Novel Intelligent Approaches for Analyzing Speech Intelligibility and Dysarthria Severity

Doctoral dissertation prepared and presented by

Sally ISMAIL

Thesis directed by **Mehdi AMMI** and co-supervised by **Margarita ANASTASSOVA**, **Mehdi BOUKALLEL**, and **Christian BOLZMACHER**

Prepared at **CEA-LIST**
With l'École Doctorale 224 - **Cognition, Langage, Interaction (CLI), Université Paris 8 Vincennes - Saint-Denis**

Thesis Jury:
Dr. Ioana VASILESCU
CNRS Research Director, rapporteure
Prof. Ramiro VELAZQUEZ
Full Professor at Universidad Panamericana, rapporteur
Stéphane BOUILLAND
R&D Engineer at fondation Hopale
Dr. Margarita ANASTASSOVA
Head of the LISA lab, CEA
Prof. Mehdi AMMI
Professor at Universié PARIS 8

Date of defense: 25 August 2025

Contents

Contents 2

List of Figures 6

Abstract 10

Résumé 12

Résume Étendu 14

1 Introduction 26
1.1 An Overview of the Subject's Relevance . . . 26
1.2 Contributions of the Thesis . . . 27
1.3 Structure of the Thesis . . . 28

I State of the Art 30

2 Exploring Speech Pathology: Neurological Insights, Human Perception, and Clinical Dimensions 31
2.1 Introduction . . . 31
2.2 Speech Production Process . . . 32
2.3 Neurological disorders . . . 34
2.3.1 The nervous system . . . 34
2.3.2 Dysarthria . . . 37
2.3.3 Cerebral Palsy . . . 38
2.3.4 Parkinson's Diease . . . 40
2.4 Assessing Speech Impairments: A Focus on Dysarthria . . . 42
2.4.1 The Frenchay Dysarthria Assessment . . . 42
2.4.2 The Batterie d'évaluation clinique de la Dysarthrie . . . 43
2.5 Conclusion . . . 44

3 Signal, Features, and Models: A Technical Perspective 45
3.1 Introduction . . . 45
3.2 Basics of Sound and Audio Signals . . . 46
3.2.1 Pitch . . . 46
3.2.2 Intensity and power . . . 46
3.2.3 Timbre . . . 47

3.3 Digital Representation of Audio . . . 47
3.3.1 Sampling . . . 48
3.3.2 Quantization . . . 48
3.3.3 Audio Features . . . 48
3.4 Signal Processing Techniques for Speech Analysis . . . 50
3.4.1 Fourier Transform . . . 50
3.4.2 Discrete Fourier Transform . . . 50
3.4.3 Short-Time Fourier Transform . . . 51
3.4.4 Spectrogram . . . 52
3.4.5 Mel Spectrogram . . . 52
3.4.6 Mel-Frequency Cepstral Coefficients . . . 53
3.5 Machine Learning . . . 54
3.5.1 From AI to ML to DL: Understanding the Evolution of Intelligent Systems . . . 54
3.5.2 Training Techniques . . . 54
3.5.3 Classical Machine Learning Models . . . 55
3.6 Pre-trained Models . . . 57
3.6.1 Wav2vec2.0 . . . 57
3.6.2 HuBERT . . . 60
3.6.3 Whisper . . . 65
3.6.4 Audio Spectrogram Transformer . . . 68
3.7 Conclusion . . . 71

4 A Bibliographic Review of Dysarthria Detection and Classification 72
4.1 Introduction . . . 72
4.2 Dysarthria Classification Approaches . . . 72
4.2.1 Limitations of the Traditional Approach . . . 73
4.2.2 Limitations of the end-to-end approach . . . 74
4.3 Data partition . . . 76
4.3.1 Speaker-dependent . . . 76
4.3.2 Speaker-independent . . . 76
4.4 Features For Dysarthria Severity Classification . . . 76
4.4.1 Audio-based features . . . 77
4.4.2 The Geneva Minimalistic Acoustic Parameter Set . . . 77
4.5 Dysarthria Classification Methods . . . 78
4.6 Conclusion . . . 83

II Experimental Work 84

5 Overview of Datasets: Publicly Available and Newly Collected Data 85
5.1 Introduction . . . 85
5.2 Data related challenges when considering data-driven models . . . 85
5.3 Public datasets . . . 86
5.3.1 UASpeech . . . 86
5.3.2 TORGO . . . 88
5.4 VocaPark . . . 90

5.4.1 Equipment and recording procedure . . . 90
5.4.2 Speakers Distribution . . . 93
5.5 Conclusion . . . 94

6 Is Attention All You Need for Dysarthria Severity Classification ? 97
6.1 Introduction . . . 97
6.2 Proposed Method . . . 98
6.2.1 Feature Extraction . . . 98
6.2.2 Classifier . . . 99
6.2.3 Dataset . . . 99
6.2.4 Training Settings . . . 100
6.2.5 Evaluation Metrics . . . 101
6.3 Results . . . 101
6.3.1 Speaker-Dependent Scenario . . . 101
6.3.2 Speaker-Independent Scenario . . . 102
6.4 Discussion and Comparison with Related Work . . . 102
6.5 Conclusion . . . 105

7 Leveraging Pre-Trained Models for Dysarthria Severity Detection: A Comparative Analysis 107
7.1 Introduction . . . 107
7.2 Pre-trained models . . . 108
7.2.1 Waveform-Based Models . . . 108
7.2.2 Spectrogram-Based Models . . . 109
7.3 Experimental Setup . . . 109
7.3.1 Datasets . . . 109
7.3.2 Baseline Features . . . 110
7.3.3 Pre-trained models as feature extractors . . . 110
7.3.4 Classifiers . . . 110
7.3.5 Evaluation Metrics . . . 111
7.4 Results and Discussion . . . 111
7.5 Comparison with Related Work . . . 114
7.6 Conclusion . . . 115

8 Extending the Use of Pre-Trained Models for Dysarthria Severity Classification 116
8.1 Introduction . . . 116
8.2 Pre-trained models . . . 117
8.3 Waveform-based Models . . . 117
8.3.1 Ablation study . . . 117
8.3.2 Datasets . . . 117
8.3.3 Classifier and Training Settings . . . 118
8.3.4 Results and Discussion . . . 118
8.4 Spectrogram-based Models . . . 122
8.4.1 Ablation study . . . 122
8.4.2 Datasets . . . 122
8.4.3 Classifier and Training Settings . . . 123

8.4.4 Results and Discussion . . . 123
8.5 Comparison with Related Work . . . 124
8.6 Conclusion . . . 126

9 Parkinson's Disease and Dysarthria: Adapting and Expanding Experiments on VocaPark 127
9.1 Introduction . . . 127
9.2 Baseline Experiments . . . 128
9.2.1 Feature Extraction . . . 128
9.2.2 Classifier . . . 128
9.2.3 Data Partition . . . 128
9.2.4 Training Settings . . . 129
9.2.5 Results and Discussion . . . 129
9.3 Leveraging Pre-trained Models . . . 132
9.3.1 Pre-trained Models . . . 133
9.3.2 Data Partition . . . 133
9.3.3 Classifier and Training Settings . . . 133
9.3.4 Results and Discussion . . . 133
9.4 Conclusion . . . 141

10 Exploring Alternative Approaches: Contrastive Pretraining and Autoencoders 143
10.1 Introduction . . . 143
10.2 Transformer-Based Model . . . 143
10.2.1 Using the Transformer-Based Model . . . 143
10.2.2 Improving Transformer Model Via Contrastive Pre-Training . . . 146
10.3 Auto-Encoder . . . 149
10.3.1 Proposed model . . . 150
10.3.2 Input Features . . . 151
10.3.3 Training settings . . . 151
10.3.4 Results and Discussion . . . 152
10.4 Results Summary and Comparison . . . 153
10.5 Conclusion . . . 154

11 Conclusion and Perspectives 156
11.1 Conclusion . . . 156
11.2 Perspectives . . . 158

Bibliography 161

Appendix I: BECD Manual 170

List of Figures

2.1 The respiratory, phonatory, resonatory, and articulatory speech subsystems [1]. 33
2.2 The anatomical organization of the nervous system [2]. 35
2.3 A typical multipolar neuron [2]. 35
2.4 The different brain regions involved in the process of speech production [3]. . . 37

3.1 A graphic representation of an ADSR envelope [4]. 48
3.2 An analog signal (top) is merged with a carrier wave (second from top) into an Amplitude Modulated signal (third from top) or a Frequency Modulated signal (bottom) [5]. 49
3.3 The relationship between AI, ML, and DL [**?**]. 54
3.4 Example of data seperated using SVM [6]. 56
3.5 The architecture of the transformer [7]. 57
3.6 Illustration of the wav2vec2.0 framework, which jointly learns contextualized speech representations and an inventory of discretized speech units [8]. 58
3.7 An illustration of HuBERT model [9]. 61
3.8 An illustration of the training process of HuBERT [10]. 63
3.9 An illustration of the first step (clustering) of HuBERT pre-training process [10]. 64
3.10 An illustration of the second step of HuBERT pre-training process [10]. 65
3.11 An overview of the whisper's approach [11]. 66
3.12 Whisper's training dataset statistics [11]. 68
3.13 The audio spectrogram transformer (AST) architecture [12]. 70

5.1 Picture of the equipment: MOTU, the microphone array mounted at the top of a laptop computer, and a video camera. [13]. 87
5.2 The Sony $SRS - NSR^{TM}$ microphone [14]. 91
5.3 The extract from "Pierrot", G. de Maupassant. 95
5.4 The image shown to the participants to be described. 96

6.1 A block diagram of the proposed model. 98
6.2 A block diagram of the proposed transformer-based model. T and x refer to number,of frames and feature dimension respectively. 100

7.1 Block diagram of the proposed voice pathology detection system. 111
7.2 The accuracy achieved on UASpeech (intra-test) and TORGO (inter-test), using the embeddings of layers 1 to 12. 112

8.1 Block diagram of the proposed voice pathology classification system. FE, FT refer to feature extractor and fine-tuned respectively. 118

8.2 Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The architecture of the models is being used, without any pre-trained weights. 119
8.3 Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The models are used as feature extractors, without any fine-tuning. 119
8.4 Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The models are fine-tuned to the task of Dysarthria severity classification. 120
8.5 Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The transformer part of the models is fine-tuned to the task of Dysarthria severity classification 120

9.1 The confusion matrices of the baseline results following the speaker-dependent data-partition, for (a) eGeMAPS and (b) MFCCs. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity. 131
9.2 The confusion matrices of the baseline results following the speaker-independent data-partition, for (a) eGeMAPS and (b) MFCCs. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity. 132
9.3 The confusions matrices of the predictions using the wav2vec2 model, using a speaker-dependent data partition. FE refers to feature extractor, and FT Li-Lj refer to fine tuned from layer i to layer j. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity. 136
9.4 The confusions matrices of the predictions using the wav2vec2 model, using a speaker-independent data partition. FE refers to feature extractor, and FT Li-Lj refer to fine tuned from layer i to layer j. 138
9.5 The confusions matrices of the predictions using the AST model, using a speaker-dependent data partition. FT Li-Lj refer to fine tuned from layer i to layer j. 140
9.6 The confusions matrices of the predictions using the AST model, using a speaker-independent data partition. FE refers to feature extractor, and FT Li-Lj refer to fine tuned from layer i to layer j. 142

10.1 The confusions matrices of the predictions using the transformer-based models, using (a) eGemaps, (b) MFCCs, and (c) wav2vec2.0 embeddings as inputs. Following the speaker-dependent scenario. 146
10.2 The confusions matrices of the predictions using the transformer-based models, using egemaps, MFCCs, and wav2vec2.0 embeddings as inputs. Following the speaker-independent scenario. 148
10.3 Block diagram of the contrastive learning. Tf refers to transformer, and o/p refers to output. 149
10.4 The confusions matrix of the predictions using the transformer model pre-trained using contrastive learning, wav2vec2.0 embeddings as input. 150
10.5 Block diagram of the auto-encoder model. LL refers to linear layer. 151
10.6 The confusions matrices of the predictions using the autoencoder models, (a) MFCCs and (b) wav2vec2.0 embeddings as inputs. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity. 153

List of Acronyms

AI Artificial Intelligence
ASR Automatic Speech Recognition
ASPD/SC Automatic Speech Pathology Detection and Severity Classification
ML Machine Learning
DL Deep Learning
CP Cerebral Palsy
PD Parkinson's Disease
FDA Frenchay Dysarthria Assessment
BECD Batterie d'évaluation clinique de la Dysarthrie
SLP Speech-Language Pathologist
ASR Automatic Speech Recognition
AST Audio Spectrogram Transformer
CNN Convolutional Neural Network
SOTA State-of-the-art
MFCC Mel-Frequency Cepstral Coefficient
eGeMAPS extended Geneva Minimalistic Acoustic Parameter Set
DLD Developmental Language Disorder
CNS Central Nervous System
PNS Peripheral Nervous System
SVM Support Vector Machine
AE AutoEncoders
LSTM Long Short-Term Memory
AUC Area Under the Curve
MLP Multi-Layer Perceptron
NN Neural Network
ANN Artificial Neural Network

SD Speaker-Dependent

SI Speaker-Independent

LLM Large Language Model

Abstract

Recent advances in artificial intelligence have opened new possibilities for automating clinical assessments, including those related to speech impairments such as dysarthria, a common motor speech disorder in individuals with Parkinson's disease, cerebral palsy, etc.. Accurate detection and severity classification of dysarthria are essential for timely intervention and monitoring, yet they remain reliant on expert analysis, which could limit the process. This has led to growing interest in the development of machine learning models capable of automatically identifying dysarthria and estimating its severity from speech. However, this task presents several challenges, including limited availability of data, language variability, the subtle nature of speech impairments across individuals, etc..

The first part of this work explores the construction of a transformer-based model for dysarthria detection and severity estimation, trained initially on two publicly available English-language dysarthric speech datasets. These models establish a foundational benchmark and demonstrate the potential of deep contextual speech representations in pathological speech analysis.

Building on this, the models are adapted and extended to a newly collected dataset of French-speaking Parkinsonian patients, collected in collaboration with Fondation Hopale, and specifically designed as part of this research to fill the gap in French pathological speech resources.

To further enhance the performance of the transformer-based model, the thesis then investigates the use of contrastive learning, integrating it into the training of the model to improve the representation of speech samples with varying severity levels. This method enhances model sensitivity to subtle articulatory changes, which are crucial for accurate severity classification.

In parallel, autoencoder-based models are proposed to facilitate unsupervised anomaly detection and representation learning in settings where labeled data is scarce or unreliable. These models prove effective in characterizing pathological deviations in speech and offer complementary insights to supervised methods.

Another central focus of the thesis is the use of pre-trained models under diverse conditions: different model architectures, and varying levels of fine-tuning (from feature extraction to full adaptation). These configurations are systematically evaluated to assess their effectiveness in generalizing to unseen speakers, and to different datasets.

This work contributes a novel, multi-faceted approach to dysarthria detection and severity classification, demonstrating how modern deep learning tools, including pre-training,

contrastive learning, and unsupervised representation learning, can be effectively applied and tailored for clinical speech analysis. While a fully deployable clinical tool is not yet realized, the methods and insights developed in this thesis lay a solid foundation for scalable, language-adaptable, and clinically meaningful AI systems for the detection and characterization of motor speech disorders.

Résumé

Les récentes avancées en intelligence artificielle ont ouvert de nouvelles possibilités pour automatiser les évaluations cliniques, y compris celles liées aux troubles de la parole tels que la dysarthrie, un trouble moteur de la parole courant chez les personnes atteintes de la maladie de Parkinson, de paralysie cérébrale, etc. La détection précise et la classification de la gravité de la dysarthrie sont essentielles pour une intervention et un suivi en temps opportun, mais elles dépendent encore de l'analyse d'experts, ce qui pourrait limiter le processus. Cela a conduit à un intérêt croissant pour le développement de modèles d'apprentissage automatique capables d'identifier automatiquement la dysarthrie et d'estimer sa gravité à partir de la parole. Cependant, cette tâche présente plusieurs défis, notamment la disponibilité limitée des données, la variabilité linguistique, la nature subtile des troubles de la parole chez les individus, etc.

La première partie de ce travail explore la conception d'un modèle basé sur les réseaux de neurones transformateurs pour la détection de la dysarthrie et l'estimation de sa sévérité. Ce modéle est initialement entraîné et testé sur deux bases de données de discours dysarthriques en anglais disponibles publiquement. Ces modèles établissent une référence fondamentale et démontrent le potentiel des représentations contextuelles profondes de la parole dans l'analyse de la parole pathologique.

En s'appuyant sur cela, les modèles sont adaptés et étendus à un nouvel ensemble de données collectées auprès de patients parkinsoniens francophones, recueillies en collaboration avec la Fondation Hopale, et spécifiquement conçues dans le cadre de cette recherche pour combler le manque de ressources de discours pathologique en français.

Pour améliorer davantage les performances du modèle basé sur les transformateurs, la thèse examine ensuite l'utilisation de l'apprentissage contrastif, en l'intégrant dans l'entraînement du modèle pour améliorer la représentation des échantillons de parole avec des niveaux de gravité variés. Cette méthode améliore la sensibilité du modèle aux changements articulatoires subtils, qui sont cruciaux pour une classification précise de la gravité.

En parallèle, des modèles basés sur des autoencodeurs sont proposés pour faciliter la détection d'anomalies non supervisée et l'apprentissage de représentations dans des contextes où les données étiquetées sont rares ou peu fiables. Ces modèles s'avèrent efficaces pour caractériser les déviations pathologiques dans la parole et offrent des perspectives complémentaires aux méthodes supervisées.

Un autre point central de la thèse est l'utilisation de modèles pré-entraînés dans des conditions diverses : différentes architectures de modèles et niveaux de réglage fin (de l'extraction de caractéristiques à l'adaptation complète). Ces configurations sont systématiquement

évaluées pour mesurer leur efficacité à se généraliser à des locuteurs non vus et à différents ensembles de données.
Ce travail apporte une approche novatrice et multifacette pour la détection de la dysarthrie et la classification de sa sévérité, démontrant comment les outils modernes d'apprentissage profond, y compris le pré-entraînement, l'apprentissage contrastif et l'apprentissage de représentation non supervisé, peuvent être efficacement appliqués et adaptés à l'analyse clinique de la parole. Bien qu'un outil clinique entièrement déployable ne soit pas encore réalisé, les méthodes et les perspectives développées dans cette thèse posent une base solide pour des systèmes d'IA évolutifs, adaptables aux langues et cliniquement significatifs pour la détection et la caractérisation des troubles moteurs de la parole.

Résumé Étendu

Les humains ont toujours existé en sociétés, entourés d'autres êtres humains. En tant qu'êtres sociaux et émotionnels, les gens ne peuvent pas prospérer seuls. La communication joue un rôle important dans les interactions entre les humains, la rendant indispensable à l'évolution de l'humanité. Cette communication était principalement non verbale dans les premiers jours de l'existence humaine, reposant principalement sur des gestes, des expressions faciales et des vocalisations simples pour exprimer des besoins fondamentaux ou donner des avertissements. Cela a été suivi par les peintures rupestres et les marques symboliques, utilisées pour enregistrer des événements ou raconter des histoires, marquant ainsi un pont entre les signaux purement physiques et les représentations plus abstraites des idées.

Les communautés continuaient de croître, tant en population qu'en complexité. Cette croissance s'est reflétée dans leurs moyens de communication, introduisant des schémas vocaux plus avancés. Au fil des générations, ces motifs vocaux, ou sons, ont évolué en discours cohérent, marquant un moment décisif de l'histoire humaine. Avec l'émergence des langues parlées, les humains sont devenus plus capables d'exprimer leurs besoins et préoccupations, ainsi que des pensées et connaissances plus complexes, posant les bases d'une riche tradition de communication qui continue de façonner nos vies aujourd'hui.

Cette avancée dans la parole cohérente a révolutionné la communication humaine, mais elle a également été accompagnée de ses difficultés. Une gamme de troubles affectant la voix est apparue, en raison de divers facteurs, tels que l'utilisation excessive de la voix, les facteurs environnementaux, ou même des conditions physiologiques sous-jacentes. Ces pathologies vocales, qui peuvent perturber la capacité à parler clairement et confortablement, soulignent l'équilibre délicat impliqué dans la production de la parole et le besoin crucial de comprendre, diagnostiquer et gérer de telles conditions.

Les pathologies de la parole concernent un grand nombre de personnes. En Europe, environ 5,8 millions d'enfants et d'adolescents souffrent de troubles du développement du langage [15–17]. Selon le National Institute on Deafness and Other Communication Disorders (2024), près d'un enfant sur douze (7,7%) âgé de 3 à 17 ans aux États-Unis a eu un trouble lié à la voix, à la parole, au langage ou à la déglutition au cours des 12 derniers mois [18]. Selon la National Stuttering Association, près de 3 millions d'adultes aux États-Unis bégaient [18]. Selon la National Stuttering Association, près de 3 millions d'adultes aux États-Unis bégaient [18].

L'évolution rapide de l'IA a eu un impact sur tous les domaines et industries, et les technologies de la parole n'ont pas été en reste. Ces technologies se reflètent dans des domaines

tels que la reconnaissance automatique de la parole, la synthèse vocale, la biométrie vocale / la reconnaissance de locuteur, la traduction de la parole à la parole, le clonage vocal / la conversion vocale, l'amélioration de la parole et la réduction du bruit, l'analyse des émotions et des sentiments dans la parole, etc. Cette évolution dans différentes tâches liées à la parole était principalement limitée à la parole normale, non pathologique, car les données utilisées pour l'entraînement étaient principalement non pathologiques. En conséquence, ces technologies semblent être insuffisantes / non fonctionnelles pour les personnes ayant des problèmes de parole, les privant ainsi de bénéficier de telles technologies. Cela a rendu important d'inclure la parole pathologique lors de la construction des technologies de la parole, ainsi que de bénéficier de l'IA pour accomplir des tâches spécifiques aux pathologies de la parole, telles que l'évaluation et la thérapie de l'articulation assistées par l'IA, la télépratique et la surveillance à distance, la détection automatique des pathologies de la parole et la classification de la gravité, qui est le cœur de cette thèse. Cela a rendu important d'inclure la parole pathologique lors de la construction de technologies de la parole, ainsi que de bénéficier de l'IA pour accomplir des tâches spécifiques aux pathologies de la parole, telles que l'évaluation et la thérapie de l'articulation assistées par l'IA, la télépratique et la surveillance à distance, la détection automatique des pathologies de la parole et la classification de leur sévérité, qui est le cœur de cette thèse. La détection automatique des pathologies de la parole et leur classification selon leur sévérité est une science qui consiste à utiliser des modèles d'apprentissage automatique pour détecter une pathologie de la parole et en évaluer la sévéité [19]. De tels outils offriraient plusieurs avantages, tels que l'assistance aux orthophonistes pour surveiller l'état de leurs patients, la fourniture d'un outil d'évaluation dans les zones à accès restreint aux spécialistes et/ou aux équipements médicaux, etc.
Ce domaine de recherche est confronté à plusieurs défis [20, 21], notamment : la rareté/la faible qualité des données de parole pathologique, l'hétérogénéité des populations de locuteurs, la limitation de l'interprétabilité des modèles et de la confiance clinique, les considérations de confidentialité et d'éthique, et la difficulté à atteindre une généralisation multilingue et interlinguistique, etc.

Contributions de la Thèse

Dans cette thèse, nous souhaitons contribuer au domaine de la recherche en matière de détection automatique des pathologies de la parole et de leur classification selon leur sévérité. Les contributions de cette thèse sont les suivantes:

- **Collecte de données:** Pour contribuer à l'accessibilité des données pour la recherche en orthophonie, au cours de cette thèse, en collaboration avec la Fondation Hopale, nous avons effectué une collecte de données auprès de patients français atteints de la maladie de Parkinson.

- **Transformateurs pour la classification de la dysarthrie:** Nous proposons un modèle exploitant l'architecture des transformateurs pour la classification de la gravité de la dysarthrie. Le modèle a été entraîné dans une approche supervisée, en utilisant différentes caractéristiques comme entrée. L'entraînement a été réalisé avec deux configurations de partition des données, dépendante du locuteur et

indépendante du locuteur. Le premier est plus adapté à la personnalisation, tandis que le second est plus approprié pour la généralisation.

- **Exploration de l'utilisation de modèles pré-entraînés:** Nous explorons l'utilisation de modèles pré-entraînés pour la détection de la dysarthrie et la classification de sa sévérité, afin de surmonter le problème de la rareté des données. Cette exploration est menée en utilisant quatre modèles de l'état de l'art, dans quatre configurations d'entraînement différentes pour étudier l'impact des modèles sur la tâche de classification. L'exploration a été menée en utilisant deux ensembles de données différents, afin de mettre davantage à l'épreuve les capacités de généralisation des modèles.

- **Extension de la recherche à la langue française:** En utilisant la base de données collectée, nous menons plusieurs expériences pour la détection automatique des pathologies de la parole et de leur classification selon leur sévérité, couvrant l'utilisation de modèles pré-entraînés, un modèle basé sur les transformateurs qui a été développé davantage par l'utilisation de l'entraînement contrastif, ainsi que des auroencodeurs. Ces expériences constituent une première étape dans la recherche dans ce domaine en langue française, aucun autre travail n'ayant été réalisé à partir d'un ensemble de données français.

Structure de la thèse

Cette thèse est composée d'une brève introduction, suivie de deux parties, et enfin d'une conclusion et de perspectives. La première partie, qui est dédiée à l'état de l'art est composée de trois chapitres, explore les bases de la production de la parole, les maladies neurologiques et leur impact sur la production de la parole, les outils techniques nécessaires à la recherche, et une revue de l'état actuel de la recherche. La deuxième partie, composée de quatre chapitres, détaille les contributions de la thèse.
Les points principaux de chaque chapitre sont présentés ci-après:

- ***Chapitre 2: Explorer l'orthophonie: Perspectives neurologiques, perception humaine et dimensions cliniques***
  Ce chapitre offre une introduction complète au processus de production de la parole. Nous expliquons comment la parole est produite et comment les troubles neurologiques affectent la production de la parole. Nous introduisons la dysarthrie, une pathologie de la parole, et deux principaux troubles qui entraînent la dysarthrie : la paralysie cérébale et la maladie de Parkinson. De plus, nous présentons comment les pathologies de la parole sont évaluées cliniquement.

- ***Chapitre 3 : Signal, Caractéristiques et Modèles : Une Perspective Technique***
  Ce chapitre présente les outils techniques nécessaires pour transformer le signal vocal en un signal numérisé pouvant être utilisé pour l'analyse de la parole. Dans ce chapitre, nous introduisons les concepts techniques clés pour la numérisation des signaux audio, en détaillant les processus impliqués dans la conversion des formes d'onde analogiques en données discrètes. Nous explorons également les principes

fondamentaux de l'extraction de caractéristiques à l'aide des techniques de traitement du signal, en mettant en évidence comment ces méthodes révèlent des caractéristiques significatives à partir de l'audio brut. En outre, nous présentons les principaux concepts d'apprentissage automatique, les paradigmes d'apprentissage et les modèles utilisés tout au long de cette thèse, en soulignant leur pertinence pour les expérimentations à venir. Une compréhension solide de ces notions fondamentales améliore la compréhension des défis, des méthodologies et des résultats abordés dans les chapitres suivants, assurant une progression plus fluide vers les sujets avancés à venir.

- ***Chapitre 4 : Une revue bibliographique de la détection et de la classification de la dysarthrie***
  Ce chapitre présente une revue bibliographique du domaine de recherche sur la détection et la classification de la gravité de la dysarthrie. Nous proposons une revue complète des méthodes de détection et de classification de la dysarthrie, établissant ainsi les bases nécessaires à la compréhension des complexités de l'évaluation automatisée des troubles de la parole. Nous explorons un éventail d'approches de classification, en détaillant leurs méthodologies et les hypothèses sous-jacentes, et examinons les caractéristiques essentielles à la classification de la sévérité de la dysarthrie. La discussion sur la partition des données met en lumière les différences entre les configurations dépendantes et indépendantes du locuteur, en soulignant leur impact sur la performance et la généralisabilité des modèles. De plus, l'enquête bibliographique offre un aperçu des tendances actuelles et de l'évolution des méthodes de classification de la dysarthrie. Une analyse critique des approches traditionnelles et de bout en bout révèle des limites importantes pour les deux types d'approches. Ces constats soulignent la nécessité de méthodologies innovantes et adaptatives, capables de gérer efficacement la diversité des caractéristiques propres à la parole dysarthrique.

- ***Chapitre 5 : Vue d'ensemble des ensembles de données : Données publiques et nouvelles données collectées***
  Ce chapitre présente un aperçu des principaux jeux de données utilisés pour entraîner et évaluer les modèles dans le cadre de cette thèse. En particulier, deux bases de données publiques en anglais, UASpeech et TORGO, sont introduites et expliquées, ainsi que notre propre nouvelle base de données française collectées, contenant des enregistrements audio de patients parkinsoniens francophones et de sujets témoins. En tant que l'une des premières bases de données dysarthriques en français collectées dans le but de soutenir la recherche sur les technologies liées aux troubles de la parole, cette dernière contribue de manière générale au domaine de la recherche en pathologie de la parole en fournissant des données essentielles à l'avancement des travaux et technologies sur la dysarthrie, et plus particulièrement pour les locuteurs francophones.

- ***Chapitre 6 : L'attention est-elle tout ce dont vous avez besoin pour la classification de la gravité de la dysarthrie ?***
  Ce chapitre présente notre première proposition pour la classification de la gravité de la dysarthrie, en utilisant un modèle basé sur les transformateurs. Dans ce chapitre,

nous proposons l'utilisation de l'architecture transformer comme classificateur pour la classification de la sévérité de la dysarthrie. Ce chapitre démontre le potentiel des modèles basés sur les transformers pour cette tâche, notamment dans le cadre des scénarios indépendants du locuteur. Les résultats montrent que, bien que les LSTM offrent de bonnes performances dans le cas dépendant du locuteur, les transformers les surpassent dans le cas plus complexe de l'indépendance du locuteur, en particulier lorsqu'ils sont combinés à des méthodes avancées d'extraction de caractéristiques comme wav2vec2.
Les résultats mettent en évidence l'importance du choix de l'architecture du modèle et de l'ensemble de caractéristiques pour atteindre de bonnes performances. Les embeddings de wav2vec2 apparaissent systématiquement comme le meilleur jeu de caractéristiques, reflétant leur capacité à capturer une information contextuelle riche à partir de l'audio brut. Les transformers, grâce à leur capacité à modéliser des relations globales et à exploiter des représentations de caractéristiques de haute dimension, montrent des avantages significatifs pour généraliser à des locuteurs non vus. Cependant, l'écart de performance observé entre les scénarios dépendants et indépendants du locuteur souligne la nécessité de poursuivre les recherches afin d'améliorer la généralisation des modèles. Relever des défis tels que la variabilité intra-classe et la quantité limitée de données d'entraînement sera essentiel pour faire progresser les systèmes de classification de la sévérité de la dysarthrie. En résumé, ce chapitre met en avant le potentiel transformateur des modèles de type transformer dans la classification de la sévérité de la dysarthrie, en particulier pour des applications réelles indépendantes du locuteur. Les travaux futurs devraient tirer parti de leurs atouts en misant sur une plus grande diversité des données, des stratégies d'apprentissage avancées, et une intégration poussée avec des techniques d'extraction de caractéristiques de pointe. Ce chapitre pose la question : "L'attention suffit-elle pour classifier la sévérité de la dysarthrie ?". Nos expériences montrent que, bien que l'auto-attention améliore la précision, elle ne suffit pas à elle seule. Les gains de performance apparaissent principalement lorsque l'attention est combinée à des embeddings acoustiques haute résolution (comme wav2vec2). Nous concluons donc que l'attention est une solution puissante mais partielle : une classification robuste de la dysarthrie nécessite toujours des représentations riches et une régularisation appropriée.

- ***Chapitre 7 : Exploitation des modèles pré-entraînés pour la détection de la gravité de la dysarthrie : Une analyse comparative***
  Dans ce chapitre, nous étudions l'utilisation des embeddings de modèles pré-entraînés pour la détection des pathologies vocales. Nous comparons les performances de trois modèles pré-entraînés : wav2vec2, HuBERT et Whisper. Ces modèles sont utilisés comme extracteurs de caractéristiques, dans leur état figé, et leurs embeddings sont fournis à deux classificateurs : un SVM et un MLP. Une étude comparative des performances des différentes couches de wav2vec2 et HuBERT est réalisée, et les performances des modèles pré-entraînés sont comparées à celles des caractéristiques de référence : MFCC et eGeMAPS. Tous les systèmes proposés obtiennent de meilleurs résultats que les caractéristiques de base sur le jeu de test intra-base; plus précisément, le modèle pré-entraîné le plus performant améliore la précision de test de 21% et 16% par rapport aux meilleures caractéristiques de référence, respec-

tivement pour SVM et MLP.
Des expériences inter-bases sont également menées en utilisant la base de données TORGO, et les résultats sur cette base varient selon le classificateur utilisé : les modèles pré-entraînés obtiennent les meilleurs résultats avec le MLP, ce qui n'est pas le cas avec le SVM. Cela met en évidence l'importance du choix du classificateur, puisqu'un simple changement de classificateur permet de passer de 41% à 70% de précision avec les embeddings de wav2vec2 lorsque l'on passe du SVM au MLP. Les résultats de ce travail soulignent l'efficacité des embeddings des modèles pré-entraînés pour bien généraliser à des données non vues, qu'elles proviennent ou non de la base d'entraînement.

- ***Chapitre 8 : Étendre l'utilisation des modèles pré-entraînés pour la classification de la gravité de la dysarthrie***
  Dans ce chapitre, nous proposons d'utiliser des modèles pré-entraînés pour atténuer le problème de la rareté des données dans le domaine de la pathologie vocale. À cette fin, nous utilisons quatre modèles pré-entraînés à l'état de l'art pour la classification de la sévérité de la dysarthrie : wav2vec2, HuBERT, Whisper et AST, couvrant à la fois les modèles de parole basés sur les formes d'onde et ceux basés sur les spectrogrammes, ainsi que des approches d'entraînement supervisées et auto-supervisées.
  Les quatre modèles sont évalués selon trois ou quatre modes d'entraînement : architecture seule, extracteur de caractéristiques, apprentissage partiel (partially fine-tuned) et apprentissage complet (fully fine-tuned). L'entraînement est réalisé selon une partition indépendante du locuteur. Les résultats démontrent que les modèles sont capables d'extraire des caractéristiques robustes et généralisables, offrant des performances compétitives aussi bien dans les évaluations intra-base que dans les évaluations inter-bases.

- ***Chapitre 9 : La maladie de Parkinson et la dysarthrie : Adapter et étendre les expériences sur VocaPark***
  Dans ce chapitre, nous menons différentes expériences sur notre nouvelle base de données française collectée auprès de patients parkinsoniens. Nous commençons par établir une référence en utilisant des caractéristiques traditionnelles, à savoir les MFCCs et eGeMAPS, combinées à un réseau de neurones. Les expériences sont réalisées selon deux approches de partition des données : dépendante et indépendante du locuteur.
  Une fois cette ligne de base établie, nous explorons l'utilisation de modèles pré-entraînés pour la classification de la dysarthrie. Les modèles wav2vec2 et AST sont étudiés dans deux configurations d'entraînement: en tant qu'extracteurs de caractéristiques et en mode partiellement fine-tuned. Là encore, les expériences suivent les deux types de partition : dépendante et indépendante du locuteur. Les résultats révèlent que, pour les deux types de partitions, les modèles pré-entraînés surpassent les caractéristiques de base, notamment en termes de précision par classe, avec un avantage marqué pour wav2vec2.
  Globalement, ces résultats confirment les conclusions des chapitres 7 et 8, dans une autre langue, quant au potentiel de l'utilisation conjointe d'extraction de caractéristiques robustes et de stratégies de pré-entraînement pour la classification de la

dysarthrie. Bien que la configuration dépendante du locuteur atteigne des précisions nettement supérieures, le scénario indépendant du locuteur demeure plus complexe en raison de la variabilité accrue entre les locuteurs.

- ***Chapitre 10: Explorer des approches alternatives: Préentraînement contrastif et autoencodeurs***
  Dans ce chapitre, d'autres expériences sont menées sur la nouvelle base de données collectée. Nous revisitons le modèle basé sur les transformers proposé au chapitre 6, et l'entraînons sur la nouvelle base de données française récemment collectée. Le modèle est entraîné en suivant à la fois les partitions dépendantes et indépendantes du locuteur. Afin d'améliorer les résultats dans le scénario indépendant du locuteur, nous introduisons une étape de pré-entraînement contrastif du modèle Transformer. Cette étape entraîne une amélioration de 7% de la précision globale, ainsi qu'une hausse générale des performances par classe.
  Une dernière expérience est réalisée dans le cadre du scénario indépendant du locuteur, en utilisant un autoencodeur. Cette configuration aboutit à une précision globale de 38%.
  Globalement, ces résultats confirment le potentiel de l'exploitation de l'extraction de caractéristiques robustes et des stratégies de pré-entraînement pour la classification de la dysarthrie. Bien que les configurations dépendantes du locuteur atteignent des précisions nettement plus élevées, le scénario indépendant du locuteur reste plus difficile en raison d'une variabilité plus importante entre les locuteurs. Néanmoins, les gains observés grâce au pré-entraînement contrastif et aux modèles pré-entraînés comme wav2vec2.0 indiquent une voie prometteuse vers une évaluation de la sévérité de la dysarthrie plus précise et plus généralisable.

Conclusions

L'objectif de cette thèse était de concevoir, implémenter et évaluer rigoureusement une chaîne de traitement automatisée pour la détection de la dysarthrie et la classification de sa sévérité à partir de la parole continue, dans le but à long terme d'intégrer cette technologie à la pratique clinique quotidienne et aux services de télémédecine. En s'intéressant à la fois à la détection binaire et à l'estimation de la sévérité, ce travail visait à combler le fossé entre les prototypes de recherche et les outils que les orthophonistes peuvent utiliser de manière fiable pour le diagnostic, la planification thérapeutique et le suivi des progrès. Dans cette optique, plusieurs méthodes d'automatisation de la détection et de la classification des pathologies vocales ont été proposées. Notre travail a contribué à l'état de l'art dans la classification de la sévérité de la dysarthrie à travers une série d'expériences innovantes et d'investigations méthodiques, en établissant un pont entre les techniques traditionnelles de traitement du signal et les approches modernes d'apprentissage profond. En particulier, ce travail s'est concentré sur le développement et l'évaluation rigoureuse de nouveaux modèles et stratégies capables de gérer la complexité inhérente aux données vocales dysarthriques.
Au chapitre 6, nous avons introduit un modèle basé sur les transformers pour la classification de la dysarthrie. Les transformers sont des architectures d'apprentissage profond entièrement construites à partir de couches empilées d'auto-attention et de couches feed-

forward, sans opérations récurrentes ni convolutionnelles, permettant à chaque position d'une séquence de modéliser ses dépendances avec toutes les autres en un seul passage. Pour évaluer l'applicabilité de cette architecture à la parole dysarthrique, nous avons implémenté un modèle utilisant des blocs encodeurs de Transformers existants, que nous avons entraîné avec différentes caractéristiques d'entrée: Mel-Frequency Cepstral Coefficient (MFCC), extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) et les embeddings de wav2vec2. Pour comparaison, nous avons mis en place un modèle de référence utilisant des couches Long Short-Term Memory (LSTM), entraîné sur les mêmes ensembles de caractéristiques. Nos expériences, menées dans des scénarios dépendants (Speaker-Dependent (SD)) et indépendants (Speaker-Independent (SI)) du locuteur, ont montré que le modèle basé sur les transformers surpassait le modèle LSTM. Les meilleurs résultats ont été obtenus en utilisant les embeddings CNN gelés de wav2vec2 comme entrée : dans cette configuration, le modèle a atteint une précision de 92% dans le scénario SD et de 46% dans le scénario SI. Ces résultats suggèrent que les mécanismes d'attention des Transformers améliorent les performances de classification en mettant efficacement en évidence les caractéristiques temporelles et spectrales pertinentes. Toutefois, malgré cette amélioration, la précision relativement basse dans le scénario SI indique qu'il reste nécessaire de développer des stratégies de généralisation plus robustes.
S'appuyant sur ces observations, les chapitres 7 et 8 se sont concentrés sur l'amélioration des performances et de la capacité de généralisation des modèles, notamment dans les contextes SI et d'évaluations inter-bases, en exploitant des modèles pré-entraînés. Ces chapitres ont offert une analyse comparative complète entre des modèles opérant sur les formes d'onde brutes et d'autres sur les spectrogrammes, pour la détection de la dysarthrie (chapitre 7) et la classification de sa sévérité (chapitre 8). L'idée principale était de pallier la baisse de performance observée en SI en utilisant des modèles déjà exposés à de grandes quantités de données vocales lors de leur pré-entraînement. Nous avons comparé quatre modèles de pointe (wav2vec2, HuBERT, Audio Spectrogram Transformer (AST) et Whisper) dans diverses configurations: architecture seule (sans poids pré-entraînés), extracteurs de caractéristiques, et fine-tuning partiel ou complet. Les expériences ont montré que l'utilisation de modèles pré-entraînés permet non seulement d'améliorer la précision de classification, mais aussi de mieux faire face au manque de données annotées. Les meilleures précisions obtenues pour la détection de la dysarthrie étaient de 96% (intra-base) et 70% (inter-base), contre 56% et 67% respectivement pour la classification de la sévérité. Ces résultats ont confirmé que l'usage de modèles pré-entraînés renforce significativement la robustesse face à la diversité des locuteurs et des conditions d'enregistrement.
Une autre contribution majeure de cette thèse a été la collecte de données auprès de patients parkinsoniens francophones. Ce corpus, réalisé en collaboration avec la Fondation Hopale, contient des enregistrements de 114 participants, comprenant des non-mots, mots, phrases et segments de parole libre. Cette base enrichit le réservoir de données pour la recherche en pathologie vocale en général, et pour la recherche francophone en particulier. Le processus de collecte a été décrit au chapitre 5. Le chapitre 9 a poursuivi l'analyse sur ce nouveau corpus. Des expériences de référence ont d'abord été menées, avant de reproduire celles des chapitres 7 et 8, ce qui a permis d'atteindre une précision de 72% en configuration SD, contre 47% en SI. De plus, dans le chapitre 10, nous avons exploré l'utilisation du modèle Transformer proposé au chapitre 6, et son amélioration via un pré-entraînement contrastif. Ces expériences ont démontré que les mécanismes d'attention,

associés à une stratégie contrastive, permettaient de mieux distinguer les subtilités de la parole dysarthrique, avec des gains notables, notamment en SI, où la précision est passée de 32% à 39%. Enfin, nous avons exploré une approche basée sur les autoencodeurs. Le modèle autoencodeur, doté de caractéristiques d'entrée sélectionnées avec soin et de réglages d'apprentissage optimisés, a fourni une alternative de représentation efficace, atteignant 38% de précision en SI.
Dans l'ensemble, les expériences décrites dans les chapitres 6 à 10 ont contribué à développer une méthodologie plus adaptative et axée sur les données pour évaluer la sévérité de la dysarthrie. L'intégration de mécanismes d'attention, de modèles pré-entraînés, de stratégies de pré-entraînement contrastif et d'approches autoencodeurs ouvre la voie à des systèmes plus généralisables et interprétables. Ainsi, cette étude constitue une base solide pour le développement d'outils diagnostiques performants et utiles en contexte clinique, favorisant des interventions thérapeutiques plus rapides et plus efficaces pour les personnes atteintes de dysarthrie.

