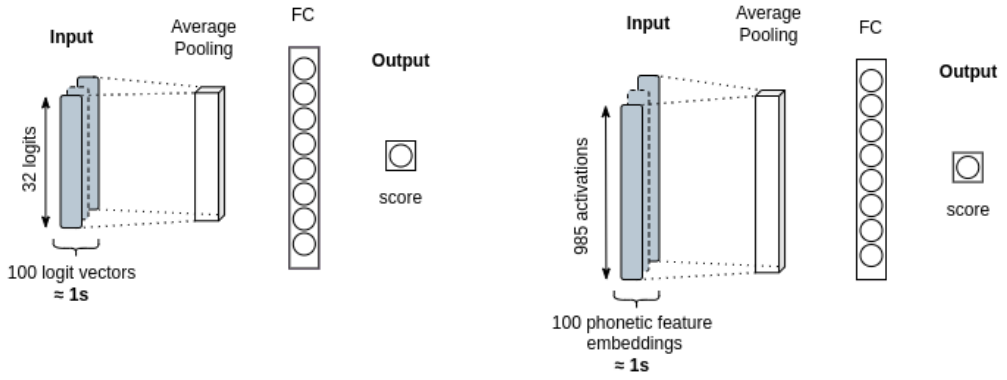
7.3 Experimental setup

In this section, we present the experimental setup including the details related to the model architecture established for the score prediction, the used datasets, and the training process.

7.3.1 Architecture of the Shallow Neural Network

The proposed model for the automatic prediction of a score assessing a given speaker's production is a shallow neural network (SNN). The structure of this model is shown in figure 7.4. To begin with, the first layer is an average pooling layer. In case we consider the logit vectors as an input to the regression task, this pooling layer takes 100 vectors each composed of 32 logits and converts them to a 32-dimensional vector (see figure 7.4a). Now, if we consider the phonetic feature embeddings as an input to the regression task, this pooling layer takes 100 vectors each composed of 985 activation values and converts them to a 985-dimensional vector (see figure 7.4b). In both cases, this transformation can be considered as passing from a frame-level representation to a segment-level representation (one-second segment). This is then fed to one fully connected layer with a ReLU activation function. The number of neurons within this layer is a hyper-parameter that we tune and fix later based on the task and input in question. However, it is worth noting that the number of neurons is constrained by the dimension of input samples depending on the type of information involved. Regarding the dimension of the logit vectors (32),

we will consider configurations limited to 16 or 32 neurons only for the unique layer. Based on the phonetic feature embeddings, which dimension is 985, larger numbers of neurons could be studied, from 64 to 256. Finally, the output layer corresponds to the final score (i.e. the assessment of the one-second input segment). To ensure that the predicted score is between 0 and 10, a bounded activation function should take place in the last regression layer of the shallow neural network. We use *sigmoid* activation function, which maps any input to a value between 0 and 1, and then we scale the output of this function to map it to the range [0, 10]. The reason for this requirement is that the perceptual measures in our possession, which will serve as the true scores for training the regression model, are evaluated within a range of 0 to 10 (see section 4.2.2 for the review of corpora used).



(a) Regression on logit vectors

(b) Regression on phonetic feature embeddings

Figure 7.4: *The structure of the Shallow Neural Network for the final score prediction*

7.3.2 Datasets and Training details

As aforementioned, the regression model is trained to predict a score assessing the quality of a one-second segment of speech represented by an input sample. It is worth recalling that our main goal is to establish a DL-based tool for the objective assessment of the speech intelligibility of HNC patients. To this end, we need an objective intelligibility score that serves as ground truth to train our model. At this stage, we raise the second research question (RQ2); *which score to adopt as an intelligibility ground-truth for the regression model, to better reach an objective intelligibility assessment?*

All along the previous chapters, we outlined the fact that the perceptual intelligibility measures, particularly in the C2SI corpus, are subjective. More specifically, we reported that intelligibility on the reading task (Intel-LEC) is even more subjective when compared to the intelligibility on the image description (Intel-DES) (i.e. an overestimation of the intelligibility due to the predictability of the text read). While the PPD-DAP is considered a more objective intelligibility measure, as previously explained, the patients in SpeeCOmco corpus were not subject to this assessment unfortunately. That is, we are unable to use this measure and answer RQ2 due to a lack of resources. In addition, we conduct another set of experiences where we train the regression model to learn the target score of severity on image description (Sev-DES). Some details about the target score distributions (i.e. Intel-DES and Sev-DES), within the training and validation

sets, are summarized in table 7.1.

The mean squared error (MSE) is taken as loss function for the score regression task. The mathematical expression of this loss is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7.1}$$

where y_i and $\hat{y}_i$ are the true and the predicted scores of the i^{th} input sample, respectively. n is the total number of samples. This function measures the average of the squared differences between the true and predicted scores. As well, we use another metric which is the mean absolute error (MAE). The mathematical expression of the MAE loss function is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7.2}$$

MAE measures the average absolute difference between the true and predicted scores. In fact, it is a commonly used metric for evaluating the performance of a model, as it provides an easily interpretable error value. The reason for this is that the error value is expressed in the same units as the target variable being predicted. It is worth mentioning that MAE and MSE measure different aspects of the errors in the model predictions. The key difference between these two metrics is that squared error penalizes larger errors to a greater extent than absolute error. In the rest of this chapter, and for the sake of simplicity, we refer to an error (i.e. the difference between a true and a predicted value) in a regression analysis as a residual.

$$Residual\ _i = y_i - \hat{y}_i \tag{7.3}$$

The proposed SNN model is trained to map an input of 100 vectors at the frame level (i.e. logit vectors or phonetic feature embeddings) to a particular score of interest. To this end, we use the datasets C2SI-SVT, C2SI-FOC, C2SI-MOD, and C2SI-SYN as input for the training process. A collection of one-second segments is issued from the different speakers' productions in these datasets (i.e. patients and HC speakers) and then, as described in 7.2.1, prepared to be an input to the regression model. We further use the dataset C2SI-LEC as a validation set to monitor the training and tune the experimental settings. As regards the test, we use SpeeCOmco dataset to evaluate the resulting model. Still in table 7.1, we report some details about the input samples to the regression task for the train, validation and test.

	Training	**Validation**	**Testing**
	(C2SI)		
Dataset	SYN & MOD & SVT & FOC	LEC	SpeeCOmco
#speakers	105	114	27
#input samples	25637	3542	867
Intel-DES (mean±std)	7.9±2.5	7.9±2.5	6.7±2.6
Sev-DES (mean±std)	6.5±2.6	6.5±2.6	5.7±2.6

The mean and standard deviation values are calculated on one-second segments.

Table 7.1: *Datasets for the training, validation, and testing of the shallow neural network*

7.4 Results

In this section, we report and discuss the results of different regression experiments. As aforementioned, these experiments are basically divided into two main sets considering whether we train the regression model on the logit vectors or on the phonetic feature embeddings. In each of these sets, we conduct multiple experiments where we modify both the model architecture and the target score. As regards the architecture, we vary exclusively the number of hidden units in the fully connected layer (FC). On the other hand, the target score is either Intel-DES or Sev-DES. In the following, we first summarize the performance of the different proposed models, then **a deeper analysis is exclusively conducted on the predictions of the best models.**

7.4.1 Regression on logit vectors

In this set of experiments, we train the regression model on the logit vectors as illustrated in figure 7.4a. We report the results in table 7.2. In the table columns, we specify the target scores for which the model was trained to make predictions, along with the number of neurons used in the hidden fully connected layer. Therefore, for each of these configurations, we provide the different loss values obtained on both the validation set (C2SI-LEC corpus) and the test set (SpeeCOmco corpus).

The first regression model is dedicated to the prediction of the severity score (Sev-DES). The MAE of 1.6 indicates that on average, the difference between the model predictions and the true score is 1.6 points on a severity scale of 0 to 10. Similarly, the MSE of 4.1 indicates that the model predictions have a higher variance, with some predictions being further away from the true score than others. By comparison, the second regression model predicting the intelligibility score reaches an MAE of 1.5. The MSE of 4.2 is slightly higher than that of the first model, indicating that the model predictions have a slightly higher variance. Overall, both models seem to be performing reasonably well, with MAE values that are within a range that can be considered acceptable. Indeed, this error is less than the difference that can be observed between the perceptual assessment of judges assessing the same patient in exactly the same conditions. On the other hand, the slightly higher MSE of the intelligibility model may indicate that predicting intelligibility scores is a slightly more challenging task than predicting severity scores, but more analysis would be needed to draw definitive conclusions.

In the next section, we check if better performance can be achieved while changing the input,

		Task			
		Sev-DES Prediction		**Intel-DES Prediction**	
		#Neurons		**#Neurons**	
		16	**32**	**16**	**32**
C2SI-LEC	**MAE**	1.64	**1.6**	1.53	**1.5**
	MSE	4.28	**4.1**	4.23	**4.2**
SpeeCOmco	**MAE**	1.79	**1.7**	1.61	**1.6**
	MSE	4.63	**4.42**	4.15	**4.13**

Table 7.2: *Results of regression on logit vectors*

on both intelligibility and severity prediction tasks.

7.4.2 Regression on phonetic feature embeddings

In this second set of experiments, we train the regression model on the phonetic feature embeddings as illustrated in figure 7.4b. We report the results in table 7.3. Similarly, we provide the errors on both C2SI-LEC and SpeeCOmco corpora while varying the target task and the number of neurons in the hidden FC of the regression model. We can observe that the best model for severity prediction is the one with 64 hidden neurons. Regarding the intelligibility prediction, the best model is the one with 256 hidden neurons. **All the analyses and comparisons below are based on these two best models.**

For the severity prediction task, the best model achieves an MAE of 1.25 and an MSE of 2.55, as average errors on the C2SI-LEC dataset. As regards the best regression model predicting the intelligibility score, an MAE of 1.21 and an MSE of 2.97 are achieved on the same data. The performance of these models using phonetic feature embeddings is the best even when compared to the previous ones trained on the logit vectors. Even though the difference is not huge, 0.23 and 0.26 of improvement on MAE for intelligibility and severity predictions respectively, this may suggest that the phonetic feature embeddings contain more informative features in terms of speech degradation, for predicting both tasks. This in turn may lead to more accurate predictions overall.

It is worth mentioning that these best models demonstrate remarkable performance on the SpeeCOmco corpus as well (test set). We can see from table 7.3 that the best regression model for severity prediction achieves an MAE equal to 1.4 and an MSE equal to 2.97 on the SpeeCOmco dataset. As regards intelligibility prediction, the best model achieves even better results with an MAE of 1.32 and an MSE of 2.97 on the same data. Despite having relatively few examples to learn from (25K one-second segments, see table 7.1), the models are able to accurately predict scores for another set of HNC patients, that were never seen in the training and validation phases. Importantly, this sheds light on the ability of the resulting models to generalize well to a completely different set of patients and confirms that they are not subject to overfitting on the patients of C2SI corpus.

To complete our analysis, we plot below the scatter plots of the mean predicted severity

		Task					
		Sev-DES Prediction			**Intel-DES Prediction**		
		#Neurons			**#Neurons**		
		64	**128**	**256**	**64**	**128**	**256**
C2SI-LEC	**MAE**	**1.25**	1.28	1.26	2.13	1.3	**1.21**
	MSE	**2.55**	2.74	2.62	10.73	3.36	**2.97**
SpeeCOmco	**MAE**	**1.4**	1.44	1.4	3.29	1.45	**1.32**
	MSE	**2.97**	3.22	3.05	17.58	3.57	**2.97**

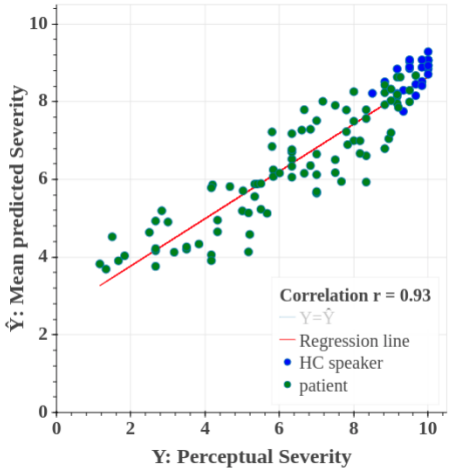
Table 7.3: *Results of regression on phonetic feature embeddings*

(resp. intelligibility) vs. the true perceptual severity (resp. intelligibility) of C2SI-LEC and SpeeCOmco speakers. We organize the analysis based on the target task.

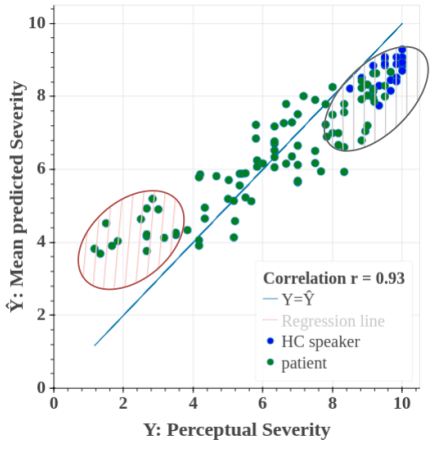
Analysis of the severity prediction

The scatter plot of the mean predicted severity vs. the true perceptual severity of C2SI-LEC speakers is depicted in figure 7.5. For the sake of clarity, we plot exactly the same scatter in figures 7.5a and 7.5b, but with different highlights. In the left figure, we highlight the best fit line between the mean predicted scores and the perceptual score, while in the right we highlight the line $Y = \hat{Y}$ to visualize any possible pattern in the errors. HC speakers and patients are distinguished with blue and green colors, respectively.