Perspectives

Malgré les avancées présentées dans cette thèse et dans la littérature, le déploiement efficace des méthodes basées sur l'intelligence artificielle pour le diagnostic et la rééducation reste un défi majeur. Cette section propose plusieurs axes de recherche prometteurs visant à surmonter les limites actuelles et à développer des systèmes de classification de la dysarthrie plus complets, adaptatifs et exploitables en pratique clinique.

Extension des modèles basés sur les Transformers
À partir des résultats prometteurs obtenus grâce à l'introduction d'un pré-entraînement contrastif dans les modèles de type transformer, les travaux futurs pourraient se concentrer sur l'extension de cette approche à l'aide de jeux de données plus larges et plus diversifiés. L'augmentation du volume et de la variété des données d'apprentissage, en intégrant une plus grande diversité de langues, d'accents, de dispositifs d'enregistrement, de conditions de bruit de fond et de degrés de sévérité de la dysarthrie, permettrait aux mécanismes d'attention de mieux capter les indices temporels, prosodiques et spectraux subtils. Cela pourrait améliorer la spécificité diagnostique, renforcer la détection de changements cliniquement significatifs au fil du temps, et accroître la tolérance du système aux variations acoustiques inhérentes aux sessions de télésoin. Toutefois, une telle richesse dans l'espace d'entrée peut dépasser la capacité d'un transformer peu profond; des améliorations architecturales seront donc nécessaires (élargissement des couches feed-forward, convolutions groupées, augmentation du nombre de couches d'encodage) afin de maintenir les performances sans augmenter excessivement la latence ou l'usage mémoire. Ce passage à l'échelle devra cependant être réalisé avec précaution, afin d'éviter d'amplifier les déséquilibres de classes ou les biais démographiques. De plus, des modules d'explicabilité compréhensibles par les orthophonistes resteront essentiels pour instaurer la confiance. Enfin, puisque la majorité des nouvelles données proviendront probablement de plateformes de télépratique, des stratégies d'apprentissage continu respectueuses de la vie privée seront nécessaires pour permettre une intégration fluide et sécurisée dans les

soins quotidiens.

Intégration de la parole pathologique dans les modèles pré-entraînés
Les modèles pré-entraînés ont montré un fort potentiel pour l'extraction de caractéristiques dans les tâches de pathologie vocale, mais leurs performances peuvent être limitées par la nature des données utilisées lors du pré-entraînement, souvent exclusivement composées de parole saine. Un axe prometteur consiste à intégrer de la parole pathologique directement dans les phases de pré-entraînement. Cela permettrait au modèle d'apprendre des indices spécifiques aux troubles, de mieux quantifier les niveaux de sévérité, et de faciliter le transfert vers des applications comme l'évaluation automatique de l'intelligibilité ou le suivi de l'évolution en télésoin. Cette stratégie soulève néanmoins plusieurs défis. Les corpus pathologiques sont rares, souvent fragmentés entre différents centres hospitaliers, et protégés par des réglementations strictes sur la vie privée. Leur centralisation nécessitera donc une collaboration interinstitutionnelle, l'adoption de techniques d'anonymisation avancées, et la mise en œuvre de protocoles d'apprentissage fédéré, où les données restent localisées derrière les pare-feu des hôpitaux. Même après la levée de ces obstacles, les cas les plus sévères comportent souvent des segments silencieux ou inintelligibles, ce qui peut perturber les objectifs contrastifs, à moins d'introduire des mécanismes de masquage ou des pondérations de perte spécifiques. Enfin, pour une adoption clinique effective, les modèles devront fournir un retour explicatif, des scores de confiance, des visualisations de saillance, et des alertes lorsque les prédictions sortent de leur domaine de validité.

Fusion de modèles et approches multimodales
Les résultats obtenus suggèrent que différents modèles peuvent exceller dans la classification de types particuliers de dysarthrie. Une piste de recherche consiste donc à combiner plusieurs modalités ou à fusionner différents modèles dans une architecture unifiée. Le signal acoustique reste indispensable, car il contient des indices spectraux et de qualité vocale, mais il peut être enrichi par : (1) une vidéo faciale à faible coût (par exemple via une webcam) permettant d'analyser les mouvements articulatoires, (2) des capteurs inertiels ou des ceintures respiratoires pour suivre la durée des groupes respiratoires ou les mouvements du thorax, notamment utiles dans les cas hypokinétiques, et (3) des caractéristiques lexico-linguistiques extraites à partir de transcriptions ASR, telles que les substitutions ou les pauses, indicatrices d'une parole scannée ataxique. Ces modalités sont déjà disponibles sur les appareils utilisés quotidiennement par les patients, ce qui facilite leur intégration. Une approche multimodale permettrait donc de construire un outil de classification robuste, performant dans des contextes cliniques variés, y compris en télésoin ou lors d'exercices à domicile non supervisés.

Exploration des invites (prompts) et des grands modèles de langage
Les approches basées sur les prompts et les grands modèles de langage offrent de nouvelles opportunités pour l'amélioration des systèmes de classification en pathologie vocale. Une stratégie envisageable consisterait à intégrer les Large Language Model (LLM)s dans la chaîne de traitement, en s'appuyant sur le fait que les patients lisent souvent des textes connus à l'avance (scripts). En comparant l'alignement entre le script de référence et

la transcription obtenue via reconnaissance automatique, il est possible d'extraire les divergences (substitutions, ajouts, hésitations, pauses anormalement longues) et de les associer aux caractéristiques acoustiques. Ce couplage pourrait alors alimenter un modèle, potentiellement enrichi par un LLM, pour estimer le niveau de sévérité de la dysarthrie. Une telle méthode permettrait d'exploiter la structure linguistique de manière explicite et ciblée.

Expérimentations multilingues et multipathologiques
Pour garantir la généralisabilité des modèles, il est crucial d'évaluer leur performance dans des contextes multilingues et sur un large éventail de pathologies neuro-motrices (dysarthries ataxiques, spastiques, flasques, mixtes, parkinsoniennes, ou liées à la SLA ou à la paralysie cérébrale). Les différences phonétiques et prosodiques entre langues, ainsi que les mécanismes physiopathologiques variés, constituent des défis importants. Une architecture prometteuse est celle du mélange hiérarchique d'experts, où un encodeur partagé apprend des représentations universelles de la parole, et des modules experts plus légers sont spécialisés par langue ou pathologie. Cette approche permet un déploiement en temps réel sur des tablettes, tout en assurant une extensibilité modulaire pour intégrer de nouveaux profils de patients. La poursuite de telles expérimentations contribuera à faire évoluer les prototypes actuels vers des outils équitables et cliniquement exploitables à grande échelle.

Modèles personnalisés pour le suivi continu
Une alternative aux grands modèles généralistes consiste à concevoir des modèles personnalisés, adaptés aux caractéristiques individuelles de chaque patient. De tels modèles peuvent évoluer au fil du temps, suivre les modifications de la parole, et fournir un retour précis et rapide sur des appareils personnels. Plus compacts et interprétables que les architectures à grande échelle, ces modèles sont particulièrement adaptés au suivi continu et à la rééducation autonome. Cette direction met l'accent sur la flexibilité et la pertinence individuelle, avec un potentiel fort pour renforcer l'autonomie des patients entre les séances cliniques.

Planification thérapeutique adaptative et interventions ciblées
Une autre voie prometteuse est l'adaptation des modèles développés à la planification de traitements individualisés. Tandis que les modèles personnalisés se concentrent sur l'évaluation continue, cette perspective vise à adapter dynamiquement les exercices de rééducation selon les besoins spécifiques de chaque patient. En intégrant les données personnelles dans le modèle, il devient possible de générer des protocoles thérapeutiques ciblés, ajustés en temps réel en fonction des progrès observés. Cela pourrait s'appuyer sur des systèmes à règles liés aux sorties des modèles, ou sur des LLMs capables d'associer les erreurs observées à des stratégies d'intervention validées par la littérature. En bouclant la chaîne entre évaluation et traitement, ces outils soutiendraient les cliniciens dans la mise en œuvre de soins personnalisés et efficaces.

Dans leur ensemble, ces pistes tracent une feuille de route vers des systèmes d'IA plus personnalisés, explicables et équitables pour la prise en charge de la dysarthrie, avec

l'ambition de dépasser le cadre du laboratoire pour un véritable impact clinique à grande échelle.

Chapter 1

Introduction

1.1 An Overview of the Subject's Relevance

Humans have always existed in societies, surrounded by other human beings. As social and emotional beings, people cannot thrive all by themselves. Communication plays an important role in the interactions between humans, making it indispensable for the evolution of humankind. This communication was primarily nonverbal in the earliest days of human existence, relying mainly on gestures, facial expressions, and simple vocalizations to express basic needs, or to give warning. This was followed by cave paintings and symbolic markings, used to record events or tell stories, thus marking a bridge between purely physical signals and more abstract representations of ideas.

Communities continued to grow, both in population and in complexity. This growth was reflected in their means of communication, introducing more advanced vocal patterns. Over generations, these vocal patterns, or sounds, evolved into coherent speech, marking a pivotal moment in human history. With the emergence of spoken languages, humans became more able to express their needs and concerns, as well as more complex thoughts and knowledge, laying the groundwork for a rich communication tradition that continues to shape our lives today.

This advancement in coherent speech revolutionized human communication, but it also was coupled with its difficulties. A range of disorders affecting the voice emerged, due to various factors, such as overuse of voice, environmental factors, or even underlying physiological conditions. These voice pathologies, which can disrupt one's ability to speak clearly and comfortably, underscore the delicate balance involved in producing speech and the critical need for understanding, diagnosing, and managing such conditions.

Speech pathologies account for a large number of people. In Europe, about 5.8 million children and adolescents have Developmental Language Disorder [15–17]. According to the National Institute on Deafness and Other Communication Disorders (2024), nearly 1 in 12 (7.7%) U.S. children ages 3-17 has had a disorder related to voice, speech, language, or swallowing in the past 12 months [18]. According to the National Stuttering Association, nearly 3 million adults in the U.S. stutter [18].

The rapid evolution of Artificial Intelligence (AI) has had its impact on all domains and industries, and speech technologies has had its fair share. These technologies are reflected in domains such as Automatic Speech Recognition (ASR), Text-to-Speech (TTS),

voice biometrics / speaker recognition, speech-to-speech translation, voice cloning / voice conversion, speech enhancement and noise reduction, emotion and sentiment analysis in speech, etc. . This evolution in different speech related tasks was mostly restricted to normal, non pathological speech, since the data used for training was mostly non pathological. As a result, these technologies seem to be lacking / non functional for people with speech problems, depriving them from benefiting from such technologies. This made it important to include pathological speech when building speech technologies, as well as benefiting from AI to accomplish tasks special for speech pathologies, such as AI-Assisted articulation assessment and therapy, telepractice and remote monitoring, automatic speech pathology detection and severity classification, which is the core of this thesis.
Automatic Speech Pathology Detection and Severity Classification (ASPD/SC) is the science of using automated Machine Learning (ML) models for the detection of a speech pathology, and the classification of the severity of the pathology for the patient [19]. Such tools would offer several advantages, such as assisting the speech pathologists to monitor the status of their patients, offering an assessment tool in areas with restricted access to specialists and/or medical equipment, etc.
This research domain is faced with several challenges [20, 21], including: the scarcity/low quality of pathological speech data, heterogeneity of speaker populations, limited model interpretability and clinical trust, privacy and ethical considerations, and difficulty in achieving multilingual and cross-linguistic generalization, etc.

1.2 Contributions of the Thesis

In this thesis, we aim to contribute to the research domain of ASPD/SC on different aspects. The contributions of this thesis are the following:

- **Data Collection:** To contribute to the data accessibility for speech pathology research, during this thesis, in collaboration with Fondation Hopale, we did a data collection from French Parkinson's patients.

- **Transformers For Dysarthria Classification:** We propose a model exploiting transformer architecture for Dysarthria severity classification. The model was trained in a supervised approach, using different features as input. The training was conducted with two data partition settings, speaker-dependent and speaker-independent. The former is more suitable for personalization, while the latter is more suited for generalization.

- **Pre-trained Models for ASPD/SC:** We explore the use of pre-trained models for dysarthria detection/severity classification, to overcome the problem of data scarcity. This exploration is conducted using four State-of-the-art (SOTA)) models, in 4 different training settings to study the impact of the models on the classification task. The exploration was conducted using two different datasets, to further challenge the generalization capabilities of the models.

- **Exploration of ASPD/SC in French Language:** Using the collected database, we conduct several experiments for ASPD/SC, covering the use of pre-trained mod-

els, transformer-based model that was further developed by the use of contrastive training, and finally using AutoEncoders (AE). These experiments constitute a first step in the research for ASPD/SC since no other work has been conducted using a French dataset.

1.3 Structure of the Thesis

This thesis is composed of a brief introduction, followed by two parts, and finally a conclusion and perspectives. The first part, which is dedicated to the state of the art and composed of three chapters, delves into the background of speech production, neurological diseases and their impact on speech production, the technical tools necessary for the research, and a review of the current research state. The second part, composed of four chapters, details the contributions of the thesis.
The main points of each chapter are presented in the following:

- ***Chapter 2: Exploring Speech Pathology: Neurological Insights, Human Perception, and Clinical Dimensions***
  This chapter provides a comprehensive introduction to the speech production process. We explain how speech is produced and how neurological disorders affect speech production. We introduce dysarthria, a speech pathology, and two main disorders that result in dysarthria: Cerebral Palsy (CP) and Parkinson's Disease (PD). Additionally, we present how speech pathologies are clinically assessed.

- ***Chapter 3: Signal, Features, and Models: A Technical Perspective***
  This chapter lays out the technical tools necessary to transform the speech signal into a digitalized signal that can be used for speech analysis. We also present key signal processing techniques used for speech analysis, along with relevant machine learning concepts.

- ***Chapter 4: A Bibliographic Review of Dysarthria Detection and Classification***
  This chapter present a bibliographic review of the dysarthria detection and severity classification research domain.

- ***Chapter 5: Overview of Datasets: Publicly Available and Newly Collected Data***
  This chapter offers an overview of the datasets used in this thesis. We start by describing two public datasets (UASpeech and TORGO), then proceed to present our own collected dataset.

- ***Chapter 6: Is Attention All You Need for Dysarthria Severity Classification ?***
  This chapter introduces our first proposal for dysarthria severity classification, using a transformer-based model. The model is tested using two data partition settings: the speaker-dependent setting and the more challenging speaker-independent setting.

- ***Chapter 7: Leveraging Pre-Trained Models for Dysarthria Severity Detection: A Comparative Analysis***
  We explore the use of pre-trained models for dysarthria severity detection. Three different models were used as feature extractors, and their ability to detect the presence of dysarthria was tested. Additionally, two different datasets were used to further test the generalizability of the trained models.

- ***Chapter 8: Extending the Use of Pre-Trained Models for Dysarthria Severity Classification***
  In this chapter, we extend the study of the use of pre-trained models for dysarthria severity detection. Four different models were studied in four different settings, and their impact on the classification task is highlighted. Additionally, two different datasets were used to further test the generalizability of the trained models.

- ***Chapter 9: Parkinson's Disease and Dysarthria: Adapting and Expanding Experiments on VocaPark***
  In this chapter, several experiments are conducted on the newly collected database. We begin by setting a baseline, then reapply some of the experiments from Chapter 8 to this dataset.

- ***Chapter 10: Exploring Alternative Approaches: Contrastive Pretraining and Autoencoders***
  In this chapter, further experiments are conducted on the newly collected database. We now reapply the experiments from Chapter 6 to this dataset. We also apply contrastive learning to improve the performance of the model. Finally, we conduct an additional experiment based on AE for dysarthria severity classification.

Part I

State of the Art

Chapter 2

Exploring Speech Pathology: Neurological Insights, Human Perception, and Clinical Dimensions

2.1 Introduction

The origin and evolution of speech remain a subject of profound scientific interest and interdisciplinary inquiry. While historically dismissed as overly speculative, recent advances in neuroscience, genetics, and cognitive science have shed new light on the mechanisms and processes that underlie human language development.

Speech production evolved gradually over millions of years, transforming from primitive forms of communication into the complex linguistic systems seen in modern humans. This progression relied on key milestones in both anatomy and cognition [22]. Anatomically, the descent of the human larynx and the development of the vocal tract enabled a broader range of phonemes [22], while the FOXP2 gene, or speech gene, emerged as a critical genetic factor in enabling speech fluency [23]. Cognitively, the expansion of the hominid brain, particularly areas such as Broca's and Wernicke's regions, played an essential role in language production and processing [22].

Theories of speech origin often debate the relative importance of vocal versus gestural precursors. Gesture-based communication, supported by evidence of mirror neurons and **motor control circuits** in the brain, suggests an early linkage between physical movements and symbolic representation. However, vocalizations may have gained prominence due to their efficiency in diverse environments and the ability to free the hands for other tasks. Mirror neuron research provides critical insight into the **neurological** foundations of imitation and language learning, bridging the gap between motor functions and speech.

Further complexity arises from the integration of syntax, symbolic thought, and recursion in language. The ability to produce and understand hierarchical structures is unique to humans, pointing to evolutionary advancements in cognitive and neurological capacities. Despite advances in imaging and modeling, questions remain about the exact pathways through which symbolic and syntactic capabilities emerged. The transition to speech as the primary communication modality coincided with cultural and technological advancements, as evidenced by archaeological findings of art and burial rituals approximately 50,000 years ago. This shift likely reflected a threshold of cognitive sophistication that enabled humans to externalize abstract thought and convey it effectively to others [22].

Although the origin of speech is still a subject of debate, it is a fact that speech is one of the main ways of communications between humans in the modern world. This fact highlights the profound impact speech-language pathology could have on an individual's everyday life. According to a 2012 study, approximately 7.7% of U.S. children aged 3 to 17 have experienced a disorder related to voice, speech, language, or swallowing [24]. In addition, around 17.9 million adults reported having voice problems in 2012 [25]. In Europe, approximately 5.8 million children and adolescents have Developmental Language Disorder (DLD) which impairs their communication skills [26, 27].
In this chapter, we will explain how speech is produced, examine the role of the nervous system and the impact of neurological disorders on speech production, and finally, discuss how speech pathologies are typically assessed.

2.2 Speech Production Process

Speech production is the process, through which the inhaled air is transformed into speech sounds. In order to produce clear and intelligible speech, four to five subsystems, as demonstrated in figure 2.1, need to work and coordinate together, these subsystems are: respiration, phonation, resonance, articulation, and prosody [28–30]. While some linguistic and neurological models consider prosody a separate subsystem due to its unique role in speech processing, traditional speech pathology models consider it as a speech component, rather than a standalone subsystem. Each subsystem serves a specific function, by the use of different components. In the following, each subsystem of speech production —respiration, phonation, resonance, articulation, and prosody— is detailed to establish a foundation for understanding how disruptions in these areas can lead to speech disorders. This background is crucial as it provides the physiological basis for understanding how neurological disorders, such as dysarthria, impact speech. The discussion will then shift towards specific neurological conditions that affect these subsystems, setting the stage for exploring how automated methods can assist in diagnosing and classifying dysarthria.

1. **Respiration:** Involves the process of breathing, which provides the airflow necessary for speech production, also known as the aerodynamic energy. It consists of two phases:

    - **Inhalation (inspiration):** The diaphragm lowers, which increases the volume in the thoracic cavity. This results in negative pressure inside the lungs compared to the atmospheric pressure, which enables the air to flow from the outside the body into the oral and nasal cavities, through the trachea and into the lungs.
    - **Exhalation (expiration):** After inhalation, the volume of thoracic cavity decreases, resulting in positive pressure inside the lungs. In order to equalize the inside and the outside pressure, the air will rush outside the lungs if the airway is open.

2. **Phonation:** After the air has been inhaled during inspiration, the larynx acts on the exhaled air stream in order to create voice. More specifically, the voice is produced by the vibration of two vocal folds (vocal cords) in the larynx. During

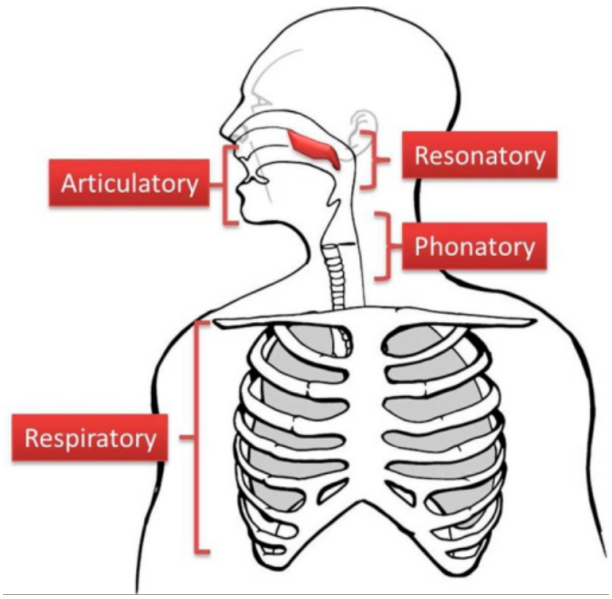
normal breathing, the vocal folds are separated from each other and the air flows easily. Once the vocal folds vibrate to and fro, the air stream passing through is released in small air puffs, which results in the production of the laryngeal tone or the vocal sound.

3. **Resonance:** The acoustic changes made to the original vocal sound, while traveling through the vocal tract. The vocal tract acts as a resonating chamber that shapes the sound into speech. Resonance contributes to the overall quality, timbre and tone of the sound, it also contributes to the uniqueness of the voice.

4. **Articulation:** It is the process through which the resonated sound is shaped into specific sounds, e.g. vowels and consonants, by the movement of the articulators. The precise and coordinated movement between the different articulators alter the shape of the vocal tract, resulting in the production of the different speech sounds.

5. **Prosody:** It is the music of the language. Prosodic features include pitch, loudness, tempo, etc. These elements work together to provide the melodic and rhythmic aspects of speech, which helps in conveying emotions, highlighting important information, or differentiating between a question, a statement and an exclamation.

Figure 2.1: The respiratory, phonatory, resonatory, and articulatory speech subsystems [1].



From the above, we conclude that speech production is a complex motor act, involving the coordination of approximately 100 muscles in the different speech subsystems [31]. Subsequently, any disorder that affects a person's ability to control their muscles could result in a speech disorder.

2.3 Neurological disorders

Neurological diseases are conditions that impair how the nervous system functions. These disorders affect the body's ability to control movements, process sensory information, and perform other critical functions such as cognition and emotion. Since the nervous system acts as the body's command center, disruptions in its communication network lead to a wide variety of symptoms, ranging from motor dysfunction to **speech problems**.

2.3.1 The nervous system

The nervous system is a highly specialized and complex structure. It is an information processing system that regulates all physiological functions of the organism, in addition to operating functions, such as consciousness and memory, that operate independently of other systems in the body [2]. To better understand the work of the nervous system, it is important to understand how it is organized functionally and anatomically. This is demonstrated in the following subsections.

2.3.1.1 Anatomy

The nervous system is divided into two main parts: the Central Nervous System (CNS) and Peripheral Nervous System (PNS). Each part plays a distinct role in managing both voluntary and involuntary actions, which is crucial for understanding how neurological diseases affect movement and speech. An overview of the anatomy of the nervous system is provided in figure 2.2:

1. The central nervous system: includes the brain, and the spinal cord. This system reads signals from the nerves, processes them and responds to the sensory information.

2. The peripheral nervous system: made up of a network of nerves, that branches out from the spinal cord. This system relays information from the CNS to the different organs and parts of the body, and vice versa. It could also be divided into two subsystems:

    - The somatic nervous system: guides the voluntary movements.
    - The autonomic nervous system: regulates the involuntary movements.

2.3.1.2 Mechanism

The nervous system contains two main types of cells: the nerve cells, or the neurons, and the glial cells. Neurons are the main communicators, while glial cells take role in support, immune surveillance, and homeostasis [2].

- Neurons have a special structure, mainly the soma, dendrites, and the axon. The basic structure of a neuron is shown in figure 2.3. The axon is what mainly differentiate between a neuron and a glia. This structure of neurons allow them to send signals rapidly and precisely to other cells, via membrane-to-membrane junctions called

Figure 2.2: The anatomical organization of the nervous system [2].

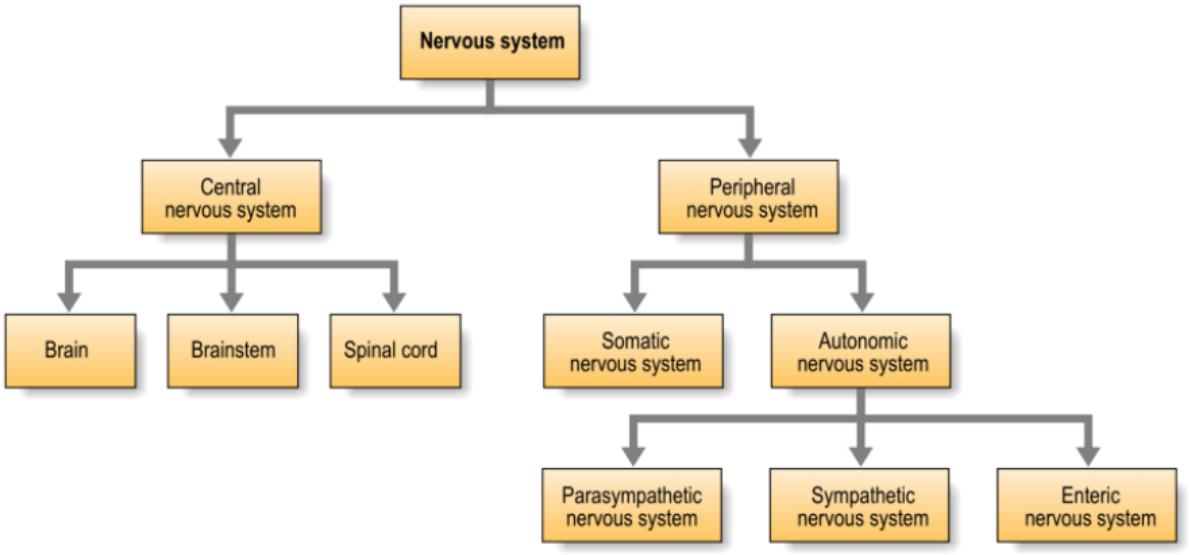
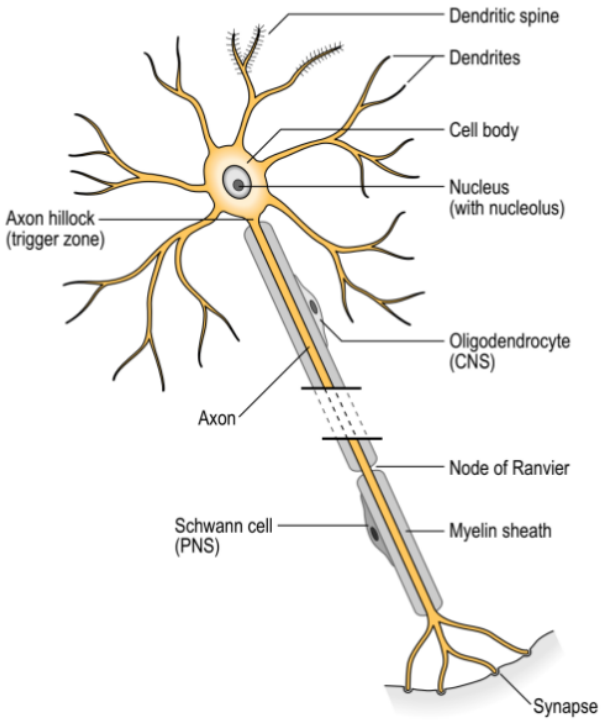


Figure 2.3: A typical multipolar neuron [2].



synapses. These signals, which take the form of electrochemical impulses, travel along the axons of the neurons, to the neighboring cells, either directly through electrical synapses, or by causing chemicals called neurotransmitters to be released at chemical synapses.
There are three types of neurons [32]:

1. Motor neurons: Carry the signals from the brain and the spinal cord to the muscles. They control the movement of the body, and coordinate the muscles and ensure that the arms and the legs move together.
2. Sensory neurons: Carry the signals from the senses to the brain. They inform the brain about the surrounding environment, and provide feedback to the muscles and joints in order to enable precise and carefully choreographed movements.
3. Inter-neurons: Communicate between motor and sensory neurons. They act as signal controllers within the body, relaying important information from one end of the nervous system to the other.

- Glial cells are non-neuronal cells that provide support and nutrition, maintain homeostasis, form myelin and participate in signal transmission in the nervous system [33]. Glia cells include astrocytes, Schwann, oligodendrocytes, and microglia [2]:
  - Astrocytes: Provide structural support, they contribute to the formation of blood-brain barrier and the nervous system's scar response, trophism and homeostasis.
  - Oligodendrocytes: Specialized cells responsible of the production of a lipid-rich sheath of the CNS covering certain axons, called myelin.
  - Schwann: They are the equivalent of oliodendrocytes in the PNS.
  - Microglia: Play crucial roles in immunological surveillance, inflammatory reactions, and the stability of neuronal circuits during development.

2.3.1.3 A Focus on Speech Production

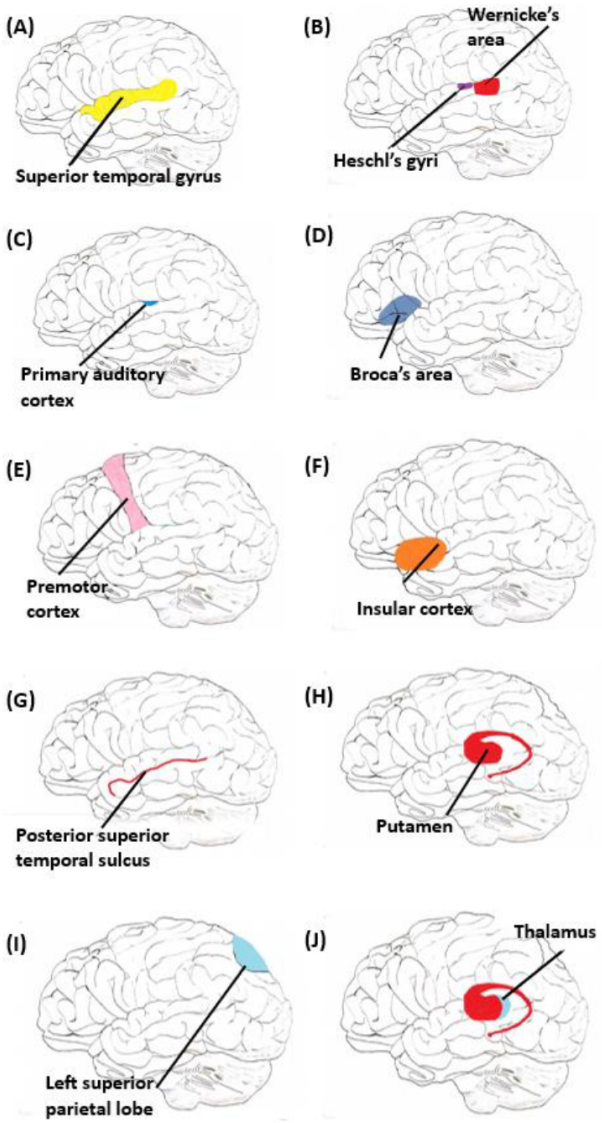
Now that we have explained how the nervous system function, in this section, we will briefly explain its role in speech production, and coordination between different muscles of speech production system.
Speech production is a complex cognitive and motor function that involves multiple brain regions working in coordination [3]. The different regions are demonstrated in figure 2.4. The process begins in the brain with neuro-motor planning, where linguistic representations are prepared and motor nerves transmit signals to the vocal muscles, enabling articulation. The speech production process in the brain follows the following steps [3]:

- Conceptualization and language encoding: Which begins in Wernick's area 2.4(B) for word selection and grammatical structure, followed by some phonological refinements in the superior temporal gyrus (STG) (fig.2.4(A)).
- Motor planning: The Broca's area (fig.2.4(D)) coordinates speech motor commands, while the premotor cortex (fig.2.4(E)), basal ganglia (Putamen) (fig.2.4(H)), and cerebellum fine-tune articulation and movement sequencing.
- Execution: The primary motor cortex sends signals to the vocal muscles. Meanwhile, the anterior insula (fig.2.4(F)) ensures smooth articulatory movements, and the putamen (fig.2.4(H)) and thalamus (fig.2.4(J)) regulate motor precision.

- Auditory feedback and correction: The superior temporal gyrus (fig.2.4(A)) monitors speech output, and the posterior superior temporal sulcus (fig.2.4(G)), along with the cerebellum, assists in error detection and self-monitoring.

Understanding the basic anatomy and function of the nervous system is critical to comprehending how neurological disorders like PD and CP disrupt motor and speech functions. This dysfunction is usually referred to as dysarthria. The following section provides more details about dysarthria: definition, symptoms and types of dysarthria. We then go into more detail about the exact definition and symptoms of PD and CP.

Figure 2.4: The different brain regions involved in the process of speech production [3].



2.3.2 Dysarthria

Dysarthria is an umbrella term grouping all motor speech disorders resulting from disturbed muscular control of the speech mechanism, due to a damage of the peripheral or

central nervous system causing weakness, incoordination, or paralysis of speech musculature [34]. Dysarthria could be developmental, due to brain damage during fetal development or at birth, like the case of cerebral palsy, usually accompanied with developmental dysarthria. Alternatively, it could be acquired as a result of a brain damage later in life, following for example a stroke, or as a result of Parkinson's disease.

2.3.2.1 Dysarthria Symptoms

Dysarthria is mainly manifested in the speech, it symptoms may include, but not restricted to: slurred speech, slow speech, rapid speech, nasal, raspy or strained voice, monotone speech, uneven or abnormal speech rhythm, speaking softly or in a whisper, difficulty chewing, swallowing, or controlling the tongue, drooling [35, 36].

2.3.2.2 Dysarthria Types

Dysarthria could be classified into one of six types, each type is characterized by the affected part of the nervous system: [37]

1. Flaccid dysarthria: results from damage to the lower motor neurons of the peripheral nervous system. It renders the speech breathy and nasal.

2. Spastic dysarthria: results from damage to the upper neurons, of the central nervous system, on one or both sides of the brain. It renders the speech strained or harsh.

3. Ataxic dysarthria: results from damage to the cerebellum, that is the part of the brain responsible of muscle movement coordination. It makes it difficult to pronounce some vowels and consonants, and to place emphasis on the right parts of a word.

4. Hypokinetic dysarthria: results from damage to the basal ganglia, that is a structure inside the brain that helps the muscles to move. It results in slow, monotone, rigid sounding speech.

5. Hyperkinetic dysarthria: also results from damage to the basal ganglia, but causes fast sounding and unpredictable speech.

6. Mixed dysarthria: it is a mix of two or more of the above mentioned types of dysarthria. It is the most common type of dysarthria.

Dysarthria is often secondary to various neurological conditions, including, but not limited to, cerebral palsy and Parkinson's disease. In this thesis, the experiments were conducted using speech samples from patients with CP and PD. Therefore, the following sections will provide a comprehensive overview on these two conditions.

2.3.3 Cerebral Palsy

Cerebral palsy is a static neurological condition resulting from brain injury that occurs before cerebral development is complete [38]. It is a group of permanent disorders of movement, posture, and motor function, due to a non-progressive interference, lesion,

or abnormality of the developing / immature brain [39]. Brain development starts in the womb, and continues up to two years after birth, this is why cerebral palsy could occur during the prenatal, perinatal, or postnatal periods. In 70% to 80% of cases, CP is acquired prenatally from unknown causes, while birth complications, such as asphyxia, account for 6% of cases, and 10% to 20% of cases are acquired postnatally due of brain damage from bacterial meningitis, viral encephalitis, etc. [38].

2.3.3.1 Clinical Presentation of Cerebral Palsy

There exist many classification of CP, the most frequently used are the ones proposed by Ingram in 1955 [40], and Hagberg in 1976 [41].
According to Ingram, classification of CP is done according to neurological syndrome, its location, and the severity of symptoms. This leads to the following possible classifications [39]:

1. Diplegia: spastic paresis occurs mainly in lower, three or four limbs.

2. Hemiplegia: unilateral spastic paresis (right- or left-sided), with predominance of the upper or lower limb.

3. Bilateral Hemiplegia (tetraplegia): spastic tetraparesis with the predominance of upper limb paresis. It is the most severe type of CP in terms of motor disability severity, and co-existing problems.

4. Ataxia: reduced muscle tension, accompanied by hand-eye coordination disorders. It could be bilateral, or with the predominance of one side of the body.

5. Dyskinesia: dystonic, athetosic, choreic type of CP, accompanied by trembling, or changes in muscle tone. It could occur in one side of the body, or in one, three, or four limbs.

6. Mixed: various combinations of the above mentioned characteristics.

As for Hagberg, classification is done according to spastic syndromes, extrapyramidal (dyskinetic) syndromes, and ataxia [39]:

1. Spastic syndromes: resulting from the damage of brain centers, and tracks controlling a given activity (monoparesis, hemiparesis, triparesis, tetraparesis, diplegia spastica).

2. Extrapyramidal (dyskinetic) syndromes: related to subcortical damage, characterized by involuntary movements, and muscles stiffness with scarce movements.

3. Ataxia: resulting from cerebellum damage. Characterized by hypotension, trembling, and motor coordination disorders.

Another simplified classification was proposed by Surveillance of Cerebral Palsy in Europe (SCPE), aiming to standardize CP classification. It separated patients into three major subgroups [39]:

1. Spastic: characterized by enhanced muscle tension, hyperreflexia, and pathological reflexes.

2. Dyskinetic: characterized by involontary, uncontrolled, repetitive, and stereotypical movements. Muscle tension can be increased or decreased, and it frequently changes over time.

3. Ataxic: characterized by motor coordination loss, resulting in ataxia, movements smoothness, and trembling. Lower muscle tension is predominant.

In addition to the motor manifestations of CP, other medical problems are noticed in CP patients, including, but not limited to, epilepsy, mental retardation, malnutrition and gastrointestinal complications, and speech problems.

2.3.3.2 Speech Problems in CP

Motor speech disorders, i.e. dysarthria, are associated with all CP types (spastic, dyskinetic, and ataxic). Children with different types of CP could manifest similar speech characteristics, this is why speech is not an indicator of the CP type [42]. Dysarthria in children with CP affects all five of speech subsystems, i.e. respiration, phonation, resonance, articulation, and prosody [43]:

- Their shallow breathing make them speak on short bursts of air, thus making their voices quiet [44].

- Slow or irregular vibration of vocal folds, resulting in low pitched, monotone, and breathy voices [45].

- Harsh voice, or rapidly varied in pitch.

- Reduced control of soft palate, resulting in nasalized speech, and reduced control of tongue and lip muscles resulting in reduced range of consonants and vowels [**?**].

- Difficulties in controlling the vocal tract could manifest in imprecision of speech in words and phrases, or in complete inability to produce intelligible speech [43].

Spastic dysarthria is the most common in CP, but other types (ataxic, hyperkinetic) or mixed forms can also be present.

2.3.4 Parkinson's Diease

Parkinson's disease is the most common neurodegenerative movement disorder [46]. It causes a specific part of the brain, the basal ganglia, to deteriorate. More specifically, it is due to the loss of dopamine-producing neurons in an area of the brain called substantia nigra, which is a part of the basal ganglia. Dopamine is one of the most important neurotransmitters that control how brain cells communicate with each other, and with different body parts. The basal ganglia plays a critical role in regulating movement, and the loss of dopamine disrupts this regulation, leading to the hallmark motor symptoms of the disease.

2.3.4.1 Pathophysiology

The underlying mechanisms that drive the progression of PD involve complex changes at the cellular and molecular levels, primarily affecting the brain's motor control pathways:

- Dopamine Depletion: Dopamine is a neurotransmitter that facilitates communication between nerve cells in the brain to coordinate smooth and controlled movements. In Parkinson's, the progressive degeneration of dopamine-producing neurons in the substantia nigra leads to dopamine deficiency.
- Lewy Bodies: Abnormal protein aggregates called Lewy bodies, primarily composed of alpha-synuclein, are often found in the brains of Parkinson's patients. These inclusions may play a role in the pathogenesis of the disease.
- Altered mitochondria: Changes to mitochondria, responsible of the creation of most of the body's energy, could cause cell damage. These changes are often observed with PD [47].

2.3.4.2 Symptoms of PD

PD often presents with a variety of motor and non-motor symptoms. The following subsections provide a non-exhaustive list of symptoms.

Motor Symptoms

Movement-related, or motor symptoms of PD include the following [47, 48]:

- Bradykinesia: Or slowed movements, it happens because of muscle control problems, and makes simple tasks more difficult.
- Tremor: Rhythmic shaking of muscles, even when at rest.
- Rigidity or stiffness: Muscles may feel tense and painful. Increased muscle tone causing stiffness in the limbs and neck, which can lead to a reduced range of motion.
- Unstable posture or walking gait: PD causes a hunched over, or stooped stance. This is usually more obvious in later stages of the disease.

Non-motor Symptoms

PD also affects non-motor functions, leading to symptoms such as [47, 48]:

- Cognitive Changes: Memory problems, difficulty concentrating, or dementia in later stages.
- Mood Disorders: Depression, anxiety, or apathy.
- Sleep Disturbances: Insomnia, rapid eye movements (REM) sleep behavior disorder, or excessive daytime sleepiness.
- Autonomic Dysfunction: Problems with blood pressure regulation, constipation, or urinary incontinence.
- Speech and Swallowing Difficulties: Dysarthria (speech impairment) and dysphagia (swallowing difficulties) can also develop.

2.3.4.3 Speech Problems in PD

Patients with Parkinson's disease typically suffer from hypokinetic dysarthria, which is the most common type associated with this neurological disorder. It is usually characterized by the following [48]:

- Impaired speech and decreased voice intensity: PD often leads to hypophonia, a reduced voice volume or soft speech, making it difficult for patients to be heard clearly.
- Monotony of pitch and loudness: This is due to the reduced range of movement in the muscles responsible for voice modulation.
- Reduced stress: The speech of PD patients may lack the normal variations in emphasis or stress, resulting in a flat or expressionless voice.
- Short rushes of speech: Speech may come out in brief, rapid bursts followed by pauses. This symptom reflects a disruption in the normal pacing of speech due to impaired motor control.
- Imprecise consonants: There can be a lack of clear articulation, making consonants sound blurred or indistinct, as fine motor control over the speech muscles is reduced.
- Breathy and harsh voice: The voice may sound breathy due to incomplete closure of the vocal folds or harsh because of increased muscle tension.

2.4 Assessing Speech Impairments: A Focus on Dysarthria

Assessing dysarthria is a crucial step in understanding the nature and severity of speech impairments resulting from neurological conditions. It involves evaluating various aspects of speech production, to identify specific motor deficits and guide appropriate treatment. Different standardized tools and techniques are used in clinical settings to measure the impact of dysarthria on communication, providing valuable insights into both diagnosis and therapeutic planning. One of the most used standardized tool is the Frenchay Dysarthria Assessment (FDA). Another assessment tool is the Batterie d'évaluation clinique de la Dysarthrie (BECD), which is a French-language clinical assessment tool.

2.4.1 The Frenchay Dysarthria Assessment

The FDA [49] is designed to evaluate the nature and severity of dysarthria. It evaluates several aspects of speech production, usually affected by dysarthria, by assessing 28 relevant perceptual dimensions of speech, grouped into 8 categories:

- Reflexes: Cough, swallow, dribble.
- Respiration: Breathing control at rest and during speech.
- Lips: The range of movement, strength, and coordination.

- Jaw: Both at rest and during speech.
- Palate: Soft palate function during speech.
- Laryngeal function: Voice quality, pitch, and loudness.
- Tongue: Its movement and coordination.
- Intelligibility: How understandable is the speech.

Each area is scored based on the observed abilities and impairments, and that is used to determine the type and severity of dysarthria, and inform the treatment approach.

2.4.2 The Batterie d'évaluation clinique de la Dysarthrie

THE BECD is a french-language assessment tool, that examines various aspects of speech production. It evaluates different speech-related functions, including:

- Respiration: Assesses breathing patterns and control during speech.
- Phonation: Evaluates voice quality, pitch, loudness, and other vocal characteristics.
- Articulation: Measures the precision and clarity of consonants and vowels.
- Resonance: Looks for signs of hypernasality or hyponasality, which can affect speech quality.
- Prosody: Assesses the rhythm, stress, and intonation patterns of speech.

It includes tasks designed to assess each speech production area, such as reading aloud, repeating sentences, or performing oral motor tasks. For each task, a score is given according to the observed performance, providing a profile of the patient's speech abilities and identifying specific areas of impairment. The BECD helps clinicians determine the appropriate therapeutic interventions by providing a detailed assessment of the speech impairments. More details on the BECD could be fined in Appendix I.
Speech-Language Pathologist (SLP)s are in short supply, in 20230, there were 60.8 certified SLPs for every 100,000 residents in the U.S. [50], and they typically carry heavy caseloads, limiting the amount of one-on-one time they can devote to each client. Yet effective dysarthria rehabilitation demands frequent, structured practice, patients must perform speech exercises regularly to achieve meaningful improvement. For these home sessions to be truly beneficial, an SLP still needs to track the patient's progress and adjust therapy plans accordingly. Relying on the expertise of SLP for assessing dysarthria can introduce subjectivity, depending on the clinician's experience. Furthermore, some traditional assessment methods, such as stroboscopy, can be expensive or invasive, and access to SLPs or specialized tools may be limited in remote areas. Together, the scarcity of SLPs, the necessity of guided home practice, and the limitations of traditional assessment tools underscore the pressing need for automated, objective solutions that can support both patients and clinicians wherever they are.

2.5 Conclusion

This chapter introduced the process of speech production, the muscles involved, and the role of the nervous system in the process. Neurological diseases often disrupt the function of the nervous system, leading to various issues, including speech problems. Hence, we introduced two neurological diseases that will be the focus of the experiments in later chapters, cerebral palsy and Parkinson's, both of which can cause dysarthria. We finished that chapter by discussing two of the methods used by speech language pathologists to assess dysathria. In the following chapters, we will explore key techniques necessary for automating speech pathology assessment, and machine learning based solutions for automatic dysarthria classification models.

Chapter 3

Signal, Features, and Models: A Technical Perspective

3.1 Introduction

Automating dysarthria severity assessment supplies objective, reproducible scores that lessen the routine rating burden on already over-extended SLPs, let them track patients' at-home practice remotely, and sharply cut the variability, and resulting misdiagnoses, that arise when evaluations rely solely on clinician's subjectivity. Automation can standardize the assessment process, ensuring consistent evaluation of speech characteristics across different patients and settings. Additionally, automated systems can quickly analyze large datasets [51], providing immediate feedback and facilitating the identification of subtle changes in speech patterns over time, which is particularly beneficial for monitoring disease progression in neurodegenerative conditions like PD. Integrating automated assessments into telehealth platforms can further improve access to care for patients in remote areas, enabling timely interventions and personalized treatment plans. Overall, automating dysarthria severity assessments not only enhances diagnostic precision but also optimizes clinical workflows, ultimately leading to better patient outcomes. Yet those gains come with non-trivial challenges. Robust performance demands large, well-annotated speech corpora that capture the full diversity of languages, accents, recording conditions, and severity levels, such datasets are still scarce. Machine learning models can inherit bias from skewed training data, producing less reliable scores for minority dialects or atypical voice qualities. Everyday factors, such as background noise, microphone quality, variable internet bandwidth, can degrade acoustic input and compromise accuracy. Clinicians also need transparent algorithms and clear confidence metrics to trust automated ratings and to know when to override them. Beyond technical obstacles, developers must address data-privacy regulations, obtain regulatory clearance, and embed these tools seamlessly into existing electronic health record and telepractice workflows. Without careful attention to these issues, automated assessments risk adding new sources of variability instead of removing them.

Automated systems leverage advances in machine learning and signal processing to analyze speech characteristics, enabling the identification of patterns associated with different types of dysarthria. This approach can provide significant advantages in terms of

efficiency, consistency, and scalability compared to traditional assessment methods, which often rely on subjective evaluation by experts.
In this chapter, we explore the fundamental principles and techniques of signal processing as they relate to speech analysis. Starting with the basics of sound and audio representation, we delve into key concepts such as spectral analysis, feature extraction, and the mathematical transformations essential for processing and analyzing speech signals in automated systems. We then proceed to define machine learning concepts, which constitutes the basis for the classifiers used in automated systems.

3.2 Basics of Sound and Audio Signals

Sound originates when an object's vibrations cause nearby air molecules to move back and forth. These oscillations lead to collisions between molecules, creating alternating areas of high pressure (compression) and low pressure (rarefaction). This fluctuation in pressure travels outward as a wave. Since sound is a mechanical wave, it requires a medium—such as air, water, or solid materials—to propagate, as it transmits energy by deforming the medium. As the wave moves, it pushes the air molecules together and then apart, carrying the sound energy to our ears, where we perceive it as sound. A complex sound can be represented by a **waveform**, which is a graph showing how air pressure deviates from the normal atmospheric pressure over time. The waveform captures various characteristics of the sound, such as frequency, intensity, and timbre, etc. .
Sound is a complex phenomenon with various qualities that shape how we perceive and interpret it. These qualities, or attributes, define the uniqueness of different sounds and influence our auditory experience. By examining the key characteristics of sound, we can better understand how factors such as frequency and energy contribute to what we hear and how we perceive it. The following sections explore these fundamental attributes, including the nuances of pitch, the power behind intensity, the perception of loudness, and the distinctive quality known as timbre.

3.2.1 Pitch

Frequency is an objective measure of a sound's oscillations per second, but our perception of frequency is subjective and is known as pitch. We perceive pitch in a logarithmic, rather than linear, manner. This means that two frequencies are perceived as similar in pitch if one is a multiple of the other by a factor of 2, which corresponds to an octave [52]. For example, a frequency of 440 Hz and 880 Hz sound like different versions of the same pitch, one octave apart, because 880 Hz is double 440 Hz.

3.2.2 Intensity and power

Sound power, measured in watts (W), represents the total energy emitted by a sound source in all directions. Sound intensity, measured in watts per square meter (W/m^2), is the sound power distributed over a given area. Humans can perceive a wide range of sound intensities, from as faint as 10^{-12} W/m^2, known as the threshold of hearing (TOH), to as intense as 1 W/m^2, known as the threshold of pain (TOP) [53]. This vast hearing range makes it practical to use the intensity level, measured in decibels (dB), to

describe sound intensity. The intensity level is a logarithmic scale that represents the ratio between the actual sound intensity and a reference intensity. While power and intensity are objective measurements, loudness is a subjective perception that describes how we experience the intensity of a sound. Measured in phons, loudness is influenced by various factors, including the sound's frequency, duration, and the listener's age [54]. As a result, sounds with the same intensity may be perceived as louder or softer depending on these conditions.

3.2.3 Timbre

Timbre, often described as the color of sound, is what distinguishes two sounds that share the same frequency, duration, and intensity. For example, when a guitar and a piano play the same note at the same volume, the sounds are still uniquely different because of their distinct timbres. Although both instruments may be equally in tune and have the same amplitude, each produces a characteristic tone that sets them apart. Another example is when two people say the word 'hello' at the same pitch and loudness, their voices still sound unique. One voice might have a bright, resonant quality due to stronger high-frequency harmonics, while another might have a darker, richer sound because of more pronounced low-frequency components. These subtle variations in harmonic content and the way each person's vocal cords and vocal tract shape the sound give their voices a distinct timbre. Timbre is shaped by several key features, including:

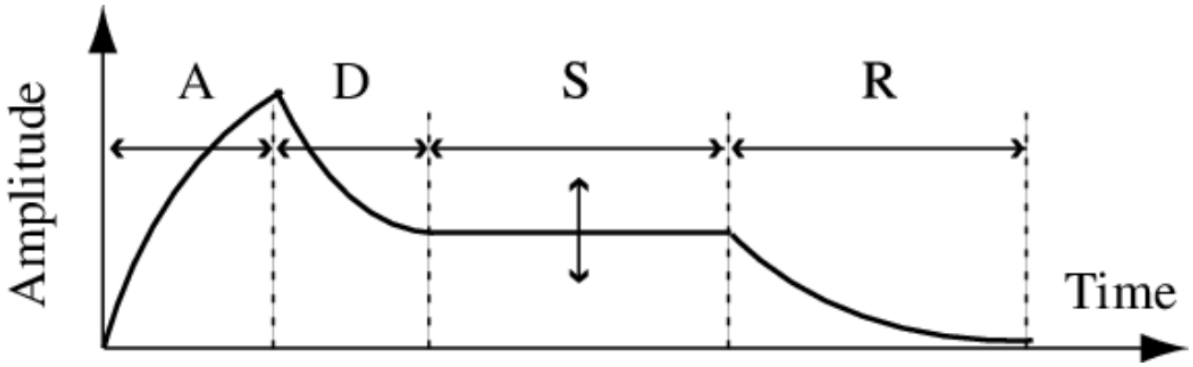
- Sound envelope: This describes how a sound evolves over time, typically represented using the Attack-Decay-Sustain-Release (ADSR) model, which outlines the different stages of a sound's lifespan, as shown in figure 3.1
- Harmonic content: A sound is composed of multiple sinusoidal waves, called partials. The lowest partial is the fundamental frequency, while harmonic partials are frequencies that are integer multiples of this fundamental frequency. The combination and relative strengths of these partials define the sound's unique character.
- Frequency / amplitude modulation: These methods encode information by varying the frequency or amplitude of a carrier wave, altering the timbre in distinct ways, as demonstrated in figure 3.2.

Together, these factors give each instrument or sound source its unique tonal quality.

3.3 Digital Representation of Audio

An audio signal is a digital representation of sound that encodes all the information needed to reproduce it. While sound is naturally an analog signal—continuous in both time and amplitude—it must be converted into a digital signal, which consists of discrete values, in order to be stored or processed digitally. This conversion is achieved through the process of analog-to-digital conversion (ADC), which involves two main steps: sampling and quantization. Sampling captures the sound at regular time intervals (sampling period T), while quantization maps the amplitude values to a finite set of levels, allowing the continuous analog signal to be approximated as a sequence of digital values.

Figure 3.1: A graphic representation of an ADSR envelope [4].



3.3.1 Sampling

Sampling involves capturing data points from a sound wave at specific intervals over time. These intervals are evenly spaced and correspond to the sampling period (T). An important aspect of sampling is the sampling rate, which is the inverse of the sampling period and indicates the number of samples taken per second. A lower sampling rate results in higher error and loss of information, as fewer data points are used to represent the original sound. Conversely, a higher sampling rate reduces the error by more accurately capturing the details of the sound wave.

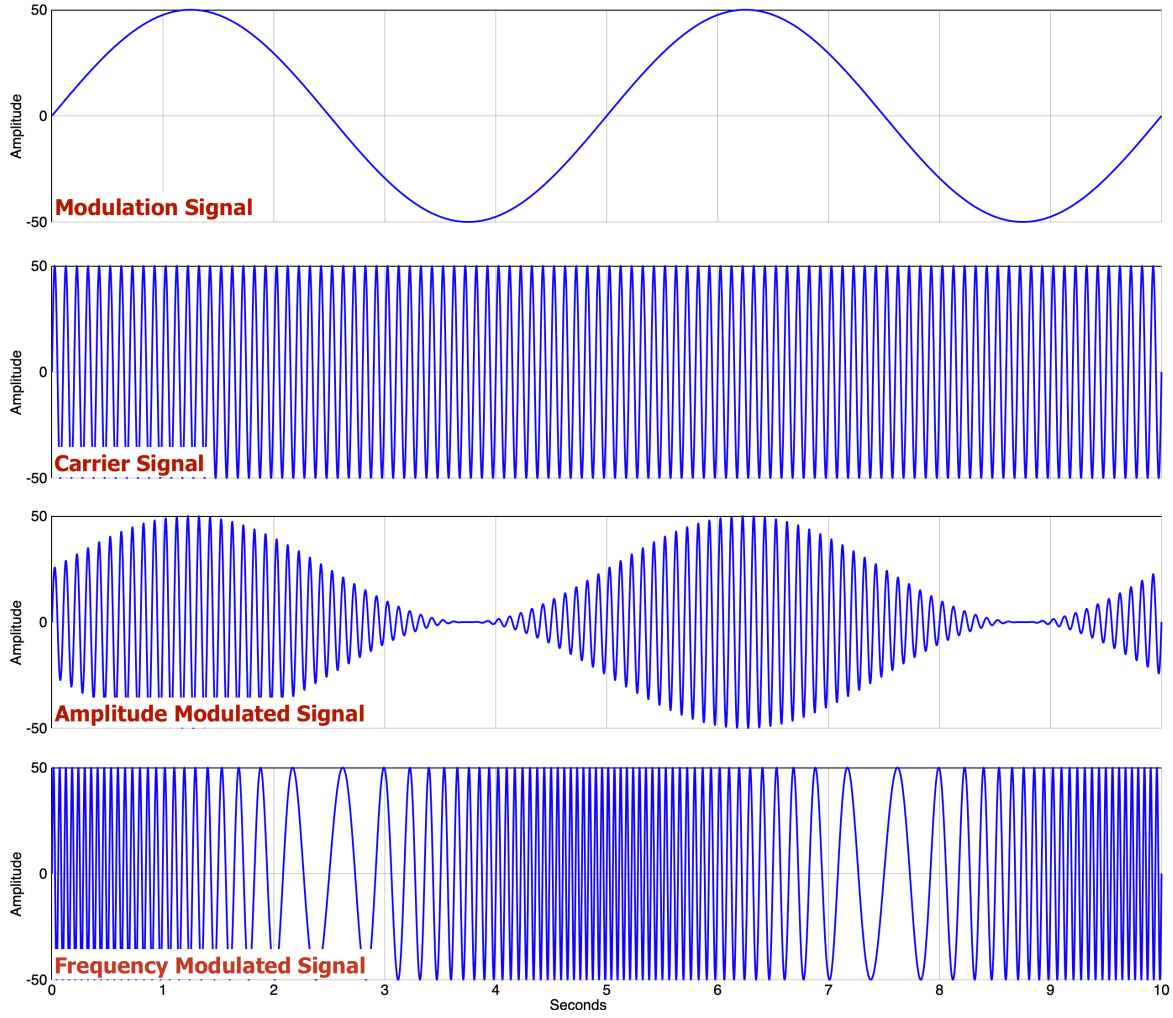
3.3.2 Quantization

Quantization is the process of approximating the amplitude of each sampled data point to the nearest value in a defined set of discrete levels. The resolution of quantization, determined by the number of available levels, affects the accuracy of this approximation. Higher resolution, meaning more levels in the array, reduces quantization error and results in a more precise representation of the original sound. Conversely, lower resolution increases the quantization error, leading to a less accurate reproduction.

3.3.3 Audio Features

Audio features are measurable characteristics that describe various aspects of sound. Each feature captures a specific quality of the audio signal, such as its frequency content, amplitude variations, or temporal patterns. These features are essential for training machine learning models, as they provide the data needed for the algorithms to learn patterns and make predictions relevant to the task at hand. By selecting appropriate audio features, we can improve the system's ability to recognize, classify, or analyze different sound-related tasks, such as speech recognition, music genre classification, emotion detection, or, as in our case, dysarthria severity classification. Audio features can be categorized according to different strategies: level of abstraction, temporal scope, signal domain, ML approach, etc. . In the following, we will introduce categories according to the signal domain:

Figure 3.2: An analog signal (top) is merged with a carrier wave (second from top) into an Amplitude Modulated signal (third from top) or a Frequency Modulated signal (bottom) [5].



- Time domain features: These features are extracted directly from the waveform, which represents the amplitude of the audio signal as it varies over time.
- Frequency domain features: These features are derived from the signal's spectrum, after it has been transformed into the frequency domain using Fourier transform, which will be explained in subsequent sections.
- Spectral domain: Spectral domain features are obtained from a time-frequency representation of the audio signal, which provides information about how the signal's frequency content changes over time. This representation is often visualized as a spectrogram, which will also be explained in subsequent sections.

In order to be able to extract the aforementioned features across various domains, the transition from one domain to another must be explained. This process of transition between time, frequency, and spectral domains is presented in the following sections.

3.4 Signal Processing Techniques for Speech Analysis

In the following sections, we introduce key signal processing techniques that are crucial for understanding the transition from the time domain to the frequency domain, as well as for extracting features and applying feature extraction methods used in this thesis.

3.4.1 Fourier Transform

FTshort=FT,long= Fourier transform The **FT!** (**FT!**) decomposes a complex sound into its frequency components, by passing from the time domain into the frequency domain. This means that instead of representing a signal as variations in amplitude over time, it breaks the signal down into a sum of sinusoidal waves, each characterized by a specific frequency, amplitude, and phase. These sinusoidal components reveal how much of each frequency is present in the original signal. In essence, the Fourier transform converts a time-based signal, where changes are plotted against time, into a frequency-based signal, where the intensity of each frequency is plotted. This transformation is fundamental in various fields, such as audio processing, image analysis, and telecommunications, as it allows for the analysis of signals in terms of their constituent frequencies. The formula for the continuous Fourier Transform is:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \cdot e^{-j\omega t}\, dt \tag{3.1}$$

where X(ω) is the frequency-domain representation of the signal, x(t) is the original time-domain signal, ω is the angular frequency, j is the imaginary unit ($j^2 = -1$), and $e^{-j\omega t}$ represents the complex exponential function.

Inverse Fourier transform
The Inverse Fourier Transform (IFT) is used to reconstruct the original time-domain signal from its frequency-domain representation. It essentially reverses the process of the Fourier transform by combining the frequency components back into a time-varying signal. This allows for a seamless transition between the time and frequency domains, making it possible to analyze a signal in the frequency domain, manipulate it if necessary, and then revert to the time domain. The formula for the continuous Inverse Fourier Transform is:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) \cdot e^{j\omega t}\, d\omega \tag{3.2}$$

where x(t) is the reconstructed time-domain signal, X(ω) is the frequency-domain representation of the signal, ω is the angular frequency, j is the imaginary unit, and $e^{j\omega t}$ is the complex exponential function.

3.4.2 Discrete Fourier Transform

The Discrete Fourier Transform (DFT) is a numerical version of the Fourier transform, used for processing signals that are sampled at discrete intervals. It decomposes a finite

sequence of equally spaced time-domain samples into its frequency components, allowing the signal to be represented as a sum of discrete sinusoidal waves. The DFT converts a discrete time-domain signal into a frequency-domain representation, showing how much of each frequency is present in the signal. This is essential in digital signal processing, such as audio analysis and image processing. The formula for the DFT is:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi}{N}kn} \tag{3.3}$$

where x[n] is the time-domain signal at sample n, X[k] is the frequency-domain representation at frequency index k, N is the total number of samples, j is the imaginary unit, and $e^{-j\frac{2\pi}{N}kn}$ represents the complex exponential function.

Inverse Discrete Fourier transform
For signals sampled at discrete intervals, the Inverse Discrete Fourier Transform (IDFT) reconstructs the original time-domain signal from its discrete frequency-domain representation. The IDFT performs the inverse operation of the DFT, combining the sinusoidal components to obtain the original sequence of time-domain samples. The formula for the IDFT is:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{j\frac{2\pi}{N}kn} \tag{3.4}$$

where x[n] is the reconstructed time-domain signal at sample n, X[k] is the frequency-domain representation at frequency index k, N is the total number of samples, j is the imaginary unit, and $e^{j\frac{2\pi}{N}kn}$ represents the complex exponential function.

3.4.3 Short-Time Fourier Transform

The Short-Time Fourier Transform (STFT) extends the Fourier transform to analyze localized frequency content over time. It works by segmenting the signal into overlapping time windows and applying the Fourier transform to each segment. This allows for the observation of how frequency components change over time, making it useful for non-stationary signals such as speech or music. The STFT provides a time-frequency representation of the signal, with each segment capturing the frequency content within a specific time interval. The formula for the STFT is:

$$\text{STFT}(m, \omega) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n-m] \cdot e^{-j\omega n} \tag{3.5}$$

where x[n] is the original signal, w[nm] is a window function centered at time m, determining the segment of the signal to be analyzed, ω is the angular frequency, and m is the time index indicating the start of each window.

Inverse Short-Time Fourier Transform
The Inverse Short-Time Fourier Transform (ISTFT) reconstructs the original time-domain signal from the STFT. It essentially combines the inverse Fourier transforms of the segments, using overlapping windows to ensure smooth reconstruction. The ISTFT is useful

in applications like audio signal processing, where modifications in the time-frequency domain are reverted back to the original sound signal. The ISTFT formula is similar to the inverse Fourier Transform applied to each segment, with appropriate overlap and summing, given as:

$$x[n] = \sum_{m} (\text{STFT}(m, \omega) \cdot w[n - m]) \cdot e^{j\omega n} \tag{3.6}$$

where x[n] is the reconstructed time-domain signal, STFT(m,ω) is the time-frequency representation, and w[nm] is the window function used during the analysis. Together, the Fourier Transform and its inverses enable the switching between the time and frequency domains, facilitating analysis, processing, and reconstruction of signals.

3.4.4 Spectrogram

A spectrogram is a visual representation of the STFT, showing how the signal's frequency content evolves over time. It displays time on the horizontal axis, frequency on the vertical axis, and the amplitude (or power) of each frequency component as color intensity. By taking the magnitude of the STFT, the spectrogram reveals the energy distribution across frequencies for each time segment, making it useful for analyzing complex signals such as speech, music, or seismic data. The spectrogram is calculated as:

$$\text{Spectrogram}(m, \omega) = |\text{STFT}(m, \omega)|^2 \tag{3.7}$$

where $STFT(m, \omega)$ represents the magnitude of the STFT at a given time and frequency, an the squared magnitude emphasizes the power or intensity of the frequency components. Together, these tools —FT, DFT, STFT, and spectrograms— enable the analysis of signals in both time and frequency domains, providing insights that are not apparent in the original time-domain representation.

3.4.5 Mel Spectrogram

A Mel spectrogram is a time-frequency representation of a signal, where the frequency axis is mapped to the Mel scale, which better reflects the human ear's perception of pitch. In a Mel spectrogram, the frequency resolution is higher at lower frequencies and lower at higher frequencies, aligning more closely with how humans perceive sound. This makes Mel spectrograms particularly useful in applications such as speech recognition, audio classification, and music analysis, where the goal is to capture perceptually meaningful features. The process to compute a Mel spectrogram involves the following steps:

1. Compute the Short-Time Fourier Transform (STFT) of the signal to obtain a spectrogram.

2. Apply a Mel filter bank, which consists of overlapping triangular filters spaced according to the Mel scale, to the spectrogram. This maps the linear frequency bins to Mel frequency bins.

3. Optionally, take the logarithm or use a power scaling on the filtered output to obtain a log Mel spectrogram, which more closely matches the human perception of loudness.

The Mel scale is defined as follows:

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.8}$$

where f is the frequency in Hertz, and Mel(f) is the corresponding frequency on the Mel scale.
The Mel spectrogram emphasizes the perceptually relevant parts of the signal by mapping the frequency axis according to the Mel scale, resulting in a representation that is more aligned with how humans perceive sound differences, especially in speech and music applications.

3.4.6 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are features used in speech and audio signal processing to represent the short-term power spectrum of a sound. They capture perceptually meaningful aspects of the audio signal by combining the Mel scale with cepstral analysis, making them a popular choice for tasks like speech recognition, speaker identification, and audio classification. MFCCs aim to model the human auditory system's response to sounds by emphasizing the frequency ranges that are most important for speech perception. The extraction of the coefficients is done according to the following steps:

1. Pre-emphasis: The audio signal is first passed through a pre-emphasis filter to amplify higher frequencies. This compensates for the natural decline in amplitude at higher frequencies.

2. Framing: The signal is divided into overlapping frames, typically 20–40 milliseconds long, to capture the signal's local properties. Overlapping is done to ensure continuity between frames.

3. Windowing: Each frame is multiplied by a window function, such as a Hamming window, to minimize spectral leakage. The Hamming window is defined as: $w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$.

4. STFT: The Fourier transform is applied to each windowed frame to obtain the frequency spectrum, transforming the signal from the time domain to the frequency domain.

5. Mel Filter bank: The power spectrum is then passed through a set of triangular bandpass filters, spaced according to the Mel scale. This step emphasizes frequencies that are more important for human hearing. The Mel scale is defined by formula (3.8).

6. Logarithm of the Mel Spectrum: Take the logarithm of the Mel-scaled filter bank output. This step simulates the human ear's sensitivity to loudness.

7. DCT: Apply the DCT to the log Mel spectrum to obtain the Mel-frequency cepstral coefficients. The DCT decorrelates the coefficients and compacts the information, making it easier to retain only a few coefficients (typically 12–13) that capture the most important information.

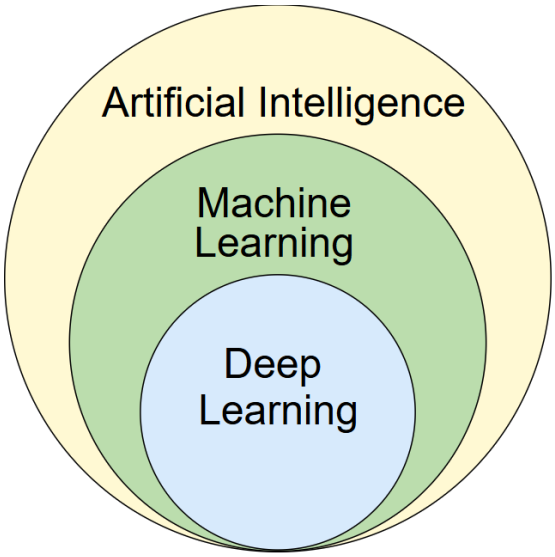
The resulting MFCCs provide a compact representation of the audio signal that reflects perceptual characteristics of sound, making them well-suited for machine learning tasks in audio analysis.

3.5 Machine Learning

3.5.1 From AI to ML to DL: Understanding the Evolution of Intelligent Systems

Artificial intelligence (AI) is a field of computer science, focused on developing systems capable of mimicking human intelligence, encompassing reasoning, learning, problem solving, perception, decision making, using language, etc.. **Machine learning** (ML) is a subset of AI that enable systems to learn directly from data, and generalize to unseen data, as stated in the 1950s by AI pioneer Arthur Samuel "the field of study that gives computers the ability to learn without explicitly being programmed". **Deep learning** (DL) is a subset of ML that uses a specific model of ML, multilayered neural networks, called deep neural networks, to simulate the complex decision-making power of the human brain. The difference between ML and DL is that ML uses simple neural networks with one or two computational layers, while DL uses three or more layers, typically hundreds or thousands of layers [**?**]. To summarize, as shown in figure 3.3, DL is a subset of ML, and ML is a subset of AI, each narrowing in focus while advancing in complexity.

Figure 3.3: The relationship between AI, ML, and DL [**?**].



3.5.2 Training Techniques

As explained in the aforementioned section, ML and DL models are trained to perform a task using data. Training the ML/DL models to accomplish a specific task is usually done following certain learning approach. Several learning approaches may be used, in the following, we present the learning approaches used in this thesis:

- **Supervised learning:** Uses labeled datasets to train models to identify the underlying patterns and relationships between input features and outputs [**?**].
- **Self-supervised learning (SSL):** Firstly introduced in robotics, this term described training models using unlabeled data, by leveraging relations between input signals [55]. This concept was then further developed by the ML community, and self-supervised learning became predicting one part of the input from another part of this input. Thus, it obtains labels from the data itself, removing the need for manually labeled examples.
- **Contrastive learning:** Contrastive Learning is a self-supervised representation learning by training a model to learn useful representations by distinguishing between similar ("positive") and dissimilar ("negative") sample pairs [56]. It recognizes similarities between occurrences that are similar and differences between cases that are dissimilar. Contrastive learning pushes dissimilar examples farther apart and brings comparable instances closer together in feature space.
- **Transfer learning:** Transfer learning is not a standalone training paradigm; rather, it is an approach that leverages knowledge learned in one setting to improve performance in another, typically smaller or related, task. It uses existing trained models, and then reuses or adapts them (by fine-tuning the models) to new tasks or domains. The fine-tuning is often done in a supervised manner.

3.5.3 Classical Machine Learning Models

In the following sections, we will introduce the models used in this thesis.

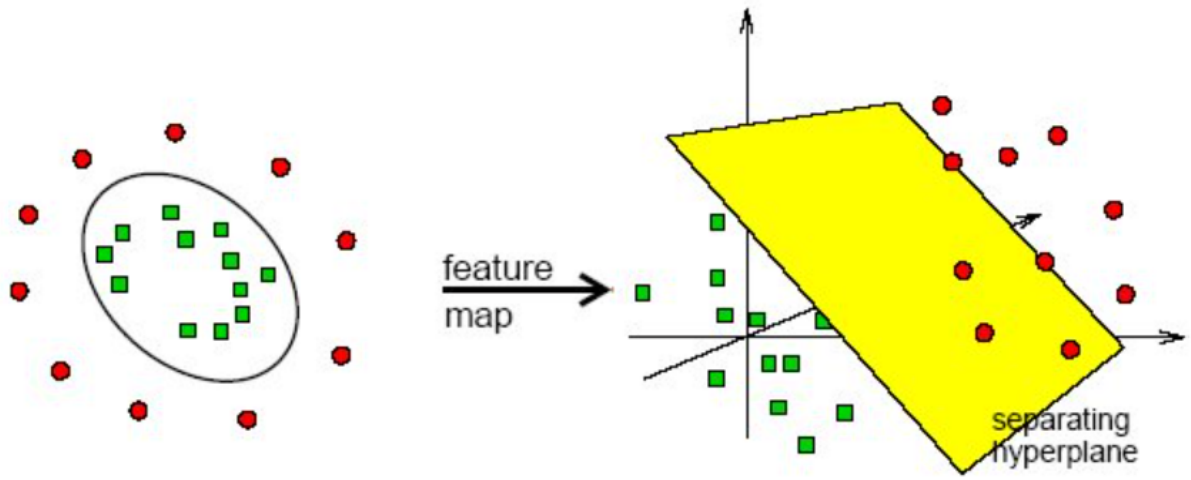
3.5.3.1 Support Vector Machine

Support Vector Machine (SVM) [57] is a supervised ML algorithm, commonly used in classification tasks. SVM works by identifying an optimal hyperplane that separates data points belonging to different classes, maximizing the margin between them. If a hyperplane in the original space is unable to separate the data, a non-linear function will be used to project the data to a higher-dimensional space where separation is possible. The most widely used options for the kernel function are the Gaussian radial basis function, sigmoid, or polynomial. Figure 3.4 provides an example where a line cannot be used to separate x values. A kernel is applied, and a hyperplane is now able to divide the data once it has been projected into its new location.
Its ability to handle high-dimensional spaces and employ kernel functions for non-linear classification makes SVM a robust choice for various applications, including speech and signal processing tasks.

3.5.3.2 Neural Networks

Neural Networks encompass a diverse set of models that have been instrumental in advancing modern machine learning. Their adaptive learning capabilities and ability to model complex non-linear relationships have led to their widespread application in classification tasks. Within this category, several key architectures are noteworthy:

Figure 3.4: Example of data seperated using SVM [6].



Multi-layer Perceptron

Multi-layer Perceptrons (MLP) are feedforward neural networks consisting of an input layer, one or more hidden layers, and an output layer. They use non-linear activation functions to learn intricate mappings between inputs and outputs. During training, the network adjusts its weights and biases using the backpropagation algorithm to minimize a loss function.

MLPs form the foundation of many neural network approaches and are typically applied in tasks ranging from regression to classification, including initial experiments in feature extraction.

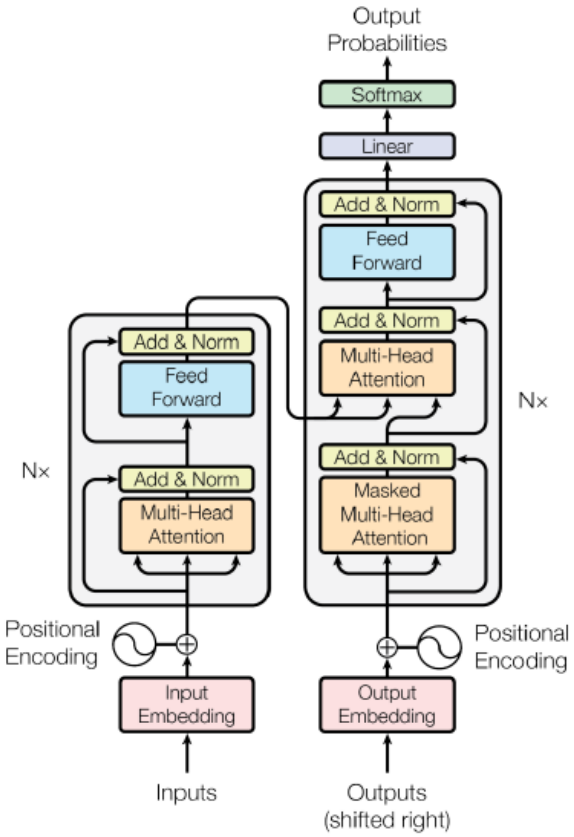
Convolutional Neural Network

Convolutional Neural Networks are designed to process data with a grid-like topology, such as images and spectrograms. By applying convolutional filters and pooling operations, CNNs automatically learn hierarchical features that capture spatial or temporal patterns. This characteristic renders them particularly effective for applications in audio processing and speech recognition, where local patterns and invariances are significant.

Transformers

Transformer is a ML architecture introduced in 2017 by Vaswani et al. [7]. The transformer has an encoder-decoder structure shown in Figure 3.5 while using the self-attention mechanism. Transformers introduce a paradigm shift with their self-attention mechanism, allowing models to weigh the relevance of different parts of the input sequence dynamically. Unlike recurrent or convolutional architectures, transformers process data in parallel, which enhances their ability to capture long-range dependencies and contextual relationships. Their success in natural language processing [58, 59] has spurred interest in applying transformers to other domains, including speech processing and classification tasks, where they can leverage contextual information to boost performance.

Figure 3.5: The architecture of the transformer [7].



3.6 Pre-trained Models

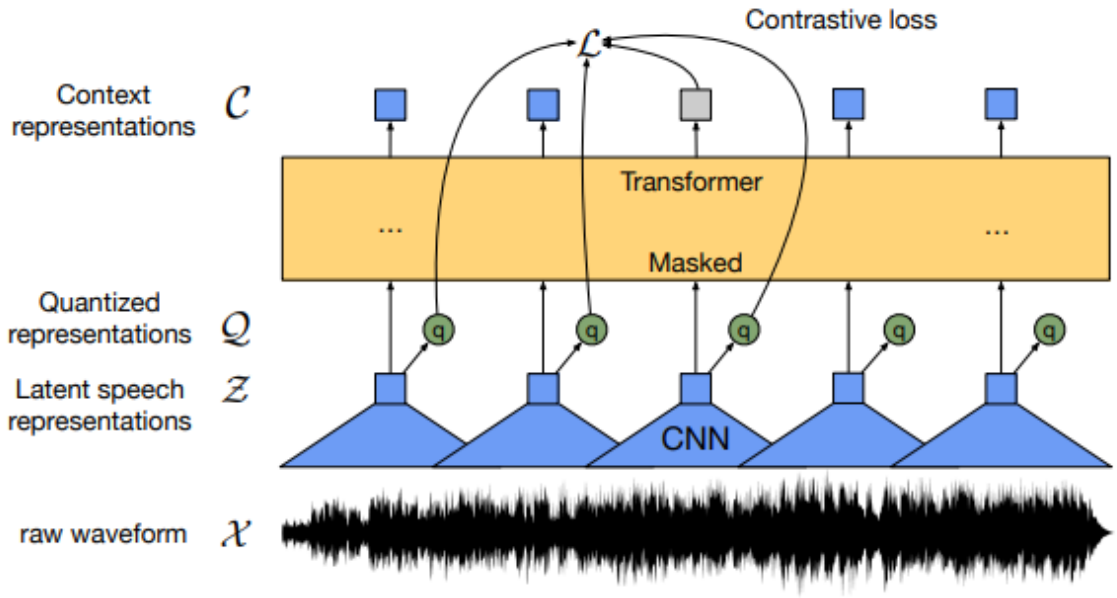
As explained in section 3.5.2, transfer learning reuses pre-trained models on new tasks, with limited data. This approach was used in this thesis, using the following pre-trained models.