First, it is worth mentioning that the range of the mean predicted severity is [3.7; 9.3] which means that this score is reduced and does not cover the complete range of severity [0; 10]. As regards figure 7.5a, we can see that a positive strong relationship exists between Y and $\hat{Y}$. This is confirmed by a high Pearson correlation, equals to 0.93, between the predicted and perceptual severity values. This may indicate that the model is able to capture some of the underlying patterns in the phonetic feature embeddings. However, it is important to note that a high correlation does not necessarily imply high accuracy or precision in the predictions. Even if the model is able to capture some of the overall trends in the data, it may still be making significant errors in individual predictions, which could lead to incorrect conclusions. To this end, we highlight in figure 7.5b the line $Y = \hat{Y}$. From this perspective, we can see that the regression model actually underestimates high severity scores (i.e. the upper right area hashed in grey) and overestimates low severity scores (i.e. the bottom left area hashed in red). Consequently, this may suggest that the model has a systematic bias in its predictions. Specifically, the model may be *"flattening"* the predicted scores towards the mean, rather than capturing the full range of variation in the target variable. In other words, this indicates that there is room for improvement in the regression



(a) Highlight of the best fit regression line



(b) Highlight of the line $Y = \hat{Y}$

Figure 7.5: *Scatter plot of the mean predicted severity vs. the true perceptual severity of C2SI-LEC speakers (exactly the same scatter plot on right and left with a difference in the line highlighted).*

model we proposed in order to tackle this specific behavior. Now moving to the model prediction analysis on the test set, we would like to discard any possibility of misleading conclusions from the previous analysis due to the fact that it was conducted on the validation set. Figure 7.6 depicts the scatter plot of the mean predicted severity vs. the true perceptual severity on SpeeCOmco patients. First, the scatter plot shows exactly the same trends as the one in figure 7.5 conducted on C2SI-LEC speakers, with a variation of the mean predicted severity in the range [3.4; 8.7]. The model bias previously observed towards underestimating high severity scores and overestimating low severity scores is still noticeable. In addition, we add examples of regression plots per second on three patients in order to have visibility on the model behavior at the one-second segment level. The selection of patients is performed based on their perceptual severity levels ("PFG13" having the greatest perceptual severity equal to 10, "CMS19" having a medium perceptual severity equal to 6, and "CMH25" having the least perceptual severity equal to 0.5). The X-axis of these plots represents the seconds of the speech production, which number depends on the time each patient takes to read the same text. The Y-axis represents the severity range. The horizontal blue line is the perceptual severity level of the patient in question, while the blue

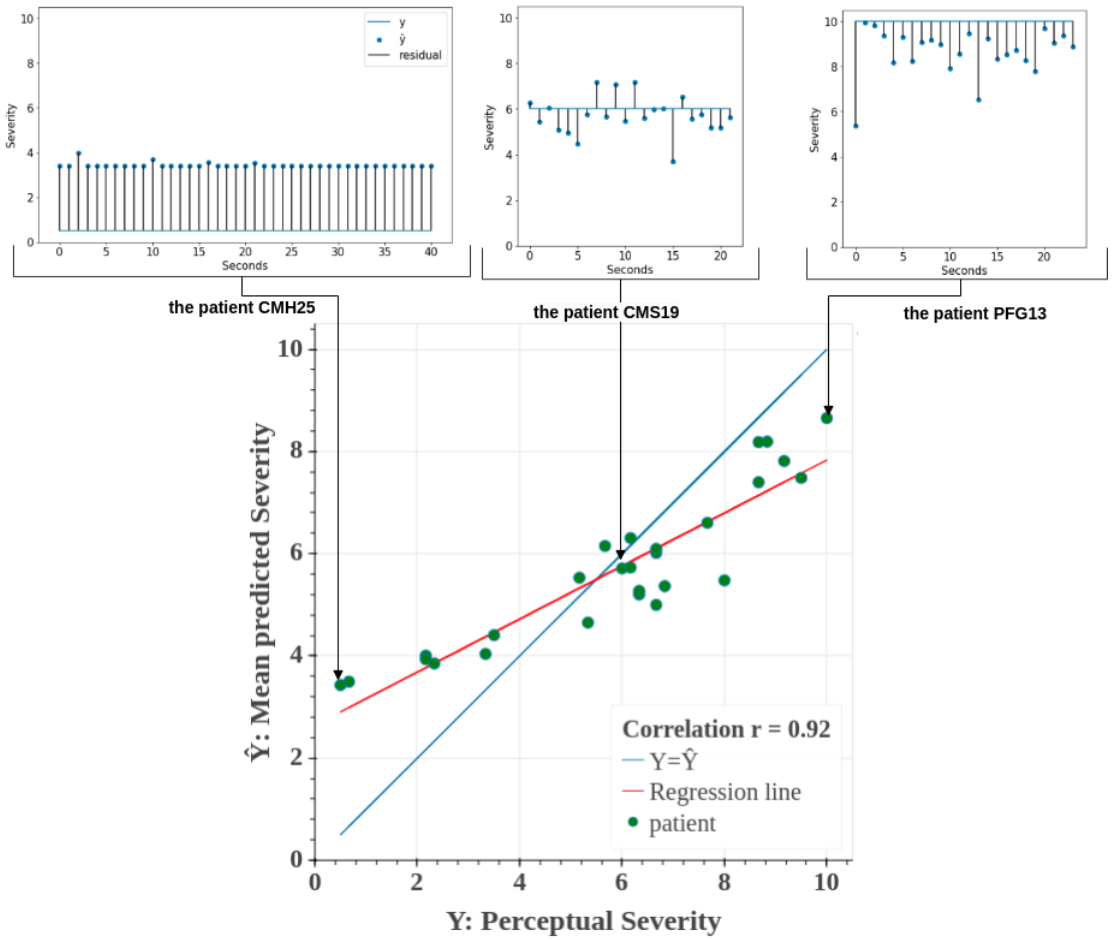


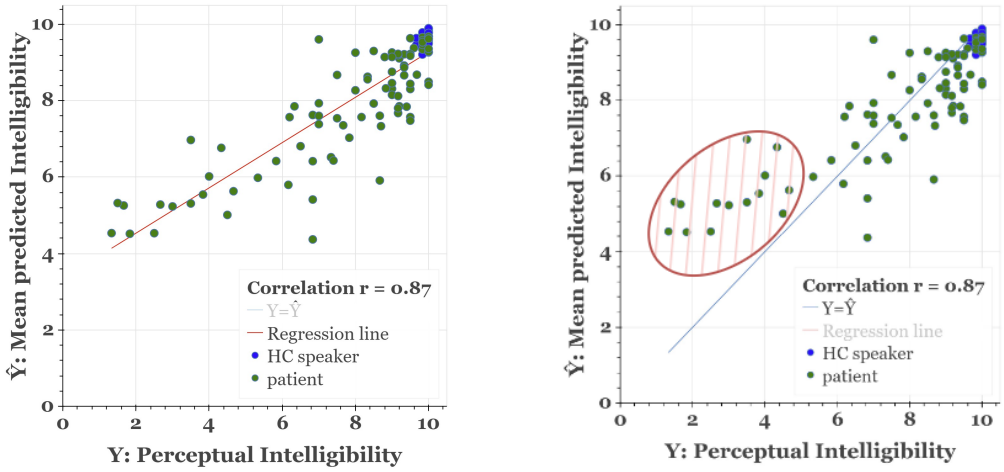
Figure 7.6: *Scatter plot of the mean predicted severity vs. the true perceptual severity on SpeeCOmco patients. Examples of the regression per second for three patients.*

dots are the predicted severity scores at the one-second segment level. The vertical black lines are the residuals at each second. While the model predictions for the patient "CMH25" at the one-second segment level show a similar behavior all along the X-axis with an overestimation of the severity, it is obvious that this is not the case for the two other patients. Indeed, we can see that the model predictions vary largely depending on the one-second segment in question, for patients "CMS19" and "PFG13". Further analyses need to be conducted in order to study the particularity of one-second segments for which the model overestimates/underestimates the severity score.

Analysis of the intelligibility prediction

The scatter plot of the mean predicted intelligibility vs. the true perceptual intelligibility of C2SI-LEC speakers is depicted in figure 7.7. Similarly to figure 7.5, we plot exactly the same scatter in figures 7.7a and 7.7b, but with different line highlights.

It is worth noting that the range of the mean predicted intelligibility varies between [4.4; 9.9]. This range is indeed slightly higher than the range of mean predicted severity. Moreover, the Pearson correlation between the mean predicted intelligibility and the true perceptual intelligibility is less than the correlation calculated on severity, but still very important (r=0.87). Additionally, as shown in figure 7.7b, we observe that the model exhibits a clear bias towards overestimating low scores in the prediction of speech intelligibility, consistent with the previously noted bias in the prediction of speech severity. However, unlike the bias observed in the prediction of speech severity, the bias towards underestimating high scores is not readily apparent in the case of speech intelligibility prediction.



(a) Highlight of the best fit line

(b) Highlight of the line $Y = \hat{Y}$

Figure 7.7: *Scatter plot of the mean predicted intelligibility vs. the true perceptual intelligibility of C2SI-LEC speakers (exactly the same scatter plot on right and left with a difference in the line highlighted).*

Similarly, we observe the predictions of intelligibility on the second corpus SpeeCOmco. Figure 7.8 depicts the scatter plot of the mean predicted intelligibility vs. the true perceptual intelligibility on SpeeCOmco patients. In addition, we choose the same patients as those selected

in the analysis of severity predictions, to analyze their intelligibility predictions at the one-second segment level. First, the variation of the mean predicted intelligibility of SpeeCOmco patients is in the range [4.3; 9.7]. Obviously, the trends and observations described for the intelligibility predictions on C2SI-LEC speakers remain valid. On the other hand, it is worth mentioning that the detailed predictions at the one-second segment level reveal that the model exhibits a high degree of confidence and consistently makes the same decision for all seconds of patient "PFG13". As regards the two other patients, similar findings to those uncovered in the severity analysis are observed. This behavior observed for the prediction of intelligibility score, including, on the one hand, stable scores for both patient "FPG13", close to the "normal" speech, and patient "CMH25", exhibiting very poor intelligibility scores over the one-second segments, and, on the other hand, more varying scores for patient CMS19 with moderate intelligibility score, tends to be coherent and expected. Indeed, with nearly "normal" speech like with patient "FPG13", it would be expected a very few altered one-second segments as observed. Conversely, with very severe impairment like with patient "CMH25", it would be expected that almost all one-second segments would be altered as observed. Finally, more variation between one-second segments should be expected with moderate impairment, with some "normal" speech segments, and others more altered. This behavior is less visible with the analysis of the severity scores, especially for the patient "FPG13". Still, additional analyses are necessary to investigate the characteristics of one-second segments and to better understand this difference in terms of behaviors between both measures.

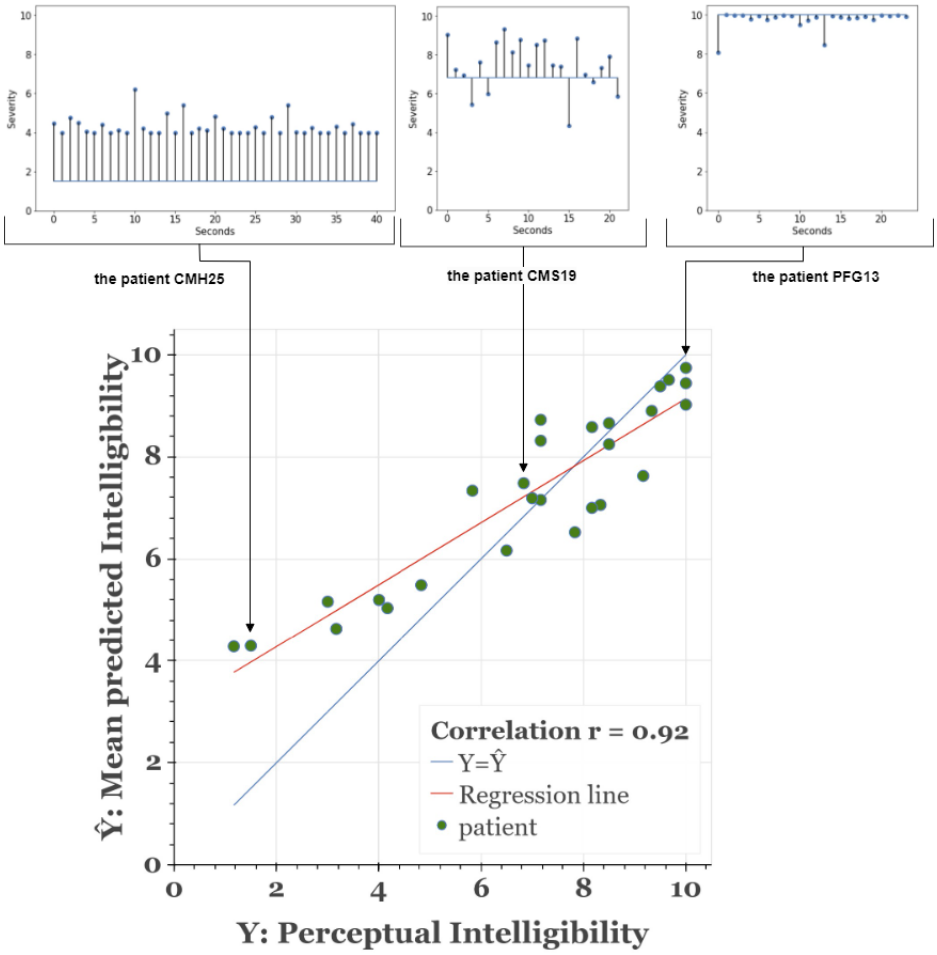


Figure 7.8: *Scatter plot of the mean predicted intelligibility vs. the true perceptual intelligibility on SpeeCOmco patients. Examples of the regression per second for three patients.*

7.5 An end-to-end application of our proposed methodology: A case study on SpeeCOmco dataset

This section is dedicated to answering RQ3 raising the question of how the overall methodology proposed throughout this thesis can be used in clinical practice. To this end, we use SpeeCOmco dataset which has not been used in any of the training or validation of regression models, nor used as a reference in any of the previously implemented steps. Let us consider the three patients involved in the SpeeCOmco dataset we have highlighted in previous sections 7.4.2 and 7.4.2: "PFG13", "CMS19", and "CMH25". These patients were rated by the experts 10, 6.8, and 1.5 in terms of intelligibility respectively.

Considering now the automatic prediction of intelligibility scores based on the Phonetic Feature Embeddings seen in this chapter, we get, for the same three patients, the prediction scores of 9.7, 7.4, and 4.2 respectively. Thanks to Step 2 of our proposed methodology, we can associate these different predicted intelligibility scores with a deeper analysis based on the altered phonetic features as depicted in figure 7.9. Indeed, this figure reports the local ANPS scores per phonetic features for both consonants and vowels for all the patients of the SpeeCOmco dataset (heatmaps), sorted according to their perceptual intelligibility scores. The three patients "PFG13", "CMS19", and "CMH25" are specifically highlighted in the figure with their local ANPS scores surrounded. Comparing these three patients, we can clearly see different configurations of local ANPS scores, showing a consistent deterioration of score values compared to the prediction intelligibility scores associated with the patients, especially regarding the consonant phonetic features. This association between the predicted intelligibility score and the heatmaps exhibiting ANPS scores should permit clinicians to directly link a score with phonetic feature alterations at time t, but also to compare different pairs scores/heatmaps for the same patient in a longitudinal way to measure the efficiency of a rehabilitation program or of a specific prosthesis.