3.6.1 Wav2vec2.0

Wav2vec2.0 is a framework for SSL that learns representations from raw audio data [8]. It employs a transformer-based neural network architecture, which is an architecture that has demonstrated notable success in natural language processing (NLP), especially when coupled with SSL. This architecture is also starting to gain popularity in the field of speech processing. By pre-training on large datasets of unlabeled audio, wav2vec2.0 can effectively learn general features, which can then be fine-tuned for downstream tasks, such as speech recognition, using limited labeled data.
Wav2vec2.0 is pre-trained by first encoding the speech audio using a multi-layer convolutional neural network. Then, on one hand, parts of the latent speech representations are masked, before being fed to a transformer network to build contextualized representations. And on the other hand, the latent speech representations are quantized into discrete speech units, via a gumbel softmax. The model is finally pre-trained via a contrastive task, where the true discrete latent is to be distinguished from distractors. This process

is illustrated in figure 3.6.

Figure 3.6: Illustration of the wav2vec2.0 framework, which jointly learns contextualized speech representations and an inventory of discretized speech units [8].



3.6.1.1 Model Architecture

The model is composed of three main parts: a multi-layer convolutional feature encoder, a transformer, and a quantization module.

1. **Feature Encoder:** The raw audio signal is fed into the feature encoder, which outputs latent speech representations for T time-steps. The feature encoder consists of seven blocks of temporal convolutional, a layer normalization and a GELU activation function. Each convolution have 512 channels with strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,3,3,2,2).

2. **Transformer:** The representations are then passed to a context network following the Transformer architecture, which outputs contextualized representations that captures information from the entire sequence. A relative positional embedding -a convolutional layer followed by a GELU and a layer normalization- is used instead of a fixed one. There exists two configurations of the Transformer part:

   a) Base model: composed of 12 transformer blocks, with a model dimension of 768, an inner dimension of 3072, and 8 attention heads.

   b) Large model: composed of 24 transformer blocks, with a model dimension of 1024, an inner dimension of 4096, and 16 attention heads.

3. **Quantization module** The latent speech representations are also passed to the quantization module, to get the targets for the self-supervised objective. This module discretize the feature encoder output into a finite set of speech representations

via product quantization, that is choosing a quantized representations from multiple codebooks and concatenating them.

3.6.1.2 Model Pre-training

Pre-training of wav2vec2.0 model is done using masked prediction, i.e. some time steps of the latent feature encoder representations are masked, and the model is trained with the objective of identifying the correct quantized latent audio representation from a set of distractors. In the following, we elaborate on the masking step and the training objective.

Masking
Masking of parts of the output of the feature encoder is done by randomly sampling a certain proportion p of all time steps to be starting indices, and then masking the following M consecutive time steps from each sampled index. The masked parts are substituted with a trained feature vector shared across all masked time steps.

Objective
Speech representations are learned by solving a contrastive task, which requires identifying the true quantized latent speech representation for a masked time step, within a set of distractors. The objective function is composed of two main loss components: the contrastive loss and the diversity loss. The overall loss is defined as:

$$L = L_m + \alpha L_d \tag{3.9}$$

where:

- L_m: the contrastive loss, which drives the model to correctly match context and quantized latent speech representations.
- L_d: the diversity loss, which encourages usage of diverse entries in the quantized codebook.
- α: a tunable hyperparameter that balances the contributions of L_m and L_d.

The contrastive loss L_m, is designed to teach the model to identify the correct quantized latent speech representation for a masked timestep by distinguishing it from other distractor candidates. For each masked timestep t, the model aims to maximize the similarity between the context representation c_t and the correct quantized latent representation q_t while minimizing the similarity with incorrect candidates. The contrastive loss is defined as::

$$L_m = -\log \frac{\exp(\text{sim}(c_t, q_t)/\kappa)}{\sum_{\tilde{q} \sim Q_t} \exp(\text{sim}(c_t, \tilde{q})/\kappa}} \tag{3.10}$$

where:

- $\text{sim}(a, b)$: the cosine similarity between vectors a and b, calculated as:

$$\text{sim}(a, b) = a^T b / \|a\| \|b\|$$

- c_t: the context representation for the masked timestep t.

- q_t: the true quantized latent representation for the masked timestep.
- Q_t: a set of distractor quantized representations sampled for timestep t.
- κ: a temperature parameter that controls the sharpness of the similarity distribution.

The diversity loss, L_d, encourages the model to use a wide range of quantized representations, promoting codebook diversity. This prevents the model from overusing a small subset of codebook entries, thereby improving representation quality. The diversity loss is defined as:

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\tilde{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \tilde{p}_{g,v} \log \tilde{p}_{g,v} \tag{3.11}$$

where:

- G: the number of codebooks.
- V: the number of entries in each codebook.
- $H(\tilde{p}_g)$: the entropy of the probability distribution $\tilde{p}_g$ over codebook entries, defined for each codebook g as:

$$H(\tilde{p}_g) = -\sum_{v=1}^{V} \tilde{p}_{g,v} \log \tilde{p}_{g,v}$$

- $\tilde{p}_{g,v}$: the probability of selecting the v-th entry in codebook g, encouraging the model to distribute its usage across all entries.

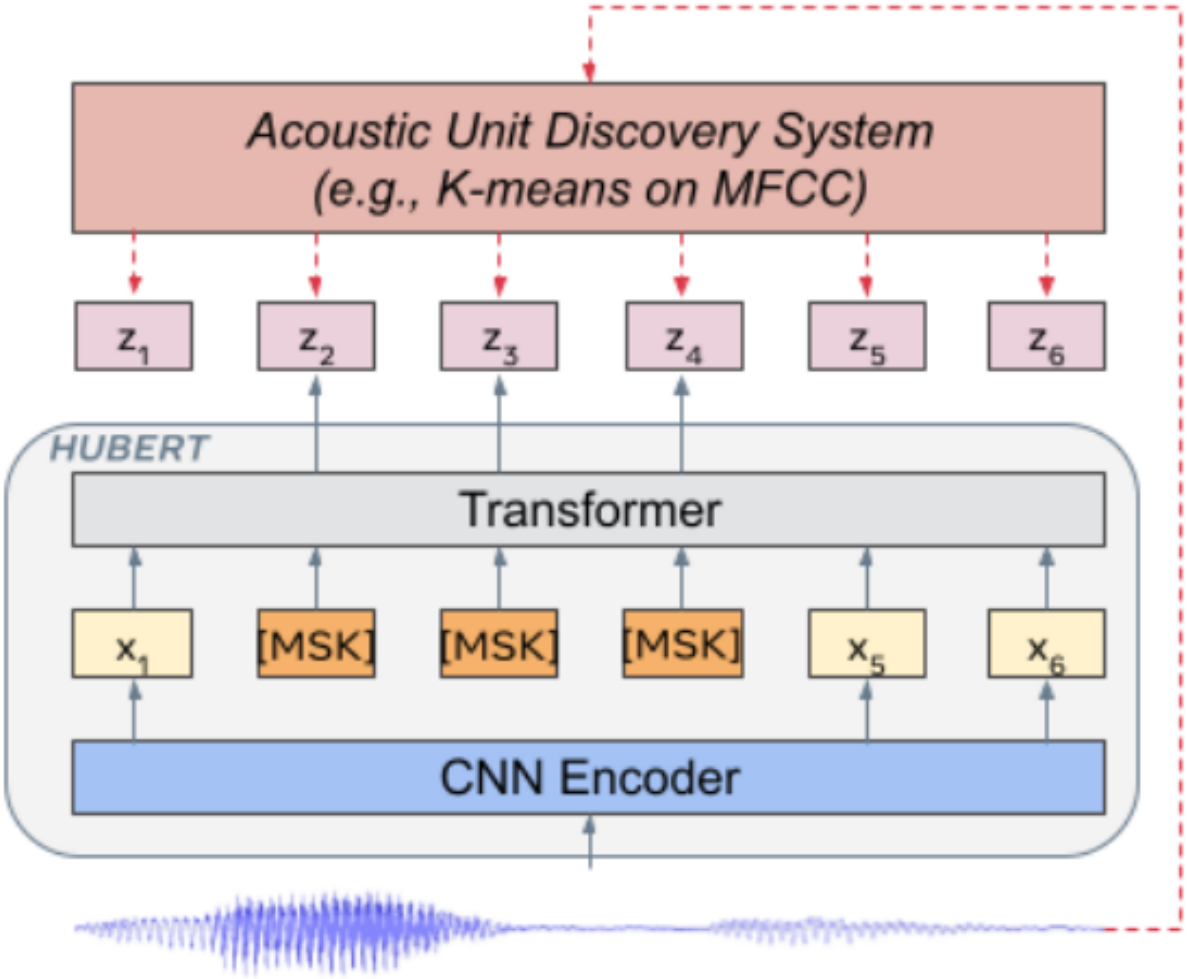
By incorporating both L_m and L_d, wav2vec 2.0 learns rich and diverse speech representations that are highly informative for downstream tasks, such as speech recognition. The contrastive loss enables it to correctly predict masked representations, while the diversity loss ensures that it utilizes a broad range of quantized codebook entries.

Training Dataset
The models are trained using the Librispeech corpus [60] without transcriptions, with 960 hours of audio, or the LibriVox dataset, which amounts to 53.2K hours of audio after pre-processing.

3.6.2 HuBERT

Hidden Unit BERT (HuBERT) is a self-supervised model that allows the powerful natural language model BERT, to be applied to speech. This is done by discretizing the audio input into hidden unit, allowing thus the model to be applied. This is demonstrated in figure 3.7.

Figure 3.7: An illustration of HuBERT model [9].



3.6.2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a bi-directional self-supervised NLP model based on the encoder block of the transformer architecture. It learns contextualized word embeddings by training on large corpora of text from the internet, while employing a bidirectional approach. This allows it to capture richer contextual information, benefiting from looking at both the right and the left context of a word simultaneously.
BERT is trained in a self-supervised manner, using unlabeled data, by following two mechanisms: masked language modeling, i.e. masking a percentage of the input, and training the model to predict them, and next sentence prediction, i.e. feeding the model two sentences, and training it to learn if they are subsequent in the original document.

3.6.2.2 Model Architecture

HuBERT architecture follows the wav2vec2.0 architecture for the feature encoder followed by the transformer context network, but differs in how the targets are derived. The HuBERT model is composed of the following parts:

1. **Feature encoder** The raw audio signal is fed into the encoder, which outputs latent speech representations for T time-steps. The feature encoder consists of seven blocks of temporal convolutions, each having 512 channels, with strides (5,2,2,2,2,2,2), and kernel widths (10,3,3,3,3,2,2), similar to wav2vec2.0's encoder.

2. **BERT encoder:** Which takes the masked feature vectors from the convolution encoder, after passing through a projection layer to match the model's configuration, to produce hidden unit representations. There exists three configurations:

   a) Base model: composed of 12 transformer blocks, with a model dimension of 768, an inner dimension of 3072, and 8 attention heads.

   b) Large model: composed of 24 transformer blocks, with a model dimension of 1024, an inner dimension of 4096, and 16 attention heads.

   c) X-large model: composed of 48 transformer blocks, with a model dimension of 1280, an inner dimension of 5120, and 8 attention heads.

3. **Projection layer:** Used to convert the output of BERT encoder to match the embedding dimension of the clustering step. The dimension are 256, 768, and 1024 for base, large, and x-large respectively.

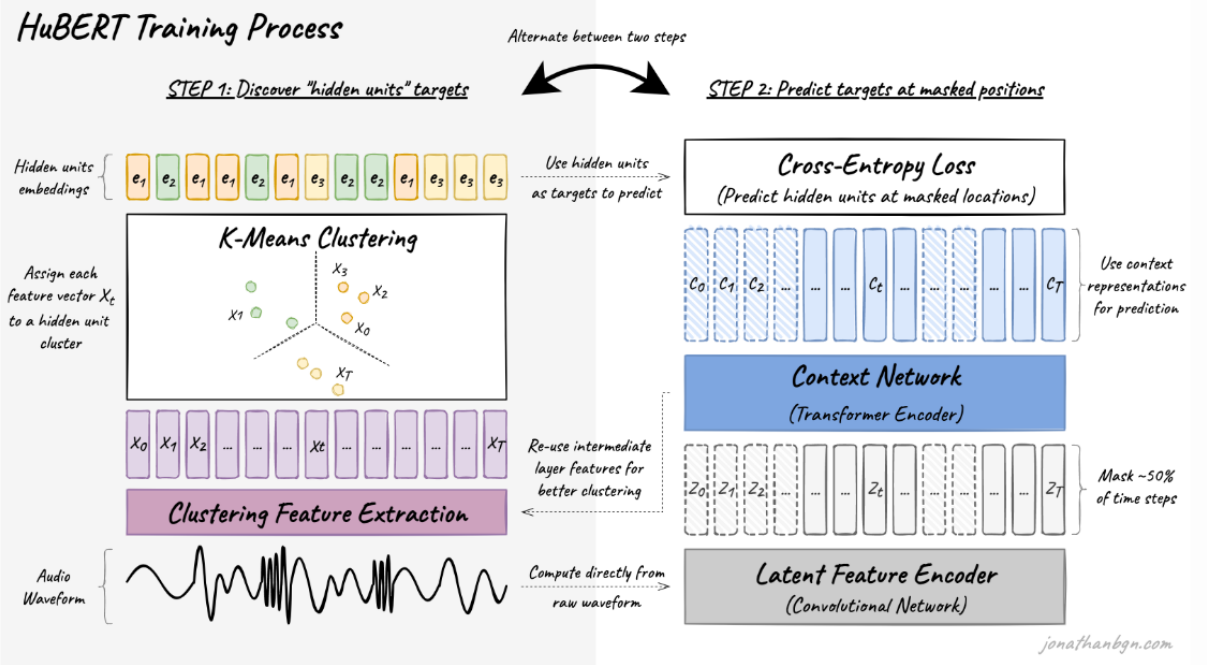
4. **Code embedding layer:** Used to convert the output of the clustering step -used to generate hidden units from the audio inputs- to its embedding vector (hidden units embeddings). The dimension are 256, 768, and 1024 for base, large, and x-large respectively.

5. **Acoustic unit discovery system:** Used to provide frame-level targets, or hidden units. It is based on ensemble k-means models, with different codebook sizes, to create targets of different granularity. The cluster assignments are refined during training, it starts by using MFCCs features from the raw audio inputs during the first iteration, and then uses the learned latent representations instead during the second iteration.

3.6.2.3 Model Pre-training

The objectibve of the pre-training is to learn the hidden unit representations. It is done for two iterations for the base model, and for one iteration for the large and X-large models. Since the training of Large and X-Large takes input from the base model, it is practically seen as 3 iterations. Each iteration consists of two steps, as demonstrated in figure 3.8: generating the hidden units by clustering, and masked language modeling. More details on each Step are provided in the following sections.

Figure 3.8: An illustration of the training process of HuBERT [10].



Step 1: Discovering hidden units targets through clustering

The first step of each iteration is to extract hidden units, to be used as (pseudo-)targets, from the raw audio inputs. This is done by first extracting features from each audio frame, and second assigning each segment of the audio into one of K clusters, which is done through K-means clustering. All audio frames will then be labeled according to the cluster to which they were assigned, and these become the hidden units. These units are then converted into embedding vectors, using the code embedding layer, to be used in the second step.

For the first iteration, the extracted features are 39-dimensional MFCCs -13 coefficients, in addition to the first and second derivatives-, with 100 clusters for the K-means clustering. As for the second and third iterations, the extracted features are the latent features extracted from the 6-th and the 9-th layers respectively, of the HuBERT model pre-trained in the previous iteration, with 500 clusters fro the K-means clustering.

This step is demonstrated in figure 3.9

Step 2: Masked language modeling

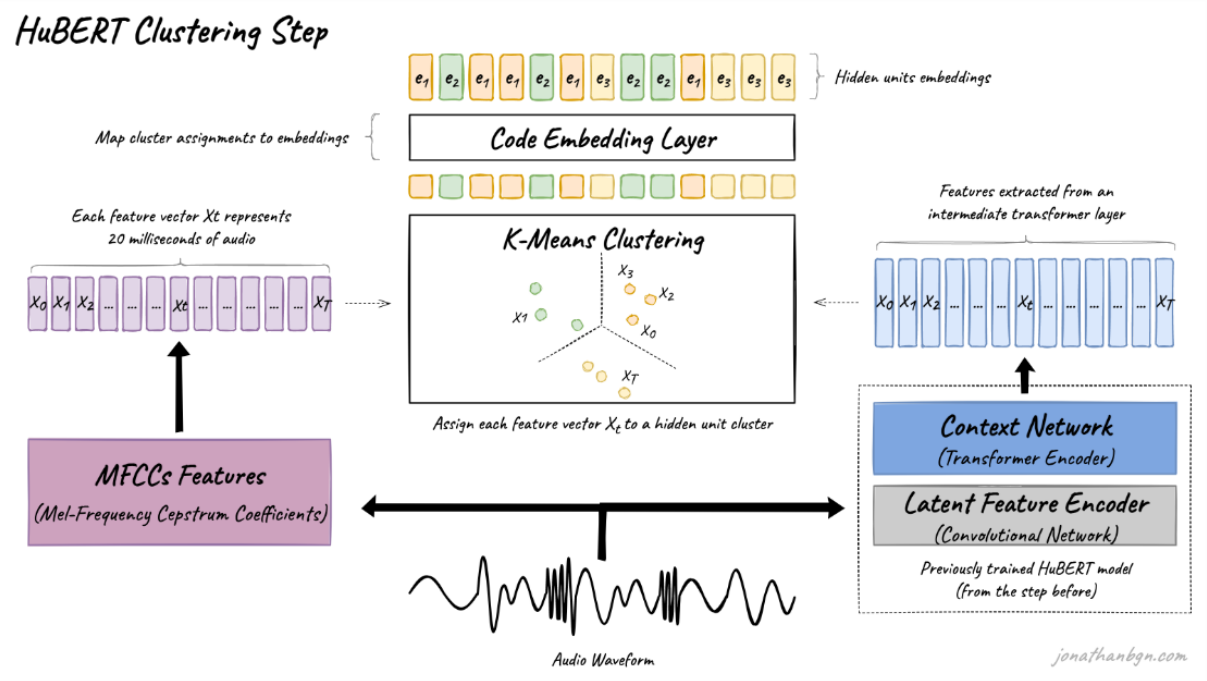
The second step, which is the same as for the original BERT, i.e. training with the masked modeling objective, is demonstrated in figure 3.10.

The masking is done following the same masking process demonstrated for wav2vec2.0. As for the loss calculation, it is computed as a weighted loss of the cross-entropy loss, computed over masked (L_m) and unmasked (L_u) timesteps in the audio sequence. The final loss is calculated as:

$$L = \alpha L_m + (1 - \alpha) L_u \tag{3.12}$$

where:

Figure 3.9: An illustration of the first step (clustering) of HuBERT pre-training process [10].



- L_m: the cross-entropy loss over masked timesteps (i.e., frames that have been hidden during training to encourage prediction).
- L_u: the cross-entropy loss over unmasked timesteps (i.e., frames that are visible to the model).
- α: a weighting parameter that balances the contribution of masked and unmasked loss terms.

Each loss term L_m and L_u can be further defined as follows:

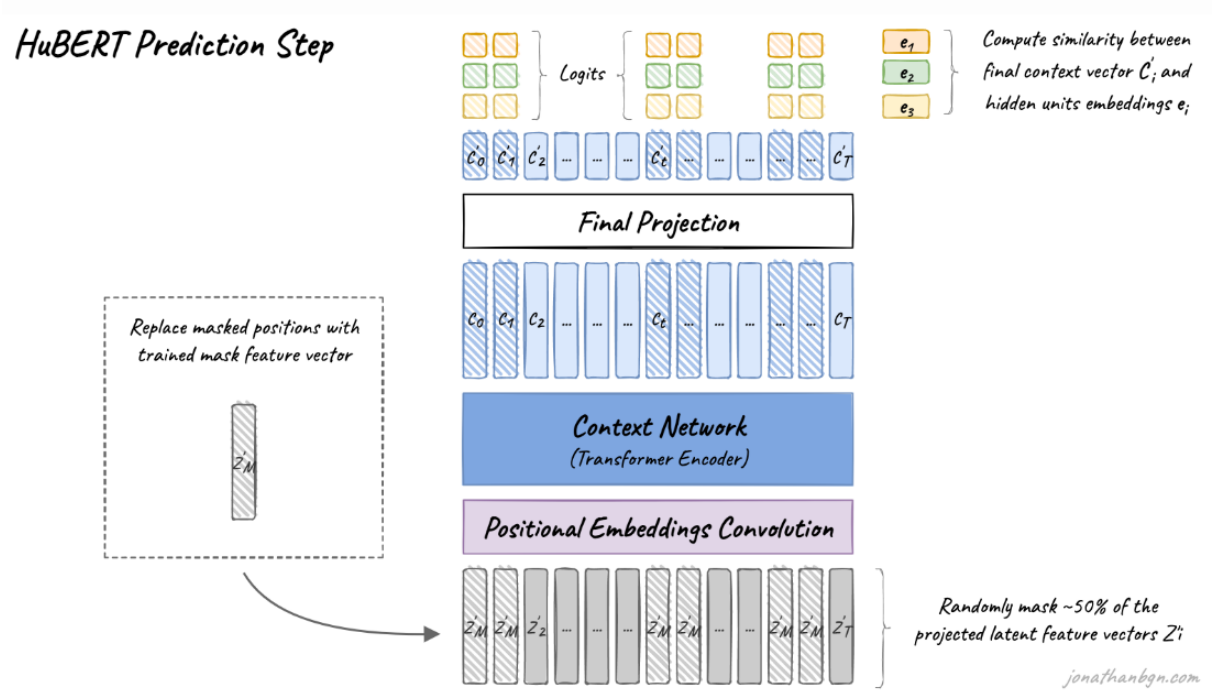
1. Masked loss (L_m): This term is the cross-entropy loss over the masked timesteps, calculated as:
$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f(z_t \mid \tilde{X}, t) \tag{3.13}$$
where:
   - M: the set of masked timesteps.
   - z_t: the true hidden unit label for timestep t.
   - $p_f(z_t \mid \tilde{X}, t)$: the probability of predicting z_t correctly, given the masked input $\tilde{X}$ at time t.
2. Unmasked loss (L_u): This term is the cross-entropy loss over unmasked timesteps, defined as:
$$L_u(f; X, M, Z) = \sum_{t \notin M} \log p_f(z_t \mid \tilde{X}, t) \tag{3.14}$$

where:

- $t \notin M$: unmasked timesteps that are used to provide additional context

The total loss L thus combines predictions from both masked and unmasked segments, allowing the model to learn effectively from both hidden and visible parts of the audio input.

Figure 3.10: An illustration of the second step of HuBERT pre-training process [10].



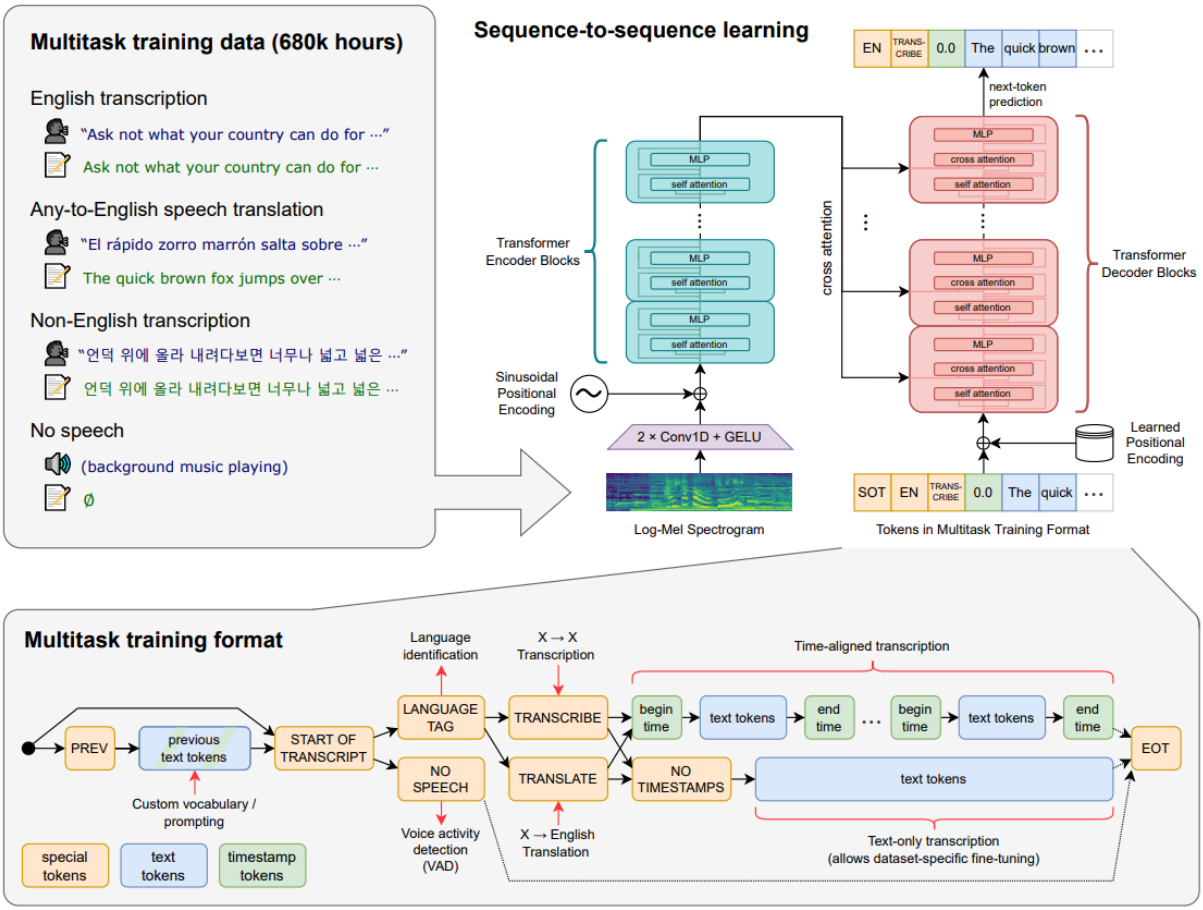
Training Dataset
The base model is trained on the 960 hours of LibriSpeech audio for both iterations. The first iteration is trained for 250k steps, while the second step is trained for 400k steps using labels generated by clustering the 6th transformer layer output of the first iteration model. As for HuBERT large and X-large, it is trained for one iteration, on 60k hours of Libri-light audio, for 400k steps, using labels generated from the 9-th trandformer layer of the second iteration of HuBERT base.

3.6.3 Whisper

Whisper [11] is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. In their work, Radford et al. explain that while unsupervised pre-training techniques, such as those used in wav2vec2.0 and HuBERT, were key elements for progressing speech recognition technologies, their impact has been limited to the encoder component. The pre-trained audio encoders are effective at learning high-quality speech representations, but an equally performant pre-trained decoder remains lacking. As a result, fine-tuning is required to

adapt these pre-trained encoders to various task, which also necessitates elaborating a dataset-specific fine-tuning protocol, limiting thus the robustness and generalization of these models. To address these limitations, Radford at al. proposed Whisper, an encoder-decoder model pre-trained using scaled weakly supervised, multilingual and multitask learning. They demonstrate that by adopting this model and learning strategy, their model is able to effectively generalize to existing datasets in a zero-shot manner, without the need for any dataset-specific fine-tuning.
An overview of the architecture, and the training approach used with Whisper is provided in figure 9.3.

Figure 3.11: An overview of the whisper's approach [11].



3.6.3.1 Model Architecture

Whisper model follows an encoder-decoder Transformer architecture, as shown in figure 9.3. The encoder learns latent speech representations, while the decoder generates text from the latent representations. The audio recording is split into 30-second chunks, and converted into a log-Mel spectrogram. It then passes through two convolutional layers, with a GELU activation function, and sinusoidal position embeddings are added to the output The resulting output is then processed by a series of encoder blocks. The decoder uses learned position embeddings and tied input-output token representations, meaning

the same embeddings are used for input tokens during decoding and the output vocabulary. The encoder and decoder share the same width and number of transformer blocks. Several variations of the Whisper model (Tiny, Base, Small, Medium, and Large), differing in the number of transformer layers and attention heads, have been proposed by the authors. Details of their architecture are provided in table 3.1.

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 3.1: Architecture details of Whisper Models [11].

3.6.3.2 Model Pre-training

Training of Whisper model follows a multitask training, i.e. the model is trained to perform multiple task. These tasks are: transcription, translation, voice activity detection, alignment, and language identification.
Since there are many possible task for the model to perform, a task specification is necessary. This is done using special tokens, that are fed as part of the input to the model's decoder. Consequently, the input to the model's decoder becomes a sequence of tokens, as demonstrated in figure 9.3. In addition, the previous text is also included in the input to the decoder, for it to learn to use long-range text context to resolve ambiguous audio. The decoder's input sequence of tokens includes the following:

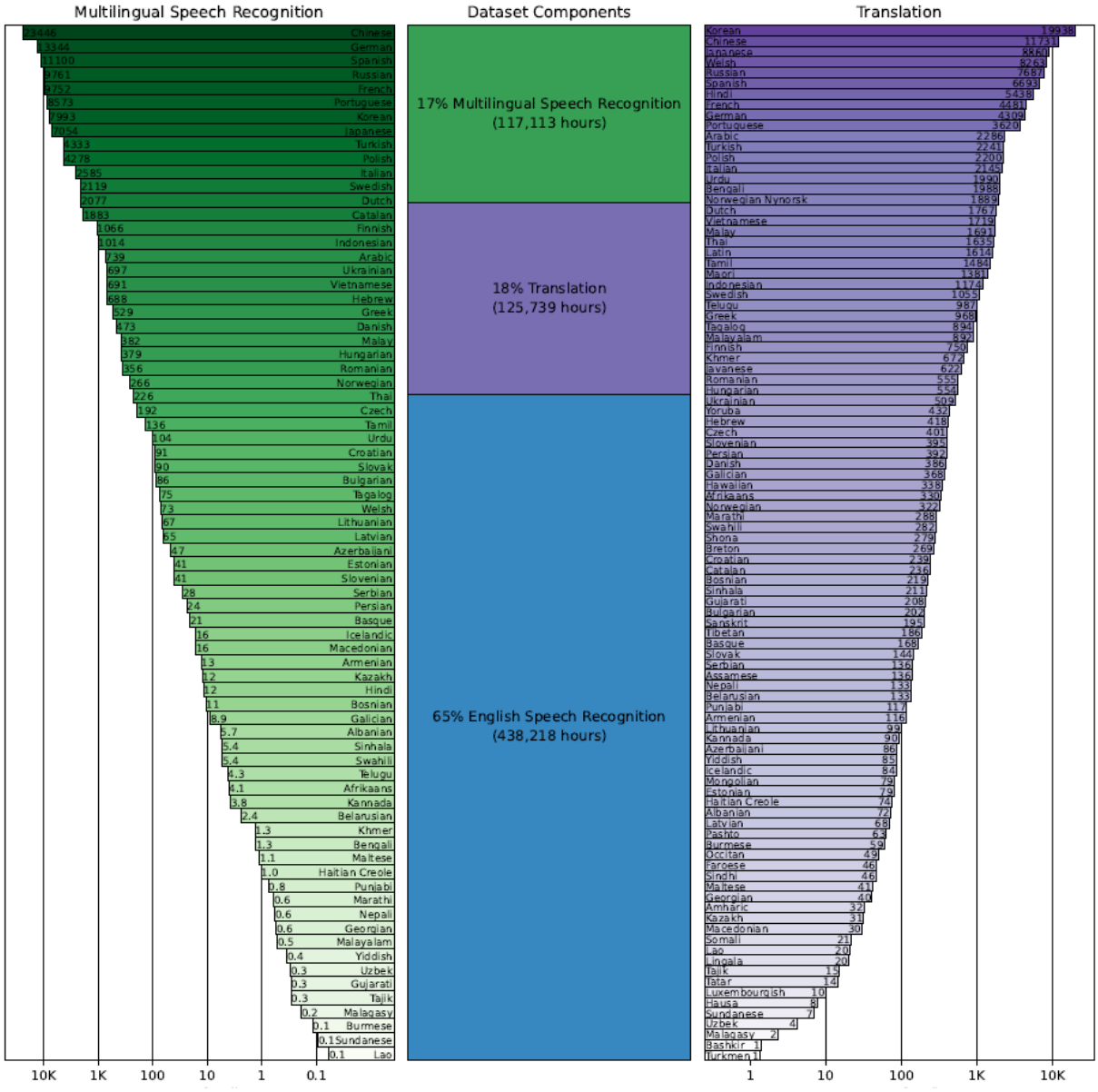
- $<|startoftranscript|>$ token: indicates the beginning of prediction.
- $<|language|>$ token: predicts the spoken language, $<|nospeech|>$ in case no speech is in the audio segment.
- $<|task|>$ token: speifies whether to perform transcription ($<|transcribe|>$) or translation ($<|translate|>$).
- $<|notimestamps|>$ token: to specify whether or not to predict timestamps. When added, timestamps are not predicted.
- $<|texttoken|>$: the text content from the previous sentence.
- $<|endoftranscript|>$ token: indicates the end of prediction.

Using this sequence of tokens and the latent features from the encoder, the decoder predicts a corresponding sequence of tokens for the current audio segment, including text, task-specific outputs, and timestamps when applicable.

Training Dataset

Whisper is trained using a dataset composed of audio paired with scripts collected from the internet. The dataset is highly diverse, covering a broad range of audio from various environments, recording setups, speakers, and languages. It consists of 68,000 hours of data in 99 different languages. Statistics on the training dataset are provided in figure 9.5.

Figure 3.12: Whisper's training dataset statistics [11].



3.6.4 Audio Spectrogram Transformer

Audio spectrogram transformer (AST) [12] is the first convolution-free, purely attention-based model, for audio classification. In almost all end-to-end models deployed for audio

classification, a CNN used to learn representations. Even with the rise of use of attention models for sequential data, it was used in conjunction with a CNN, where a self-attention mechanism is added on top of the CNN. CNN-attention models, as the ones earlier introduced (wav2vec2, HuBERT, and Whisper), achieved SOTA in many audio classification tasks. Nevertheless, purely attention-based models in vision domains (Vision Transformer ViT) are able to achieve competitive results, which is what motivated Gong et al. [12] to investigate the performance of purely attention-based models for audio related tasks. AST is inspired from ViT, and its pre-training benefited of knowledge transfer from ViT.

3.6.4.1 Model Architecture

The AST architecture is demonstrated in figure 9.6. It is composed of the following parts:

- **Linear projection/patch embedding layer:** The t seconds long audio input is first converted into 128*100t spectrogram, by computing a sequence of 128-dimensional log Mel filterbank features, with a 25ms Hamming window every 10ms. The spectrogram is then split into N 16*16 patches, with an overlap of 6 in both time and frequency dimension. The projection layer is used to flatten each 16*16 patch to a 1D patch embedding of size 768.

- **Positional embedding layer:** A trainable layer, also of size 768, is used to add positional embeddings to the output of the projection layer. This allows the model to capture the spatial structure of the 2D audio spectrogram.

- **Transformer encoder:** A [CLS] token is appended at the beginning of the sequence, before being inputted to a classic transformer encoder. It has a dimension of 768, 12 layers, and 12 heads.

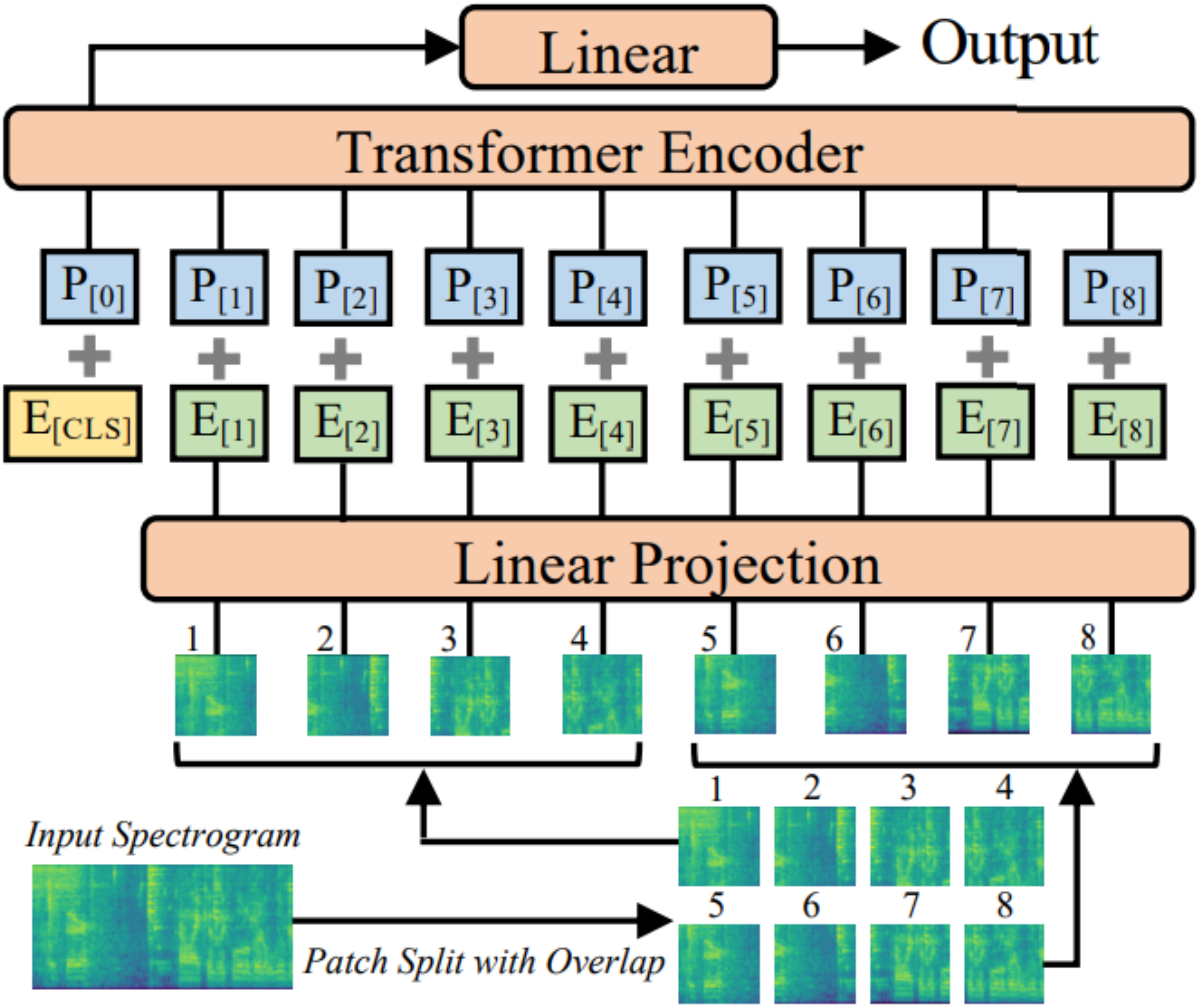
- **Classification layer:** A linear layer with sigmoid activation, to map the output of the encoder to labels for classification.

3.6.4.2 Model Pre-training

In [61], Dosovitskiy et al. show that the transformer is able to outperform CNNs when training data exceed 14 million for image classification tasks. However, such amount of audio data is not available, which motivated Gong et al. in [12] to apply cross-modality transfer learning to AST, i.e. transfer learning from vision tasks to audio tasks. They did that by adapting an off-the-shelf pretrained Vision Transformer (ViT)-a data-efficient image Transformer (DeiT) [62]-to AST.
ViT and AST have similar architectures, but they are not identical, this is why some adjustments were necessary before adaptation. In particular:

- **Input differences:** ViT model takes 3-channel image as input, while AST models takes a single-channel spectrogram. To mitigate this difference, they take the average of the wights of the ViT patch embedding layer across the three input channels, and use them for the AST patch embedding channel.

Figure 3.13: The audio spectrogram transformer (AST) architecture [12].



- **Input shape differences:** While ViT has fixed input shape, AST has variable length audio spectrograms. Even though the transformer is able to deal with variable length input, the positional embedding needs to be processed. The authors propose a cut and bi-linear interpolate method for adaptation.

- **Classification layer:** The ViT's classification layer is dropped, and a new one is used for the AST model.

By taking into account these differences, the authors were able to use the pre-trained weights of DeiT, which is trained with CNN knowledge distillation, 384 × 384 images, has 87M parameters, and achieves 85.2% top-1 accuracy on ImageNet [63] 2012. In addition to this adaptation, the AST was also pre-trained on AudioSet [64], a collection of over 2 million 10-second audio clips excised from Youtube, and labeled with the sounds that the clip contains from a set of 527 labels.

3.7 Conclusion

In this chapter, we introduced key technical concepts for digitizing audio signals, detailing the processes involved in converting analog waveforms into discrete data. We also explored the fundamentals of feature extraction using signal processing techniques, highlighting how these methods uncover meaningful characteristics from raw audio. Furthermore, we presented the primary machine learning concepts, learning paradigms, and models employed throughout this thesis, underscoring their relevance to subsequent experimentation. A solid grasp of these foundational ideas will enhance comprehension of the challenges, methodologies, and findings discussed in the following chapters, ensuring a smoother progression through the more advanced topics that lie ahead.

Chapter 4

A Bibliographic Review of Dysarthria Detection and Classification

4.1 Introduction

Clinical evaluation and automated identification of dysarthria, a motor speech problem brought on by neurological abnormalities, are extremely difficult. A thorough bibliographic assessment of dysarthria detection and categorization is given in this chapter, which also explores the development of techniques and highlights both conventional and contemporary approaches.
The rest of the chapter is organized as follows: we start by describing the classification strategies used in the literature. We continue with addressing the challenges inherent in data partitioning, specifically contrasting speaker-dependent with speaker-independent approaches, as these factors significantly influence model generalizability and performance. The chapter then synthesizes recent bibliographic findings on dysarthria classification methods, providing insights into current trends and innovations in the field. In a critical analysis, limitations of both traditional and end-to-end approaches are examined, shedding light on issues such as the reliance on manually engineered features, data scarcity, and the interpretability challenges associated with deep learning models.

4.2 Dysarthria Classification Approaches

When examining prior research on automating dysarthria classification, two primary approaches emerge:

1. **Traditional approach:** This approach typically involves two main stages. First, relevant features for dysarthria classification are extracted from the speech signal or its frequency/spectral representation. These extracted features are then input into a classifier, which is trained to categorize the audio input according to the severity level of dysarthria.

2. **End-to-end approach:** This approach utilizes deep learning models for both feature extraction and classification of speech. The raw speech signal, or its (mel-)

spectrogram, is directly provided to the deep learning model, which learns to derive meaningful features from the input to classify speech based on dysarthria severity.

4.2.1 Limitations of the Traditional Approach

Several limitations of the traditional approach have been documented in [65, 66]:

Reliance on Manually Engineered Features
Traditional approaches rely on manually engineered features, such as spectral, prosodic, or phonetic characteristics, which are designed based on prior knowledge about speech and linguistic patterns. While these features can be effective, they may not capture all the nuances of dysarthric speech, especially in complex or subtle cases. Impairment-related cues may be missed because the hand-crafted features do not account for all possible variations in dysarthria symptoms across different speakers, dialects, and speaking environments.

Feature Selection Challenges
After feature extraction, a feature selection process is often used to identify the most relevant features for dysarthria classification. Although feature selection can improve model efficiency by reducing dimensionality, it also runs the risk of discarding features that may be significant for capturing dysarthric patterns. This step can also be labor-intensive and computationally expensive, requiring extensive experimentation and domain expertise to identify which features best represent dysarthria severity across different cases [67, 68].

Separation of Classifier Training and Feature Extraction
In the traditional approach, classifier training and feature extraction are separate processes. As a result, the classifier is trained based on predefined features, and there is limited opportunity to refine or adapt the features during training. This separation may lead to suboptimal results, as the classifier cannot influence feature extraction to better capture dysarthric speech characteristics. Ideally, features would adapt dynamically based on classifier feedback to maximize classification performance, but traditional methods lack this flexibility.

Language and Dialect Constraints
Hand-crafted features are often developed for a specific language or dialect, requiring extensive adaptation and validation to ensure they work across different languages or populations. Each language may have unique phonetic and prosodic characteristics that affect dysarthric speech differently, making it challenging to generalize the features effectively.

Controlled vs. Real-World Conditions
Additionally, many traditional methods are trained and tested on datasets recorded in controlled environments [69], with consistent audio quality and recording setups. As a result, the model's performance may degrade significantly in real-world or varied recording conditions, where background noise, varying microphone quality, or different recording setups could affect the input.

Labor-Intensive Development
Developing, testing, and refining hand-crafted features requires significant manual effort and domain expertise, making it both time-consuming and resource-intensive. For dysarthria, understanding how specific speech impairments affect acoustic features demands close collaboration with clinical experts and extensive testing on diverse datasets. This requirement for specialized knowledge and prolonged development time can slow down the research and application process for practical dysarthria classification tools.

Sensitivity to Data Variations
Since traditional approaches often depend on a limited set of features and specific datasets, they can be sensitive to variations in the speech data. Changes in speaker demographics, dialects, or even different dysarthria subtypes can reduce the effectiveness of pre-defined features, resulting in lower classification accuracy. This sensitivity makes it difficult to deploy traditional models across broader populations or adapt them to new datasets without significant re-engineering.
These limitations underscore the challenges inherent in the traditional approach to dysarthria classification and highlight the need for more adaptive, data-driven methods, such as end-to-end deep learning models, that can overcome some of these constraints.

4.2.2 Limitations of the end-to-end approach

Although end-to-end methods have achieved state-of-the-art results in various classification tasks, their application remains limited due to specific requirements. End-to-end models, particularly deep learning architectures, require extensive datasets to achieve high performance and generalization. Since these models learn directly from raw input data, such as speech signals or spectrograms, they need to capture a wide variety of speech patterns, linguistic nuances, and impairment-related features. However, collecting large datasets in the context of dysarthria classification is challenging. Medical speech data is often scarce due to privacy restrictions, difficulties in obtaining dysarthric speech samples across different severities and subtypes, and limited availability of affected individuals for study participation.

Labeling Challenges
Supervised end-to-end methods, a prominent type of deep learning, require labeled data where each sample (e.g., speech segment) is annotated with the corresponding dysarthria severity level or type. Labeling medical data, such as dysarthric speech samples, is complex and requires domain expertise from speech-language pathologists or clinicians. This process can be both time-intensive and costly, as it often involves nuanced assessments of speech impairments or multiple experts to ensure labeling accuracy [70]. Manual labeling also introduces variability and subjectivity, potentially impacting model performance if inconsistencies occur across labels.

Data Variability and Overfitting Risks
End-to-end models are highly sensitive to the characteristics of the training data. For medical applications like dysarthria classification, there is significant variability across

speakers (e.g., age, dialect, impairment severity, and co-existing conditions). If the training data does not represent this variability, the model may struggle to generalize to unseen cases, limiting its practical utility. For instance, a model trained on dysarthric speech from a specific demographic or dialect may fail to accurately classify samples from other groups, potentially introducing bias into performance. Moreover, with limited and homogeneous datasets, end-to-end models are at a high risk of overfitting. Overfitting occurs when the model performs well on training data but fails to generalize to new data [71], which is especially problematic in medical applications where misclassification (e.g., underestimating dysarthria severity) could impact clinical decision-making. To counteract overfitting, researchers often employ techniques like data augmentation [72] or regularization, but these approaches may not be sufficient to overcome limitations posed by small datasets.

Computational Intensity and Resource Demands
End-to-end deep learning models, especially those working with raw speech signals or large spectrograms, are also computationally intensive. Training and deploying these models require substantial hardware resources (e.g., high-performance GPUs) and long training times, which may not be feasible in all research or clinical settings [70].

Sensitivity to Data Quality and Interpretability Challenges
The complexity of end-to-end models makes them harder to interpret and troubleshoot, an important consideration in medical fields where understanding the model's decision, making process is essential for clinical validation [70]. Unlike traditional models, which often use engineered features to reduce noise and variability, end-to-end models process raw input directly, making them more sensitive to inconsistencies in data quality. In dysarthria classification, factors such as background noise, recording device quality, and room acoustics can affect model performance. Because end-to-end models lack preprocessing steps to minimize these effects, they may perform poorly on data collected outside controlled environments, such as in real-world clinical settings or patients' homes.

Interpretability and Trust Issues
End-to-end models are often described as 'black boxes' due to the challenge of interpreting their internal decision-making processes. In medical applications like dysarthria classification, understanding which features or speech patterns influence classification is critical for clinical validation and for gaining trust from practitioners. Without interpretable insights, clinicians may hesitate to rely on model predictions, especially when decisions impact patient care. This limitation has driven interest in explainable AI (XAI) methods [73], though these are still developing and may not yet provide adequate interpretability for complex models.

High Development and Maintenance Costs
Finally, building and optimizing end-to-end models demands considerable expertise in machine learning, deep learning architectures, and data engineering. For dysarthria classification, developing a robust model often requires close collaboration between clinicians, speech scientists, and data scientists, leading to high initial development costs. Additionally, ongoing maintenance, retraining, and tuning are often needed to adapt the model to new data, conditions, or languages, further increasing the operational burden.

These limitations underscore the challenges that end-to-end methods face in medical applications like dysarthria classification. While they offer promising performance and automation potential, their effectiveness depends heavily on overcoming data, interpretability, and resource-related challenges that are often inherent in healthcare applications.

4.3 Data partition

Data partition for training and evaluation models for dysarthria severity classification can be either speaker-dependent or speaker-independent. More details about both splits are provided in the following subsections.

4.3.1 Speaker-dependent

In the speaker-dependent data split, the partition is done following the general data partition method used in machine learning, i.e. from all the available audio data, a certain percentage is used for training, while the remaining audios are used for testing. This split makes it easier to train and test the model using data from the same people, though the testing phase uses different audio content. Despite the fact that this approach facilitates model training since voice patterns are consistent, its use in a clinical setting is limited because the model is not validated against unseen patients, which is a basic prerequisite for an automated diagnosis tool [74].

4.3.2 Speaker-independent

Unlike the speaker-dependent data split, in the speaker-independent data split setup, data from distinct speaker groups are used to train and evaluate the model. This guarantees that the model can generalize to voices that are foreign to it [74].

4.4 Features For Dysarthria Severity Classification

In dysarthria classification, diverse feature types, spanning audio, video, text, and image data, play an essential role in capturing the distinctive speech impairments associated with this motor speech disorder. Dysarthria, caused by neurological impairment, affects speech production in various ways, such as articulation, prosody, and facial movements, leading to unique challenges in classification. Audio features, which capture acoustic and prosodic aspects, are widely used to detect and quantify disruptions in speech signals. Video features, such as facial movements and lip patterns, offer complementary visual information about articulation and muscle control [54]. Text features, often derived from speech-to-text processes, provide insights into the linguistic and phonetic characteristics of dysarthric speech, highlighting issues like phoneme substitution or word repetition [75]. Lastly, image features, often extracted from (mel-)spectrograms and visual representations of sound, provide detailed frequency and temporal information for advanced analysis [76, 77]. Examining these different feature types and their contributions to dysarthria classification can deepen our understanding of the disorder and improve the development of robust classification models. Audio derived features are the most studied and used

form of features when developing dysarthria detection / classification models, this is why in the following we provide more details on the audio-based features.

4.4.1 Audio-based features

The number of features is too large to be listed, in the following we categorize the different types of audio-based features, and we give some examples for each category.

1. **Prosodic features:** They reflect the rhythmic, stress and intonational aspects of speech. These features play an important role in conveying emotions, emphasis, and the structure of spoken communication. These features include: the fundamental frequency ($F0$), the loudness, the root mean square energy (RMS Energy), zero-crossing rate (ZCR), etc. .

2. **Spectral features:** They analyze the frequency distribution of speech energy, representing its structure in different frequency bands. These features include: flux, entropy, band energy 250-650 Hz and 1-4 kHz, spectral centroid, spectral bandwith, specrtal roll-off, formant frequencies, etc. .

3. **Cepstral features:** Cepstral features capture the shape and characteristics of the vocal tract by analyzing the frequency spectrum. These features effectively separate source (e.g., pitch) from filter (vocal tract shape) information, which is useful in tasks like phoneme classification and speaker recognition. These features include: MFCCs, linear predictive cepstral coefficients (LPCCs), etc. .

4. **Voice quality features:** Voice quality features focus on specific vocal attributes that reflect the physiological state of the speaker and qualities of the vocal fold vibration. These features capture aspects like breathiness, hoarseness, and the presence of noise in the signal. These features include: jitter, shimmer, (logarithmic) Harmonics-to-Noise ratio ((log)HNR), etc. .

4.4.2 The Geneva Minimalistic Acoustic Parameter Set

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [78], is a minimalistic feature set proposed by a group of researchers as an attempt to provide a common baseline for evaluation of research in various areas of automatic voice analysis, and eliminate differences caused by varying parameter sets or even different implementations of the same parameters. The selected features were carefully selected, based on the following criteria:

1. The potential of the parameter to index physiological changes in voice production during affective processes.

2. The frequency and success with which the parameter has been used in the literature.

3. The theoretical significance of the parameter.

Eyben et al. proposed two versions of the set: the first (GeMAPS) includes prosodic, excitation, vocal tract, and spectral descriptors, while the second (extended GeMAPS - eGeMAPS) is an extension of the first, adding a small set of cepstral descriptors.
The **minimalistic parameter** set contained the following 18 Low-level descriptors (LLD):

- Frequency related parameters: pitch, jitter, formant frequency 1,2, and 3, formant 1 bandwidth.
- Energy / Amplitude related parameters: shimmer, loudness, HNR.
- Spectral (balance) parameters: alpha ratio, hammarberg index, spectral slope 0-500 Hz and 500-1500 Hz, fromant 1, 2 and 3 relaive energy, harmonic difference H1-H2, harmonic difference H1-A3.

Two functionals (arithmetic mean, and coefficient of variation) are applied to each one of the aforementioned LLDs, over the voiced regions only. In addition, 8 other functionals are applied to loudness and pitch parameters, also over the voiced regions only. This leads to a total of 52 parameters. The arithmetic mean of the alpha ratio, the Hammarberg index, and the spectral slopes from 0–500 Hz and 500–1500 Hz over all unvoiced segments are also added to the set, along with 6 temporal features (rate of loudness peaks, mean length and the standard deviation of both continuously voiced regions and unvoiced regions, and the number of continuous voiced regions per second). All this lead to a total of 62 parameters in the Geneva minimalistic standard parameter set.
As for the **extended GeMAPS**, it extends the first set with the following 7 LLDs:

- Spectral parameters: MFCC 1-4, spectral flux.
- Frequency related parameters: formant 2-3 bandwidth.

To the aforementioned LLDs are applied the arithmetic mean and the coefficient of variation, over voiced and unvoiced regions together, except for the formant bandwidths to which the functionals are applied only in voiced regions. To these are also added the arithmetic mean of spectral flux in unvoiced regions, the arithmetic mean and coefficient of variation of the spectral flux and MFCC 1–4 in voiced regions only, and the equivalent sound level. This results in a 26 extra parameters, leading to a total of 88 parameters in the extended Geneva minimalistic acoustic parameter set.
Throughout this thesis, the eGeMPAS was chosen as a baseline feature set, providing a standardized foundation for evaluating models performance.

4.5 Dysarthria Classification Methods

In this section, we will provide a discussion on the most relevant work done for dysarthria severity classification using UASpeech and TORGO datasets.
In [79], Joshy et al. performed a comparative study on the classification of dysarthria severity levels using different deep learning (DL) techniques and acoustic features. Specifically, three experiments were conducted encapsulating three features sets, and different DL techniques, a deep neural network (DNN), convolutional neural network, gated recurrent units and long short-term memory network were used. In the first experiments, MFCCs and constant-Q cepstral coefficients were used as input features. In the second one, speech-disorder specific features computed from prosody, articulation, phonation and glottal functioning are evaluated on DNN models. Finally, low-dimensional feature based i-vectors, were used as inputs to the DNN models. Evaluation was done using both UASpeech and TORGO, in speaker-dependent data partition for both, in addition to a

speaker-independent scenario data partition using UASpeech. The best results were reported using the MFCC-based i-vectors with DNN classifier, amounting to 93.97% and 49.22% for the speaker-dependent and the speaker-independent scenarios respectively. This study underscores the performance improvement achieved by the use of Deep Learning (DL) models compared to traditional ML models, yet it highlights the importance of choosing the appropriate features in combination with the DL model. Examining both the speaker-dependent and the speaker-independent cases showed the striking difference in results for each case, and further emphasizes the importance of evaluation following the speaker-independent data split, in order to design robust, generalizable models. However, further studies should be conducted to validate the findings of this work, specifically, cross-dataset evaluation should be conducted to study the robustness of the models.

In [80] Chandrashekar et al. propose the use of different spectral and cepstral features with temporal and Fisher vector encoding, coupled with two classifiers, probabilistic linear discriminant analysis (PLDA) and ANNs. Experiments were conducted using both UASpeech and TORGO databases, with both speaker-dependent and speaker-independent data partitions for both UASpeech and TORGO. Results revealed an average accuracy ranging from 72.2% to 99.9%, using the different proposed features, and 45.6% on UASpeech using the best performing features, for the SD and SI scenarios, respectively. For TORGO, training was separated using words and sentences. For words, the average accuracy ranged from 55.5% to 80.4% using the different proposed featurs, for the SD scenario, vs an average accuracy of 57.3% for the SI scenario, using the best performing features. For sentences, the average accuracy ranged from 62.7% to 86.8% for the SD scenario, using the different proposed features, vs an average accuracy of 62.4% for the SI scenario. The study's findings demonstrated that in intelligibility classification tests, Fisher vector encoding consistently performed better than temporal encoding. It is more useful for evaluating dysarthric speech intelligibility because it records more intricate, combined spectral-temporal changes. However, compared to temporal encoding, it requires more calculation, which can make real-time use in clinical contexts difficult. Additionally, results showed the model struggled to correctly classify low and mid intelligible speakers, compared to other intelligibility levels, highlighting the need for improvement to better classify all severity levels.

In [81], Kachhi et al. studied the use of Teager Energy Cepstral Coefficients (TECC) for dysarthric speech classification. Three deep learning architectures were compared, a CNN, a Light-CNN (LCNN), and Residual Networks (ResNet). For feature comparison, STFT, MFCC, and Linear Frequency Cepstral Coefficients (LFCC) were used. Experiments were conducted using UASpeech, following a speaker-dependent data partition. Results showed that the best performance of 98.02% was achieved using ResNet, with an improvement of 1.69% compared to MFCC. Results showed that TECCs outperformed traditional features in classifying dsyarthric speech. However, this was validated using only UASpeech dataset, with a speaker-dependent data split. Further tests, such as using the speaker-independent data split and cross-dataset testing, should be conducted to validate these results.

In [82], Gurugubelli et al. investigated the use of analytic phase features for both dysarthria detection and severity classification. Their approach estimates instantaneous frequency components using a single frequency filtering technique. The extracted analytic phase features capture the fine temporal variations in the phase of the speech signal and

are compared against conventional spectral features. An i-vector with PLDA classifier was used. Experiments were conducted using UASpeech. Their work resulted in a classification accuracy of 95.61% for dysarthria detection and 64.47% for dysarthria classification. The authors specify the a leave-one-speaker-out validation is used, but it is not clear if that was repeated for the speakers. Further studies should be conducted to validate the generalizability of the model, among this dataset, as well as using cross-dataset testing. Additionally, results showed that the extracted features showed an overlap between the very low - low severity classes and medium - high severity classes, highlighting the difficulty of seperating between consecutive classes.

Vyas et al. [83] studied the use of specific prosodic features selected using genetic algorithm for dysarthria detection and severity classification. A SVM was used as classifier, and experiments were conducted UASpeech dataset following a speaker-dependent data partition approach. The proposed method achieved an accuracies of 98% and 87% for detection and classification respectively. This study highlighted the importance of prosodic features for dysarthria classification, however, their approach using a speaker-dependent data split limits the generalizability of the results. In addition, they limited their work to the isolated words recordings od the dataset, which might not adequately capture the complexity of continuous speech, where intelligibility may be impacted by prosody and context.

In [84], Narendra et al. studied the use of glottal parameters for dysarthria severity classification. Two sets of glottal parameters were studied: time- and frequency-domain parameters, as well as parameters based on principal component analysis. A two-stage framework was proposed: first, the input is classified into one of two possible classes (low–mid versus mid–high) using an SVM trained under a speaker-independent data partition. Second, a linear regression model is employed to estimate a more precise intelligibility rating using a speaker-dependent partitioning of the data. Experiments were conducted using UASpeech. Results showed that the proposed method outperformed baseline method, with an accuracy of 68.51% for stage one, and a root mean square error of 6.33 between the ground truth and the predicted class in stage two. This study demonstrated the glottal parameters have a strong correlation with the subjective intelligibility ratings, suggesting that these features are useful in assessing dysarthric speech. However, relying just on glottal source parameters, which primarily reflect the vocal fold vibration aspect, disregards the influence of other aspects of speech production, such as prosody and articulation, on the speech intelligibility.

In [85], Javanmardi et al. Study the use of pre-trained models for dysarthria detection and severity level classification. Three pre-trained models, wav2vec2-BASE, wav2vec2-LARGE, and HuBERT, were used as feature extractors, with the pre-trained weights frozen. The extracted features served as input to two classifiers: a Support Vector Machine (SVM) and a CNN. Training and testing were conducted on both UASpeech and TORGO datasets, following a speaker-independent data split based Leave-One-Speaker-Out (LOSO) cross-validation. Results showed that, when trained on UASpeech, wav2vec2-LARGE combined with the CNN classifier outperformed the other models on the detection task with an accuracy of 95.79%, while HuBERT outperformed wav2vec2 in the classification task, achieving an accuracy of 48.01%. When trained on TORGO dataset, the best results were achieved using HuBERT for both detection and classification tasks, in particular, it scored an accuracy of 76.12% and 49.83% for detection and classification

respectively. While this work demonstrated the potential of use of pre-trained models for dysarthria classification by evaluation on two datasets, one of its limitations is the fact that the training and testing was performed separately on each dataset instead of cross-dataset evaluation, which further tests the robustness of the models. Additionally, the study primarily used pre-trained models without focusing on fine-tuning for dysarthric speech. Fine-tuning the models should be studied to see how it affects the performance of the models.
In another study, Javanmardi et al. [86] further studied the effect of fine-tuning pre-trained wav2vec2 model for cross-database dysarthria detection. They used both wav2vec2-BASE and wav2vec2-XLSR, comparing their performance when frozen and when fine-tuned. For evaluation, three scenarios were considered: 1- In-database testing, where the same dataset was used for both training and testing; only fine-tuning the models was studied. 2- Cross-database language-dependent testing, where training and testing datasets were different but in the same language. 3- Cross-database language-independent, where training and testing datasets were in different languages. The datasets used where UASpeech and TORGO for the language-dependent testing, and an Italian dataset, EasyCall, for the language-independent testing. Results showed that fine-tuning the models did improve the generalizability of the models for both language-dependent and language-independent scenarios, compared to using the frozen models. This work underscores the importance of cross-database and cross-language evaluation in developing robust models for dysarthria detection. Nevertheless, this work should be extended to severity classification, to see whether similar conclusions could be observed in severity classification task, and what other complications might rise from the multi-class classification case.
In [87], Stumpf et al. developed a transformer framework incorporating a multi-task learning objective and contrastive learning for speaker-independent multi-class dysarthria severity classification. They used the wav2vec2-Base pre-trained model, and fine-tuned it to perform dysarthria severity classification, in addition to an auxiliary task, a contrastive task. The contrastive task is designed to separate the word embeddings of speakers from the same severity level. This task pushes the model to learn speaker-agnostic embeddings, making it more focused on severity-specific characteristics instead of overfitting the speakers and learning speaker-specific characteristics. The training was performed on the UASpeech dataset, following a speaker-independent data split. Results showed that using their method, they were able to achieve an accuracy of 70.48% outperforming the previous benchmark of AI-based dysarthria classification models. This study sets a new benchmark for AI-driven speech analysis in dysarthria and demonstrates the potential of transformer-based models in healthcare applications. However, the limited data in the "Low" and "Mid" severity categories restricted the model's ability to accurately distinguish between these severity levels. This highlights the need for larger and more diverse dysarthric datasets. Additionally, the work should be extended to other datasets and other languages, to study the robustness of the model across different recording settings, as well as different languages.

A wide range of methods has been proposed in these articles, from traditional handcrafted acoustic features to advanced machine learning models, including deep learning techniques and transformer-based architectures. Reviewing this bibliography on dysarthria severity classification reveals several key research challenges. In the following, we highlight

the ones addressed in this thesis:

- Data Scarcity: Over the past years, considerable efforts have been made to collect dysarthric speech data to advance research. However, the available databases remain limited in terms of both the number of participants and the variability of speakers with respect to severity, age, and gender.

- Speaker-dependent vs. Speaker-independent data partition: Studies that performed both the speaker-dependent and the speaker-independent data splits demonstrated significant differences in performance. Specifically, the speaker-independent split presents an increased challenge, underscoring the need for future research to incorporate this split to develop more robust and generalizable models. It is also crucial to standardize training and testing strategies to allow for fair comparisons between different solutions.

- Cross-database Testing: The speaker-independent data split has been proved to be important for building robust generalizable models. Cross-database testing allows further generalization of these models, by minimizing the impact of recording environment-specific factors, or even language-specific factors, making the models more adaptable across different conditions.

- Feature Extraction: A central theme across the studies is the selection of features that effectively capture the nuances of dysarthric speech. Traditional features, such as MFCC and eGeMAPS, have been the cornerstone of speech recognition for years and remain widely used in dysarthria detection and classification. These features focus primarily on spectral and prosodic characteristics such as pitch, energy, and formants. While effective in many speech recognition tasks, these features struggle to generalize well, especially in the speaker-independent scenario. This limitation has prompted researchers to explore more advanced features, such as TECC or features extracted by pre-trained models, like HuBERT or wav2vec2.

- Machine Learning Approaches: Another notable trend across studies is the shift from traditional ML models, like Support Vector Machine (SVM), to more sophisticated DL models, like Convolutional Neural Network (CNN) and wav2vec2. This shift reflects the growing recognition of deep learning's ability to model complex speech patterns and improve classification accuracy.

These are the main challenges addressed in this thesis. To tackle these issues, we collected a dysarthric speech dataset from French patients, which included a larger number of participants compared to existing datasets. We also ensured that our experiments prioritized the speaker-independent data split, as well as cross-database testing, to enhance the robustness and generalizability of the models. Additionally, we experimented with advanced feature extraction using pre-trained deep learning models. For classification, we explored employing transformers, AE, and pre-trained models as classifiers. These approaches allowed us to leverage powerful feature representations and improve classification performance for dysarthric speech severity.

4.6 Conclusion

In this chapter, we provided a comprehensive review of dysarthria detection and classification methods, setting the foundation for understanding the complexities of automated speech impairment assessment. We explored a range of classification approaches, detailing their methodologies and underlying assumptions, and examined the features critical to dysarthria severity classification. The discussion on data partitioning highlighted the differences between speaker-dependent and speaker-independent setups, emphasizing their impact on model performance and generalizability. Additionally, the bibliographic survey offered insight into current trends and the evolution of dysarthria classification methods. A critical analysis of both traditional and end-to-end approaches uncovered significant limitations for both approaches. These findings underscore the necessity for innovative, adaptive methodologies that can effectively handle the diverse characteristics inherent in dysarthric speech.

Part II

Experimental Work

Chapter 5

Overview of Datasets: Publicly Available and Newly Collected Data

5.1 Introduction

Machine learning is a data-driven domain, this implies that the data used for training the model plays an important role in its efficiency once trained and deployed. The quality, relevance and size of a dataset are central for the success of the ML model, in terms of efficiency, accuracy and complexity [88]. Since data is the backbone of any ML algorithm, it is important to choose adequate, representative, unbiased, and good quality data for the training process.
In this chapter, we provide a comprehensive overview of the datasets utilized in this thesis. Two public datasets, UASpeech and TORGO, were used for the experimental work. In addition, we introduce a newly collected dataset, VocaPark. This dataset was collected during this thesis, from French speakers with dysarthria secondary to Parkinson's disease, in collaboration with Fondation HOPALE. The datasets were selected based on their relevance with the objectives of this thesis, and their prevalence in SOTA research. By providing these information, this chapter provides a solid ground for understanding the experimental setup, and the findings presented in later chapters.

5.2 Data related challenges when considering data-driven models

When dealing with data derived from every day life, several challenges arise. These challenges are not just related to the technical part of collecting, managing, and processing the data, but also to the ethical, legal and strategic considerations. More details on these challenges are provided in the following [89–91]:

- Data quality: Inaccurate, noisy, duplicates, or incomplete data can lead to flawed models, and unreliable predictions.
- Data quantity: Data collection is one of the biggest barriers to implementing AI solutions. Some solutions for this are: data augmentation, transfer learning, synthetic

data generation, active learning, etc. .

- Imbalanced data: An imbalanced distribution of data could lead to biased models.
- Data privacy and security: collecting and exploiting data should follow regulations on data privacy, that protect user privacy. It should also implement necessary precautions to ensure secure storage and transmission of data.
- Data labeling and annotation: Labeling large datasets can be time-consuming, and expensive. It might also include subjective decisions, leading to inconsistencies in the labeling process, which ultimately would affect model's learning.
- Scalability: The infrastructure used to hold and manage the data should be scalable, ensuring the possibility to handle the increased computational demands.

5.3 Public datasets

During the entirety of this work, two main datasets -UASpeech and TORGO- were utilized for training and validation of the proposed models. In the following, we will dive in depth on the content of these datasets.

5.3.1 UASpeech

The Universal Access Speech (UASpeech) corpus [13], is a dataset collected by Kim et al., with the goal of developing automatic speech recognition (ASR) systems for dysarthric speech. Since its collection and distribution, it has become a vital resource for developing assistive technologies for people with motor disabilities [92].

5.3.1.1 Equipment and recording procedure

Two types of recordings, audio and video, were collected in this dataset.
For audio recordings, eight microphones, of 6 mm of diameter each, were arranged in an array, and spaced by 1.5 inches. Attached to this array were preamplifiers, with their output being sent to a multi-channel Firewire audio interface (MOTU 828mkII) through cables. The MOTU recorded eight audio channels, at a sampling rate of 48 kHz. One channel of the MOTU was reserved for recording DTMF tones, which were used to segment words and save each word as a separate wav file. The array was mounted at the top of the laptop computer screen, as shown in Figure 5.1.
As for video recordings, a Canon ZR500 camera was used to capture the visual features of speech.
Recordings were taken with the speakers seated in front of the laptop. Each speaker was asked to read an isolated word, displayed on the laptop's screen. Words were divided into three blocks, and speakers had a break between each block, and whenever a break was needed. Each block consisted of 255 words, divided as follows:

1. **155 common words between the three blocks:**
   a) 10 digits (zero to nine)

Figure 5.1: Picture of the equipment: MOTU, the microphone array mounted at the top of a laptop computer, and a video camera. [13].



b) 26 radio alphabet ('Alpha', 'Bravo', 'Charlie', etc.)

c) 19 computer commands (e.g. 'backspace', 'delete', 'enter')

d) 100 of the most common words in the Brown corpus of written English (such as 'it', 'is', 'you')

2. **100 distinct words for each block:** selected from children's novels digitized by Project Gutenberg, using a greedy algorithm that maximized token counts of infrequent bi-phones.

To summarize, each speaker recorded 765 isolated words, including 465 distinct words (three repetitions of the 155 words), and 300 distinct uncommon words. The audio from the seven channels of speech were stored in a separate wav file, for each word.
Only data from 15 dysarthric participants, and 13 controls were available for use.

5.3.1.2 Speakers distribution

The characteristics of the participants are summarized in table 5.1. 19 English speakers (4 female, and 15 male) with dysarthria, secondary to CP, participated in this dataset. Most of participants have spastic dysarthria, except for four (2 athetoid, and one mixed). In addition, data was collected from 13 controls, with no speech problems.

Speech intelligibility of each participant was calculated based on word transcription tasks by human listeners. More specifically, 200 distinct words were selected from the second block (10 digits, 25 radio alphabet letters, 19 computer commands, 73 words from the common words, and 73 words from the uncommon words. In addition, 25 words out of the 200 were arbitrarily chosen and repeated twice in the list. So, in total, each listener rated 225 speech files. For each speaker, five naive speakers were recruited, and each one of them listened to one word at a time, and was then asked to write down the word, with number from 0 (not sure at all) to 2 (completely sure) for each word, to indicate the degree of certainty of their choice. For each listener's transcription, the percentage of

Speaker	Age	Speech Intelligibility (%)	Dysarthria Diagnosis
M01	>18	very low (10%)	Spastic
M04	>18	very low (2%)	Spastic
M05	21	mid (58%)	Spastic
M06	18	low (39%)	Spastic
M07	58	low (28%)	Spastic
M08	28	high (93%)	Spastic
M09	18	high (86%)	Spastic
M10	21	high (93%)	Mixed
M11	48	mid (62%)	Athetoid
M12	19	very low (7.4%)	Mixed
M13	44	not obtained yet	Spastic
M14	40	High (90.4%)	Spastic
F02	30	low (29%)	Spastic
F03	51	very low (6%)	Spastic
F04	18	mid (62%)	Athetoid
F05	22	high (95%)	Spastic
M01	>18	very low (19%)	Spastic
M02	>18	high (92%)	Spastic
M03	>18	low (29%)	Spastic
F01	>18	low (19%)	Spastic

Table 5.1: Summary of speaker information: The letter M and F in speaker code specifies a participants' gender.

correct responses was calculated. The correct percentage was then averaged across five listeners to obtain each speaker's intelligibility. The speech intelligibility score was then used to assess dysarthria severity of each participant.

5.3.2 TORGO

The TORGO database [93] of dysarthric articulation consists of aligned acoustics and measured 3D articulatory features from English speakers with dysarthria secondary to either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), and matched controls.

5.3.2.1 Equipment and recording procedure

The collection of movement data and time-aligned acoustic data is carried out using the 3D AG500 electro-magnetic articulograph (EMA) system [94] with fully-automated calibration. This system allows for 3D recordings of articulatory movements inside and outside the vocal tract, thus providing a detailed window on the nature and direction of speech related activity [95]. Acoustic data is recorded simultaneously using two microphones: the first records audio at 44.1 kHz, and is placed facing the participant at a distance of 61 cm, and the second is a head-mounted microphone, which records audio at 16 kHz.

All participants read English text from a 19-inch LCD screen, placed 60 cm in front of them. The read text belongs to one of the following categories:

1. **Non-words:** used to account for the baseline abilities of the dysarthric speakers. These include the following:
   - 5 to 10 repetitions of /iy-p-ah/, /ah-p-iy/, and /p-ah-t-ah-k-ah/, respectively.
   - High-pitch and low-pitch vowels maintained over 5 s (e.g., "Say 'eee' in a high pitch for 5 s").
2. **Short words:** These include the following:
   - Digits (one to ten)
   - international radio alphabet
   - 'yes', 'no', 'up', 'down', 'left', 'right', 'forward', 'back', 'select', 'menu' .
   - 50 words from the word intelligibility section of the standardized FDA.
   - 360 words from the word intelligibility section of the Yorkston-Beukelman assessment of intelligibility of dysarthric speech [96].
   - The 10 most common words in the British national corpus [97].
   - All phonetically contrasting pairs of words from Kent et al. [98].
3. **Restricted sentences:** This category includes the following:
   - Preselected phoneme-rich sentences.
   - The Grandfather passage from the Nemours database [99].
   - 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech.
   - The 460 TIMIT-derived sentences used as prompts in the MOCHA database [100, 101].
4. **Unrestricted sentences:** Participants were asked to spontaneously describe 30 images of situations taken randomly from among the cards in the Webber Photo Cards: Story Starters collection [102].

5.3.2.2 Speakers distribution

8 English speakers (3 female, 5 male), with dysarthria secondary to CP (7 participants) or ALS (1 participant), participated in this dataset. The subjects were between the ages of 16 and 50. In addition, 7 English speakers (3 female, and 4 male), with no speech problems, served as controls. The controls were matched according to age and gender.
The motor functions of each subject were assessed according to the standardized FDA by a SLP.
Not all data from each participant, is freely available for use. We have only used the freely available data.
The severity level distributions of the dysarthric participants, used in the experiments conducted during this thesis, for both UASpeech and TORGO, are shown in table 5.2

Severity	UA-Speech	TORGO
Very Low	F05, M08, M09, M10, M14	F03, F04, M03
Low	F04, M05, M11	F01, M05
Medium	F02, M07, M16	M01, M02, M04
High	F03, M01, M04, M12	-

Table 5.2: Dysarthria severity levels of participants in the UA-Speech and TORGO datasets

5.4 VocaPark

The VocaPark dataset was collected within the framework of the EU-funded OASEES project, which aims to develop an open, decentralized, intelligent, and programmable framework for swarm software architectures and applications. Grounded in the paradigm of decentralized autonomous organizations (DAOs), the project incorporates "Human-in-the-Loop" (HITL) processes to enhance decision-making efficiency. OASEES seeks to provide a fully open-source, decentralized, and secure programming environment for European swarm-based systems, fostering the development of decentralized AI-driven services. The capabilities of OASEES are meant to be demonstrated across six relevant use cases. The VocaPark dataset comes in the context of the Use Case 1: Development of a self-guided speech rehabilitation tool for Parkinson's disease patients. This tool aims to extend rehabilitation sessions beyond the therapist's presence by incorporating advanced voice and speech analysis algorithms powered by deep learning. It will deliver personalized exercises tailored to patient needs within a motivating, interactive, and multi-user environment.

5.4.1 Equipment and recording procedure

The recording sessions took place in a quiet, isolated room, without echo. Data collection for each participant was conducted in an one-hour session, during which regular breaks were offered to the participant. Patients were recorded while they were in the "ON" state, i.e. one to two hours after receiving the daily dose of Levodopa equivalent. During this state, the patient experiences improved motor function following the administration of dopaminergic medication, which aims to replenish the brain's dopamine levels, which are deficient in individuals with PD. During this time, the patient's symptoms are better controlled. They experience enhanced mobility, reduced rigidity, and fewer tremors, making it an optimal period for assessing motor and speech functions.
During the session, a speech therapist conducted the BECD. The participants were shown cards, on which the element to repeat is written. Each element said by the participant was recorded, stored, and scored directly by the therapist. Recordings were collected using a Sony $SRS - NS7^{TM}$ microphone, shown in figure 5.2. This omnidirectional microphone was chosen to limit the effect of variable distance between the microphone and the participant's position, and because it follows the following specifications [103, 104]:

- Positioned 4 to 10 cm from the lips and at an angle of 45° to 90° to the front of the mouth to ensure the same sensitivity whatever the direction of the sound source.

- Although the signal-to-noise ratio is lower than with a unidirectional microphone, the proximity effect will have less of an impact on measures based on spectral and sound pressure level measurements.
- The frequency response varies by less than 2 dB.
- The frequency range is between 50 and 8000 Hz.
- The noise level is at least 10 dB lower than the sound level of the weakest vocal sound.
- The upper limit of the dynamic range is greater than the sound level of the loudest phonations.

After the session, a second speech therapist listened to the recordings, and gave another score for each element. Based on the average of the individual scores, the participant was assigned to one of four possible classes: no dysarthria/control, mild, moderate or severe dysarthria. Some PD patients were classified as having no dysarthria because the disease had not yet manifested in their speech. Additionally, some control participants were found to exhibit dysarthria.

Figure 5.2: The Sony $SRS - NSR^{TM}$ microphone [14].



The recordings belonged to one of the following categories:

1. **Intelligibility**: This category consists of words and sentences that the participant had to read in order to assess their intelligibility. The participant read 10 words and 10 sentences randomly selected from two lists of 50 words and two lists of 50 sentences, resulting in a total of 20 recordings. Details about the lists of words and sentences could be found in Appendix I - Annexe 1.

2. **Phonetic intelligibility:** This category is based on the identification of phonetic alterations leading to incorrect phoneme identification and thus reduced intelligibility. It is a multiple-choice test of 52 bi-syllabic words, representing 13 sets of phonetic contrasts. Each contrast is tested 8 times. For each word read, 4 words are presented: the target word and three distractors. The words are presented alternately as targets and distractors, so that each series is presented 4 times. In each series the words are contrasted in pairs by one or two phonetic contrasts. As each word is read by the participant, the therapist selects, from the 4 possible choices, the word they believe to be said, or, failing that, the word that comes closest to what was understood. This resulted in a total of 52 recordings. Details about the lists of words could be found in Appendix I - Annexe 2.

3. **Phonetic analysis:** This category consists of three tests of increasing complexity, to identify phonetic disturbances using phoneme, syllable and word repetitions (details to be found in Appendix I - Annexe 4).

    a) Isolated phonemes: consists of 30 phonemes to be read by the patient. It allows the therapist to check the patient's articulatory system.

    b) Simple words: consists of 37 words selected because of their simple phonetic structure, their frequency, and their length.

    c) Complex words: consists of 25 complex words, containing semi-vowels and one or more consonant clusters of two, three or four consonants.

4. **Motricity:** The motor examination grid evaluates how well the lungs, larynx, velopharynx, tongue, lips, cheeks, and jaws work together to produce speech. It also includes observation of posture and face, as well as alternative movements. In this dataset, we only performed recordings regarding the motricity of larynx (phonemes /a/, /s/, /z/), and alternating movements (Diadochokinesis, i.e. repetitions of /o/-/e/, /pa/, /ta/, /ka/, /kala/, /pataka/).

5. **Perceptual Analysis:** Allows the therapist to note the predominant perceptual characteristics of the patient. The participant read a paragraph (extract from "Pierrot", G. de Maupassant, see figure 5.3), and was also shown the image shown in figure 5.4 and asked to freely describe it. For both cases, five metrics were evaluated by the therapist.

For more details about the exact content of each category, and the scoring system, refer to Appendix I.
In total, each participant recorded around 175 recordings. In addition to the recordings, the following metadata was collected: age, gender, height, weight, ethnic background,

tobacco consumption in packs/year, time since quitting smoking in years if applicable, medication (Levodopa/Levodopa equivalent prescribed, daily dose, and time since last administration), and time since diagnosis in months.

5.4.2 Speakers Distribution

The VocaPark dataset consists of data collected from 50 PD patients, and 50 control participants. The patients were recruited from the "Centre Calvé", at the "Fondation Hopale" in Berck-sur-Mer, France. Controls were matched to the PD patients on the basis of age, sex and body mass index (BMI) and recruited from other patients without neurological disorders (orthopaedic conditions). The selected participants were subjected to the following inclusion/exclusion criteria:

1. Inclusion criteria:

    - For patients: a clinical diagnosis of Parkinson's disease (PD) confirmed by a neurologist.
    - Native French speaker.
    - Adult subject.
    - Ability to cooperate, understand, and follow simple instructions.
    - Patient affiliated with the French social security system.
    - Voluntary participation.