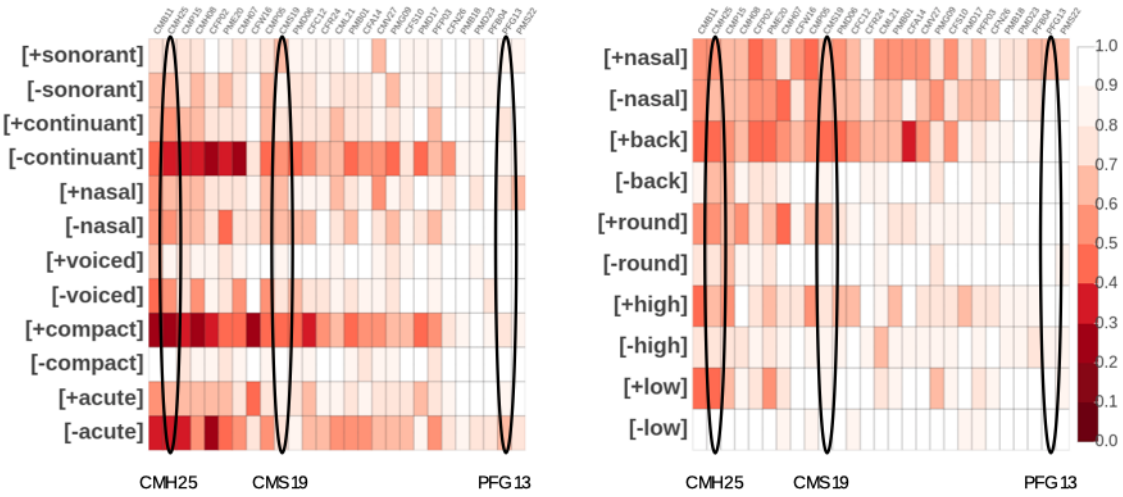


Figure 7.9: *Heatmaps (outcome of Step2) showing local ANPS scores per phonetic features for both consonants (on left) and vowels (on right) for all the patients of the SpeeCOmco dataset. Patients are sorted according to their perceptual intelligibility scores, from most intelligible speaker (on right - e.g. patient "PFG13") to least intelligible speaker (on left - e.g. patient "CMH25")*

.

7.6 Discussion

In this chapter, we carry out the third and final step of our proposed methodology dedicated to the prediction of an intelligibility score. By implementing this step, the missing piece of the puzzle is placed into position and we can clearly see the purpose of this study (see figure 7.10). Throughout this chapter, we utilized various techniques and examined multiple input types in order to attain our ultimate assessment objective. In summary, promising results have been obtained regarding the prediction of speech intelligibility and severity of disordered speech due to head and neck cancer. While these findings show great promise, we believe that further improvements are necessary to enhance the reliability and generalizability of the models. Moving forward, we recommend a few techniques and provide some suggestions to improve the outcome of our current implementation.

1. **Data collection and representativeness:** The relatively small sample size of disordered speech datasets as well as the potential biases characterizing their assessment could have a negative effect on the regression model performance. Indeed, the resulting model will likely have limited generalizability and may perform poorly when faced with speech impairments outside the training set. We have shown that this is absolutely not the case with our proposed models on the corpora used here, although they were trained on a very limited dataset. Nonetheless, we believe that gathering additional data from varied populations can further enhance the reliability and generalizability of the models, especially if we consider the high diversity of disordered speech.

2. **Training on variable segment lengths:** Training a regression model on variable segment

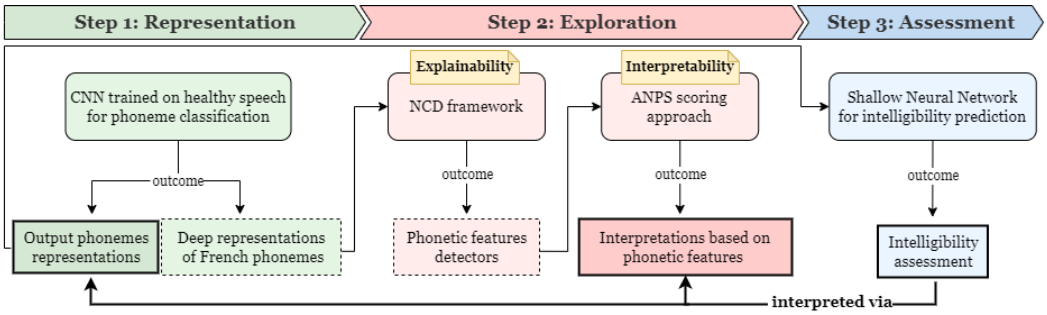


Figure 7.10: *A step forward in the achievement of the proposed methodology: the accomplishment of step 3*

lengths could improve the prediction of intelligibility compared to training it exclusively on one-second segments. Indeed, we can assume that training on variable segment lengths would better capture more speech disorder-related alterations and assign them to the perceptual measures given by experts on the overall speech records. Therefore, the model can learn to rely on certain features that are specific to that segment length. Additionally, in real-world scenarios, speech segments can vary in length depending on the speaker, the task, or the context. By training on variable segment lengths, the model can learn to generalize better and perform well on a wider range of speech segments.

3. **Training on an objective intelligibility:** One major issue in the perceptual assessment of speech disorders is the subjectivity of human perception. We outlined this problem throughout this study, but more importantly when we had to fix the perceptual assessment measure that serves as a ground truth for intelligibility to train the regression model. We expressed this concern in the second research question of this chapter. Due to a lack of resources, we did not have the opportunity to train the proposed model on a more objective perceptual intelligibility measure as the PPD-DAP score. However, we believe that once available, the implementation of a model predicting this measure will significantly improve the intelligibility prediction.

4. **Attention mechanism:** An attention mechanism can potentially improve the performance of the regression model by enabling the focus on the most relevant frames and features for the prediction. This technique is very used in the speaker recognition field [Okabe et al., 2018], where it has been shown that some frames are more unique and important for discriminating speakers than others, for a given utterance. In speech intelligibility assessment of patients with speech disorders, some frames of speech may contain more critical information for understanding the intended message than others. For example, frames containing consonants or vowels that are frequently mispronounced due to speech disorder may be more important for understanding the message than frames containing more easily recognizable sounds. Therefore, an application of a frame-level attention mechanism can force the model to automatically focus on these meaningful frames, and thus, produce an utterance-level representation that is more reflective of the speech intelligibility level. As a result, we can obtain a more accurate assessment of speech intelligibility.

5. **Explainability of the final score based on the feature embedding vectors:** A possible

further analysis could be to examine the input features used by the regression model and their importance in predicting the target score. For instance, consider the model trained on phonetic feature embeddings for the task of intelligibility prediction. An application of the SHAP (SHapley Additive exPlanations) framework can be used to explain the predicted score of the model for a specific input by attributing a contribution value to each element in the embedding. This provides insight into which features are driving the predictions of the model and how they are influencing the final intelligibility score.

Conclusions and Perspectives

Conclusions

Throughout this thesis, we investigate the contribution of deep learning and interpretability tools in achieving an objective intelligibility assessment of disordered speech. Particularly, the central research question addressed in this study is whether it is possible to develop such a tool that incorporates the advantages of deep learning methods while overcoming the limitations of current assessment tools. To this end, we introduce an overall methodology composed of three steps. Each step is specially designed to tackle specific limitations of current assessment tools and thus answers a specific research question. We summarize graphically the overall methodology in figure 7.11.

In the first step, we tackle a major issue in the current automatic tools dedicated to disordered speech assessment which is *the limited insight into the relationship between speech disorders and the resulting assessment*. To this end, we implement a DL-based model (CNN), trained on healthy speech and dedicated to an intermediate task which is French phoneme classification. By requiring the transition of speech signal through the intermediate and understandable dimension of phonemes, this simple classification task allows, in successive steps, the assessment of intelligibility to be subsequently linked to the specific linguistic units that affect it. On the other hand, by training the CNN exclusively on healthy speech, we address another challenge that arises from both the nature of deep neural networks and the fact that it is applied in a speech pathology task. Indeed, in speech pathology, the amount of data available is often not sufficient to learn reliable models given the large variability of the patterns in interest. This problem is further compounded when the model in question is based on deep learning, a field that is mainly known for its high data requirements. Consequently, training solely on healthy speech was a compelling alternative, as it enabled us to include a large dataset and gain insights into normal speech patterns, which is certainly valuable for future consideration of speech disorders. In one fell swoop, the methodological choice taken in this first step not only prepares for a detailed final assessment by providing insightful information at the phoneme level but also addresses the issue of *limited data availability in speech pathology* while also meeting the *high data requirement of deep learning* applications.

Moving on to **the second step**, we focus *on one major aim of this work which ensures that the developed solution is interpretable and reliable*, to be accepted within clinical practice. We, therefore, investigate the capacity of the CNN-based phoneme classifier to yield relevant knowledge related to the characteristics of speech pathology. Our contributions in this step are noteworthy since we have proposed a variety of original methods that are tailored to our particu-

lar context, while also having the capability to handle a range of other applications. Particularly, we design and propose the framework **Neuro-based Concept Detector (NCD)**, a general analytic framework for the explainability of hidden neurons/layers of a DL-based model performing a classification task. By applying NCD for the proposed CNN explainability, we bring to light an extra-interpretable dimension of great relevance in the clinical phonetics context which is the phonetic features. Subsequently, we propose a scoring approach **Artificial Neuron-based Phonological Similarity (ANPS)** to retrieve fine-grained interpretations of the speech impairment based on the emergent dimension of phonetic features. In an overall view of the proposed methodology, we hit two targets with one shot through this step. Indeed, not only do we actively take steps to mitigate the impact of the black-box nature of DL models and alleviate the mistrust among experts in a clinical context, but also we ensure an additional granularity level (i.e. phonetic features) with which we can link and interpret the final intelligibility assessment. More interestingly, the interpretation of this extra-dimension is of great practical relevance in the clinical phonetics context since it allows the establishment of a clearer connection between the final intelligibility assessment and the physiologic characteristics of impaired speech.

As we progress in our proposed methodology, we have laid the groundwork for the last and third step. **This third step** is dedicated to the prediction of a final score assessing the speech production of a speaker in the context of speech disorders. We explore multiple techniques and tasks, while we are always based on the outcome of the previous steps. Promising results have been obtained regarding the prediction of speech intelligibility and severity of disordered speech due to Head and Neck Cancer. Finally, we propose a first attempt of end-to-end application of the overall framework implying a few patients, demonstrating how the resulting outputs can be used in clinical practice. More globally, the results obtained in this step reflect the great interest of the overall proposed methodology.

This study sheds light on a relatively unexplored area which is deep learning interpretability for speech disorders assessment and characterization. To the best of our knowledge, no prior work has explored and explained the hidden representation inside a DL speech model to provide a deeper understanding and interpretation of the final assessment of the disordered speech. By examining this speech in terms of production at the phonemes and phonetic features levels, clinicians can gather more useful information to monitor the progress of therapy and evaluate the effectiveness of different treatments. In other words, the identification of the specific linguistic units that affect intelligibility from an acoustic point of view could enable clinicians to develop tailored rehabilitation protocols that improve the patient's ability to communicate effectively, and thus, his/her quality of life.

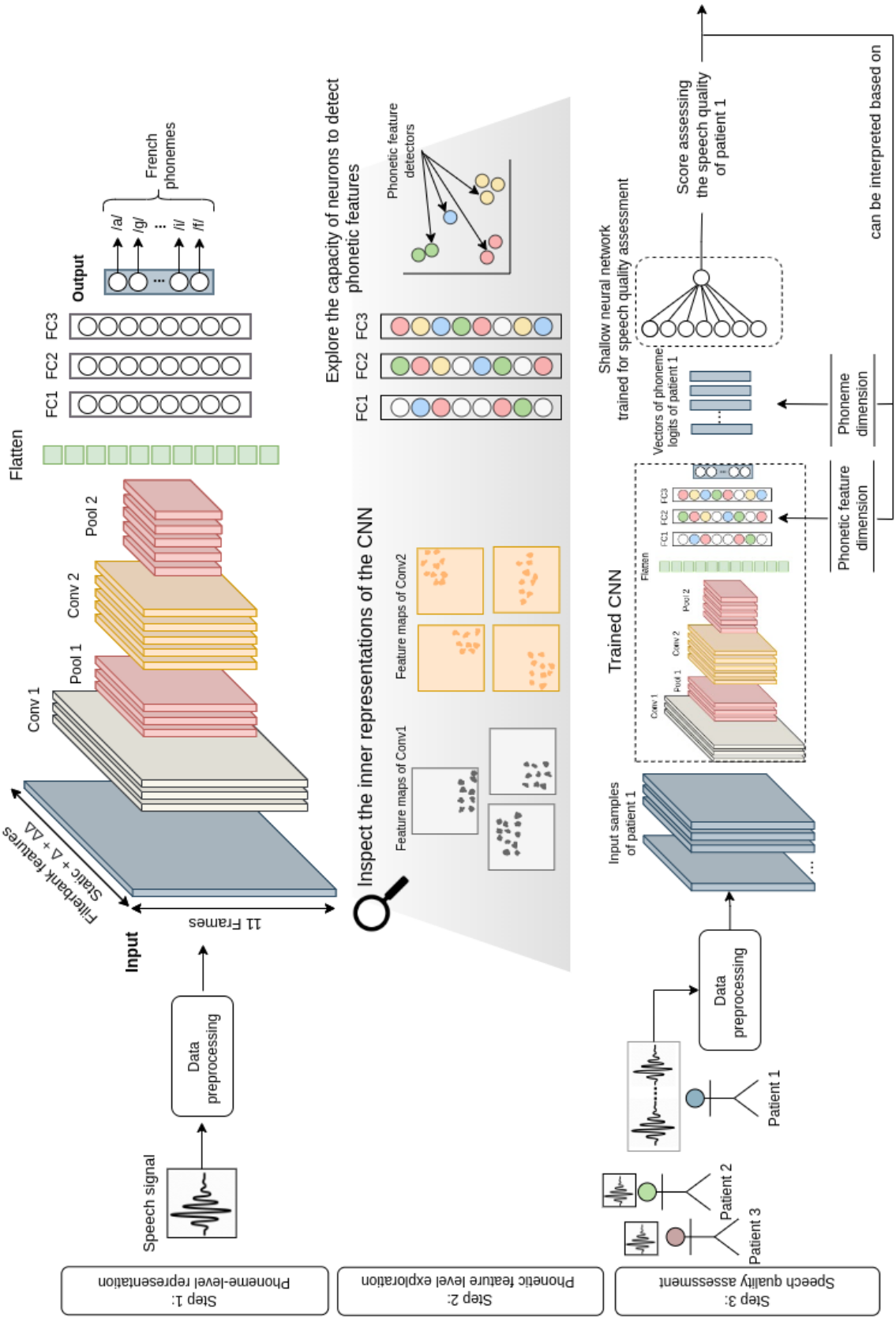


Figure 7.11: *An overview of the global methodology proposed in this study*

Perspectives

The results obtained in this study demonstrate the potential of deep learning in predicting and interpreting speech intelligibility of disordered speech. Furthermore, they offer a promising direction for future research in this field. We already dedicated a section for future work in each chapter of our contributions. In this final section, we introduce more global perspectives related to the entire work.