2. Exclusion criteria:

    - Parkinsonian syndrome without diagnosis of idiopathic Parkinson's disease.
    - Respiratory, gastroesophageal, auditory system, or vocal cord disorders.
    - For patients: advanced clinical symptoms of Parkinson's disease (PD), such as visual hallucinations, frequent falls, cognitive impairment, or the need for institutional care.
    - Pregnant or breastfeeding subjects.
    - Visual, cognitive, or auditory impairment at a level that would interfere with the measurements.

5.4.2.1 Statistics

The VocaPark database currently consists of the data of 114 participants, of which 40 are female participants and 74 are male participants. We present in table 5.3 the control/patient, male/female distribution of participants, the average age, and the average weight. In addition, the severity level distributions of participants with PD is presented in table 5.4.

Category	Number of participants	Average age (years)	Average weight (Kg)
Female Controls	12	70.5	70
Female Patients	28	74.9	69.4
Male Controls	21	71.9	88.4
Male Patients	55	72.8	79.2

Table 5.3: Some statistics of VocaPark.

Severity	Number of participants	IDs of participants	Number of recordings
No dysarthria	21 F	PP36, CP03, CP05, CS01, CS02, CS03, CS05, PP10, PS02, PS03, PS07, PS08, CS08, CS13, CP17, CP25, CP14, PP24, CP27, PP26, CP36	3551
	36 M	PP35, CP02, CP07, CS04, CS06, PP06, PP08, PP11, PS01, PS04, PS05, CP09, CS09, CS12, CS15, PP14, PS09, PS10, PS11, PS12, CS10, PS13, PS14, CP12, CS17, CS18, CS19, PP19, PP22, CP32, PP25, PP27, PP28, PP29, PP31, PP34	6218
Low	13 F	CP04, PP04, PS06, CS11, CP21, CP23, CP24, CP15, PP20, CP28, CP29, PP32, CP41	2224
	21 M	CP08, CS07, PP02, PP03, PP05, PP07, PP12, CS14, PP13, CS16, PP18, CP18, CP13, PP21, CP33, CP35, PP30, CP37, CP38, CP42, PP39	3632
Medium	4 F	PP15, CP19, CP20, CP31	621
	10 M	CP06, PP09, CP10, CP11, PP16, PP17, PP23, CP26, CP34, PP42	1720
High	2F	CP39, PP44	347
	7 M	CP01, CP16, CP30, PP33, CP40, PP38, PP41	1106

Table 5.4: Dysarthria severity levels of participants in VocaPark dataset

5.5 Conclusion

This chapter provided an overview of the main datasets that were utilized to train and evaluate the models for this thesis. In particular, two publicly available English databases, UASpeech and TORGO, were presented and explained, along with a newly collected French database that contained audio recordings of French Parkinson patients and controls. Being one of the first French dysarthric databases gathered with the intention of supporting research on speech pathology technologies, the latter contributes to the field of speech pathology research in general by offering data that is crucial to the advancement of dysarthric speech research and technologies, and to French speaking in particular.
UASpeech and TORGO were mainly used in experiments presented in chapters 6 and 7, whereas VocaParK was used to conducted the experiments presented in chapter 9.

Figure 5.3: The extract from "Pierrot", G. de Maupassant.

Les deux femmes habitaient une petite maison à volets verts, le long d'une route en Normandie, au centre du Pays de Caux.

Comme elles possédaient, devant l'habitation un étroit jardin, elles cultivaient quelques légumes.

Or, une nuit, on vola une douzaine d'oignons.

Dès que Rose s'aperçut du larcin, elle courut prévenir Madame, qui descendit en jupe de laine.

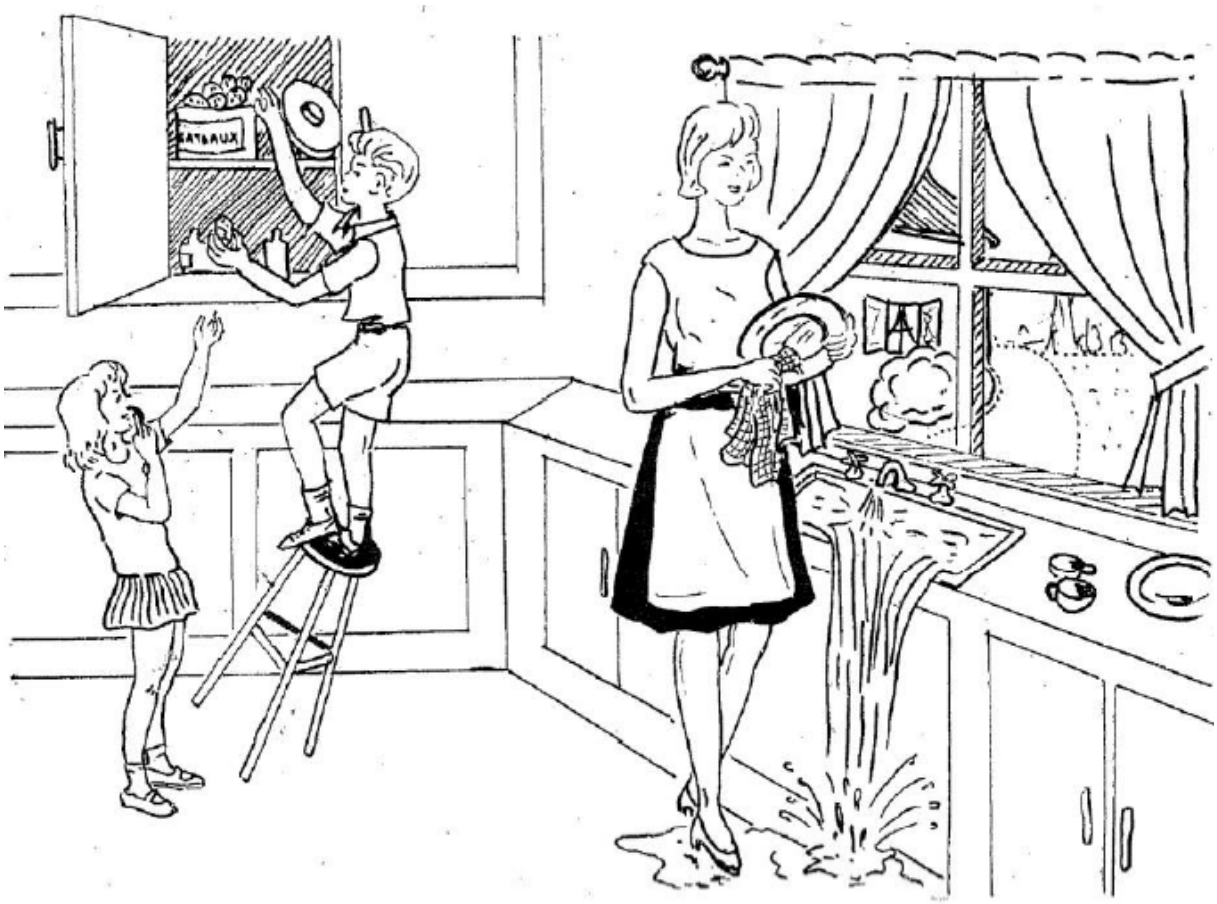
Ce fut une désolation et une terreur. On avait volé, volé Madame Lefebvre ! Donc, on volait dans le pays, puis on pouvait revenir.

Et les deux femmes, effarées, contemplaient les traces de pas, bavardaient, supposaient des choses : « tenez, ils ont passé par là. Ils ont mis leurs pieds sur le mur ; ils ont sauté dans la plate-bande. »

Et elles s'épouvantaient pour l'avenir. Comment dormir tranquillement maintenant.

Extrait de « Pierrot »,
G. de Maupassant

Figure 5.4: The image shown to the participants to be described.



Chapter 6

Is Attention All You Need for Dysarthria Severity Classification ?

6.1 Introduction

Earlier research on dysarthria severity classification was based on a two-steps scheme: firstly, well studied features are extracted from the recorded audios. Secondly, these features are used to train a classifier to classify the audio input according to its dysarthria severity. This era resulted in the construction of different training features, such as MFCCs, prosodic measures (e.g., pitch and intensity), and spectral features. It also resulted in the use of various classifiers, ranging from earlier statistical models, such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), to more advanced machine learning models, such as Support Vector Machines (SVMs) and deep neural networks.
In the last few years, the transformer architecture have been achieving state of the art results in various domains. After its introduction by Vaswani et al. in their work "Attention is All you Need" [7], it quickly became one of the main used architectures in some of the most popular deep learning models, such as BERT [105], GPT [106], and Vision Transformers [107].
Despite the advantages and the powerful performance of transformers, it is yet to be used in the problem of dysarthria severity classification. Even though some work has been done using pre-trained models based on transformers, to our knowledge, no work have investigated training transformer-based models from scratch for dysarthria severity classification.
In this chapter, we propose a model following the traditional approach. The main point addressed in this chapter is the **classifier** used for dysarthria classification. Specifically, we propose the use of transformers, a neural network architecture based on attention mechanisms, as classifiers for dysarthria classification. Transformer is a prominent sequence-to-sequence model architecture that have been widely adopted in various fields, such as computer vision (CV), natural language processing (NLP), and speech processing [108]. It was first introduced for machine translation [7], but quickly replaced recurrent neural networks (RNNs) and its variations, i.e. LSTM [109] and gated recurrent networks (GRU) [110] neural networks, which were the state of the art approaches in sequence

modeling and transduction problems [7] such as language modeling and machine translation [111–113]. In this chapter, we study how this transformer architecture could benefit research on dysarthria severity classification. As features to this proposed transformer-based model, we use two traditional features, MFCC and eGeMAPS. In addition, another important point addressed in this chapter is the use of more advanced **features**, extracted using wav2vec2 model. The performance of the proposed model is compared to a LSTM model.
The rest of the chapter is organized as follows: Section 6.2 introduces the proposed model, along with all the technical and experimental settings, including the extracted features, the classifiers used in the work, the datasets used for training, the training settings, as well as the evaluation metrics. Results and discussion are presented in Sections 6.3 and 6.4 respectively, for both speaker-dependent and speaker-independent data partition scenarios. And finally, Section 6.5 concludes the chapter.

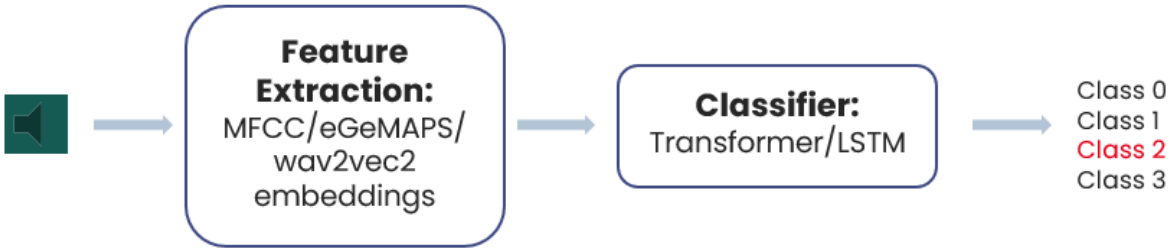
6.2 Proposed Method

We propose a dysarthria severity classification model following the traditional two-step approach, as demonstrated in figure 6.1:

1. Feature extraction: Features are extracted from audio inputs. In the current work, we experiment with three different features: MFCCs, eGeMAPS, and wav2vec2 embeddings. Further details are provided in later sections.

2. Classifier: A classifier is trained using the extracted features, to classify the audios according to dysarthria severity. Our primary focus is a model incorporating the transformer architecture, which forms the core research question of this work. For comparison, we also train a classifier using LSTM layers. Additional details are provided in later sections.

Figure 6.1: A block diagram of the proposed model.



6.2.1 Feature Extraction

Three feature sets were used in this study: MFCCs, eGeMAPS, and wav2vec2 speech representations.

- **MFCCs:** In this work, we extracted 13 coefficients from each audio. This resulted in a feature set of shape (n,T,13), where T is the number of frames per audio, and n is the number of audios in a batch.
- **eGeMAPS:** In this work, we utilized the low level descriptors of the set, resulting in a total of 25 features. The resulting feature set has a shape of (n,T,25).
- **wav2vec2 speech representations:** As explained in Chapter 3, the wav2vec2 architecture consists of two main parts: a CNN feature extractor, and a transformer classifier. In this work, we used the output of the CNN feature encoder, i.e. the latent speech representations. So, in the following, when wav2vec2 embeddings is used, it is to refer to the output of the CNN network. This resulted in a feature set of shape (n,T,768).

To obtain a fixed T for the different audio lengths, the average length of all audios is calculated. Each audio is limited to this average value, longer audios are trimmed, while shorter audios are padded with 0.

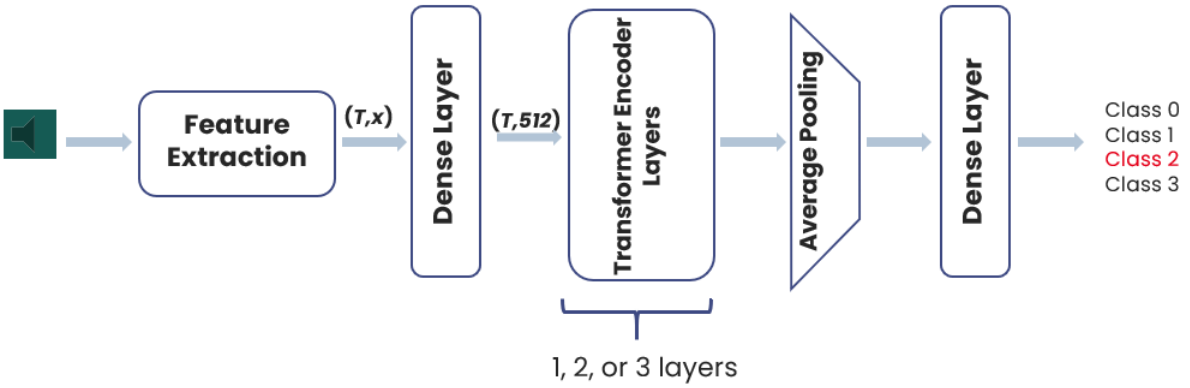
6.2.2 Classifier

In this work, we propose a classifier based on transformer, to leverage its capabilities for dysarthria severity classification. To this end, we only use the encoder part of the transformer architecture, as the task is a classification task, and does not require sequence generation. We experiment with transformer encoders comprising one to three layers, each with a hidden dimension of 512 and 8 attention heads. These were followed by a linear layer with an input size of 512, and output size of 4, corresponding to the four severity classes. A block diagram of the classifier is provided in figure 6.2.
To evaluate the transformer's performance, we also train a classifier using LSTM layers as an alternative. Similar to the transformer classifier, we experiment with architectures comprising one to three layers. The LSTM layers have a hidden size of 128, and we use the final hidden state as the input to a linear layer. This layer has an input size equal to 128, and output size of 4, corresponding to the four possible severity levels.
The input size of the models corresponds to the feature size (13 for MFCCs, 25 for eGeMAPS, and 768 for wav2vec2 embeddings).
For each set of features, we train separate classifiers using both Transformer and LSTM models, experimenting with different layer configurations for each model.

6.2.3 Dataset

In this work, we used the UASpeech dataset to train and test the proposed models. Training and testing of the models were performed following both speaker-dependent approach and speaker-independent approach scenarios. This dual evaluation allowed us to compare performance across both settings and emphasize the importance of considering both approaches when assessing model effectiveness. It also highlights the need to select the most appropriate evaluation strategy based on the intended application of the model. For the speaker-dependent scenario, audios from blocks 1 and 3 are used for training, while those from block 2 for testing (cf. 5.3.1). For the speaker-independent scenario, audios from 2 participants per severity class were used for training, while audios from the

Figure 6.2: A block diagram of the proposed transformer-based model. T and x refer to number,of frames and feature dimension respectively.



remaining participants were used for testing. The exact data distributions are provided in tables 6.1 and 6.2 for the speaker-dependent and speaker-independent cases respectively.

Table 6.1: Data distribution across train and test sets for the speaker-dependent data split.

Severity	**Train (nb of audios)**	**Test (nb of audios)**
Very Low	2550	1275
Low	1519	762
Medium	1530	765
High	2016	1020

Table 6.2: Participants' distribution across train and test sets for the speaker-independent data split.

Severity	**Train**	**Test**
Very Low	F05, M09	M08, M10, M14
Low	F04, M05	M11
Medium	F02, M16	M07
High	F03, M04	M01, M12

6.2.4 Training Settings

The models are trained using the Adam optimizer, which was chosen for its adaptive learning rate properties and ability to converge efficiently in deep learning tasks. We use cross-entropy loss, a standard loss function for multi-class classification, as it effectively measures the difference between predicted and actual class probabilities.
To optimize training, we experiment with multiple learning rates (0.00001, 0.0001, 0.001

and 0.01) to assess their impact on model convergence and performance. A lower learning rate (e.g., 0.0001) allows for more stable updates but may slow convergence, while a higher learning rate (e.g., 0.01) speeds up training but risks overshooting optimal weights. The learning rate is chosen based on validation performance.
We train the models for 10 epochs using a batch size of 64. Early stopping is applied to prevent overfitting, with training terminating if validation loss does not improve for 5 consecutive epochs.
All experiments are conducted under both speaker-dependent and speaker-independent settings to ensure robustness across different evaluation scenarios.

6.2.5 Evaluation Metrics

To assess the performance of the proposed model, we calculate the following metrics: accuracy, f1-score, precision, and recall.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.3}$$

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

Where TP, TN, FP, and FN correspond to true positive, true negative, false positive, and false negative respectively.

6.3 Results

6.3.1 Speaker-Dependent Scenario

Results for the speaker-dependent case are summarized in Table 6.3. The best-performing models for each feature set are reported across all layer configurations.

Table 6.3: A comparative table of the results achieved by each model, with a speaker-dependent data partition.

Model	**Feature set**	**Accuracy (%)**	**F1-score**	**Precision**	**Recall**	**Number of layers**
Transformer	eGeMAPS	53	0.55	0.58	0.53	2
	MFCCs	88	0.89	0.89	0.88	2
	wav2vec2	**92**	0.92	0.92	0.92	2
LSTM	eGeMAPS	50	0.49	0.49	0.49	2
	MFCCs	**90**	0.9	0.9	0.9	2
	wav2vec2	**90**	0.9	0.91	0.9	1

For the transformer-based model, **wav2vec2** features achieved the highest accuracy of **92%** and an F1-score of **0.92**, followed by **MFCCs** with an accuracy of **88%** and an F1-score of **0.89**. The **eGeMAPS** feature set lagged significantly behind, with an accuracy of **53%** and an F1-score of **0.55**. Notably, for all three feature sets, the best-performing models used **two transformer encoder layers**, suggesting that this configuration was effective in learning speaker-specific patterns.
For the LSTM-based model, the results followed similar trends. Both **MFCCs** and **wav2vec2** features achieved the highest accuracy of **90%**, with F1-scores of **0.90**. Meanwhile, the **eGeMAPS** feature set performed the worst, with an accuracy of **50%** and an F1-score of **0.49**. The optimal model configuration consisted of **two LSTM layers** for eGeMAPS and MFCCs, while **one layer** sufficed for wav2vec2.

6.3.2 Speaker-Independent Scenario

Results for the speaker-independent case are summarized in Table 6.4. Compared to the speaker-dependent scenario, performance dropped significantly for all models and feature sets, highlighting the challenges of generalizing across unseen speakers.

Table 6.4: A comparative table of the results achieved by each model, with a speaker-independent data partition.

Model	Feature set	Accuracy (%)	F1-score	Precision	Recall	Number of layers
Transformer	eGeMAPS	27	0.28	0.3	0.26	2
	MFCCs	44	0.38	0.4	0.37	2
	wav2vec2	**46**	0.37	0.37	0.36	2
LSTM	eGeMAPS	34	0.32	0.32	0.27	3
	MFCCs	**38**	0.4	0.37	0.34	1
	wav2vec2	37	0.39	0.34	0.34	1

For the transformer-based model, **wav2vec2** features achieved the highest accuracy of **46%** and an F1-score of **0.37**. **MFCCs** followed closely with an accuracy of **44%** and an F1-score of **0.38**, while **eGeMAPS** lagged behind with an accuracy of **27%** and an F1-score of **0.28**. The optimal transformer configuration for all feature sets consisted of **two encoder layers**.
For the LSTM-based model, **MFCCs** achieved the highest accuracy of **38%** and an F1-score of **0.40**, followed by **wav2vec2** with an accuracy of **37%** and an F1-score of **0.39**. The **eGeMAPS** feature set scored the lowest, with an accuracy of **34%** and an F1-score of **0.32**. The best model configurations involved **three LSTM layers** for eGeMAPS, and **one layer** for both MFCCs and wav2vec2.

6.4 Discussion and Comparison with Related Work

In this chapter, we proposed a dysarthria severity classification system, following the traditional approach. In chapter 4, we discussed several articles proposing a traditional dysarthria severity classification system. In particular, we will compare our work to four

references.
In [79], the MFCC + i-vector + DNN pipeline gives pretty good results, 93.97% and 49.22% for SD and SI scenarios, respectively. However, one drawback of this method is data hunger, building the 512-component universal background model and total-variability matrix that generate those vectors requires hours of balanced, high-quality speech, a luxury rarely available in clinical corpora. A second drawback is that the pipeline is multi-stage and somewhat fragile (MFCC extraction, expectation–maximization (EM) training of two generative models, i-vector extraction, then discriminative learning), so every change of microphone, language or patient population demands costly re-engineering. Third, squeezing the entire utterance into one vector discards within-word temporal detail that often betrays articulatory breakdowns in dysarthria. Fifth, performance is fragile with respect to the chosen descriptor: substituting CQCCs for MFCCs in the original study cut accuracy by more than 25 percentage points, showing heavy reliance on hand-tuned features, which in turn rely heavily on the provided data. Finally, any infusion of new data triggers a full round of EM retraining, incurring substantial computational overhead.
In [80], The Fisher-vector (FV) follows a similar generative approach, starting with hand-crafted spectral/cepstral features computed frame-wise. Each corpus then undergoes three additional steps: first, a Gaussian-mixture model must be estimated; Second, frame posteriors are accumulated into a high-dimensional FV that collapses the utterance to a single fixed-length vector; And third, the vectors are passed through linear LDA/PLDA projections or a one hidden layer Artificial Neural Network (ANN) for classification . This pipeline reaches an accuracy ranging between 72.2% and 99.9% in the SD partition, against 45.6% in the SI partition, with the mid intelligibility class often entirely missed, for UASpeech. A similar trend appears in TORGO: word-level FV ranges between 55.5% and 80.4% for the SD scenario, vs an accuracy of 57.3% for the SI scenario, but still fails to recover mid-class talkers and needs separate GMMs for every new database. Even though this method is less computationally expensive compared to that of Joshy et al.'s method, since it uses uses only 2 Gaussian components instead of 512 Gaussian mixtures, it still inherits several limitations of generative front-ends: speaker leakage, hand-tuned descriptors, multi-stage complexity, temporal information loss, class-imbalance sensitivity, retraining overhead and dataset-specific parameterization.
In [81], Kachhi et al. ditch the generative modeling and craft TECCs, that emphasize vocal fold non-linearities. These coefficients are used to train a deep 22-layer ResNet, which attains an accuracy of 98.02% on UASpeech following the SD data split. While the paper confirms TECC's ability to exploit glottal non-linearities and to work with very short test segments, the approach still inherits classic handcrafted front-end drawbacks: it needs task-specific signal processing, separate tuning of filter-bank and coefficient dimensions. The model was never tested using other datasets or following the SI data split, so no evidence of robustness outside that setting. A full network retraining for every new corpus might thus be needed.
Finally, Gurugubelli et al. [82] craft an analytic-phase feature set (SFFB-IFCC) that captures instantaneous frequency micro fluctuations, convert it into 100-dimensional i-vectors and classify with PLDA, obtaining 64.47% accuracy for dysarthria severity classification, using LOSO validation (it should be noted that , and 95.61% for dysarthria detection. These results were obtained at the cost of two EM loops, utterance-level temporal flat-

tening.
The method we presented in this chapter, i.e. the frozen wav2vec 2.0 combined with a light multi-layer transformer resolves, point by point, the limitations presented by the aforementioned pipelines. First, it eliminates fragile feature design: instead of MFCCs, CQCCs, TECCs or analytic-phase cepstra that must be re-tuned for every microphone or language, wav2vec2 delivers the same high-quality embeddings directly from raw waveforms, so no front-end redesign is ever needed. Second, its multi-head self-attention gives the model true long range temporal reasoning, whereas i-vector, Fisher-vector and ResNet systems either flatten time or rely on fixed, local receptive fields. Third, that same attention mechanism dynamically down-weights speaker-specific frames and up-weights dysarthria-relevant segments, sharply reducing the speaker "leakage" that makes handcrafted systems excel on familiar voices yet stumble on unseen ones. Fourth, because the encoder is frozen, training cost collapses: we optimize only the head parameters, instead of running several expectation–maximisation loops for a Gaussian mixture and a total-variability matrix or re-training a deep 22-layer CNN. Fifth, the method thrives on modest labeled data, when wav2vec2 embeddings are used, whereas generative front-ends need hours of balanced in-domain audio to estimate their background models reliably. Finally, maintenance is trivial: adapting to a new clinic or language requires at most a short fine-tune of the transformer (or unfreezing wav2vec2), not a wholesale rebuild of the feature extractor and back-end. Together, these properties let the wav2vec2 + transformer architecture match handcrafted pipelines' real-time speed while offering far greater robustness, lower engineering overhead and a clear path to scalable, cross-speaker dysarthria assessment.
Focusing the discussion on our results, they reveal stark contrasts between the speaker-dependent and speaker-independent scenarios, with transformers showing notable strengths, especially in handling the complexities of dysarthria severity classification.

Benefits of Transformers in Speaker-Independent Classification
The speaker-independent scenario posed significant challenges due to intra-class variability and limited training data, leading to substantial performance drops for both models. However, transformers demonstrated clear advantages, particularly with MFCCs and wav2vec2 features. In these cases, transformers outperformed LSTMs by achieving accuracy gains of 6% and 9%, respectively. This indicates that transformers' ability to model global dependencies makes them more adept at generalizing across unseen speakers, a critical requirement for real-world applications.

Transformers vs. LSTMs
While LSTMs showed strong performance in the speaker-dependent scenario, their reliance on temporal modeling limited their ability to generalize in the speaker-independent case. Transformers, on the other hand, excelled due to their capacity to model global relationships and capture high-level patterns, which are essential for the intricate task of dysarthria severity classification. This difference underscores the transformative potential of the transformer architecture for complex speech processing tasks.

Feature Set Analysis: wav2vec2 and Transformers
wav2vec2 features consistently provided the best results in both scenarios, particularly

when used with transformers. The combination of transformers and wav2vec2 achieved a 46% accuracy in the speaker-independent case, outperforming all other configurations. This highlights the synergy between advanced feature extraction methods like wav2vec2 and the transformer architecture, which excels at leveraging high-dimensional, contextual representations.

Practical Implications and Future Directions
The advantages of transformers, particularly in the speaker-independent scenario, highlight their potential for building robust and scalable models for dysarthria severity classification. Future research should focus on:

- Exploring transformer-based models with larger, more diverse datasets.
- Leveraging techniques like domain adaptation or transfer learning to further enhance generalization.
- Investigating the integration of transformers with advanced features like wav2vec2 to maximize performance.

In conclusion, while both transformers and LSTMs achieved strong results in the speaker-dependent scenario, transformers stood out as the more promising architecture for generalizing across unseen speakers. Their ability to handle complex relationships and capitalize on advanced feature extraction makes them a key asset for future developments in dysarthria severity classification.

6.5 Conclusion

In this chapter, we proposed the use of transformer architecture as classifier for dysarthria severity classification. This chapter has demonstrated the potential of transformer-based models for dysarthria severity classification, particularly in addressing the challenges of speaker-independent scenarios. The results showed that while LSTMs performed competitively in the speaker-dependent case, transformers outperformed them in the more complex speaker-independent case, especially when combined with advanced feature extraction methods like wav2vec2.
The findings highlight the importance of model architecture and feature set selection in achieving high performance. wav2vec2 embeddings consistently emerged as the best feature set, reflecting its ability to capture rich contextual information from raw audio data. Transformers, with their ability to model global relationships and leverage high-dimensional feature representations, showed significant advantages in generalizing across unseen speakers.
However, the observed performance gap between speaker-dependent and speaker-independent scenarios underscores the need for further research into improving model generalization. Addressing challenges such as intra-class variability and limited training data will be critical for advancing dysarthria severity classification systems.
In summary, this chapter emphasizes the transformative potential of transformer models in dysarthria severity classification, particularly for real-world, speaker-independent applications. Future work should focus on leveraging their strengths through improved data

diversity, advanced training strategies, and integration with cutting-edge feature extraction techniques.

This chapter asked: "Is attention all you need for dysarthria-severity classification?". Our experiments show that while self-attention improves accuracy, it is not sufficient on its own. Performance gains materialize mainly when attention is paired with high-resolution acoustic embeddings (wav2vec2). We therefore conclude that attention is a powerful but partial solution, robust dysarthria classification still needs rich representations and regularization. Building on these findings, the next chapter turns to state-of-the-art attention-based pre-trained models. We evaluate and compare them in different settings. By doing so, we ask another question: Is large-scale pretraining the missing ingredient, or can the architecture alone close the remaining performance gap? Answering this question will clarify whether attention is a necessary and sufficient component, or merely one part of a broader representation-learning pipeline.

Chapter 7

Leveraging Pre-Trained Models for Dysarthria Severity Detection: A Comparative Analysis

7.1 Introduction

In the previous chapter, we trained a transformer-based model from scratch and demonstrated that it outperformed an LSTM model for dysarthria severity classification. Notably, the Transformer model performed best when features were extracted from the frozen CNN part of the wav2Vec2 model, compared to traditional features like MFCC and eGeMAPS. These findings highlight the potential of transformer-based architectures, particularly when paired with features derived from pre-trained models, to improve performance in speech-related tasks with limited labeled data.
Motivated by these results, we aim to further explore the use of pre-trained models in this chapter. A key challenge in speech-related tasks, particularly for dysarthric speech, is the scarcity of labeled data. Acquiring labeled data, such as transcribed speech samples, is often impractical, as seen in speech recognition systems that require thousands of hours of transcribed audio to achieve satisfactory performance [114]. Similarly, medical data such as dysarthric speech samples is challenging to obtain due to the intricate nature of the data and strict privacy regulations [115]. This scarcity of diverse and representative data makes it difficult to develop robust models capable of extracting meaningful clinical insights [116, 117].
To address this issue, transfer learning offers a promising solution. Transfer learning is a subfield of deep learning, that aims to improve target learners' performance on target domains, through the transfer of knowledge from various but related source domains. It leverages the knowledge encoded in pre-trained models for new, similar tasks with only a small amount of additional labeled data. By doing this, the reliance on an abundance of target-domain data can be lessened for creating target learners [118]. Motivated by these characteristics, we chose to investigate the application of pre-trained models for dysarthria severity detection. By using pre-trained models, we aim to exploit the power of deep models, despite the shortage of dysarthric speech data. Similar to other models, the pre-trained models could be pre-trained following one of the different learning approaches (supervised, self-supervised, etc.). Most of the state of the art models for speech tasks are pre-trained using supervised or self-supervised learning. In addition, in the con-

text of speech tasks, pre-trained models can operate on either waveform or spectrogram inputs. In the former, the raw waveform is used as input to the model in order to train it to perform a certain task. Whereas the latter uses the spectrogram of the audio signal as input.
In this chapter, we study the use of different pre-trained models for dysarthria severity detection, i.e. differentiating between healthy and pathological audios. We cover both waveform-based (wav2vec2 and HuBERT) and spectrogram-based (Whisper) models, as well as supervised and self-supervised pre-training approaches.
The contributions presented in this chapter can be summarized as follows:

- We leverage the use of state-of-the-art pre-trained models for dysarthria detection, compared to traditional features.
- We study the use of both spectrogram-based and waveform-based models.
- We conduct the study following a speaker-independent approach.
- We perform two levels of generalization test: in-database and cross-database.

The remainder of the chapter is structured as follows: we begin by introducing the different pre-trained models used in this chapter. Then, the experimental setup is presented: we present the data distribution and the baseline features used for comparison. We then explain how the pre-trained models are used, and what classifiers are used for dysarthria detection. This is followed by the presentation of the found results, and a discussion. Finally, Section 7.6 concludes this chapter.

7.2 Pre-trained models

As presented in the introduction, in this chapter, three different pre-trained models are studied. Two of the models, wav2vec2.0 and HuBERT, are waveform-based, self-supervised models. While the remaining one, whisper, is a spectrogram-based, supervised model.

7.2.1 Waveform-Based Models

In this chapter, we use two SOTA waveform-based models, wav2vec2.0 and HuBERT. Both models are pre-trained using a self-supervised learning approach.
Even though both wav2vec2.0 and HuBERT use the same convolutional network, followed by a transformer encoder, their training processes are very different. In the following, we highlight some of these differences:

- Loss: HuBERT used the cross-entropy loss, while wav2vec2.0 uses a combination of contrastive loss and diversity loss.
- Targets learning: Wav2vec2.0 learns targets simultaneously while training the model, while HuBERT builds targets via a seperate clustering process.
- Updating targets: Wav2vec2.0 only uses the output of the convolutional network, while HuBERT re-uses embeddings from the BERT encoder to improve targets.

7.2.2 Spectrogram-Based Models

In this chapter, we use Whisper, a SOTA spectrogram-based models. The model is trained following a supervised learning approach.
Unlike waveform-based models, spectrogram-based models, such as Whisper, use spectrograms of the audio signal as input for training. The use of spectrograms is interesting because:

- They effectively capture both temporal and frequency aspects of audio signals.
- They could be directly used as input when using models based on transformers, without having to use a specialized architectures for extracting useful features from raw audios (e.g. the feature encoder part in wav2vec2.0 and HuBERT).
- They provide SOTA for many audio tasks [119, 120].

7.3 Experimental Setup

In this work, we are evaluating the ability of the models to distinguish between non pathological and pathological audios. To this end, the three aforementioned pre-trained models, wav2vec2.0, HuBERT, and Whisper, were used as feature extractors.
In transfer learning, a pre-trained model could be used in one of two states: feature extraction, or fine-tuning. Both techniques benefit from the knowledge of the pre-trained model on a very large source task to improve performance on a target task, but differ in how much the pre-trained model's parameters are modified during training on the target task. Feature extraction directly uses the representations learned by the pre-trained model, so the pre-trained model's parameters are kept frozen during training on the target task. While in fine-tuning, the pre-trained model's parameters are further trained on the target task's dataset along with the additional layers.
Consequently, in our experiments, the pre-trained models were frozen, and the embeddings of the transformer layer were used as input to train a classifier for dysarthria detection. In the following subsections, more details about the data, features, and classifiers are provided in the following subsections.

7.3.1 Datasets

In order to verify how generalizable are the pre-trained embeddings, we perform two levels of testing: in-database and cross-database. More details about the in-database and cross-database testing are provided in section 7.3.5. For the in-database testing, we test on data that was held from the database used for training the classifiers, as for cross-database testing, the test is performed on a second dataset, that was not introduced during the training. More specifically, in our case, the UASpeech dataset is used for training and in-database testing, while TORGO is used for cross-database testing. As for the train/test partition, we followed the speaker-independent approach, where no audios from the same speaker are present in both train and test sets. The exact data distribution, in terms of number of participants per class, is provided in Table 7.1.

Class	Train	Test	Cross-test
Non-pathological	9	4	7
Pathological	9	6	8

Table 7.1: The train, test, and cross-test data distribution for dysarthria detection models.

7.3.2 Baseline Features

For comparison purposes, two baseline features were extracted: MFCCs and eGemaps.

- **MFCCs:** We extract 13 coefficients, along with their first and second derivatives, which results in a 39-dimensional feature vector.
- **eGemaps:** This set consists of 88 features.

7.3.3 Pre-trained models as feature extractors

The pre-trained models were used as features extractors, i.e. the pre-trained weights were frozen and the models were used to extract embeddings from the input audios.
For both wav2vec2.0 and HuBERT, we chose the Base architecture. This choice was motivated by the size of the models, considering that our dataset is limited, we wanted models with less parameters.
Both wav2vec2.0 and HuBERT generate T*768 embeddings per audio, where T depends on the length of the audio recording. To fix the dimension of the features vector, we decided to use the average and standard deviation over time of each feature, which resulted in a feature set of dimension 2 * 768.
In addition, to test whether the last layer of the model gives the best features, or ones from earlier layers are more adapted, different models were trained using the output of each one of the 12 layers of the transformer part, and their performance was compared. In total, we got 12 sets of features per model, of dimension 2 * 768 each.
As for Whisper, the features' dimension generated depend on the architecture used, and it is fixed regardless of the audio input's length. The input to the model is fixed to 30s, so the audios are trimmed or padded accordingly. We used the small configuration, which resulted in features of dimension 1*1500*768. We then calculated the average along the time dimension, which results in a feature dimension of 1*768.

7.3.4 Classifiers

Two classifiers were used for the dysarthria detection task: SVM and Multi-Layer Perceptron (MLP). The SVM is well-suited for tasks with clear separability and high-dimensional data, offering robustness and interpretability, especially with linear decision boundaries. In contrast, the MLP is capable of capturing complex, non-linear relationships and learning complex features through its multiple hidden layers.

- **SVM:** SVM with radial basis kernel basis is used, with the soft margin C and the kernel width γ parameters optimized using grid search cross-validation.

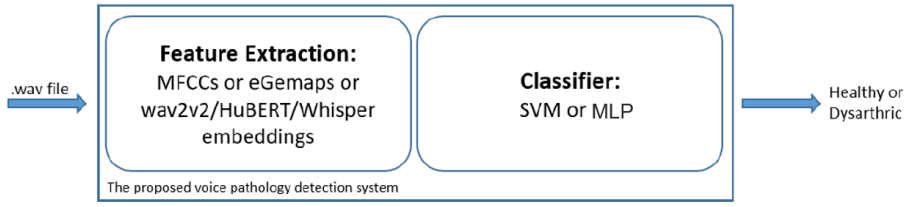


Figure 7.1: Block diagram of the proposed voice pathology detection system.

- **MLP:** The network is used more like a projector to map the high-dimensional features from the pre-trained models to a lower-dimensional space suitable for the classification task in hand, because we don't want the network to drastically change the information captured by the pre-trained models. This is why a simple feed forward neural network is used, composed of an input layer of size equal to the dimension of the input, two hidden layers of size 384 and 192 respectively, with ReLU activation function, and an output layer of size 2 (healthy/pathological). The adamW optimizer and cross-entropy loss function are used. The training was performed for 10 epochs.

A block diagram of the proposed model is provided in Figure 7.1.

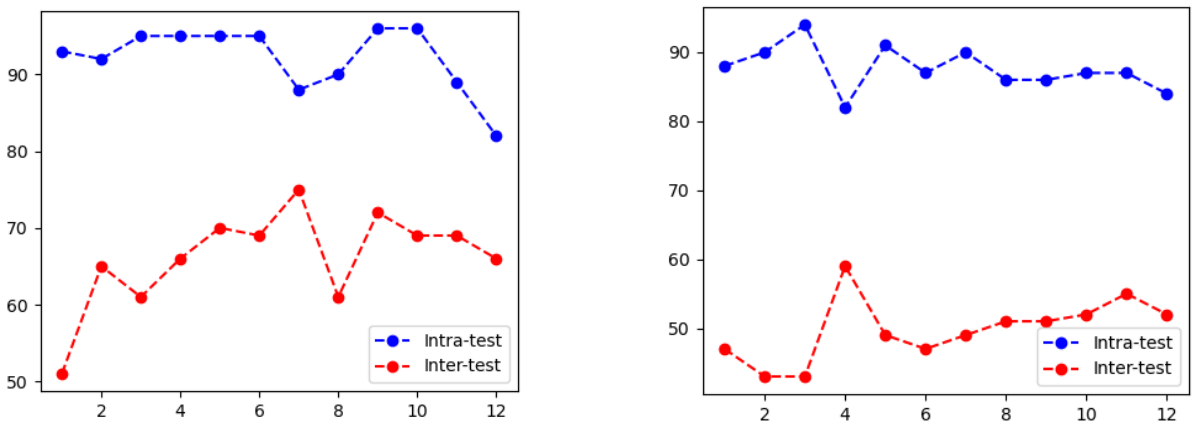
7.3.5 Evaluation Metrics

To evaluate the performance of the trained models, the metrics presented in section 6.2.5 are used, specifically the accuracy, precision, recall, F1-score.
In addition, we calculate the Area Under the Curve (AUC) score, and more specifically the Receiver Operating Characteristic (ROC) curve. The AUC score is a single number that summarizes how well a model separates the positive class from the negative class, regardless of the specific decision threshold, i.e. the cutoff value that decides how a model's predicted probability is turned into a class label.
As for the testing strategy, as explained earlier, we performed two levels of testing: in-database and cross-database. In-database testing involves training and testing the model on different subsets of the same dataset, usually using a train/test split or cross-validation. This method evaluates how well the model learns and generalizes within a controlled and consistent data environment. In contrast, cross-database testing assesses a model's generalization capabilities across different datasets: the model is trained on one dataset and tested on a completely separate one, often collected under different conditions. This approach is more challenging but realistic, as it reveals how robust the model is to domain shifts like variations in background noise, microphone quality, speaker accent, or recording environment. While in-database testing can showcase high accuracy, cross-database testing provides a better measure of real-world performance and generalizability.