A main contribution of this study is the development of the *Neuro-based Concept Detector* framework. As previously highlighted, NCD is a general analytic framework for the explainability of hidden neurons/layers of a DNN performing a classification task. In our specific context, the application of NCD revealed the emergence of the concept of phonetic features in the deep representations of the CNN-based phoneme classifier. Subsequently, we set up a scoring approach, *Artificial Neuron-based Phonological Similarity*, to retrieve relevant interpretations from the phonetic feature detector. Applied to the context of speech disorders in which we are involved, these interpretations reflect the phonetic feature alteration of patients, which allows the design of specific rehabilitation protocols by clinical experts. An interesting perspective that can be considered is the application of these phonetic feature detectors and related ANPS scores to the speech of second language (L2) speakers. The resulting interpretations can help identify the phonetic features that have to be improved for a given speaker, in order to communicate more effectively in the language he/she is learning. The identification of phonetic feature alterations can be used in an e-learning platform dedicated to the improvement of speech realization by L2 speakers. In addition, we can imagine such fine-grained interpretations in the identification of regional accent characteristics. This identification can help take into account the specificity of these regional accents in some applications.

Beyond that, we believe that the capacity of the NCD explainability framework can go beyond identifying detectors of phonetic features within the proposed CNN and even in totally different application domains implying speech analysis. For instance, in a DNN dedicated to speaker characterization, the application of NCD could reveal neurons detecting certain acoustic features, such as pitch and tone, which can be indicative of a typical speaker's emotional, or physiological state.

Another relevant perspective that can be considered is the usage of end-to-end models. Indeed, these models operate directly on the raw audio waveform instead of extracted features. In our case, it would be very interesting to include this option since the CNN architecture, particularly, has shown its performance on raw speech [Passricha and Aggarwal, 2018], with a first convolutional layer able to act as a feature extractor.

Now from a clinical point of view, one interesting perspective would be to confirm the promising results that we have obtained within a longitudinal study. This type of study involves following a group of individuals with speech disorders for a period of time and collecting data at multiple time points. This study is of great importance for speech disorder assessment because it allows clinicians to observe and track changes in speech production over time. In other words, if a longitudinal study can confirm the effectiveness of the proposed methodology and demonstrate consistent and meaningful interpretations, it can provide clinicians and researchers

with a deeper understanding of how speech disorders impact individuals. This can inform better treatment approaches and improve the overall management of speech disorders.

Furthermore, in many head and neck cancers, patients can be treated with a glossectomy (i.e. surgical removal of all or part of the tongue). In a very new clinical practice, prostheses can be provided to those patients in order to compensate for the disorders caused by the partial ablation of the tongue. We could expect that our proposed methodology provides knowledge about the effectiveness of the prosthesis in improving speech intelligibility. Indeed, this latter can be used to evaluate and interpret the intelligibility of patients before the usage of the prosthesis, during the design of the prosthesis (based on a lot of adjustment tests with the patient), and in regular follow-up appointments. Consequently, the knowledge provided would be really interesting for clinicians to ensure that the tongue prosthesis is fitting properly and no necessary adjustments are needed.

Appendices

Appendix A

Clinical Texts

A.0.1 La chèvre de Monsieur Seguin

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.

A.0.2 Le Cordonnier

Dans un petit village de la montagne, il y a un pauvre cordonnier, tout vieux et tout cassé. Les villageois lui apportent des chaussures à réparer. Mais il ne travaille pas vite. Tous les soirs, il mange tout seul, bien tristement. Ce soir, il a devant lui, un gros tas de souliers et de guêtres à recoudre.
-"Jamais je ne pourrai les réparer. Je suis trop âgé et trop malade."
Près de lui, la grosse horloge fait: tic tac, tic tac. Le pauvre vieux, tout découragé, s'endort. Aussitôt, l'horloge s'ouvre, et deux petits lutins sautent sur le plancher. L'un s'appelle Tic, l'autre s'appelle Tac.
- "Rangeons les étagères, réparons les souliers, recousons le linge", dit Tic.
- "Préparons un gâteau, mettons du gui au plafond, changeons ces vieux
rideaux", ajoute Tac.
Minuit sonne! Les deux vaillants petits lutins rentrent dans la pendule. Le lendemain, le pauvre cordonnier s'éveille:
- "O joie! Qui a préparé ce bon gâteau? Qui donc a rangé la maison?"
- "Tic tac! Tic tac!", dit la vieille horloge.

Appendix B

Extracts from the GDPR and AI act

Contents

B.1 GDPR

B.1.1 Art. 15 GDPR: Right of access by the data subject

The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

(a) the purposes of the processing;

(b) the categories of personal data concerned;

(c) the recipients or categories of recipients to whom the personal data have been or will be disclosed, in particular recipients in third countries or international organisations;

(d) where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period;

(e) the existence of the right to request from the controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing;

(f) the right to lodge a complaint with a supervisory authority;

(g) where the personal data are not collected from the data subject, any available information as to their source;

(h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

B.1.2 Art. 22 GDPR: Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   (c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

B.2 AI act

B.2.1 Extract from Art. 13: Transparency and provision of information to users

1. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of this Title.

B.2.2 Extract from Recital 38:

Actions by law enforcement authorities involving certain uses of AI systems are characterised by a significant degree of power imbalance and may lead to surveillance, arrest, or deprivation of a natural person's liberty as well as other adverse impacts on fundamental rights guaranteed

in the Charter. In particular, if the AI system is not trained with high-quality data, does not meet adequate requirements in terms of its accuracy or robustness, or is not properly designed and tested before being put on the market or otherwise put into service, it may single out people in a discriminatory or otherwise incorrect or unjust manner. Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defense and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable and documented. It is therefore appropriate to classify as high-risk a number of AI systems intended to be used in the law enforcement context where accuracy, reliability, and transparency are particularly important to avoid adverse impacts, retain public trust, and ensure accountability and effective redress.

Appendix C

Extra approaches explored in Step 2

Contents

In this appendix, we showcase the additional methods we have explored concurrently with our work on step 2. Each method is thoroughly explained in its own section and operates independently from the other methods. As the reader moves through the main chapter 6 of step 2, a redirection to these sections is performed.

C.1 Explainability based on Class Selectivity Index

As introduced in chapter 6, several methods were proposed to investigate the information content of a unit in a neural network. One of these metrics is the Class Selectivity Index (CSI), which is a property reflecting the degree to which a neuron is selective for one specific class. This metric identifies neurons with the same class tuning properties within a layer, and how these properties evolve through layers to improve the distinctiveness of final classes. In the following section, we describe this approach, explain how we adjust it to fit our particular application and finally report the results and conclusion.

C.1.1 Approach description

As introduced by [Rafegas et al., 2020], one method for calculating the class selectivity index consists of choosing input samples that maximize the activation of a given neuron, and then identifying the classes to which they correspond. In other words, the selectivity of a neuron is assessed by measuring the variability of its responses across different classes of data samples.

In their study, authors retrieved activations from individual neurons of a CNN trained on image classification task after presenting a dataset for explainability purposes according to the following steps. Firstly, a normalized activation is calculated for each neuron by dividing the activation values of that neuron for different input images of the dataset by the maximum of these values reached by the same neuron over all the images in the dataset. Next, for each neuron, a ranking of images based on their corresponding normalized responses from the highest to the lowest value is performed. With such a normalization, authors were able to detect neurons having flattened behavior, signifying they were activated by most of the images vs. neurons being highly activated for only a subset of images. The next step consists of selecting, for each neuron, the first $N = 100$ images having the highest normalized activation values by setting a threshold activation value greater than 70% of the maximum activation. The purpose of such constraints is to discard neurons which activation values were not strong enough. The final step consists of quantifying the selectivity of a particular neuron. To this end, the set of class labels corresponding to the N images previously selected is examined. Then a frequency measurement f_c for each of the classes present among the selected images is calculated in order to weight the significance of each class taking into account the normalized activation values associated with this class. f_c value for each class c for the i^{th} neuron at layer L, $n^{i,L}$ is defined as follows :

$$f_c(n^{i,L}) = \frac{\sum_j^{N_c} w_{j,i,L}}{\sum_l^{N} w_{l,i,L}} \tag{C.1}$$

where N_c is the number of images belonging to class c among the N selected images for the neuron n^i, $w_{j,i,L}$ is the normalized activation of the $j-th$ selected image issued from the neuron $n^{i,L}$. Finally the Class Selectivity Index for a neuron $n^{i,L}$ is defined as:

$$\gamma(n^{i,L}) = \frac{N - M}{N - 1} \tag{C.2}$$

where M is the number of classes that describe the selectivity of a neuron $n^{i,L}$. Therefore, the class selectivity index gives insight into how strong the contribution of a neuron to a single class is. Notably, a low class selectivity index indicates a poor contribution of the neuron to the classification of a single class whereas a strong value, $M = 1$ in the best case, indicates that the neuron is highly selective for a single class.

C.1.2 Adjusting the approach to fit our particular case

In this analysis, we adapted the approach of class selectivity index to explain the hidden neurons of our CNN performing the task of phoneme classification. Hence, images correspond to the frames associated with phoneme labels. All the parameters mentioned above were kept unchanged except the N parameter involved in equation C.2. Indeed, here this parameter has to refer to the maximum number of phonemes potentially represented in the set of selected frames (i.e. 31 phonemes vs 1000 image classes compared to [Rafegas et al., 2020]) in order to have a range of CSI values from 0 to 1.

C.1.3 Results

For each neuron, the phoneme labels of the top 100 associated frames having the strongest activation values, and at the same time, achieving at least 70% of the maximum activation of that neuron were identified. As we expected, not all neurons are highly stimulated by at least 100 frames, depending on the layers observed. In this study, we are interested in analyzing only neurons satisfying these constraints. As we move towards the output layer in the fully connected layers, the outcome of these constraints application is an increase in the number of retained neurons. Indeed, while only 44% of the FC1 neurons are considered as having responses strongly enough to be taken into account in the selectivity analysis, 93% of FC2 neurons and all neurons of FC3 were selected.

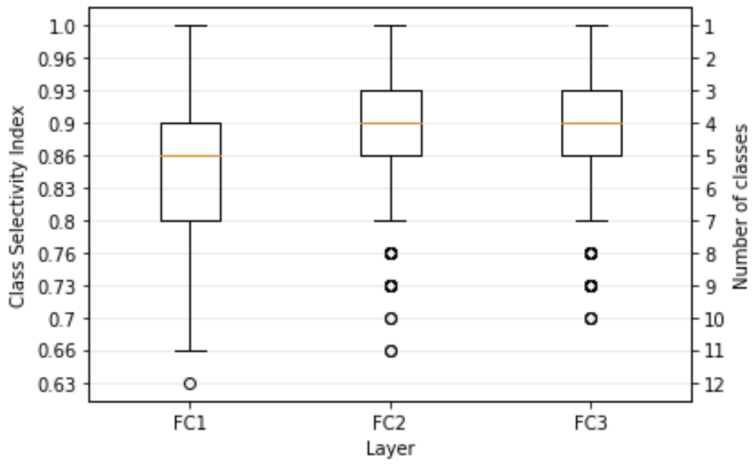


Figure C.1: *Boxplot per layer representing the range of Class Selectivity Index (CSI) values computed (and associated number of selected classes) for retained neurons respecting constraints.*

Additionally, the global analysis of CSI values computed per neuron on the three layers shows layer-dependent distributions as illustrated in figure C.1. We can observe that there is a notable decrease in the interquartile range of CSI from FC1 to FC3, and both FC2 and FC3 exhibit higher CSI values. In a more detailed manner, neurons of the first layer have a class selectivity index varying between 0.66 and 1, meaning that they are considered selective for up to 11 phonemes. On the other hand, the two succeeding layers converge towards a similar behavior, where the class selectivity values of neurons are spread out on a smaller interval with a minimum of 0.8 in most cases (i.e. the majority of neurons in these two layers are considered as selective for up to 7 phonemes). This highlights that, as we go deeper into the neural network, an increasingly abstract representation of phoneme features is performed in order to enhance the separation of final phoneme classes.

C.1.4 Summary

It is worth mentioning that CSI is not a perfect measure of the encoding properties of a given neuron. In fact, this index reflects the selectivity of a neuron for a single phoneme which is not necessarily the case. For instance, let's consider a neuron selective for the six following

phonemes - /p/, /t/, /k/, /b/, /d/, /g/ - it will have a relatively low CSI value equal to 0.83. However, such a neuron is very interesting since it reflects a selectivity for a specific phonetic class which is the stop consonants. Now let's consider another case where similarly the neuron has a distinctive response for the stop consonants as the case of neuron 98 of the FC2 in figure C.2. Yet, the CSI considers only the 100 frames maximizing the activation of the neuron (i.e. those circled in red). Consequently, neuron 98 will be identified as detecting the phoneme /g/ to which belong these frames. To conclude, neurons selectivity will be either underestimated or reflect incomplete information, while actually in both cases, the neuron encodes a very important distinctive feature related to the manner of articulation of consonants. To overcome these limits, we propose a more adequate explainability approach, Neuro-based Concept Detector NCD, that we detail in the main chapter 6.

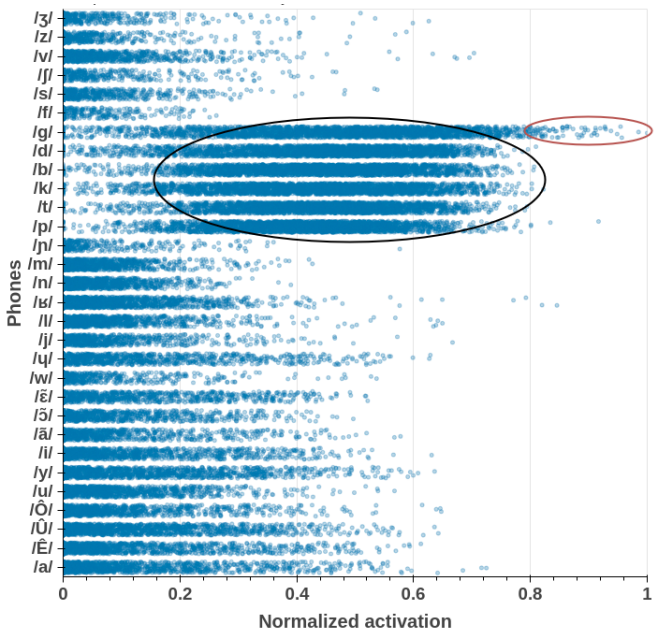


Figure C.2: *Jitter plot visualizing the normalized activations for unit 98 of FC2 on BREF-Int dataset (the distinctive response for stop consonants is circled in black)*

C.2 Ablation Study on the Phonetic Feature detectors

Ablation study is a surgical procedure that was first developed in the early 19th century to understand the role of different components of the brain [Carlson et al., 2009]. An ablation study involves removing a specific part of the brain and observing any resulting behavioral changes. That is, it allows uncovering the specialized regions for certain behaviors and the relative contribution of these regions to the overall function. Similarly, an ablation study in DL involves evaluating a model's performance after removing one or more of its components. In this line, many researchers have included this study in their research work [Morcos et al., 2018, Zhou et al., 2018, Meyes et al., 2019, Sheikholeslami et al., 2021].