7.4 Results and Discussion

We start by comparing the performance of the embeddings of each one of the 12 layers of wav2vec2.0 and HuBERT. This is done by comparing the results obtained by using the

neural network classifier. Results are reported in Figure 10.1. We refer to the testing done on UASpeech test set as intra-test, and the testing on TORGO dataset as inter-test.



(a) Wav2vec2.0 (b) HuBERT

Figure 7.2: The accuracy achieved on UASpeech (intra-test) and TORGO (inter-test), using the embeddings of layers 1 to 12.

Table 7.2: Intra- and inter-testing results using different features+classifier combinations.

Combination	Intra-testing					Inter-testing				
	Accuracy [%]	Precision	Recall	F1-score	AUC	Accuracy [%]	Precision	Recall	F1-score	AUC
eGemaps+SVM	71	0.71	0.72	0.71	0.72	45	0.44	0.44	0.43	0.44
MFCC+SVM	75	0.74	0.75	0.75	0.75	**54**	0.56	**0.56**	**0.54**	**0.56**
wav2vec2+SVM	95	0.95	0.95	0.95	0.95	41	0.49	0.49	0.39	0.49
HuBERT+SVM	94	0.93	0.94	0.94	0.94	42	**0.59**	0.53	0.37	0.53
Whisper+SVM	**96**	**0.96**	**0.96**	**0.96**	**0.96**	42	0.57	0.52	0.37	0.52
eGemaps+MLP	76	0.85	0.73	0.78	0.77	38	0.35	0.87	0.50	0.49
MFCC+MLP	80	0.80	0.80	0.80	0.80	36	0.18	0.50	0.26	0.50
wav2vec2+MLP	**96**	**0.98**	**0.95**	**0.97**	**0.96**	**70**	0.63	**0.66**	**0.65**	**0.70**
HuBERT+MLP	87	0.87	0.89	0.86	0.89	55	0.59	0.59	0.55	0.59
Whisper+MLP	90	0.90	0.92	0.90	0.90	61	**0.64**	0.65	0.61	0.65

Figure 10.1(a) shows the results when embeddings of the wav2vec2 model are used as features to the classifier. We see an intra-testing accuracy ranging between 82% and 96%, with layers 9 and 10 achieving the highest score, and an inter-testing accuracy ranging between 51% and 75%, with layer 7 scoring the highest. In Figure 10.1(b), embeddings of the HuBERT model are used as features to the classifier. The intra-testing accuracy is ranging between 84% and 94%, with layers 3 achieving the highest score, and an inter-testing accuracy ranging between 43% and 59%, with layer 4 scoring the highest. For both models, we see that in the case of intra-testing, the results from the majority of layers are fairly close, with the exception of a few layers where the accuracy tends to change relatively a lot.

We now compare the performance of the pre-trained trained models, i.e. wav2vec2, HuBERT and Whisper, to that of some traditional features. The results of our experiments are shown in Table 7.2. As stated earlier, two sets of traditional features are used, i.e.

MFCCs and eGemaps, and two classifiers are used for the classification, i.e. SVM and a MLP. For wav2vec2 and HuBERT, we report the results of the layer that we consider to give the best compromise between intra-test accuracy and inter-test accuracy.
For both SVM and MLP, the best results are achieved using the pre-trained models, for all metrics, for the intra-test set. When SVM is used as the classifier, the three models generally achieve very similar results for all metrics, with a difference of 1-3% between the three models. The use of these models improve the performance of the classifier of a minimum of 20%, 0.21, 0.2, 0.2, and 0.2 for accuracy, precision, recall, f1-score and AUC score, respectively, when compared to the traditional features. Conversely, the difference between the models becomes more apparent when MLP is used as the classifier, with the accuracy going from 87%, to 90%, to 96% for HuBERT, Whisper, and wav2vec2, respectively. Similarly, the AUC score goes from 0.89, to 0.90, to 0.96 for HuBERT, Whisper, and wav2vec2, respectively. With these results, the pre-trained models surpass the results of the best performing baseline features by a minimum of 7% and 9% for accuracy and AUC score respectively.
As for the inter-test, i.e. the testing on TORGO dataset, the pre-trained models also achieve the best results compared to the baseline features when MLP is used as classifier. A comparable difference is noticed between the pre-trained models, with wav2vec2 achieving the highest scores for all metrics except for precision. More specifically, wav2vec2 scored 70% for both accuracy and AUC score, followed by Whisper, with an accuracy and AUC score of 61% and 0.65 respectively, and finally HuBERT with an accuracy of 55% and AUC score of 0.59. The pre-trained models surpass the baseline features by a minimum of 17% and 10% in terms of accuracy and AUC score. Unlike when the MLP is used as a classifier, the pre-trained models perform worse on the inter-test set compared to the baseline features when the SVM classifier is used. Their performance is similar to each other, falling behind the MFCC features by 12-13% in accuracy and 0.3-0.7 in AUC score. These results underline the importance of the choice of the classifier, as we noticed a drop ranging between 13% and 29% in terms of accuracy when the SVM was used instead of the MLP.
Overall, these results highlight the significant influence that the choice of classifier has on the performance of the models. While the pre-trained models (wav2vec2, HuBERT, and Whisper) demonstrate strong performance on the intra-test set, particularly when using the MLP classifier, their performance on the inter-test set (TORGO dataset) is more variable and dependent on the classifier used. The substantial improvement observed with MLP, especially for the pre-trained models, suggests that these models are capable of capturing complex, non-linear patterns in the data, which are better handled by neural networks compared to traditional SVM classifiers.
The finding that the pre-trained models outperform traditional features (MFCC and eGeMAPS) by significant margins in both accuracy and AUC score on the intra-test set indicates that leveraging pre-trained models in speech tasks provides considerable advantages. However, the performance drop on the inter-test set when using the SVM classifier raises important questions about generalization and model robustness. The difference between the intra-test and inter-test results may reflect the challenge of domain shifts, where the pre-trained models are more suited to the specific domain of the UASpeech dataset but face difficulties adapting to the different characteristics of the TORGO dataset, particularly when using simpler classifiers like SVM.

In conclusion, the experiments suggest that while pre-trained models can significantly enhance performance, particularly in terms of accuracy and AUC, the choice of classifier plays a critical role in determining the generalization capability of these models. Future work should explore fine-tuning the pre-trained models further and investigating hybrid approaches to better adapt them to diverse datasets, ensuring robust performance across various domains.

7.5 Comparison with Related Work

In this chapter, we proposed the use of three pre-trained models, wav2vec2, HuBERT, and whisper, for dysarthria detection. The models were used as feature extractors, and their ability to extract meaningful feature to differentiate between normal and pathological speech was tested. We further validated the robustness of our method by performing cross-database testing. Results showed that the best performing model was able to achieve an intra-test accuracy of 96% and an inter-accuracy if 70%. In this section, we will be comparing our work to some of the articles we cited in chapter 4.
In [83], Vyas et al. proposed a system for dysarthria detection and severity classification, following the traditional approach. Using recordings from UASpeech dataset, they extract a rich set of time- and frequency-domain prosodic features, and use a genetic algorithm to select the best performing features to be used for dysarthria detection and severity classification. A SVM is trained using the selected features, using a 3-fold cross-validation method, the proposed model achieved a detection accuracy of 98%. Despite the high result achieved using this method, the SI data split was not adopted, so the generalizability of this method was not challenged. In addition, only 20 utterances were used to train the diagnostic SVM. Such small, homogeneous samples make the model vulnerable to over-fitting and offer no guarantee it will scale to new speakers, dialects or languages. As for the genetic algorithm, it can converge to different feature subsets when the training set changes slightly, which is risky with such few training utterances; retraining on a larger or noisier corpus might yield a completely different optimal subset. Our method was able to achieve a comparable performance 96%, following a SI data split, and was additionally tested on a second database, thus further challenging the robustness of the model.
In [85], Javanmardi et al. study the use of three pre-trained models, wav2vec2-Base, wav2vec2-Large, and HuBERT for dysarthria detection and severity classification. Their study came out in 2024, which was around the time we performed our experiments. They used the models as features extractors, and trained a SVM and a CNN network to detect the presence of dysarthria, and to classify the pathological audios according to dysarthria's severity. They trained the models using UASpeech and achieved a detection accuracy of 95.76% using wav2vec2-Large and the CNN. They also trained a model using TORGO and achieved an accuracy of 76.12%, using HuBERT and the CNN. Their results validate the ones we obtained about the ability of the pre-trained models to distinguish between healthy and pathological speech. Comparing the results, our method matches their UA-Speech score while approaching their TORGO performance without any retraining on TORGO, demonstrating stronger cross-corpus robustness.
One limitation of the present work is that, like the above studies, it is still restricted to English datasets; future work should assess multilingual transfer.

7.6 Conclusion

In this chapter, we investigated the use of embeddings of pre-trained models for voice pathology detection. We compared the performance of three pre-trained models: wav2vec2, HuBERT, and Whisper. The models were used as feature extractors, in their frozen state, and we fed their embeddings to two classifiers: SVM and MLP. A comparative study of the performance of the different layers of wav2vec2 and HuBERT was performed, and the performance of the pre-trained models was compared to that of the baseline features: MFCC and eGeMAPS. All proposed systems achieved results better than those of the baseline features on the intra-test set, more specifically, the best performing pre-trained model improved the test accuracy by 21% and 16% compared to the best performing baseline features, for SVM and MLP, respectively.
Cross-database experiments were also carried out using TORGO database, and results on this database were different between SVM and MLP: the pre-trained models performed the best when MLP was used as classifier, but it was not the case with SVM. This highlighted the importance of choice of classifier, which allowed us to go from 41% to 70% of accuracy using wav2vec2 embeddings when MLP was used instead of SVM. The results of this work highlight the efficiency of pre-trained models' embeddings to generalize well on unseen data, whether it be from the training database or not.
Motivated by these results on the dysarthria detection problem, we proceed, in the following chapter, to examine the performance of the pre-trained models for dysarthria severity classification.

Chapter 8

Extending the Use of Pre-Trained Models for Dysarthria Severity Classification

8.1 Introduction

In the previous chapter, we demonstrated the effectiveness of pre-trained models, particularly when used as feature extractors, for dysarthria detection. The promising results observed in this context highlighted the potential of leveraging pre-trained models to enhance model performance, even with limited labeled data. Motivated by these findings, we now extend this work to dysarthria severity classification, where the task is to assess the degree of severity in dysarthric speech.
In this chapter, we examine four different configurations of pre-trained models to further explore their impact on dysarthria severity classification. Specifically, we investigate the models in four distinct states: 1- using only their architecture without the pretrained weights, 2- using them as feature extractors, 3- fine-tuning the models on the target task, and 4- partially fine-tuning the models. By comparing these configurations, we aim to understand the influence of pre-training and fine-tuning on model performance in the context of dysarthria severity classification. This study builds upon the insights from the previous chapter and seeks to uncover the most effective approach for utilizing pre-trained models in this new task.
The contributions presented in this chapter can be summarized as follows:

- We leverage the use of state-of-the-art pre-trained models for dysarthria severity classification, compared to traditional features.
- We study the use of both spectrogram-based and waveform-based models.
- We conduct an ablation study to analyze the impact of different model settings on performance.
- We conduct the study following a speaker-independent approach.
- We perform two levels of generalization test: in-database and cross-database.

The remainder of the chapter is structured as follows: we begin by citing the different pre-trained models used in this chapter. The chapter is then divided into two parts: one pre-

senting the work using the waveform-models, and the other using the spectrogram-based models. For each part, we present the performed ablation study, the data distribution, followed by the results and a discussion. Finally, Section 8.6 concludes this chapter.

8.2 Pre-trained models

In this chapter, we used the same models seen in the previous chapter, i.e. wav2vec2, HuBERT, and Whisper. In addition, a forth model, AST, a spectrogram-based supervised pre-trained model is used.

8.3 Waveform-based Models

8.3.1 Ablation study

An ablation study was conducted to study the impact of the pre-trained models on the training process. To this end, four different training modes were evaluated and their results were compared. The modes are as follows:

1. **Architecture:** In this mode, the architecture of the models (i.e., Wav2Vec2 and HuBERT) was used without any pre-trained weights. This setup allows us to isolate and assess the contribution of the models' architecture to the classification task.

2. **Feature Extractor (FE):** In this mode, the pre-trained models were used with their weights frozen. This configuration enables us to evaluate the impact of the pre-trained features on the classification task.

3. **Fine-tuning transformer part (FT-part2):** In this mode, the pre-trained models were used with the CNN part weights frozen, while fine-tuning the transformer layers being for the dysarthria severity classification. This case was studied to address the challenge of training the large models on the relatively small dataset.

4. **Fine-tuning (FT):** In this mode, the pre-trained models were fine-tuned for the dysarthria severity classification task. This approach highlights the combined contributions of the pre-training and the fine-tuning process.

8.3.2 Datasets

Just like the case of dysarthria detection, we perform two levels of testing: in-database and cross-database. UASpeech was used for training and in-database testing, whereas TORGO dataset was used for cross-database testing. As for the train/test partition, we followed the speaker-independent approach, where no audios from the same speaker are present in both train and test sets. The exact data distribution, in terms of participants per class, is provided in table 8.1. The training data was further divided in 90%/10% for training / validation respectively.

Table 8.1: Participants distribution across train, test and cross-test sets.

Severity	Train	Test - UASpeech	Cross-test - TORGO
Very Low	F05, M09	M08, M10, M14	F03, F04, M03
Low	F04, M05	M11	F01, M05
Medium	F02, M16	M07	M01, M02, M04
High	F03, M04	M01, M12	-

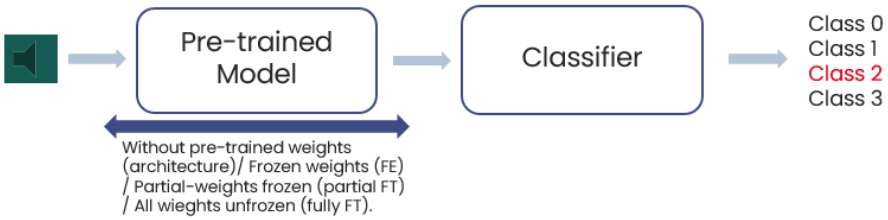


Figure 8.1: Block diagram of the proposed voice pathology classification system. FE, FT refer to feature extractor and fine-tuned respectively.

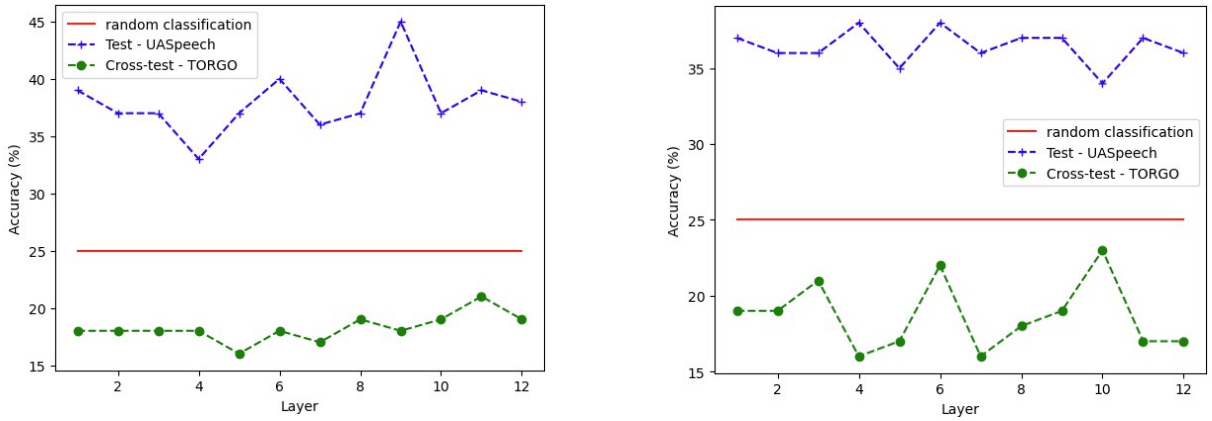
8.3.3 Classifier and Training Settings

For both models, wav2vec2 and HuBERT, the Base configuration was used. The performance of each one of the transformer layers was studied, by using the embeddings of each layer as input to a classification head. The embeddings of each layer are of dimension T*768, where T depends on the length of the recording. To fix the input dimension, we take the average of the embeddings across time, which results in an input dimension of 768. This input is then fed to a classification head.
For dysarthria severity classification, we only add a classification head to the pre-trained models, i.e. a fully connected feed forward Neural Network (NN), with input size 768 and output size 4, was added to the models.
The classifier is trained with Adam optimizer, cross entropy loss, a batch size of 64, 50 epochs with early stopping, and learning rates of $10^{-6}, 10^{-5}$, and 10^{-4}.
As explained earlier, the pre-trained models were used in four different modes, so the training process was repeated for each mode, and for both models.
A block diagram of the proposed model is provided in Figure 8.1.

8.3.4 Results and Discussion

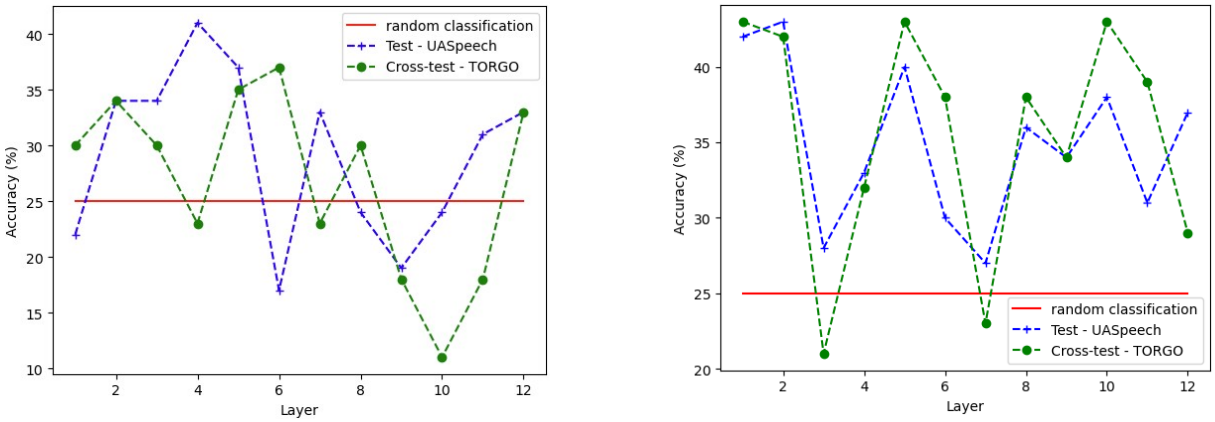
The results of the testing and cross-testing accuracies are reported in figures 10.2, 10.4, 10.6 and 8.5 for the architecture, FE, FT, and FT-part2 modes, respectively. We also plot a line corresponding to the random classification accuracy (25%).
To compare the performance of the two models, wav2vec2 and HuBERT, we report, in table 8.2, the highest test and cross-test accuracy across all layers, even if these highest values do not come from the same layer for each case. Additionally, we report the average test accuracy, cross-test accuracy, and the test to cross-test gap across all layers.

The performance of Wav2vec2 and HuBERT was evaluated across four scenarios: using the models as architectural baselines, as feature extractors without fine-tuning, with



(a) wav2vec2 (b) HuBERT

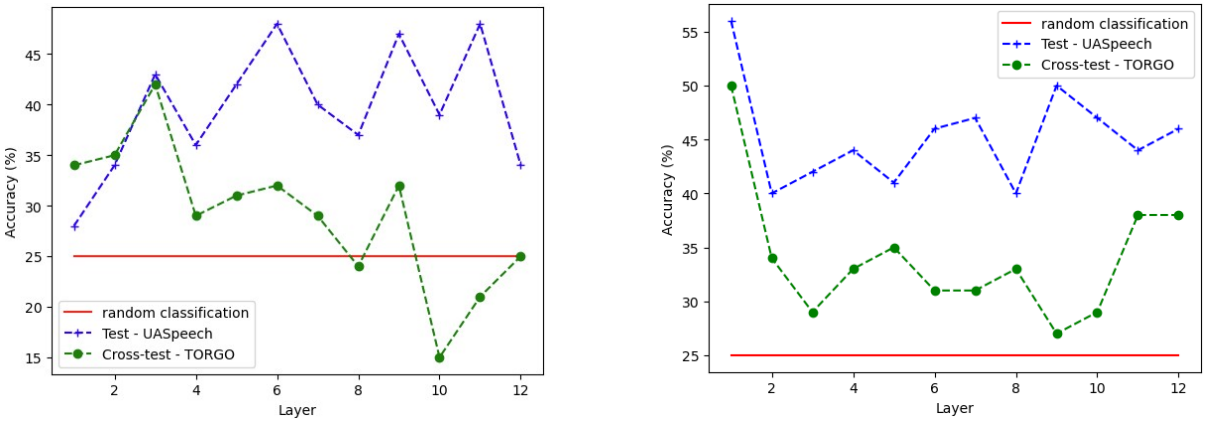
Figure 8.2: Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The architecture of the models is being used, without any pre-trained weights.



(a) wav2vec2 (b) HuBERT

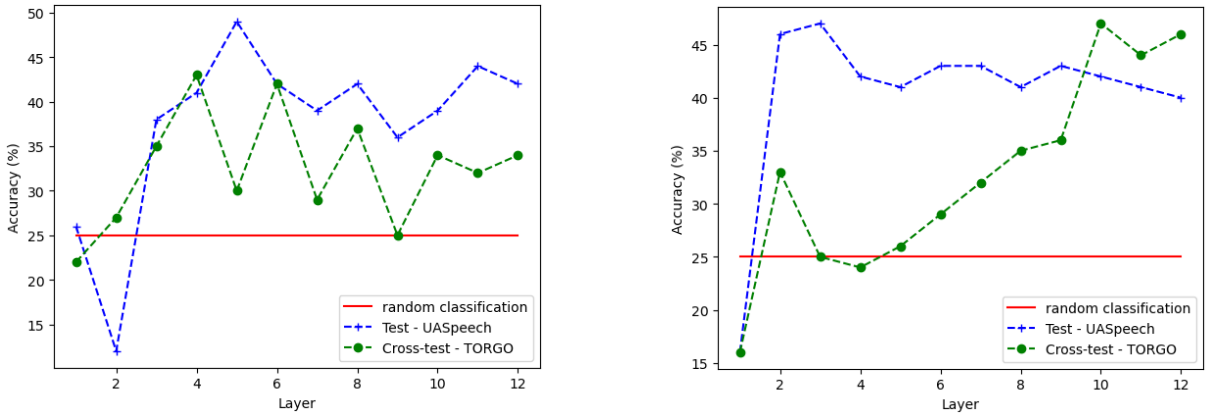
Figure 8.3: Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The models are used as feature extractors, without any fine-tuning.

full fine-tuning, and fine-tuning only the transformer component.
In the first scenario, we noticed that the performance on UASpeech (test accuracy) outperformed the performance on TORGO (cross-test accuracy), for both models. wav2vec2 slightly outperformed HuBERT on UASpeech, with an average test accuracy across layers of 37.92% vs 36.42% for HuBERT. As for TORGO, no comparable difference in performance between models was noticed, with the average cross-test accuracy across layers being 18.25% and 18.67% for wav2vec2 and HuBERT, respectively. But HuBERT did outperform wav2vec2 on the test to cross-test gap across layers, with an average gap of 17.75% vs 19.67% for wav2vec2. Notably, for both models, the cross-test accuracy failed to surpass the random classification accuracy of 25%, indicating that the models likely learned dataset-specific patterns rather than generalizable features. This limitation may arise from differences in recording environments and content between UASpeech and



(a) wav2vec2 (b) HuBERT

Figure 8.4: Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The models are fine-tuned to the task of Dysarthria severity classification.



(a) Wav2vec2 (b) HuBERT

Figure 8.5: Plots showing the variation of the test and cross-test accuracies, with respect to the layer from which the embeddings are used. The transformer part of the models is fine-tuned to the task of Dysarthria severity classification

TORGO.
However, when the models are used as feature extractors, i.e. with the pre-trained weights, we notice that test to cross-test gap diminishes considerably. For wav2vec2, the gap dropped to an average of 2.25%, while for HuBERT, it dropped to only 0.5%. This reduction reflects an improvement in cross-dataset generalization, as the average cross-test accuracy increased to 26.83% and 35.42% for wav2vec2 and HuBERT, respectively. HuBERT outperformed wav2vec2 in this scenario, indicating that its pre-trained weights allowed it to extract superior generalizable features. These findings emphasize the importance of pre-training on large datasets in enabling models to capture meaningful, generalizable speech representations. Despite this improvement, the models' performance on TORGO remains modest, suggesting potential room for further enhancements.

Table 8.2: A comparative table of the best results achieved by each model, and the average test accuracy, cross-test accuracy, and test to cross-test gap. Arch, FE and FT refer to architecture, feature extractor, and fine-tuned respectively, which describe the training mode of the model

Model	**Training Mode**	**Best Test Accuracy**	**Best Cross-test Accuracy**	**Average Test Accuracy**	**Average Cross-test Accuracy**	**Average test to cross-test gap**
Wav2vec2	Arch	45	21	37.92	18.25	19.76
	FE	41	37	29.08	26.83	2.25
	FT	48	42	39.67	29.08	10.58
	FT-part2	47	47	37.5	32.5	5
HuBERT	Arch	48	42	36.42	18.67	17.75
	FE	42	43	34.92	35.42	0.5
	FT	56	50	45.25	34	11.25
	FT-part2	49	43	40.42	32.75	7.67

Fine-tuning the models on UASpeech resulted in substantial improvements in test accuracy (UASpeech) and cross-test accuracy (TORGO), with HuBERT consistently outperforming wav2vec2 on both datasets. Specifically, the average test, cross-test accuracies and gap were 39.67%, 34%, and 10.58% for wav2vec2, and 45.25%, 40.42%, and 11.25% for HuBERT, respectively.

Even though this scenario did improve accuracy across both datasets, the gap between them actually increased in comparison to when the models were used as feature extractors, suggesting that the models became dataset-specific. Fine-tuning did help to enhance the models' performance for dysarthria severity classification, but it showed some form of limitation of generalization capabilities, which could be due to the fact that the dataset used for fine-tuning is relatively small compared to the size of the model. For this scenario also, HuBERT outperformed wav2vec2 across both datasets.

To compromise between the generalization capability and the performance enhancement, we explored the last case, where fine-tuning is restricted to the transformer layers. By doing so, the average test to cross-test accuracy went down to 5% and 7.67% for wav2vec2 and HuBERT respectively, suggesting improved generalization compared to full fine-tuning. However, this reduction came at the cost of slightly lower test and cross-test accuracies. For wav2vec2, the average accuracy decreased to 37.5%, while the cross-test accuracy slightly increased to 32.5%. For HuBERT, the average test accuracy dropped to 40.42%, and the cross-test accuracy decreased to 32.75%. These findings suggest that transformer-only fine-tuning may provide a viable trade-off between generalization and computational efficiency. Additionally, selecting embeddings from layers with minimal test to cross-test gaps could further optimize performance for cross-dataset applications.

Across all scenarios, HuBERT either almost matched, or outperformed wav2vec2 in both test and cross-test accuracy. This consistent trend suggests that HuBERT is better suited

for dysarthria severity classification, and cross-dataset generalization. HuBERT's superior performance may be attributed to differences in its pre-training objective, which enables it to capture more robust and transferable speech features.
The findings from this study provide valuable insights into the performance and generalization capabilities of wav2vec2 and HuBERT for dysarthria severity classification. Pre-training and fine-tuning strategies significantly influence the models' ability to generalize across datasets, with HuBERT demonstrating superior performance overall.

8.4 Spectrogram-based Models

8.4.1 Ablation study

An ablation study was conducted to study the impact of the pre-trained models on the training process. For the spectrogram-based models, i.e. Whisper and AST, three training modes were evaluated and their results were compared. We opted to focus on these three modes, while excluding the architecture mode, as these configurations align more closely with typical usage in transfer learning scenarios and are more likely to provide valuable insights into the advantages of fine-tuning pre-trained models for specific tasks. The modes are as follows:

1. **Feature Extractor (FE):** In this mode, the pre-trained models were used with their weights frozen. This configuration enables us to evaluate the impact of the pre-trained features on the classification task.

2. **Fine-tuning (FT):** In this mode, the pre-trained models were fine-tuned for the dysarthria severity classification task. This approach highlights the combined contributions of the pre-training and the fine-tuning process.

3. **Partial fine-tuning (Partial-FT):** In this mode, the pre-trained models were used with some layers' weights frozen, while fine-tuning some of the transformer layers for the dysarthria severity classification. This case was studied to address the challenge of training the large models on the relatively small dataset. In particular, the three partial-tuning cases were studied: fine-tuning the second half of the transformer layers (layers 7 to 12), fine-tuning the four last transformer layers (layers 9 to 12), and fine-tuning the last two layers (layers 11 and 12).

8.4.2 Datasets

We follow the same data partition as for the waveform-based models. We perform two levels of testing: in-database and cross-database (see section 7.3.5). UASpeech was used for training and in-database testing, whereas TORGO dataset was used for cross-database testing. As for the train/test partition, we followed the speaker-independent approach, where no audios from the same speaker are present in both train and test sets. The exact data distribution, in terms of participants per class, is provided in table 8.1. The training data was further divided in 90%/10% for training / validation respectively.

8.4.3 Classifier and Training Settings

The small configuration of Whisper was used, composed of 12 layers, and of hidden size 768. The AST also results in an input size of 768.
For dysarthria severity classification, we add a classification head to the pre-trained models, i.e. a feed forward neural network, with input size 768 and output size 4, was added to the models.
The models are trained using the Adam optimizer, with a learning rate of 0.0001. We use cross-entropy loss, a standard loss function for multi-class classification, as it effectively measures the difference between predicted and actual class probabilities.
We train the models for 50 epochs using a batch size of 64. Early stopping is applied to prevent overfitting, with training terminating if validation loss does not improve for 5 consecutive epochs.
As explained earlier, the pre-trained models were used in three different modes, so the training process was repeated for each mode, for both models.

8.4.4 Results and Discussion

Results of the experiments are reported in table 8.3. When the models were used as feature extractors (with the pre-trained weights frozen), Whisper performed similarly to AST on the intra-test (UASpeech), achieving accuracies of 42% and 41%, and macro F1-scores of 0.37 and 0.34, respectively. However, AST outperformed Whisper on the cross-test, TORGO, with an accuracy of 52% compared to 43% for Whisper.
When the models are fully fine-tuned, we noticed a slight decrease of 2% in terms of accuracy on the intra-test, against a 3% increase in accuracy for the inter-test with Whipser model. While with the AST, an accuracy increase of 2% and 3% was noticed on the intra-test and inter-test, respectively. These results show that the limited dataset does not allow for a full model fine-tuning.
As a compromise between using the models as feature extractors and fully fine-tuning, partial fine-tuning was performed. For Whisper, we noticed an increase in performance on both intra- and inter-test, for all three partial fine-tuning cases (fine-tuning last 6, 4, and 2 layers) compared to the FE and FT modes. In particular, the intra-test and inter-test increased to 46% and 64%, respectively, when the last two layers were fine-tuned, with F1-scores of 0.38 and 0.45 for intra-test and inter,-test respectively. Fine-tuning the last 4 layers maintained the performance on the intra-test for both the accuracy and f1-score, while the inter-test performance increased to 67% and 0.47 for accuracy and F1-score, respectively. Finally, with the last 6 layers fine-tuned, a drop of accuracy of 3% and 6% was noticed on both intra-test and inter-test respectively compared to when fine-tuning the last 4 layers, but it still performed better than FE and FT modes.
Conversely, a different trend was noticed for AST. Fine-tuning the last 6 layers achieved an accuracy increase of 2% on the intra-test vs and accuracy decrease of 1% on inter-test, respectively, compared to fully fine-tuning the model. While fine-tuning the last 4 and 2 layers resulted in an accuracy decrease of 3% and 7%, respectively, on the intra-test, compared to fully fine-tuning the model. As for the inter-test, fine-tuning the last 4 layers achieved the same accuracy as fully fine-tuning, while only fine-tuning the last 2 layers resulted in an accuracy increase of 4%.
These results showed that both pre-trained models, Whisper and AST, are able to extract

generalizable features, which was demonstrated by their ability to achieve comparable results on both the intra- and inter-test when used as feature-extractors. It also shows that fully fine-tuning the models requires more data than used in the current work. Partial fine-tuning did provide a good alternative, especially in the case of Whisper, where partial fine-tuning allowed increase in performance, especially on the cross-dataset. This improvement was less noticeable in the case of AST, where it even decreased in some cases. This shows that Whisper might be more suitable for training generalizable models for dysarthria severity classification, with limited data.

Table 8.3: A comparative table of the results achieved by each model. FE and FT refer to feature extractor, and fine-tuned respectively, which describe the training mode of the model

Model	**Training Mode**	**Test Accuracy (%)**	**Cross-test Accuracy(%)**
Whisper	FE	40	46
	FT	42	43
	Partial-FT L7-L12	43	61
	Partial-FT L9-L12	46	67
	Partial-FT L11-L12	46	64
AST	FE	41	52
	FT	43	55
	Partial-FT L7-L12	45	54
	Partial-FT L9-L12	40	55
	Partial-FT L11-L12	36	59

8.5 Comparison with Related Work

In this chapter, we proposed the use of pre-trained models for dysarthria severity classification. The models were tested and their performance was compared under different settings: architecture, feature extraction, and partial/total fine-tuning. Additionally, to test the generalizability and the robustness of the models, they were and tested using the SI data split, as well as using a cross-database testing. The best achieved accuracies were 56% and 67%, on the in-database and the cross-database, respectively.
In this section, we compare our experiments to some related work.
Compared to the articles discussed in section 6.4 ([79–82]), unfreezing the pre-trained models, and training a classification head on top of them instead of a lightweight SVM/MLP, preserve most of the core benefits discussed over the classic multi-stage pipelines remain: we still bypass MFCC, CQCC, TECC and other hand-tuned front-ends; the transformer's self-attention continues to model long-range timing and can down-weight speaker-specific frames, curbing the speaker-leakage that cripples i-vector and Fisher-vector systems; and adapting to a new clinic is still a single fine-tune, not a rebuild of Gaussian mixtures or filter banks. What changes is cost: fine-tuning exposes up to millions of parameters, so

training moves from minutes on a CPU to hours on a GPU, and requires stronger regularization on small corpora. Thus we surrender a slice of the compute/light-data advantage, but we keep the freedom from fragile feature design and multi-stage EM training while often gaining extra accuracy and robustness. In short, fine-tuning adds computational costs, but it does not re-introduce the fragile, hand-engineered stages that burden i-vector, Fisher-vector or TECC pipelines, and it often boosts accuracy and robustness. It retains the absence of feature-design effort, global self-attention, and speaker-agnostic weighting, while accepting higher GPU time and stricter regularization needs, a trade-off that is still simpler than the multi-stage handcrafted alternatives. All this while providing a competitive performing of 56% intra-test accuracy, in addition to a second level of testing, cross-dataset testing, with an accuracy of 67%.

As for the articles discussed in section 7.5, we saw that in [83], Vyas et al. proposed a system for dysarthria detection and severity classification, following the traditional approach. Using recordings from UASpeech dataset, they extract a rich set of time- and frequency-domain prosodic features, and use a genetic algorithm to select the best performing features to be used for dysarthria detection and severity classification. A SVM is trained using the selected features, using a 3-fold cross-validation method, the proposed model achieved a classification accuracy of 87%. Unlike our method, theirs was not validated following the SI data split, so its generalizability remains questionable. In addition, only 20 utterances were used to train the SVM. So same limitations discussed remain present: the model is vulnerable to over-fitting and offer no guarantee it will scale to new speakers, dialects or languages; the genetic algorithm can converge to different feature subsets when the training set changes slightly, which is risky with such few training utterances.

In [85], Javanmardi et al. study the use of three pre-trained models, wav2vec2-Base, wav2vec2-Large, and HuBERT for dysarthria detection and severity classification. They used the models as features extractors, and trained a SVM and a CNN network for dysarthria detection and severity classification. They trained the models using UASpeech and achieved a classification accuracy of 48.01% using HuBERT and the CNN. They also trained a model using TORGO and achieved an accuracy of 49.83%, using HuBERT and the CNN. Our work surpasses theirs, since we test the models' performance on a cross-dataset, in addition, we study how fine-tuning the models affect the performance. Our work outperforms theirs, with an intra-test accuracy of 56% and a cross-test accuracy of 67%.

In [87], Stumpf et al. further exploit the use of pre-trained models, setting a new record for SI dysarthria severity classification. In their article, publish in 2024, they used multi-task learning objective and contrastive learning to perform dysarthria classification. Using wav2vec2-Base pre-trained model, they fine-tuned it to perform dysarthria severity classification, in addition to an auxiliary task, a contrastive task, aiming to separate the word embeddings of speakers from the same severity level. They trained the model on the UASpeech dataset, and achieved an accuracy of 70.48% outperforming the previous benchmark of AI-based dysarthria classification models, including our results.

In summary, end-to-end fine-tuning of pre-trained models eliminates hand-crafted feature design and surpasses earlier fixed-embedding pipelines, though it still trails the latest multitask systems. The trade-off is higher computational cost and, like all prior work, evaluation is restricted to English; expanding to other languages and spontaneous speech

remains future work.

8.6 Conclusion

In this chapter, we proposed to use the pre-trained models to mitigate the problem of data scarcity in speech pathology domain. To this end, we used four different SOTA pre-trained models for dysarthria severity classification, wav2vec2, HuBERT, Whisper, and AST, spanning both waveform-based speech models and spectrogram models, in addition to both supervised and self-supervised pre-training approaches.
All four models were evaluated, following three or four training modes: architecture-only, feature extractor, partially fine-tuned, and fully fine-tuned. The training was performed following a speaker-independent data partition. Results demonstrated that the models were able to extract robust, generalizable features, yielding competitive performance across both in-database and cross-database evaluations.
Despite these improvements compared to chapter 6, where no pre-training was employed for the classifier, the overall accuracy remains suboptimal. Further work is needed to enhance performance. Although the use of pre-trained models reduces the amount of data required compared to training from scratch, collecting additional data would still help boost accuracy.
Moreover, incorporating pathological speech data in the pre-training phase could further enhance model performance. By incorporating pathological speech data into the pre-training phase, models could learn to better represent the unique acoustic and prosodic characteristics of disordered speech. This domain adaptation would allow the model to become more sensitive to the variability and distinct features present in pathological speech, potentially improving its ability to generalize across different types and severities of disorders.

Chapter 9

Parkinson's Disease and Dysarthria: Adapting and Expanding Experiments on VocaPark

9.1 Introduction

Throughout this thesis, it has become evident that data plays a pivotal role in developing automated deep learning models for dysarthria severity classification. The availability of high-quality, representative datasets is crucial for training robust models that can generalize effectively to real-world clinical scenarios. However, existing datasets for dysarthria and speech impairments are often limited in size, linguistic diversity, and recording conditions, restricting the progress of automated speech pathology research.
To address this gap and contribute to advancements in speech pathology, we collaborated with Fondation Hopale to collect speech data from French-speaking Parkinson's patients. This dataset serves as a valuable resource for the community, enabling the study of dysarthria in the French language, a domain that has received relatively less attention compared to English. A comprehensive description of the data collection process, including participant recruitment, recording protocols, and annotation procedures, is provided in Chapter 5.
Beyond its primary role in dysarthria severity classification, this dataset can play a role in broader applications in speech-related research tailored for impaired speech. Potential applications include automatic speech recognition (ASR) systems designed to improve accessibility for individuals with speech impairments, speaker adaptation techniques that enhance the performance of voice-assisted technologies for dysarthric speakers, etc. .
Additionally, the dataset can support the development of personalized speech therapy systems, leveraging AI to provide real-time feedback and adaptive exercises for patients undergoing rehabilitation.
By introducing this dataset, we aim to bridge the gap between machine learning and clinical speech pathology, fostering interdisciplinary research that benefits both technological innovation and patient care.
In this chapter, we present the initial experiments conducted using this dataset.
In this chapter, we present experiments focusing on pre-trained models applied to dysarthria severity classification in VocaPark dataset. Specifically, we explore the performance of pre-trained models, wav2vec2 and AST, examining their effectiveness in two different

modes, as feature extractors, and partially fine-tuned. These experiments are compared to traditional baseline features, including MFCC and eGeMAPS, with a particular focus on how these pre-trained models can improve classification performance. The rest of the chapter is organized as follows: we start with some baseline experiments in section 9.2, aimed to establish a reference for performance evaluation. We then proceed in section 9.3 to adapt some of the experiments previously performed and demonstrated in the chapters 8 on VocaPark. And finally, section 9.4 concludes the chapter.

9.2 Baseline Experiments

As baseline experiments, we train a model following the traditional approach, with MFCC and eGeMAPS as features, and a feed-forward NN as classifier. We train the models following both the speaker-dependent approach and the speaker-independent approach data splits. Further details are provided in the following subsections.

9.2.1 Feature Extraction

Two feature sets were used in this experiment: MFCC and eGeMAPS.

- **MFCCs:** In this work, we extracted 13 coefficients from each audio, along with their first and second derivatives. This resulted in a feature set of shape (n,39), where n is the number of audios in a batch.
- **eGeMAPS:** This sets results in a total of 88 features. The resulting feature set has a shape of (n,88).

9.2.2 Classifier

The model used for classification consisted of three fully connected linear layers:

1. **First layer:** Maps the input to a hidden layer of size equal to the half of the input size (19 for MFCC, and 44 for eGeMAPS), followed by a ReLU activation.
2. **Second layer:** Another fully connected layer maintaining the hidden size, followed by another ReLU activation.
3. **Final layer:** Projects the hidden representation to the output size, which corresponds to the severity levels in our datasets, i.e. four classes in this case (No dysarthria, low, medium, and high severity).