In this setting, we apply the ablation study on the trained phoneme classifier in our pos-

session. The study is viewed as multiple trials, each trial involves removing all the neurons detecting one specific phonetic feature. Through this study we address the following questions: *How important are the neurons identified through the NCD framework to the classification of each phoneme? Are these neurons exclusive to the phonetic features they were identified to detect?* Our aim is to further show the relevance of phonetic feature detectors (i.e. the outcome of our NCD framework). In the following, we illustrate the process and results of this application.

C.2.1 Classification Accuracy Drop

To start, we ablate[1] all the neurons that were identified as detectors for a particular phonetic feature by the *NCD* framework in all examined layers. In our case, the ablation is performed on both FC2 and FC3 of the trained CNN, which are concerned with the emergence of phonetic features. Thereafter, we feed *BREF-Int* dataset to the model after ablation. Consequently, we compute the resulting classification accuracy drop per phoneme. We call phoneme accuracy drop the percentage of the frames belonging to a specific phoneme that were misclassified due to the ablation process. We must note that, so far, we considered the two macro-classes of vowels and consonants separately. In the same direction, after the ablation of detectors corresponding to a vowel phonetic feature (resp. a consonant phonetic feature), a vector of phoneme accuracy drop with a dimension equal to the number of vowels (resp. consonants) is obtained.

C.2.2 Results

At this stage, we summarize the results of the ablation study per macro-class in figures C.3 and C.4, respectively for vowels and consonants. These figures are made up of a set of horizontal bar charts which number corresponds to the number of vowel phonetic features, resp. consonant phonetic features. Each of the horizontal bar charts is a visualization of the vector of phoneme accuracy drop after the ablation of neurons which are detectors of a specific phonetic feature. That is, the Y-axis shows the list of phonemes belonging to the macro-class in question. The X-axis displays the accuracy drop with positive and negative values centered around zero. The blue bars correspond to the subset of phonemes that present the specific phonetic feature in question, whereas the red bars correspond to the rest of the phonemes.

Ideally, this visualization should display a pattern where all the blue bars have a positive value of accuracy drop, and all the red bars have very low (i.e. around zero) or negative values of accuracy drop. In our case, we can see that this pattern appears in most cases. In addition, the value range of accuracy drop (X-axis) is highly different when comparing plots. Indeed, this can be explained by the difference in terms of the number of neuron detectors that were ablated for each phonetic feature. Let's take the example of the bar plot illustrating the impact of ablation of the neurons detecting *[+nasal]* phonetic feature in figure C.3. We can clearly see that the nasal vowels /ɑ̃/, /œ̃/, and /ɔ̃/ are damaged significantly after this ablation with an accuracy drop achieving almost +40% for the phoneme /ɑ̃/ On the other side, we can see that this ablation did not damage any of the oral phonemes accuracies. This finding once more confirms the important nasality-specific information carried by the neurons arising from the NCD framework as detectors of the vowel phonetic feature *[+nasal]*. In the same figure, we can see that ablating

[1] A neuron is ablated by setting its weight and bias to zero so that it will not contribute to the prediction for any input.

detectors of *[+high]* phonetic feature results in a huge drop in the accuracies of phonemes /u/, /y/, and /i/ (i.e. those presenting high phonetic feature). On the other hand, we show that the aforementioned ideal pattern is not verified for some phonetic features. Specifically, this can be seen in two sub-figures of the C.4. Indeed, ablating detectors of the consonant phonetic feature *[−sonorant]* does not tend to cause a significant drop in accuracy for the subset of obstruent phonemes (i.e. not presenting the sonorant phonetic feature). Even worse, ablating detectors of the consonant phonetic feature *[−compact]* had no impact, and even resulted in a slight accuracy amelioration of some phonemes not presenting the compactness phonetic feature (i.e. phonemes on which we were supposed to see an accuracy drop due to this ablation).

C.2.3 Summary

In this part, we aimed to show the relevance of phonetic feature detectors through an ablation study. The results of the study generally demonstrate the significance of the NCD framework's outcome, however, the level of relevance varies depending on the particular phonetic feature being considered. It is important to point out once more that we are involved in a medical context, considered a high-stakes domain and requiring a high level of trust. In light of this, the varying relevance can be viewed as varying levels of trust in the interpretations yielded by these phonetic feature detectors that will be generated considering disordered speech later on.

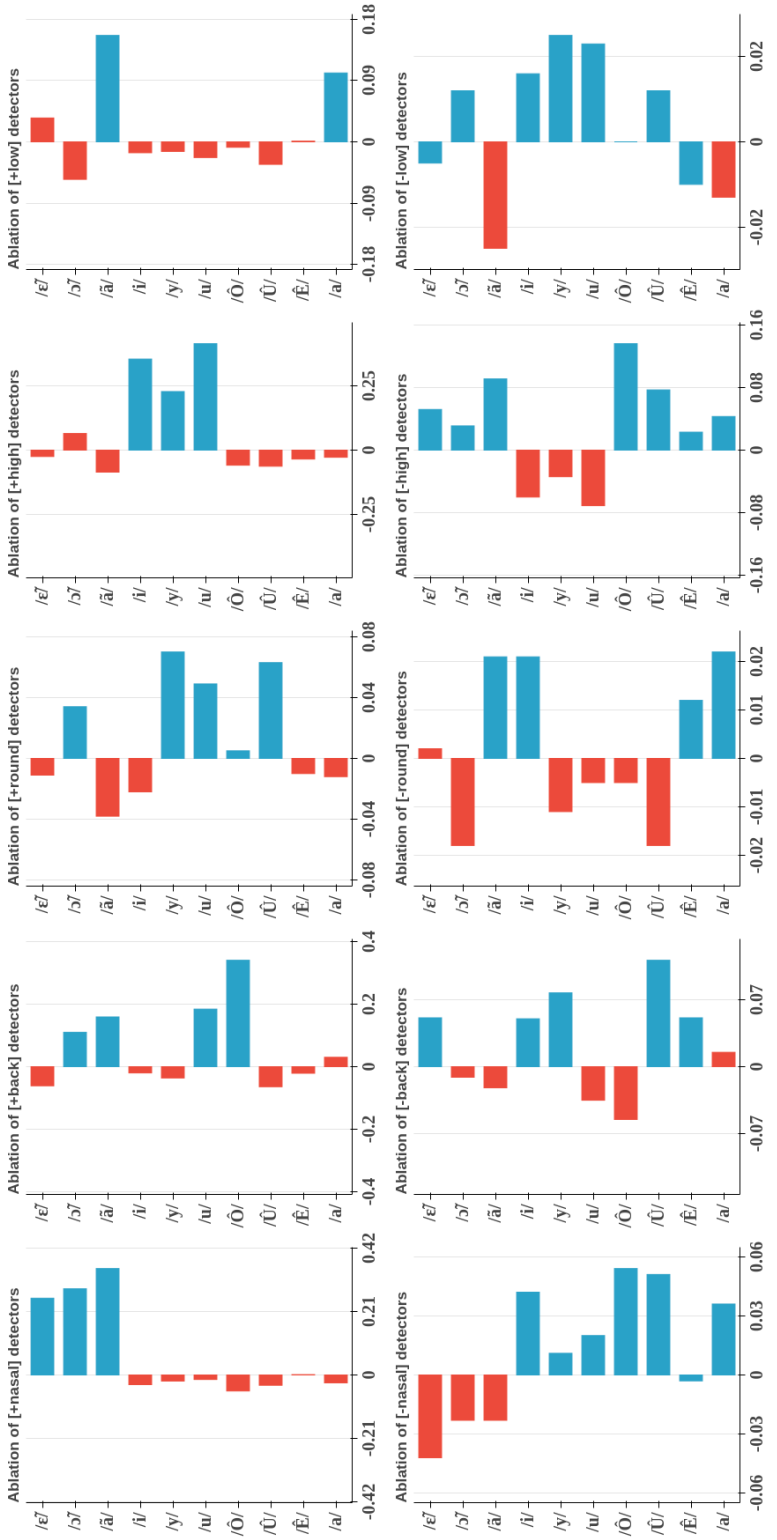


Figure C.3: *Ablation study on detectors of vowel phonetic features*

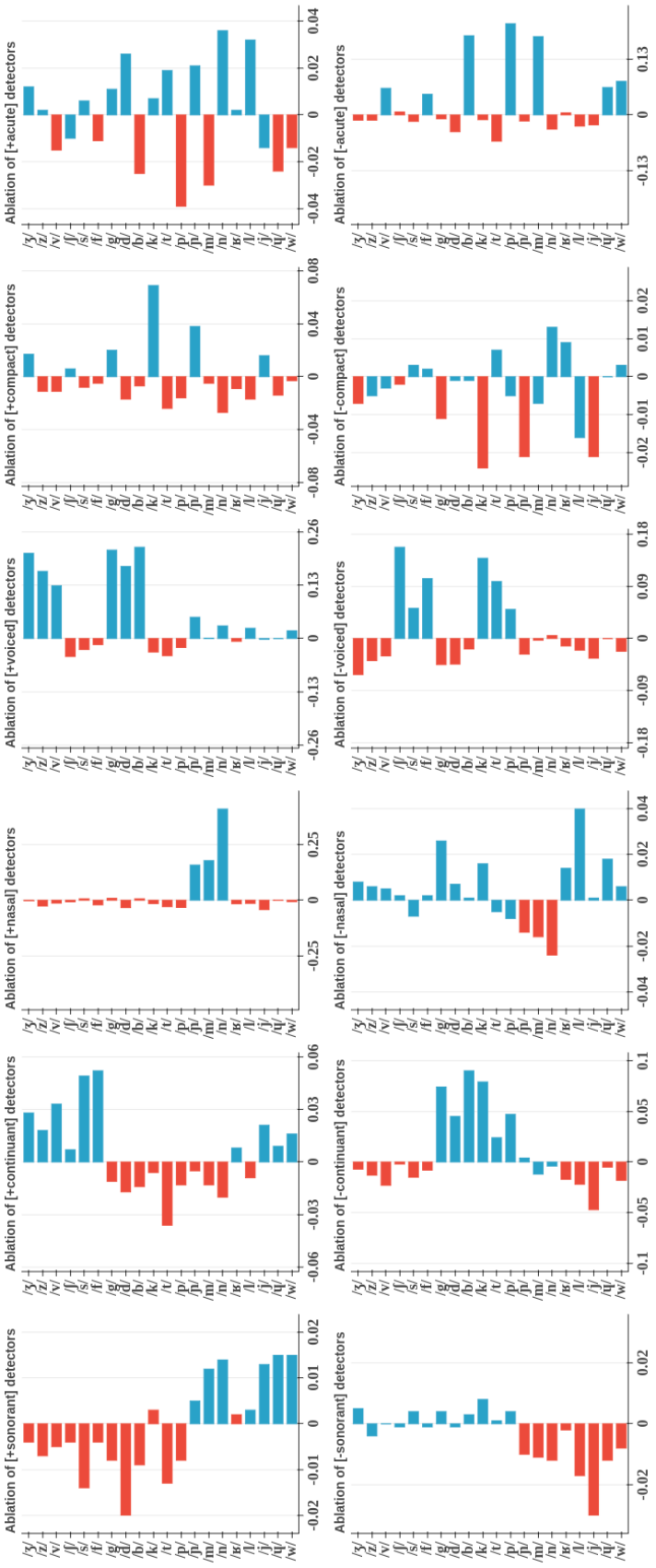


Figure C.4: *Ablation study on detectors of consonant phonetic features*

C.3 Visualization of Convolutional layers

In this section, a brief exploration of the convolutional layers is performed. In the same logic, the following analyses are performed on *Bref-Int* dataset.

C.3.1 Visualization method

Let's recall that the proposed CNN-based model is composed of two convolutional layers with a ReLU activation function, where the first layer has 32 filters and the second layer has 64 filters. For every input sample x in the *Bref-Int* dataset, the activation map $M_f(x)$ of every internal convolutional filter f is collected. In order to be able to visually compare the activation maps to the input, the activation maps are scaled up to the input dimension using bilinear interpolation as shown in figure C.5. We note $M'_f(x)$ the activation map after interpolation.

For each filter, we have a set of interpolated activation maps corresponding to the number of

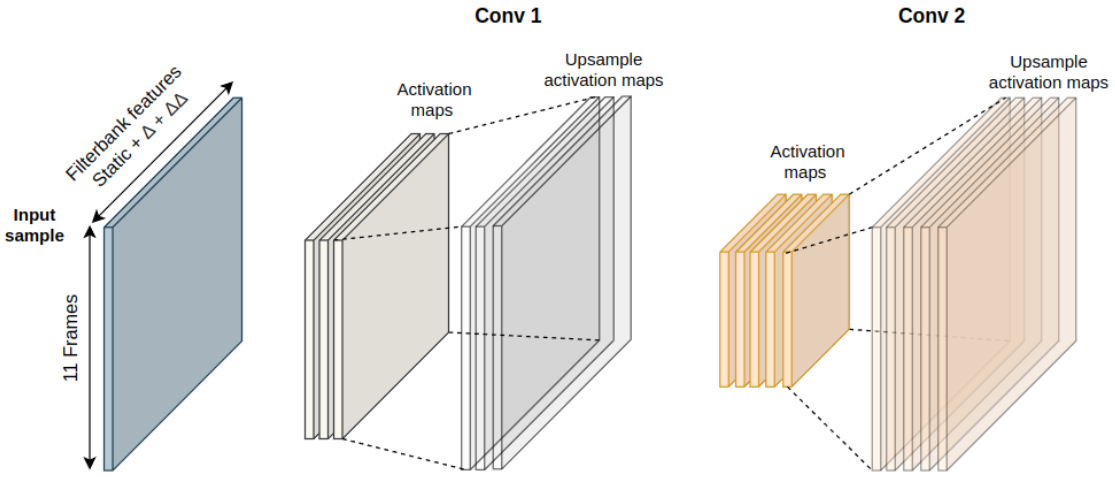
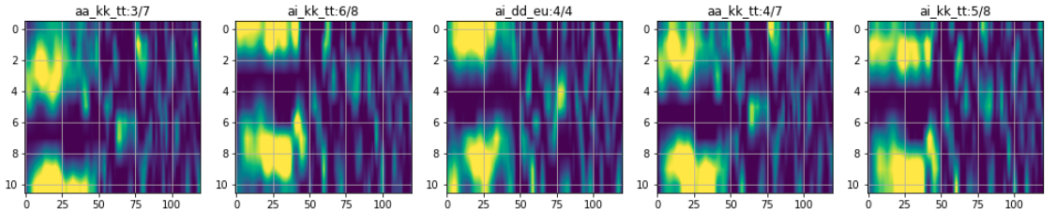


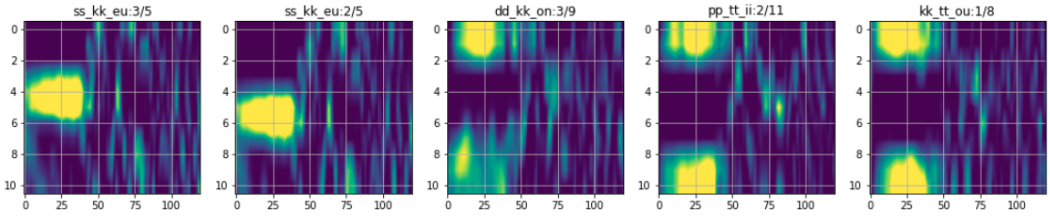
Figure C.5: *Illustration of the activation maps extraction and upsampling of one input sample*

input samples. A scalar is then calculated for each activation map corresponding to the sum of its activation values. Consequently, a sorting of the activation maps in descending order is performed based on their sum values. In order to visualize the activation maps, a Viridis color scale is used where dark blue represents zero activation value and yellow indicates the maximum activation value. For each filter, the maximum value is set equal to the highest activation value among all activation maps obtained from the particular filter in question. Figure C.6 illustrates the visualization of the top five activation maps per filter selected from the first convolutional layer. On top of each activation map, we added extra information about the sample following the pattern:
"Preceding-phoneme_Current-phoneme_Succeeding-phoneme : rank of the central frame in the current phoneme segment / length of the current phoneme segment (in number of frames)".
We are able to observe that, depending on the filter examined, some regions of the input are most activated for a given sample. It is therefore possible to explain this region based on the temporal dimension (i.e. frames in Y-axis). For instance, in C.6b, we can see in the first two activation maps that the region that is most activating filter 23 corresponds to the frames related

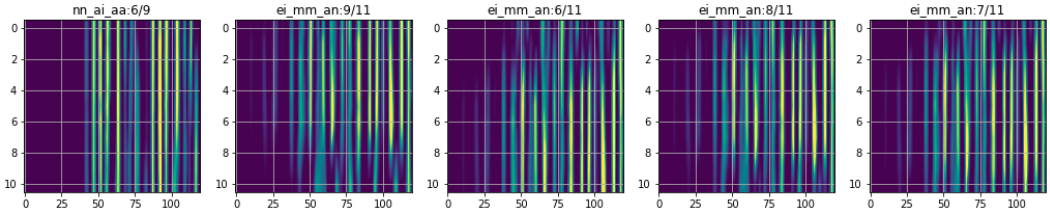
to the phoneme /k/. In the top activation map of the same filter, we mention the presence of two regions highly activated. Based on the extra information provided, we know that the upper part of the activation map corresponds to frames from the preceding phoneme /k/, and then from the central frame we have frames belonging to the phoneme /t/. Therefore, we can say that filter 23 detected the presence of the two phonemes /t/ and /k/. On the other hand, unlike filters 17 and 23, we can mention that filter 29 is less visibly interpretable since high activation regions highlight frequency bands (X-axis) and not the time domain. Indeed, this refers to the idea we introduced in chapter 4, revealing that we are not in the case of a fully interpretable input domain, to perform direct post-hoc interpretability.



(a) Filter 17



(b) Filter 23



(c) Filter 29

Figure C.6: *Top 5 activation maps from filters belonging to the first convolution layer*

Now, if we go more into detail about the top 1 activation map in filter 23, this filter did not simply detect the phoneme /t/, but it detected the frames of /t/ starting from the 4^{th} frame in the phoneme segment. So could the encoded representation be a finer characterization reflecting the phoneme production process for example? Indeed, the phoneme /t/ is a plosive, and plosives (i.e. /p/, /t/, /k/, /b/ /d/, /g/) are typically analyzed as having three phases: (1) *an occlusion/closing phase* during which the articulators are positioned, (2) *a holding phase* during which the air is blocked and (3), *an explosion phase (or burst)* which corresponds to the relaxing of the articulators and the liberation of the air stream. In our example, a possible explanation is that the filter detects the burst phase. This however remains an assumption that needs to be analyzed further if we want to take this filter into consideration for later interpretability.

C.3.2 Summary

To conclude, in this section, we performed the visualization of the activation maps issued from the convolutional layers of our CNN-based model for phoneme classification. We illustrated some examples of visualization in which we were able to observe that, depending on the filter examined, some regions of the input are most activated for a given sample. We have shown that these regions can provide valuable insights about the nature of the feature detected by the filter in question, once they appear in the interpretable dimension. It has not been acted upon thus far, but the explainability of these filters is certainly something we could consider if we want to further take advantage of the internal representations of phonemes in our clinical context.

C.4 Conclusion

In this appendix, we describe the supplementary methods that we investigated all along the second step of our proposed methodology. We elaborate on the approaches that we considered and describe the results and insights that we gained from these investigations. This appendix complements the main discussion in chapter 6.

Acronyms

AHN Aix Hôpital Neurologie.

AI Artificial Intelligence.

ALS Amyotrophic Lateral Sclerosis.

ANN Artificial Neural Network.

ANPS Artificial Neuron-based Phonological Similarity.

ASHA American Speech-Language-Hearing Association.

ASR Automatic Speech Recognition.

ASSIDS Assessment of Intelligibility of Dysarthric Speech.

BECD Batterie d'Evaluation Clinique de la Dysarthrie.

C2SI Carcinologic Speech Severity Index.

CA Cerebellar Ataxia.

CAPE-V Consensus Auditory Perceptual Evaluation of Voice.

CDSS Clinical Decision Support Systems.

CNN Convolutional Neural Network.

CSI Class Selectivity Index.

DAP Décodage Acoustico-Phonétique.

DES picture DEScription.

DL Deep Learning.

DNN Deep Neural Network.

ENT Ear, Nose & Throat.

FC Fully Connected layer.

FDA Frenchay Dysarthria Assessment.

FOC pragmatic FOCus.

GDPR General Data Protection Regulation.

GEPD Grille d'Evaluation Perceptive de la Dysarthrie.

GMM Gaussian Mixture Model.

Grad-CAM Gradient-weighted Class Activation Mapping.

GRBAS Grade, Roughness, Breathiness, Asthenia, Strain.

HC Healthy Control.

HMM Hidden Markov Model.

HNC Head and Neck Cancer.

INCa Institut National du Cancer.

Intel Intelligibility.

IPA International Phonetic Alphabet.

IUCT Institut Universitaire du Cancer Toulouse.

KNN K-Nearest Neighbors.

LEC LECture (passage reading).

LIME Local Interpretable Model-agnostic Explanations.

LRP Layer-wise Relevance Propagation.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MFCC Mel Frequency Cepstral Coefficients.

ML Machine Learning.

MOD MODality function.

MSE Mean Square Error.

NAP Neuron Activation Profile.

NCD Neuron-based Concept Detector.

NLP Natural Language Processing.

PD Parkinson Disease.

PPD Perceived Phonological Deviation.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

RQ Research Question.

Sev Severity.

SGD Stochastic Gradient Descent.

SHAP SHapley Additive exPlanations.

SLP Speech and Language Pathologist.

SNN Shallow Neural Network.

SVM Support Vector Machine.

SVT Sentence Verification Tasks.

SYN SYNtactic disambiguation.

t-SNE t-Distributed Stochastic Neighbour Embedding.

Tanh hyperbolic Tangent.

TDNN Time-Delay Neural Network.

TNM Tumor/Node/Metastasis.

UICC Union for International Cancer Control.

UPDRS Unified Parkinson's Disease Rating Scale.

WHO World Health Organization.

XAI Explainable Artificial Intelligence.

Personal Bibliography

[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. In *Proc. Interspeech 2020*, pages 2522–2526, Shanghai, China.

[Abderrazek et al., 2021] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2021). Classification de phonèmes à base d'apprentissage profond - Projection dans un contexte de parole dégradée. In *Séminaire AFCP – Phonétique Clinique*.

[Abderrazek et al., 2022a] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2022a). Interprétation des représentations profondes des traits phonétiques via l'approche NCD -Neuro-based Concept Detector : Application aux troubles de la parole. In *Journées d'Études sur la Parole - JEP*, Île de Noirmoutier, France.

[Abderrazek et al., 2022b] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2022b). Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders step 2: Contribution of the emergence of phonetic traits. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7387–7391, Singapore.

[Abderrazek et al., 2022c] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2022c). Validation of the Neuro-Concept Detector framework for the characterization of speech disorders: A comparative study including Dysarthria and Dysphonia. In *Proc. Interspeech 2022*, pages 3638–3642, Incheon, Korea.

[Abderrazek et al., 2023a] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2023a). Interpreting Deep Representations of Phonetic Features via Neuro-Based Concept Detector: Application to Speech Disorders Due to Head and Neck Cancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:200–214.

[Abderrazek et al., 2023b] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2023b). Vers une interprétation, en terme d'altération de traits phonétiques, du niveau d'intelligibilité chez des patients atteints de cancer de la tête ou du cou. In *9ème Journées de Phonétique Clinique (JPC'9)*.

Bibliography

[Jak, 1951] (1951). *Preliminaries to speech analysis*. MIT Press, Cambridge.

[Aal et al., 2021] Aal, H., Taie, S., and El-Bendary, N. (2021). An optimized rnn-lstm approach for parkinson's disease early detection using speech features. *Bulletin of Electrical Engineering and Informatics*, 10(5):2503–2512.

[Abdel-Hamid et al., 2014] Abdel-Hamid, O., Mohamed, A.-r., 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, 2019] Abderrazek, S. (2019). Intelligibility assessment of disordered speech using deep learning approaches. In *Intership report*, pages 1–65.

[Abdul et al., 2018] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–18, New York, NY, USA. Association for Computing Machinery.

[Ackermann and Ziegler, 1992] Ackermann, H. and Ziegler, W. (1992). Articulatory deficits in parkinsonian dysarthria: An acoustic analysis. *Journal of neurology, neurosurgery, and psychiatry*, 54:1093–8.

[Acs et al., 2020] Acs, B., Rantalainen, M., and Hartman, J. (2020). Artificial intelligence as the next step towards precision pathology. *Journal of internal medicine*, 288(1):62–81.

[Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

[AIA, 2021] AIA (2021). European Commission. Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act); COM (2021) 206 final; European Commission: Brussels, Belgium, 2021.

[Alhussein and Muhammad, 2018] Alhussein, M. and Muhammad, G. (2018). Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access*, 6:41034–41041.

[Alvarez Melis and Jaakkola, 2018] Alvarez Melis, D. and Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. In Bengio, S., Wallach, H., Larochelle,

H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Amann et al., 2020] Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20.

[Amari, 1993] Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.

[Amorim et al., 2020] Amorim, J. P., Abreu, P. H., Reyes, M., and Santos, J. (2020). Interpretability vs. complexity: The friction in deep neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

[An et al., 2015] An, G., Brizan, D. G., Ma, M., Morales, M., Syed, A. R., and Rosenberg, A. (2015). Automatic recognition of unified parkinson's disease rating from speech with acoustic, i-vector and phonotactic features. In *Proceedings of Interspeech'15*, Dresden, Germany.

[Antoniadi et al., 2021] Antoniadi, A. M., Galvin, M., Heverin, M., Hardiman, O., and Mooney, C. (2021). Development of an explainable clinical decision support system for the prediction of patient quality of life in amyotrophic lateral sclerosis. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 594–602, New York, NY, USA. Association for Computing Machinery.

[Arjmandi and Pooyan, 2012] Arjmandi, M. K. and Pooyan, M. (2012). An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomedical Signal Processing and Control*, 7(1):3–19. Human Voice and Sounds: From Newborn to Elder.

[Aronson and Bless, 2009] Aronson, A. and Bless, D. (2009). *Clinical Voice Disorders*. Thieme Publishers Series. Thieme.

[Auzou, 2007] Auzou, P. (2007). Définition et classifications des dysarthries. *Les dysarthries, édition Solal*, Neurophysiologie et production de la parole, Part III(31):308–323.

[Auzou and Rolland-Monnoury, 2006] Auzou, P. and Rolland-Monnoury, V. (2006). *Batterie d'évaluation clinique de la dysarthrie*. Édition Ortho.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

[Badley, 1993] Badley, E. M. (1993). An introduction to the concepts and classifications of the international classification of impairments, disabilities, and handicaps. *Disabil Rehabil*, 15(4):161–178.

[Balaguer, 2021] Balaguer, M. (2021). *Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé*. Theses, Université Paul Sabatier - Toulouse III.

[Balaguer et al., 2019] Balaguer, M., Boisguerin, A., Galtier, A., Gaillard, N., Puech, M., and Woisard, V. (2019). Factors influencing intelligibility and severity of chronic speech disorders of patients treated for oral or oropharyngeal cancer. *Eur. Arch. Otorhinolaryngol.*, 276(6):1767–1774.

[Barredo Arrieta et al., 2020] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

[Barrett et al., 2004] Barrett, W. L., Gluckman, J. L., Wilson, K. M., and Gleich, L. L. (2004). A comparison of treatments of squamous cell carcinoma of the base of tongue: surgical resection combined with external radiation therapy, external radiation therapy alone, and external radiation therapy combined with interstitial radiation. *Brachytherapy*, 3(4):240–245.

[Bau et al., 2020] Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*.

[Bernal et al., 2019] Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., and Lladó, X. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial intelligence in medicine*, 95:64–81.

[Bhat et al., 2018] Bhat, C., Das, B., Vachhani, B., and Kopparapu, S. K. (2018). Dysarthric speech recognition using time-delay neural network based denoising autoencoder. In *Proc. Interspeech 2018*, pages 451–455.

[Biran and Cotton, 2017] Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13.

[Bjerring and Busch, 2021] Bjerring, J. and Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy Technology*, 34:349–371.

[Blumstein, 1998] Blumstein, S. E. (1998). 5 - phonological aspects of aphasia. In Taylor Sarno, M., editor, *Acquired Aphasia (Third Edition)*, pages 157–185. Academic Press, San Diego, third edition edition.

[Boone et al., 2005] Boone, D., McFarlane, S., and Von Berg, S. (2005). *The Voice and Voice Therapy*. Pearson/Allyn & Bacon.

[Carlson et al., 2009] Carlson, N. R., Heth, D., Miller, H., Donahoe, J., and Martin, G. N. (2009). *Psychology: the science of behavior.* Pearson.

[Caruana et al., 2015] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA. Association for Computing Machinery.

[Carvalho et al., 2019] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8).

[Chaki and Woźniak, 2023] Chaki, J. and Woźniak, M. (2023). Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review. *Biomedical Signal Processing and Control*, 80:104223.

[Chaturvedi et al., 2011] Chaturvedi, A. K., Engels, E. A., Pfeiffer, R. M., Hernandez, B. Y., Xiao, W., Kim, E., Jiang, B., Goodman, M. T., Sibug-Saber, M., Cozen, W., Liu, L., Lynch, C. F., Wentzensen, N., Jordan, R. C., Altekruse, S., Anderson, W. F., Rosenberg, P. S., and Gillison, M. L. (2011). Human papillomavirus and rising oropharyngeal cancer incidence in the united states. *Journal of Clinical Oncology*, 29(32):4294–4301. PMID: 21969503.

[Chen et al., 2021] Chen, L., Wang, C., Chen, J., Xiang, Z., and Hu, X. (2021). Voice disorder identification by using hilbert-huang transform (hht) and k nearest neighbor (knn). *Journal of Voice*, 35(6):932.e1–932.e11.

[Chen et al., 2007] Chen, W., Peng, C., Zhu, X., Wan, B., and Wei, D. (2007). SVM-based identification of pathological voices. In *29th Annual International Conference of the IEEE*, pages 3786–3789.

[Chibelushi et al., 2002] Chibelushi, C., Deravi, F., and Mason, J. (2002). A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37.

[Chomsky and Halle, 1968] Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper and Row, New York.

[Chow, 2020] Chow, L. Q. (2020). Head and neck cancer. *New England Journal of Medicine*. PMID: 31893516.

[Chowdhury et al., 2020] Chowdhury, S. A., Ali, A., Shon, S., and Glass, J. (2020). What Does an End-to-End Dialect Identification Model Learn About Non-Dialectal Information? In *Proc. Interspeech 2020*, pages 462–466.

[Chowdhury et al., 2021] Chowdhury, S. A., Durrani, N., and Ali, A. (2021). What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis.

[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. In *Proceedings of Interspeech'12*, Portland, USA.

[Cimpian and Salomon, 2014] Cimpian, A. and Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, 37(5):461–480.

[Corny et al., 2020] Corny, J., Rajkumar, A., Martin, O., Dode, X., Lajonchère, J.-P., Billuart, O., Bézie, Y., and Buronfosse, A. (2020). A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error. *J Am Med Inform Assoc*, 27:1695–1704.

[Costa et al., 2008] Costa, S. C., Aguiar Neto, B. G., and Fechine, J. M. (2008). Pathological voice discrimination using cepstral analysis, vector quantization and hidden markov models. In *2008 8th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–5.

[Crameri et al., 2020] Crameri, F., Shephard, G., and Heron, P. (2020). The misuse of colour in science communication. *Nature Communications*, 11.

[Cutillo et al., 2020] Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K. D., Shabesta, T. E. C. K. V. R. B. N. B. C. C. G. G. J., Beck, T., Collier, E., Colvis, C. M., Gersing, K., Gordon, V., Jensen, R., Shabestari, B. J., and Southall, N. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital Medicine*, 3.

[Dahmani and Guerti, 2017] Dahmani, M. and Guerti, M. (2017). Vocal folds pathologies classification using naïve bayes networks. In *2017 6th International Conference on Systems and Control (ICSC)*, pages 426–432.

[Dahmani and Guerti, 2018] Dahmani, M. and Guerti, M. (2018). Glottal signal parameters as features set for neurological voice disorders diagnosis using k-nearest neighbors (knn). In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–5. IEEE.

[Dalvi et al., 2019] Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., and Glass, J. (2019). What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. AAAI'19/IAAI'19/EAAI'19. AAAI Press.

[Darley et al., 1969a] Darley, F. L., Aronson, A. E., and Brown, J. R. (1969a). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research*, 12:462–496.

[Darley et al., 1969b] Darley, F. L., Aronson, A. E., and Brown, J. R. (1969b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12:246–269.

[de Fine Licht and de Fine Licht, 2020] de Fine Licht, K. and de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI Soc.*, 35(4):917–926.

[Denman et al., 2019] Denman, D., Kim, J.-H., Munro, N., Speyer, R., and Cordier, R. (2019). Describing language assessments for school-aged children: A delphi study. *International Journal of Speech-Language Pathology*, 21:1–11.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

[Doyle et al., 1997] Doyle, P. C., Leeper, H., Kotler, A.-L., Thomas-Stonell, N., O'Neill, C., Dylke, M.-C., and Rolls, K. (1997). Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, 34(3):309–316.

[Du et al., 2022] Du, Y., Antoniadi, A. M., McNestry, C., McAuliffe, F. M., and Mooney, C. (2022). The role of xai in advice-taking from a clinical decision support system: A comparative user study of feature contribution-based and example-based explanations. *Applied Sciences*, 12(20).

[Durrani et al., 2020] Durrani, N., Sajjad, H., Dalvi, F., and Belinkov, Y. (2020). Analyzing individual neurons in pre-trained language models. pages 4865–4880.

[Dwivedi et al., 2009] Dwivedi, R. C., Kazi, R. A., Agrawal, N., Nutting, C. M., Clarke, P. M., Kerawala, C. J., Rhys-Evans, P. H., and Harrington, K. J. (2009). Evaluation of speech outcomes following treatment of oral and oropharyngeal cancers. *Cancer Treat Rev*, 35(5):417–424.

[Ebers et al., 2021] Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., and Steinrötter, B. (2021). The european commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and ai law society (rails). *J*, 4(4):589–603.

[Enderby, 1980] Enderby, P. (1980). Frenchay dysarthria assessment. *British Journal of Disorders of Communication*.

[Enderby, 1983] Enderby, P. (1983). Frenchay dysarthric assessment. *Pro-Ed, Texas*.

[Enderby and Palmer, 2008] Enderby, P. and Palmer, R. (2008). Fda-2: Frenchay dysarthria assessment. *(2nd ed), Pro-Ed, Tex*.

[España-Bonet and Fonollosa, 2016] España-Bonet, C. and Fonollosa, J. A. R. (2016). Automatic speech recognition with deep neural networks for impaired speech. In Abad, A., Ortega, A., Teixeira, A., García Mateo, C., Martínez Hinarejos, C. D., Perdigão, F., Batista, F., and Mamede, N., editors, *Advances in Speech and Language Technologies for Iberian Languages*, pages 97–107, Cham. Springer International Publishing.

[Fan et al., 2021] Fan, F.-L., Xiong, J., Li, M., and Wang, G. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760.

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

[Ferrier et al., 1992] Ferrier, L. J., Jarrell, N., Carpenter, T., and Shane, H. C. (1992). A case study of a dysarthric speaker using the dragondictate voice recognition system. *Journal for Computer Users in Speech and Hearing*, 8(1):33–52.

[Fonseca et al., 2005] Fonseca, E. S., Guido, R. C., Silvestre, A. C., and Pereira, J. C. (2005). Discrete wavelet transform and support vector machine applied to pathological voice signals identification. In *Seventh IEEE International Symposium on Multimedia (ISM'05)*, pages 785–789.

[Fontan et al., 2015] Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., and Ruiz, R. (2015). Relationship between speech intelligibility and speech comprehension in babble noise. *Journal of Speech, Language, and Hearing Research*, 58(3):977–986.

[Fougeron et al., 2010] Fougeron, C., Crevier-Buchman, L., Fredouille, C., Ghio, A., Meunier, C., Chevrie-Muller, C., Bonastre, J.-F., Colazo-Simon, A., Delooze, C., Duez, D., Gendrot, C., Legou, T., Lévêque, N., Pillot-Loiseau, C., Pinto, S., Pouchoulin, G., Robert, D., Vaissière, J., Viallet, F., and Vincent, C. (2010). The DesPho-APaDy project: Developing an acoustic-phonetic characterization of dysarthric speech in french. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10*, pages 2831–2838, Valletta, Malta.

[Fraile et al., 2009] Fraile, R., Saenz-Lechon, N., Godino-Llorente, J. I., Osma-Ruiza, V., and Fredouille, C. (2009). Automatic detection of laryngeal pathologies in records of sustained vowels by means of mfcc parameters and differentiation of patients by sex. *Folia phoniatrica et logopaedica, International Journal of Phoniatrics, Speech Therapy and Communication Pathology, Special issue: COST Action 2103 - A Joint European Project for Advanced Voice Assessment,*, 61(3):146–52.

[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. In *International Congress of Phonetic Sciences (ICPhS)*, pages 3051–3055.

[Galliano et al., 2005] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005). Ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of Interspeech'05*, pages 1149–1152, Lisboa, Portugal.

[Gandini et al., 2008] Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., and Boyle, P. (2008). Tobacco smoking and cancer: A meta-analysis. *International Journal of Cancer*, 122(1):155–164.

[GDPR, 2016] GDPR (2016). General Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation = GDPR), OJ 2016 L 119/1.

[Ghahremani et al., 2016] Ghahremani, P., Manohar, V., Povey, D., and Khudanpur, S. (2016). Acoustic modelling from the signal domain using cnns. In *Interspeech*.

[Ghio, 1997] Ghio, A. (1997). *Achile : un dispositif de décodage acoustico-phonétique et d'identification lexicale indépendant du locuteur à partir de modules mixtes*. Theses, Université d'Aix Marseille.

[Ghio et al., 2019] Ghio, A., Giusti, L., Blanc, E., and Pinto, S. (2019). French adaptation of the "frenchay dysarthria assessment 2" speech intelligibility test. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 137.

[Ghio et al., 2016] Ghio, A., Giusti, L., Blanc, E., Pinto, S., Lalain, M., Robert, D., Fredouille, C., and Woisard, V. (2016). Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ? In *Journées d'Etude sur la Parole, JEP'16*.

[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. In *12th Conference on Language Resources and Evaluation (LREC)*, France.

[Ghio et al., 2021] Ghio, A., Lalain, M., Rebourg, M., Marczyk, A., Fredouille, C., and Woisard, V. (2021). Validation of an intelligibility test based on acoustic-phonetic decoding of pseudo-words: overall results from patients with cancer of the oral cavity and the oropharynx. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*.

[Ghio et al., 2013] Ghio, A., Revis, J., Mérienne, S., and Giovanni, A. (2013). Top-down mechanisms in dysphonia perception: the need for blind tests. *Journal of voice : official journal of the Voice Foundation*, 27 4:481–5.

[Gilpin et al., 2018] Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. pages 80–89.

[Glackin et al., 2018] Glackin, C., Wall, J. A., Chollet, G., Dugan, N., and Cannings, N. (2018). Convolutional neural networks for phoneme recognition. In *International Conference on Pattern Recognition Applications and Methods*.

[Godino-Llorente et al., 2006] Godino-Llorente, J. I., Gomez-Vilda, P., and Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *IEEE Transaction on Biomedical Engineering*, 53(10):1943–1953.

[Gonzalez-Garcia et al., 2018] Gonzalez-Garcia, A., Modolo, D., and Ferrari, V. (2018). Do semantic parts emerge in convolutional neural networks? *International Journal of Computer Vision*, 126:476–494.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[Gordon and Ladefoged, 2001] Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406.

[Green et al., 2021] Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M. A., Tobin, J., Brenner, M. P., et al. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Interspeech*, pages 4778–4782.

[Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

[Hahm et al., 2015] Hahm, S., Heitzman, D., and Wang, J. (2015). Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization. In *6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54, Dresden, Germany.

[Hashibe et al., 2007] Hashibe, M., Brennan, P., Benhamou, S., Castellsague, X., Chen, C., Curado, M. P., Maso, L. D., Daudt, A. W., Fabianova, E., Wünsch-Filho, V., Franceschi, S., Hayes, R. B., Herrero, R., Koifman, S., La Vecchia, C., Lazarus, P., Levi, F., Mates, D., Matos, E., Menezes, A., Muscat, J., Eluf-Neto, J., Olshan, A. F., Rudnai, P., Schwartz, S. M., Smith, E., Sturgis, E. M., Szeszenia-Dabrowska, N., Talamini, R., Wei, Q., Winn, D. M., Zaridze, D., Zatonski, W., Zhang, Z.-F., Berthiller, J., and Boffetta, P. (2007). Alcohol Drinking in Never Users of Tobacco, Cigarette Smoking in Never Drinkers, and the Risk of Head and Neck Cancer: Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium. *JNCI: Journal of the National Cancer Institute*, 99(10):777–789.

[Hirano, 1981] Hirano, M. (1981). Psycho-acoustic evaluation of voice : GRBAS scale for evaluating the hoarse voice. *Clinical Examination of voice, Springer Verlag*.

[Hodge and Whitehill, 2010] Hodge, M. and Whitehill, T. (2010). *Intelligibility Impairments*, chapter 4, pages 99–114. John Wiley Sons, Ltd.

[Hoq et al., 2021] Hoq, M., Uddin, M. N., and Park, S.-B. (2021). Vocal feature extraction-based artificial intelligent model for parkinson's disease detection. *Diagnostics*, 11(6).

[Hustad, 2008] Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language and Hearing Research*, 51(3):562–573.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org.

[Janbakhshi et al., 2019] Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *ICASSP'19*, UK.

[Jany-Luig, 2017] Jany-Luig, J. (2017). *Prosodic and Paralinguistic Speech Parameters for the Identication of Emotions and Stress*. PhD thesis.

[Karlik and Olgac, 2011] Karlik, B. and Olgac, A. V. (2011). Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122.

[Keintz et al., 2007] Keintz, C. K., Bunton, K., and Hoit, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, 16(3):222–234.

[Kempster et al., 2009] Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., and Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2):124–132.

[Kent et al., 1990] Kent, R., Kent, J., Weismer, G., Sufit, R., Rosenbek, J., Martin, R., and Brooks, B. (1990). Impairment of speech intelligibility in men with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Disorders*, 55(4):721–728. Copyright: Copyright 2018 Elsevier B.V., All rights reserved.

[Kent, 1992] Kent, R. D. (1992). *Intelligibility in speech disorders: Theory, measurement and management*, volume 1. John Benjamins Publishing.

[Kent et al., 1975] Kent, R. D., Netsell, R., and Bauer, L. L. (1975). Cineradiography assessment of articulatory mobility in the dysarthrias. *Journal of Speech and Hearing Disorders*, 40(4):467–480.

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

[Kim et al., 2016] Kim, B., Khanna, R., and Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2288–2296, Red Hook, NY, USA. Curran Associates Inc.

[Kim et al., 2015] Kim, J., Kumar, N., Tsiartas, A., Li, M., and Narayanan, S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech Language*.

[Kim and Kim, 2012] Kim, M. J. and Kim, H. (2012). Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility. In *Proc. Interspeech*.

[Klumpp et al., 2022] Klumpp, P., Arias-Vergara, T., Vásquez-Correa, J. C., Pérez-Toro, P. A., Orozco-Arroyave, J. R., Batliner, A., and Nöth, E. (2022). The phonetic footprint of parkinson's disease. *Computer Speech Language*, 72:101321.

[Korzekwa et al., 2019] Korzekwa, D., Barra-Chicote, R., Kostek, B., Drugman, T., and Lajszczak, M. (2019). Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech. In *Proc. Interspeech 2019*, pages 3890–3894.

[Krug et al., 2021] Krug, A., Ebrahimzadeh, M., Alemann, J., Johannsmeier, J., and Stober, S. (2021). Analyzing and visualizing deep neural networks for speech recognition with saliency-adjusted neuron activation profiles. *Electronics*.

[Krug et al., 2018] Krug, A., Knaebel, R., and Stober, S. (2018). Neuron activation profiles for interpreting convolutional speech recognition models. In *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*.

[Kundu, 2021] Kundu, S. (2021). Ai in medicine must be explainable. *Nature Medicine*, 27:1–1.

[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. In *Interspeech*, India.

[Laaridh et al., 2015] Laaridh, I., Fredouille, C., and Meunier, C. (2015). Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on Accessible Computing*, 6(3).

[Ladefoged, 1971] Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics*. Midway reprints. University of Chicago Press.

[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*.

[Lamel et al., 1991] Lamel, L. F., Gauvain, J. L., and Eskénazi, M. (1991). BREF, a large vocabulary spoken corpus for french. In *Eurospeech'91*, Italy.

[Langer et al., 2021] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473.

[Lapuschkin et al., 2019] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10.

[LeCun and Bengio, 1998] LeCun, Y. and Bengio, Y. (1998). *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, Cambridge, MA, USA.

[LeCun et al., 2010] LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE.

[Lhoussaine, 2012] Lhoussaine, L. (2012). Première validation de la grille d'évaluation perceptive de la dysarthrie (g.e.p.d.) : effet du niveau d'expertise du jury et différenciation entre types de dysarthrie.

[Liberman et al., 1967] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74(6):431.

[Linardatos et al., 2021] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1).

[Lindblom, 1990] Lindblom, B. (1990). On the communication process: Speaker-listener interaction and the development of speech*. *Augmentative and Alternative Communication*, 6(4):220–230.

[Lipton, 2018] Lipton, Z. C. (2018). The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

[Liu and Demosthenes, 2022] Liu, F. and Demosthenes, P. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, 22(1):1–10.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

[Maier et al., 2010] Maier, A., Haderlein, T., Stella, F., North, E., Nkenke, E., Rosanowski, F., Schutzenberger, A., and Schuster, M. (2010). Automatic speech recognition system for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech and Music Processing*, 2010.

[Maier et al., 2007] Maier, A., Schuster, M., Batliner, A., Noeth, E., and Nkenke, E. (2007). Automatic scoring of the intelligibility in patients with cancer of the oral cavity. volume 2, pages 1206–1209.

[Malakar and Keskar, 2021] Malakar, M. and Keskar, R. B. (2021). Progress of machine learning based automatic phoneme recognition and its prospect. *Speech Communication*, 135:37–53.

[Malle, 2011] Malle, B. (2011). *Attribution theories: How people make sense of behavior*, pages 72–95.

[Markaki and Stylianou, 2011] Markaki, M. and Stylianou, Y. (2011). Voice pathology detection and discrimination based on modulation spectral features. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1938–1948.

[Martínez et al., 2015] Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., and Miguel, A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)*, 6(3):10.

[Meunier et al., 2016] Meunier, C., Fougeron, C., Fredouille, C., Biggi, B., Crevier-Buchman, L., Delais-Roussarie, E., Georgeton, L., Ghio, A., Laaridh, I., Legou, T., Pillot-Loiseau, C., and Pouchoulin, G. (2016). The typaloc corpus: A collection of various dysarthric speech recordings in read and spontaneous styles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC'16*, Portoroz, Slovenia.

[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. *Laryngoscope*, 114(11):1977–1981.

[Meyes et al., 2019] Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

[Middag et al., 2009] Middag, C., Martens, J.-P., Nuffelen, G. V., and Bodt, M. D. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Applied Signal Processing*, 2009(1).

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[Molnar, 2022] Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

[Montavon et al., 2017] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

[Montavon et al., 2018] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

[Montufar et al., 2014] Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27.

[Morcos et al., 2018] Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., and Botvinick, M. (2018). On the importance of single directions for generalization. In *ICLR*.

[Nagamine et al., 2015] Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. In *Proc. Interspeech 2015*, pages 1912–1916.

[Namatēvs et al., 2022] Namatēvs, I., Sudars, K., and Dobrājs, A. (2022). Interpretability versus explainability: Classification for understanding deep learning systems and models. *Computer Assisted Methods in Engineering and Science*, 29(4):297–356.

[Narendra and Alku, 2021] Narendra, N. and Alku, P. (2021). Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. *Computer Speech Language*, 65.

[Okabe et al., 2018] Okabe, K., Koshinaka, T., and Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. pages 2252–2256.

[Orozco-Arroyave et al., 2016] Orozco-Arroyave, J., Vdsquez-Correa, J., Arias-Londo, J., Vargas-Bonilla, J., Skodda, S., Rusz, J., Noth, E., et al. (2016). Towards an automatic monitoring of the neurological state of parkinson's patients from speech. In *ICASSP'16*, China.

[Palaz et al., 2013] Palaz, D., Collobert, R., and Magimai.-Doss, M. (2013). Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In *Interspeech*.

[Palaz et al., 2015] Palaz, D., Magimai-Doss, M., and Collobert, R. (2015). Analysis of CNN-based speech recognition system using raw speech as input. In *Proc. Interspeech 2015*, pages 11–15.

[Palmer and Enderby, 2007] Palmer, R. and Enderby, P. (2007). Methods of speech therapy treatment for stable dysarthria: A review. *Advances in Speech Language Pathology*, 9(2):140–153.

[Passricha and Aggarwal, 2018] Passricha, V. and Aggarwal, R. (2018). *Convolutional Neural Networks for Raw Speech Recognition*.

[Pellegrini and Mouysset, 2016] Pellegrini, T. and Mouysset, S. (2016). Inferring Phonemic Classes from CNN Activation Maps Using Clustering Techniques. In *Proc. Interspeech 2016*, pages 1290–1294.

[Ploug and Holm, 2020] Ploug, T. and Holm, S. (2020). The four dimensions of contestable ai diagnostics- a patient-centric approach to explainable ai. *Artificial Intelligence in Medicine*, 107:101901.

[Politi et al., 2013] Politi, M. C., Dizon, D. S., Frosch, D. L., Kuzemchak, M. D., and Stiggelbout, A. M. (2013). Importance of clarifying patients' desired role in shared decision making to match their level of engagement with their preferences. *BMJ*, 347.

[Poliyev and Korsun, 2020] Poliyev, A. V. and Korsun, O. N. (2020). Speech recognition using convolutional neural networks on small training sets. *IOP Conference Series: Materials Science and Engineering*, 714(1):012024.

[Pommée et al., 2021] Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., and Woisard, V. (2021). Assessment of adult speech disorders: current situation and needs in French-speaking clinical practice. *Logopedics Phoniatrics Vocology*, pages 1–15.

[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 and Communication Disorders*, 57(1):21 – 41.

[Pommée et al., 2021] Pommée, T., Balaguer, M., Pinquier, J., Mauclair, J., Woisard, V., and Speyer, R. (2021). Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech: a systematic review. *Speech, Language and Hearing*, 24(2):105–132.

[Pouchoulin et al., 2007] Pouchoulin, G., Fredouille, C., Bonastre, J.-F., Ghio, A., and Giovanni, A. (2007). Frequency study for the characterization of the dysphonic voices. In *Proceedings of Interspeech'07*, Antwerp, Belgium.

[Qian et al., 2016] Qian, P., Qiu, X., and Huang, X. (2016). Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, Austin, Texas. Association for Computational Linguistics.

[Quan et al., 2021] Quan, C., Ren, K., and Luo, Z. (2021). A deep learning based method for parkinson's disease detection using dynamic features of speech. *IEEE Access*, 9:10239–10252.

[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. In *Interspeech*, China.

[Rafegas et al., 2020] Rafegas, I., Vanrell, M., Alexandre, L. A., and Arias, G. (2020). Understanding trained cnns by indexing neuron selectivity. *Pattern Recognition Letters*, 136:318–325.

[Ragab et al., 2022] Ragab, M., Albukhari, A., Alyami, J., and Mansour, R. F. (2022). Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images. *Biology*, 11.

[Rajpurkar et al., 2022] Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. (2022). Ai in health and medicine. *Nature Medicine*, 28.

[Rebourg, 2022] Rebourg, M. (2022). *Evaluation de l'intelligibilité après un cancer ORL : Approche perceptive par décodage acoustico-phonétique et mesures acoustiques*. PhD thesis, Aix-Marseille Université - Laboratoire Parole et Langage.

[Revis, 2004] Revis, J. (2004). *L'analyse perceptive des dysphonies : approche phonétique de l'évaluation vocale*. PhD thesis, Université de la Méditerranée.

[Ribeiro et al., 2016] Ribeiro, M., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

[Riedhammer et al., 2007] Riedhammer, K., Stemmer, G., Haderlein, T., Schuster, M., Rosanowski, F., Noth, E., and Maier, A. (2007). Towards robust automatic evaluation of pathologic telephone speech. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 717–722.

[Rizvi et al., 2020] Rizvi, D. R., Nissar, I., Masood, S., Ahmed, M., and Ahmad, F. (2020). An lstm based deep learning model for voice-based detection of parkinson's disease. *Int. J. Adv. Sci. Technol*, 29(8).

[Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

[Schuster et al., 2006] Schuster, M., Haderlein, T., Noeth, E., Lohscheller, J., Eysholdt, U., and Rosanowski, F. (2006). Intelligibility of laryngectomees' substitute speech: Automatic speech recognition and subjective rating. *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, 263:188–93.

[Schuster et al., 2005] Schuster, M., Noth, E., Haderlein, T., Steidl, S., Batliner, A., and Rosanowski, F. (2005). Can you understand him? let's look at his word accuracy-automatic evaluation of tracheoesophageal speech. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/61–I/64 Vol. 1.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Sharma et al., 2009] Sharma, H. V., Hasegawa-Johnson, M., Gunderson, J., and Perlman, A. (2009). Universal access: preliminary experiments in dysarthric speech recognition. In *Proceedings of Interspeech'09*, Brighton, United Kingdom.

[Sheikholeslami et al., 2021] Sheikholeslami, S., Meister, M., Wang, T., Payberah, A. H., Vlassov, V., and Dowling, J. (2021). Autoablation: Automated parallel ablation studies for deep learning. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, pages 55–61.

[Sidi Yakoub et al., 2020] Sidi Yakoub, M., Selouani, S.-a., Zaidi, B.-F., and Bouchair, A. (2020). Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. *EURASIP J. Audio Speech Music Process.*, 2020(1).

[Siegler and Stern, 1995] Siegler, M. and Stern, R. (1995). On the effects of speech rate in large vocabulary speech recognition systems. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 612–615 vol.1.

[Sobin et al., 2011] Sobin, L. H., Gospodarowicz, M. K., and Wittekind, C. (2011). *TNM classification of malignant tumours*. John Wiley & Sons.

[Springenberg et al., 2015] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A. (2015). Striving for simplicity: The all convolutional net. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

[Sutton et al., 2020] Sutton, R., Pincock, D., Baumgart, D., Sadowski, D., Fedorak, R., and Kroeker, K. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*.

[Talkar et al., 2020] Talkar, T., Williamson, J., Hannon, D., Rao, H., Yuditskaya, S., Claypool, K., Sturim, D., Nowinski, L., Saro, H., Stamm, C., Mody, M., McDougle, C., and Quatieri, T. (2020). Assessment of speech and fine motor coordination in children with autism spectrum disorder. *IEEE Access*, PP:1–1.

[Tanner, 2006] Tanner, D. (2006). *An Advanced Course in Communication Sciences and Disorders*. Plural Pub.

[Teston, 2007] Teston, B. (2007). L'étude instrumentale des gestes dans la production de la parole ; importance de l'aérophonométrie. *Les dysarthries, édition Solal*, Evaluation, Part II(24):248–258.

[Thagard, 1978] Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, 75(2):76–92.

[Tonekaboni et al., 2019] Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Machine Learning in Health Care*.

[Trinh and O'Brien, 2019] Trinh, N. H. and O'Brien, D. (2019). Pathological speech classification using a convolutional neural network. In *IMVIP 2019: Irish Machine Vision Image Processing*.

[Tripathi et al., 2020] Tripathi, A., Bhosale, S., and Kopparapu, S. K. (2020). A novel approach for intelligibility assessment in dysarthric subjects. In *ICASSP'20*, Spain.

[Tu et al., 2017] Tu, M., Berisha, V., and Liss, J. (2017). Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *Proc. Interspeech 2017*, pages 1849–1853.

[Umapathy et al., 2005] Umapathy, K., Krishnan, S., Parsa, V., and Jamieson, D. (2005). Discrimination of pathological voices using a time-frequency approach. *IEEE Transactions on Biomedical Engineering*, 52(3):421–430.

[van der Maas et al., 1990] van der Maas, H. L., Verschure, P. F., and Molenaar, P. C. (1990). A note on chaotic behavior in simple neural networks. *Neural Networks*, 3(1):119–122.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Viualizing data using t-sne. *Journal of Machine Learning Research*.

[Vavrek et al., 2021] Vavrek, L., Hires, M., Kumar, D., and Drotár, P. (2021). Deep convolutional neural network for detection of pathological speech. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000245–000250.

[Vellido, 2020] Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, pages 1–15.

[Vilone and Longo, 2021] Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76.

[Voulodimos et al., 2018] Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018:1–13.

[Vásquez-Correa et al., 2017] Vásquez-Correa, J., Orozco-Arroyave, J. R., and Nöth, E. (2017). Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease. In *Proc. Interspeech 2017*, pages 314–318.

[Vásquez-Correa et al., 2021] Vásquez-Correa, J. C., Rios-Urrego, C. D., Arias-Vergara, T., Schuster, M., Rusz, J., Nöth, E., and Orozco-Arroyave, J. R. (2021). Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. *Pattern Recognition Letters*, 150:272–279.

[Walsh, 2005] Walsh, R. (2005). Meaning and purpose: A conceptual model for speech pathology terminology. *Advances in Speech Language Pathology*, 7(2):65–76.

[Woisard et al., 2021] Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., et al. (2021). 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(1).

[Woisard et al., 2022] Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M., Puech, M., Astesano, C., Pinquier, J., and Lepage, B. (2022). Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: The carcinologic speech severity index. *Head & Neck*.

[Wood, 1971] Wood, K. S. (1971). *Terminology and nomenclature*.

[World Health Organization, 2021] World Health Organization (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*.

[Xu et al., 2023] Xu, L., Liss, J., and Berisha, V. (2023). Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, 3(1):015201.

[Yi et al., 2019] Yi, H., Leonard, M., and Chang, E. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, 102:1096–1110.

[Yorkston and Beukelman, 1981] Yorkston, K. M. and Beukelman, D. R. (1981). Assessment of intelligibility of dysarthric speech. *Tigard, OR: C.C. Publications*.

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

[Zaidi et al., 2021] Zaidi, B. F., Selouani, S. A., Boudraa, M., and Sidi Yakoub, M. (2021). Deep neural network architectures for dysarthric speech analysis and recognition. *Neural Comput. Appl.*, 33(15):9089–9108.

[Zech et al., 2018] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):1–17.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

[Zhou et al., 2018] Zhou, B., Bau, D., Oliva, A., and Torralba, A. (2018). Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145.

[Zhou et al., 2015] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Object detectors emerge in deep scene cnns.