9.2.3 Data Partition

For the baseline experiments, we performed both speaker-dependent approach and speaker-independent approach data partitions. For the SD approach, audios from all participants are mixed, and a percentage is used for training, while the rest is used for testing. As for the SI approach, the training/testing data partition is performed according to the participants, i.e. audios from the same participant can not be present in both the train and test

sets. The exact data distribution is shown in tables 9.2 and 9.1 for the speaker-dependent and speaker-independent cases, respectively.

Severity	Train - nb of audios (nb of participants)	Test - nb of audios (nb of participants)
No dysarthria	1505 (10)	8056 (47)
Low	1530 (10)	4141 (24)
Medium	1508 (10)	689 (4)
High	1153 (8)	170 (1)

Table 9.1: Data distribution following the speaker-independent scenario.

Severity	Train - nb of audios	Test - nb of audios
No dysarthria	**1451**	**8318**
Low	**1447**	**4401**
Medium	**1445**	**886**
High	**1161**	**288**

Table 9.2: Data distribution following the speaker-dependent scenario.

9.2.4 Training Settings

The models are trained using the Adam optimizer, which was chosen for its adaptive learning rate properties and ability to converge efficiently in deep learning tasks. We use cross-entropy loss, a standard loss function for multi-class classification, as it effectively measures the difference between predicted and actual class probabilities. A learning rate of 0.001 was used.
We train the models for 50 epochs using a batch size of 64. Early stopping is applied to prevent overfitting, with training terminating if validation loss does not improve for 10 consecutive epochs. The validation was performed using 10% of the training data.

9.2.5 Results and Discussion

9.2.5.1 Speaker-dependent scenario

The results of the baseline experiments for the speaker-dependent scenario are presented in table 9.3. In addition, we visualize the corresponding confusion matrices in figures 9.1a and 9.1b for eGeMAPS and MFCCs respectively.
The results reveal interesting insights into the performance of the model. The overall accuracy highlights a significant difference between the two feature sets, with MFCCs significantly outperforming eGeMAPS (49% vs. 28%). This suggests that MFCCs may provide more informative or discriminative features for the classification task in this specific context.
Delving into the per-class accuracy further strengthens this observation. For eGeMAPS,

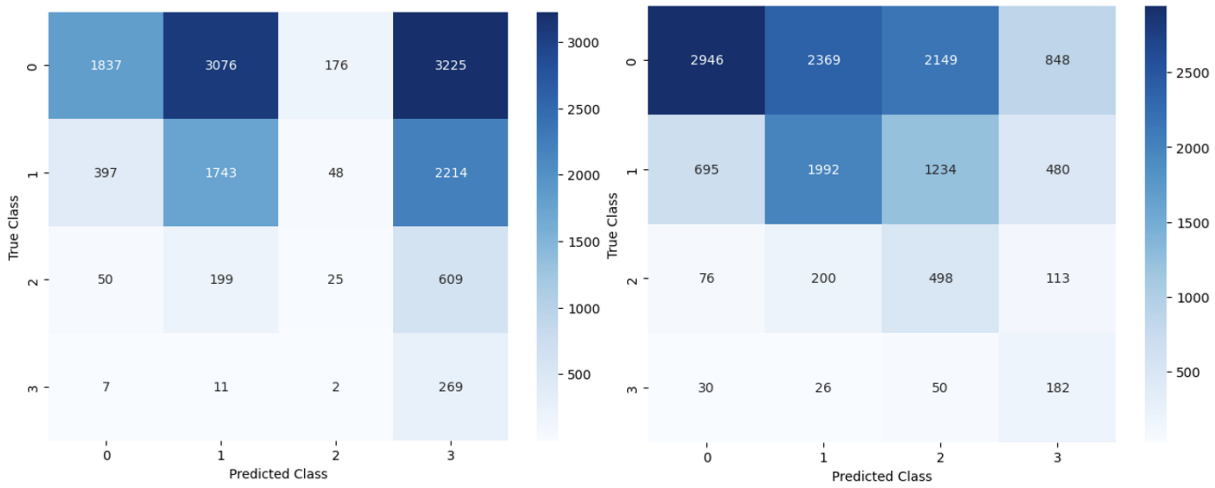
the model exhibits highly uneven performance across classes, ranging from a low of 3% for class 2 to a high of 93% for class 3. Such inconsistency indicates that while eGeMAPS may capture features relevant for certain classes (especially class 3), it fails to generalize well across others. This imbalance might result from the nature of eGeMAPS, which includes a wide range of low-level descriptors tailored more toward general affective or prosodic characteristics, which may not align well with the specific discriminative patterns needed for some classes in this task.
In contrast, MFCCs demonstrate a more balanced distribution of accuracy across all classes: 35%, 45%, 56%, and 63% for classes 0 through 3. Although not perfect, these results suggest that MFCCs enable the model to generalize more evenly across the class spectrum. MFCCs primarily capture the spectral envelope of the audio signal, which may carry more consistent cues useful for distinguishing between the different classes.
The confusion matrices (Figures 8.1a and 8.1b) would reflect these trends, particularly showing greater confusion in the eGeMAPS case, especially between classes where performance was poor, such as 0 and 2. In contrast, the MFCC-based confusion matrix shows more class separability, although still with some misclassifications.
Overall, these results imply that MFCCs are a more suitable feature representation for this speaker-dependent classification task, at least in the current experimental setup. However, the high performance of eGeMAPS for class 3 also suggests that there could be complementary information in these features that might be leveraged in future work, for instance through feature fusion or ensemble learning approaches.

Features	**Accuracy**	**F1-score**	**Precision**	**Recall**	**Class**	**Per-class accuracy**
eGeMAPS	28	0.3	0.3	0.6	0	0.22
					1	0.4
					2	0.03
					3	0.93
MFCCs	49	0.5	0.6	0.4	0	0.35
					1	0.45
					2	0.56
					3	0.63

Table 9.3: Baseline results for the speaker-dependent scenario.

9.2.5.2 Speaker-independent scenario

The results of the baseline experiments for the speaker-independent scenario are presented in table 9.4. In addition, we visualize the corresponding confusion matrices in figures 9.2a and 9.2b for eGeMAPS and MFCCs respectively.
In contrast to the speaker-dependent scenario, the overall accuracy using eGeMAPS unexpectedly outperforms MFCCs in this speaker-independent scenario. This finding suggests that eGeMAPS may be more effective at capturing speaker-invariant features, which could explain its improved generalization across different speakers.
Delving into the per-class accuracy, it becomes evident that this increase in performance



(a) eGeMAPS (b) MFCCs

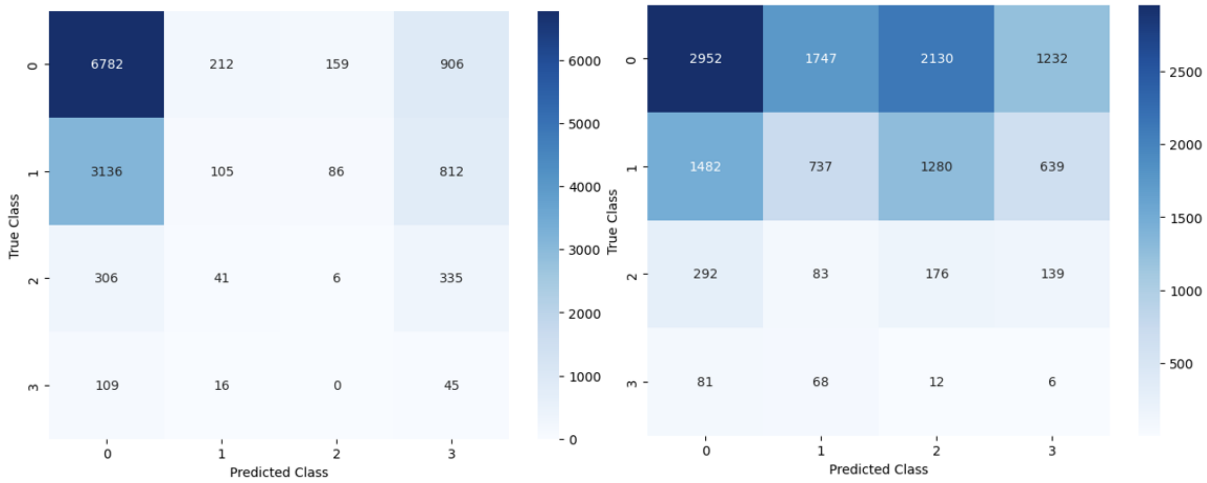
Figure 9.1: The confusion matrices of the baseline results following the speaker-dependent data-partition, for (a) eGeMAPS and (b) MFCCs. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

with eGeMAPS is primarily driven by its high accuracy in classifying class 0, which also contains the highest number of audio samples. eGeMAPS achieves 84% accuracy for class 0, but the same discrepancy in class-wise performance observed in the previous scenario persists here. The accuracy sharply drops for other classes, reaching as low as 1% for class 2. This indicates that while eGeMAPS generalizes well for the most represented class, it still struggles to provide sufficient discriminative information for the less frequent or more complex classes.
On the other hand, MFCCs demonstrate a more balanced but overall lower distribution of per-class accuracies: 37%, 18%, 26%, and 4% for classes 0 through 3 respectively. Although this suggests that MFCCs maintain a more even classification performance across classes, the overall accuracy has significantly dropped compared to the speaker-dependent case. This decline aligns with our expectations, as the speaker-independent scenario introduces greater variability between training and test data, making the classification task more challenging.
These results confirm that speaker-independent classification of dysarthria severity is a more difficult setup, with a notable impact on model performance for both feature types. While eGeMAPS shows better overall results due to strong performance on the most frequent class, MFCCs exhibit more consistent, though lower, performance across all classes. This highlights the importance of exploring methods that can better handle class imbalance and speaker variability in future work.

The results of these baseline experiments in both the speaker-dependent and speaker-independent scenarios were obtained to establish a point of reference for the subsequent experiments. These baselines serve as a foundation for evaluating the effectiveness of the approaches introduced throughout the remainder of this chapter.

Features	Accuracy	F1-score	Precision	Recall	Class	Per-class accuracy
eGeMAPS	53	0.47	0.5	0.53	0	0.84
					1	0.03
					2	0.01
					3	0.27
MFCCs	30	0.36	0.47	0.3	0	0.37
					1	0.18
					2	0.26
					3	0.04

Table 9.4: Baseline results for the speaker-independent scenario.



(a) eGeMAPS (b) MFCCs

Figure 9.2: The confusion matrices of the baseline results following the speaker-independent data-partition, for (a) eGeMAPS and (b) MFCCs. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

9.3 Leveraging Pre-trained Models

In the following section, we will be extending some of the experiments performed in chapter 7 to the VocaPark dataset. This will allow us on one hand to further challenge the performance the pre-trained models. This is because the experimentation is carried out on a dataset with patients with dysarthria from another etiology (Parkinson in this case), more patients meaning more diversity in the pathology markers, and in a different language (French). And on the other hand, it present an alternative, possibly an improvement over the baseline experiments.

We limit the experiments on two pre-trained models: waveform-based wav2vec2 model, and spectrogram-based AST model. We chose these two models, because preliminary experiments demonstrated they give better results compared to the other model of the same category. We train most of the models following both the speaker-dependent and the

speaker-independent data splits. Further details are provided in the following subsections.

9.3.1 Pre-trained Models

Two pre-trained models were used in the experiments of this section:

1. Wav2vec2 (cf. section 3.6.1): The Base configuration was used, with the average calculated along the time dimension. This resulted in a feature set of dimension (n,768), where n corresponds to the batch size.

2. AST (cf. section 3.6.4): This model resulted in a feature set of dimension (n,768).

In Chapter 8, results showed that partial fine-tuning gave the best trade-off between performance and training cost. Inspired by these results, experiments were restricted to two training modes, and their results were compared. The modes are as follows:

1. **Feature Extractor (FE):** In this mode, the pre-trained models were used with their weights frozen.

2. **Partial fine-tuning (Partial-FT):** In this mode, the pre-trained models were used with some layers' weights frozen, while fine-tuning some of the transformer layers for the dysarthria severity classification.

9.3.2 Data Partition

For most of the experiments, we perform both speaker-dependent and speaker-independent data partition. The exact data distribution is shown in tables 9.2 and 9.1 for the speaker-dependent and speaker-independent case respectively.

9.3.3 Classifier and Training Settings

A classification head is added to the pre-trained models, i.e. a feed forward neural network, with input size 768 and output size 4.
The models are trained using the Adam optimizer, which was chosen for its adaptive learning rate properties and ability to converge efficiently in deep learning tasks. We use cross-entropy loss, a standard loss function for multi-class classification, as it effectively measures the difference between predicted and actual class probabilities. A learning rate of 0.001 was used.
We train the models for 50 epochs using a batch size of 64. Early stopping is applied to prevent overfitting, with training terminating if validation loss does not improve for 10 consecutive epochs. The validation was performed using 10% of the training data.

9.3.4 Results and Discussion

9.3.4.1 wav2vec2

Speaker-dependent scenario
Results of dysarthria severity classification using wav2vec2 model, with a speaker-dependent data partition are provided in table 9.5. In addition, the confusion matrices for each

training mode are provided in figure 9.3. Each row in the table corresponds to a different fine-tuning configuration, with the associated overall accuracy, F1-score, precision, recall, and per-class accuracy listed.
Results showed a clear trend: the less layers are trained, the better is the performance of the fine-tuned model. Specifically, when layers 8 to 12 are fine-tuned, the overall accuracy is 49%, with a notable variation in per-class performance, with accuracies ranging from 14% for class 1 to 83% for class 3. The overall accuracy increases when the number of fine-tuned layers is reduced, while the discrepancy between the per-class accuracies decreases. This variation continue until achieving a maximum overall accuracy of 72% when training the final couple layers, coupled with a per-class accuracy ranging between 68% for class 2 and 89% for class 3.
Conversely, this improvement is not strictly linear. A decrease of performance is noticed when only the last layer is fine-tuned. This mode resulted in an overall accuracy of 58%, and per-class accuracies ranging from 45% for class 1, and 82% for class 3. This drop suggests that fine-tuning a single layer may not be sufficient to adapt the model to the specifics of the dysarthria severity classification task, despite retaining some of its generalization capabilities.
These results highlight the importance of choosing the best number of layers to fine-tune in order to achieve the most powerful performance.
These results also indicate that optimal performance is achieved when fine-tuning a small portion of the upper layers of the wav2vec2 model. Fine-tuning too many layers leads to a decrease in performance, this is likely due to overfitting on the speaker-specific data. Conversely, fine-tuning only the final layer limits the model's ability to adapt, resulting in suboptimal performance.
Compared to the baseline results using traditional acoustic features in the speaker-dependent scenario, wav2vec2 shows a substantial improvement in overall accuracy, reaching up to 72% when only the top two layers are fine-tuned, as opposed to the 49% achieved with MFCCs and 28% with eGeMAPS. Additionally, the per-class accuracy is more balanced with wav2vec2, showing less variation across classes. These results demonstrate the advantage of leveraging pre-trained self-supervised models over conventional handcrafted features, particularly when fine-tuned in a controlled manner.

Speaker-independent scenario

Results for the speaker-independent scenario using the wav2vec2 model are provided in table 9.6, and the confusion matrices for each training mode are provided in figure 9.4. Each row in the table corresponds to a different fine-tuning configuration, with the associated overall accuracy, F1-score, precision, recall, and per-class accuracy listed. The results also include a feature extraction (FE) baseline, where the model is used as a frozen feature extractor without any fine-tuning.
When using wav2vec2 as a frozen feature extractor, the model achieves an accuracy of 47%, with a balanced F1-score of 0.50. The per-class accuracies, however, remain uneven, ranging from 4% for class 2 to 54% for class 0. This pattern indicates that, although the model retains general discriminative capacity, its performance is still strongly affected by class imbalance.
Fine-tuning the upper layers (L8–L12) results in a drop in performance, with accuracy decreasing to 36% and per-class scores ranging from 12% to 58%. This trend continues

Mode	Accuracy	F1-score	Precision	Recall	Class	Per-class accuracy
FT L8-L12	49	0.50	0.60	0.49	0 1 2 3	66 14 47 83
FT L9-L12	55	0.58	0.71	0.55	0 1 2 3	52 52 81 72
FT L10-L12	62	0.64	0.73	0.62	0 1 2 3	59 64 75 74
FT L11-L12	72	0.74	0.76	0.72	0 1 2 3	72 73 68 89
FT L12	58	0.61	0.71	0.58	0 1 2 3	61 45 81 82

Table 9.5: Results using wav2vec2 model, for the speaker-dependent scenario. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

when less layers are included in the fine-tuning. Specifically, fine-tuning layers L9–L12 yields the lowest overall accuracy of 24%, despite a relatively high 76% accuracy for class 3. This setting shows the greatest imbalance in class-wise performance, with class 0 dropping to 20%.

Performance begins to recover when fewer layers are fine-tuned. For example, training layers L10–L12 leads to a moderate improvement, with an accuracy of 39% and a more stable per-class distribution. The best results are achieved when only layers L11–L12 are fine-tuned. This configuration reaches an overall accuracy of 47%, matching the FE baseline, but with slightly improved balance in per-class accuracies, ranging from 30% to 54%. Similarly, fine-tuning only the final layer (L12) achieves a competitive accuracy of 42%, with class-wise scores showing improved consistency.

These results confirm the same trend observed in the speaker-dependent scenario: fine-tuning too many layers harms overall performance and increases class-wise variability. In the speaker-independent setup, this effect is even more noticeable due to the added



(a) FT L8-L12.

(b) FT L9-L12.

(c) FT L10-L12.

(d) FT L11-L12.

(e) FT L12.

Figure 9.3: The confusions matrices of the predictions using the wav2vec2 model, using a speaker-dependent data partition. FE refers to feature extractor, and FT Li-Lj refer to fine tuned from layer i to layer j. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

complexity of generalizing across speakers. The best performance is again achieved when the model is either used as a frozen feature extractor or only the topmost layers are fine-

tuned. This further supports the idea that the pretrained representations of wav2vec2 are highly transferable and require minimal adaptation for the task of dysarthria severity classification across speakers.

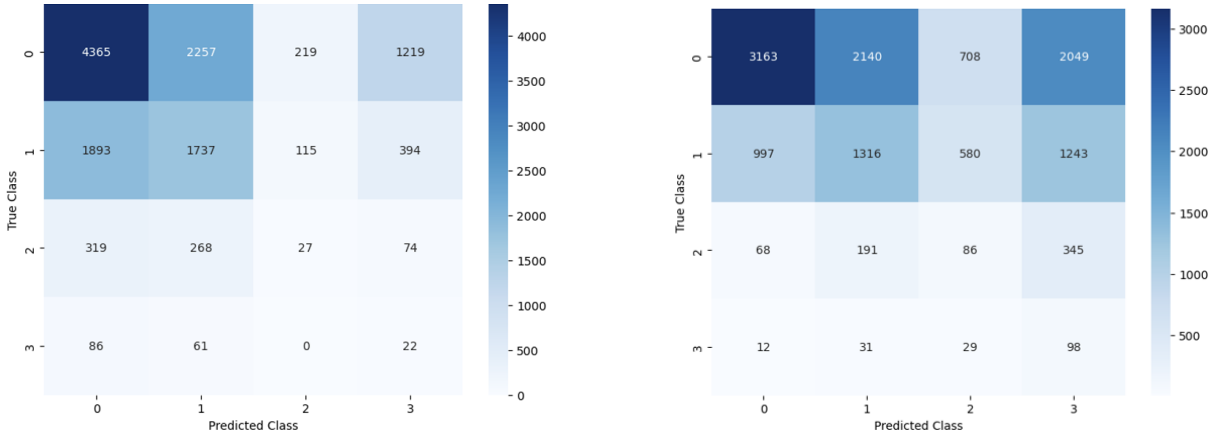
Mode	Accuracy	F1-score	Precision	Recall	Class	Per-class accuracy
FE	47	0.50	0.54	0.47	0	54
					1	42
					2	4
					3	13
FT L8-L12	36	0.43	0.58	0.36	0	39
					1	32
					2	12
					3	58
FT L9-L12	24	0.30	0.54	0.24	0	20
					1	27
					2	30
					3	76
FT L10-L12	39	0.43	0.50	0.39	0	45
					1	28
					2	28
					3	33
FT L11-L12	47	0.51	0.59	0.47	0	52
					1	38
					2	30
					3	54
FT L12	42	0.46	0.54	0.42	0	54
					1	18
					2	37
					3	47

Table 9.6: Results using wav2vec2 model, for the speaker-independent scenario. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

9.3.4.2 AST

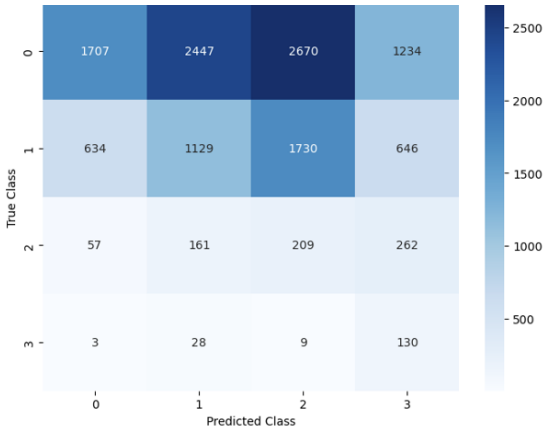
Speaker-Dependent Scenario

Results for the speaker-dependent scenario using the AST model as a pretrained backbone are presented in table 9.7. In addition, the corresponding confusion matrices are provided

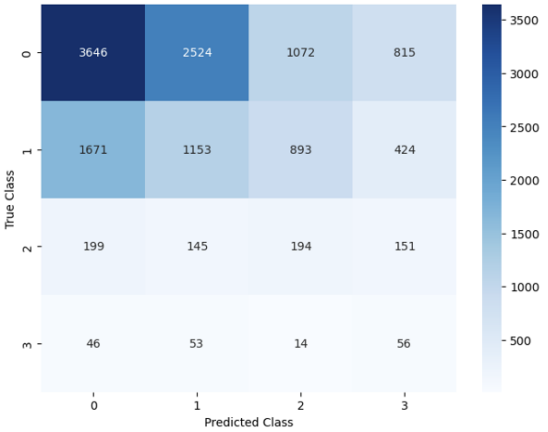


(a) FE.

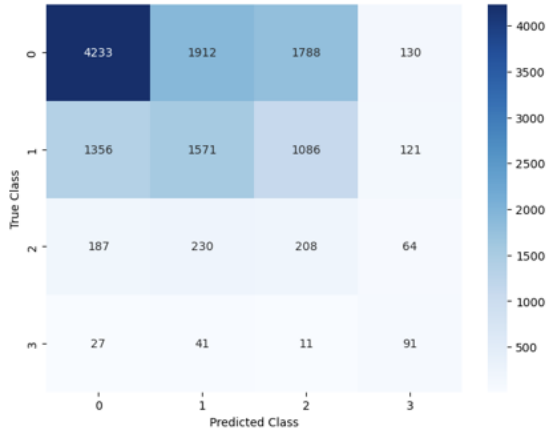
(b) FT L8-L12.



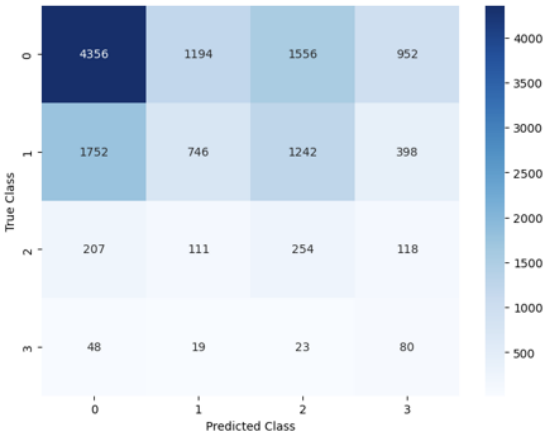
(c) FT L9-L12.



(d) FT L10-L12.



(e) FT L11-L12.



(f) FT L12.

Figure 9.4: The confusions matrices of the predictions using the wav2vec2 model, using a speaker-independent data partition. FE refers to feature extractor, and FT Li-Lj refer to fine tuned from layer i to layer j.

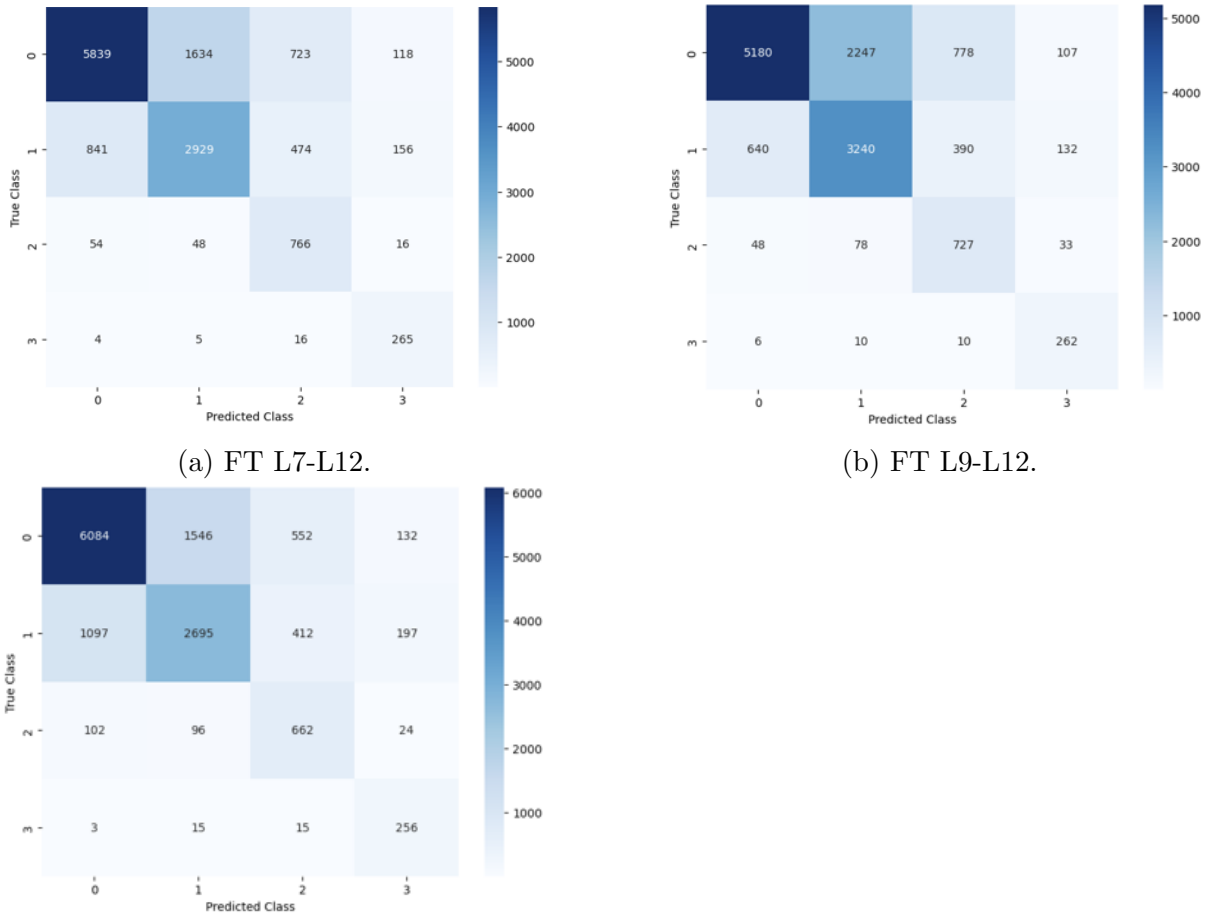
in figure 9.5. Three fine-tuning configurations were evaluated, each involving different subsets of upper transformer layers. The results include overall accuracy, F1-score, precision, recall, and per-class accuracies across the four severity classes.

Fine-tuning the top six layers (L7–L12) yields the highest overall accuracy of 71%, accompanied by an F1-score of 0.72. The per-class accuracies are well-balanced, ranging from 67% for class 1 to 91% for class 3. These results suggest that this configuration provides a good trade-off between adapting to the task and preserving the general representations learned during pretraining.
When only layers L9–L12 are fine-tuned, a slight drop in overall performance is observed. The accuracy decreases to 68%, with an F1-score of 0.69. Interestingly, this configuration achieves higher per-class accuracy for class 1 (74%) and class 2 (82%), while maintaining strong performance for class 3 (91%). However, it underperforms slightly in class 0 (62%), which brings down the overall average.
Fine-tuning only the final two layers (L11–L12) results in an overall accuracy of 70%, with an F1-score of 0.71. This setup delivers more consistent per-class accuracy, ranging from 61% to 86%, with class 3 again showing the highest performance. The relatively stable results across all classes indicate that even minimal fine-tuning of the AST model can yield strong performance in a speaker-dependent setting.
Overall, these findings show that the AST model is highly effective for dysarthria severity classification when fine-tuned on a speaker-dependent partition. In contrast to wav2vec2, AST demonstrates greater robustness to the number of fine-tuned layers, with all three configurations achieving strong and relatively balanced results. The best trade-off between accuracy and class-wise consistency appears when fine-tuning layers L7–L12.

Mode	**Accuracy**	**F1-score**	**Precision**	**Recall**	**Class**	**Per-class accuracy**
FT L7-L12	71	0.72	0.75	0.71	0	70
					1	67
					2	87
					3	91
FT L9-L12	68	0.69	0.75	0.68	0	62
					1	74
					2	82
					3	91
FT L11-L12	70	0.71	0.73	0.70	0	73
					1	61
					2	75
					3	86

Table 9.7: Results for the speaker-dependent scenario, using AST as pre-trained model. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

Speaker-Independent Scenario
Results for the speaker-independent scenario using AST as a pretrained model are pre-



(a) FT L7-L12.

(b) FT L9-L12.

(c) FT L11-L12.

Figure 9.5: The confusions matrices of the predictions using the AST model, using a speaker-dependent data partition. FT Li-Lj refer to fine tuned from layer i to layer j.

sented in table 9.8. In addition, the confusion matrices for each training mode are provided in figure 9.6.

The table reports results for a frozen feature extractor setup (FE), as well as for three fine-tuning configurations involving different upper-layer subsets. The performance is evaluated in terms of overall accuracy, F1-score, precision, recall, and per-class accuracy. When used as a frozen feature extractor, AST achieves the highest overall accuracy of 39%, with an F1-score of 0.45. Despite the modest overall performance, class-wise accuracies remain notably imbalanced, with class 3 reaching only 5%, while class 0 achieves 45%. This indicates that the model struggles to generalize across speaker identities without fine-tuning.

Fine-tuning layers L7–L12 leads to a drop in performance, with overall accuracy decreasing to 32% and a lower F1-score of 0.36. Per-class accuracy remains skewed, with class 3 again underperforming at 26%, while class 0 slightly improves to 24%. A further drop is observed when fine-tuning only layers L9–L12, resulting in the lowest overall accuracy of 28%, and a drastic variation across classes, from 25% for class 0 to 57% for class 2.

Fine-tuning just the top two layers (L11–L12) recovers some performance, achieving an accuracy of 34% and a slightly more stable class-wise distribution. However, class 3 re-

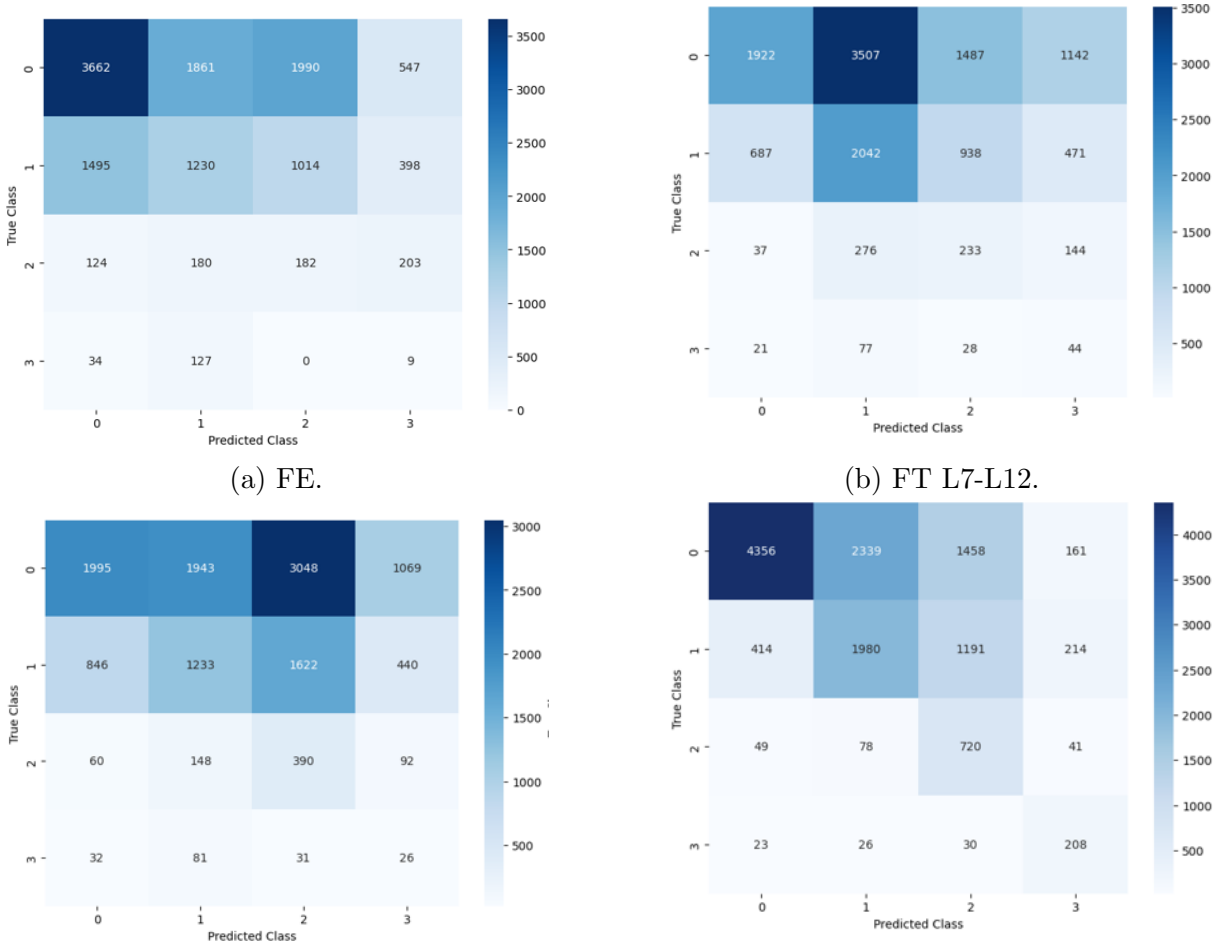
mains consistently low at 15%, indicating that AST struggles to model certain severity levels in speaker-independent conditions, regardless of fine-tuning depth.
These results confirm that the speaker-independent setting remains challenging for AST, with lower performance and noticeable per-class variability. While the model performs reasonably when used as a feature extractor, the benefit of fine-tuning remains limited. In contrast to wav2vec2, AST appears to require further task-specific adaptation or regularization strategies to reach comparable robustness across speakers.

Mode	**Accuracy**	**F1-score**	**Precision**	**Recall**	**Class**	**Per-class accuracy**
FE	39	0.45	0.54	0.39	0	45
					1	30
					2	26
					3	5
FT L7-L12	32	0.36	0.56	0.32	0	24
					1	49
					2	34
					3	26
FT L9-L12	28	0.34	0.54	0.28	0	25
					1	30
					2	57
					3	15
FT L11-L12	34	0.39	0.52	0.34	0	36
					1	28
					2	42
					3	15

Table 9.8: Results for the speaker-independent scenario, using AST as pre-trained model. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

9.4 Conclusion

In this chapter, we conducted different experiments on our newly collected french Parkinson dataset. We started by establishing a baseline, using traditional features, i.e. MFCCs and eGeMAPS, in combination with a neural network. Experiments were carried out following both speaker-dependent and speaker-independent data partition approaches.
Once the baseline was established, we explored the use of pre-trained models for dysarthria classification. Both wav2vec2 and AST were studied, in two different training setup, as feature extractors and when partially fine-tuned. Experiments were also carried out following both speaker-dependent and speaker-independent approaches. Results revealed



(a) FE.

(b) FT L7-L12.

(c) FT L9-L12.

(d) FT L11-L12.

Figure 9.6: The confusions matrices of the predictions using the AST model, using a speaker-independent data partition. FE refers to feature extractor, and FT Li-Lj refer to fine tuned from layer i to layer j.

that for both data partitions, the pre-trained models outperformed the baseline features, especially in terms of per-class accuracy, with wav2vec2 in advance.
Overall, these findings confirms the findings of Chapter 7 and 8, using another language, on the potential of leveraging the use of robust feature extraction and pre-training strategies for dysarthria classification. While speaker-dependent setup achieves significantly higher accuracies, the speaker-independent scenario remains more challenging due to greater variability across speakers.

Chapter 10

Exploring Alternative Approaches: Contrastive Pretraining and Autoencoders

10.1 Introduction

While in the previous chapter we adapted the experiments with pre-trained models, this chapter focuses on adapting and further improving the experiments from Chapter 6. Specifically, we refine the transformer-based model through contrastive pretraining and investigate the use of an autoencoder for feature learning in dysarthric speech.
In Chapter 6, we introduced a transformer-based model for dysarthria severity classification. In this chapter, we aim to enhance its performance by applying contrastive pretraining, a self-supervised technique designed to help the model learn more discriminative features by distinguishing between similar and dissimilar samples. This approach is expected to improve the transformer model's robustness, particularly in the context of dysarthria severity classification. Additionally, we explore the use of an autoencoder, a separate experiment focused on unsupervised learning, which learns efficient representations of speech data that can be leveraged for downstream tasks such as classification.
This chapter is organized as follows: Section 10.2 discusses the transformer-based model, and its enhancement using contrastive pretraining. Section 10.3 introduces the autoencoder experiment, detailing its architecture and results. Finally, Section 10.5 concludes the chapter.

10.2 Transformer-Based Model

In this section, we expand the experiments carried out in chapter 6 on the VocaPark dataset. Additionally, we work on improving the results by the use of contrastive pretraining. More details are provided in the following subsections.

10.2.1 Using the Transformer-Based Model

The transformer-based model used in the following experiments follows the structure presented in chapter 6.

10.2.1.1 Data Partition

We perform both speaker-dependent and speaker-independent data partition. The exact data distribution is shown in tables 9.2 and 9.1 for the speaker-dependent and speaker-independent case respectively.

10.2.1.2 Feature Extraction and Classifier

We used the same features and classifier presented in chapter 6. The detailed information could be found in section 6.2.

10.2.1.3 Training Settings

The same training settings presented in chapter 6 were used. Details could be found in section 6.2.

10.2.1.4 Results and Discussion

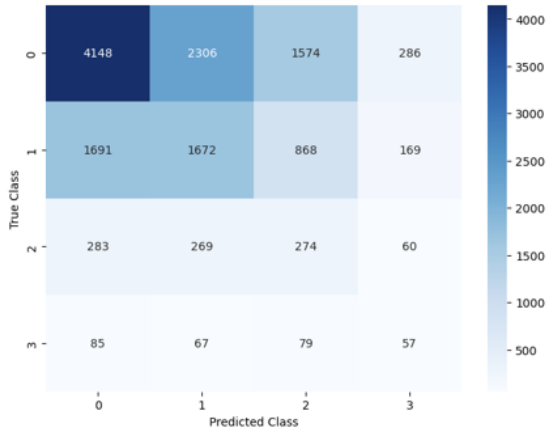
Speaker-Dependent Scenario
Results for the speaker-dependent scenario using the transformer-based model are shown in table 10.1. In addition, we visualize the corresponding confusion matrices in figure 10.1.
Three feature configurations, eGeMAPS, MFCCs, and W2v2-CNN, were paired with a varying number of encoder layers. The table reports the overall accuracy, F1-score, precision, recall, and per-class accuracy.
When eGeMAPS are used (with a 2-layer encoder), the model achieves an accuracy of 44%. Class-wise performance remains unbalanced, ranging from 50% for class 0 to just 20% for class 3. This relatively modest result suggests that two layers may not be sufficient to capture the relevant nuances of prosodic features for dysarthria severity classification.
In contrast, MFCCs (also with a 2-layer encoder) provide the strongest overall performance. The model reaches an accuracy of 73% and an F1-score of 0.74, along with balanced per-class accuracies from 69% to 86%. This indicates that two encoder layers are adequately leveraging the spectral details in MFCCs for effective severity classification.
With the W2v2-CNN features, the number of encoder layers increases to three. Despite slightly lower overall accuracy (71%) compared to MFCCs, the model maintains a strong F1-score of 0.72 and achieves particularly high performance of 95% for class 3. The additional encoder layer in this setup may help capture more detailed representations for some classes while slightly reducing overall accuracy relative to MFCCs.
Overall, these findings show that MFCCs combined with a 2-layer encoder produce the best trade-off between accuracy and class balance, outperforming both eGeMAPS and W2v2-CNN. They also highlight that, while increasing the number of encoder layers can benefit certain classes, it does not guarantee uniformly superior performance across the board.

Speaker-Independent Scenario
Results for the speaker-independent scenario using the transformer-based model are pre-

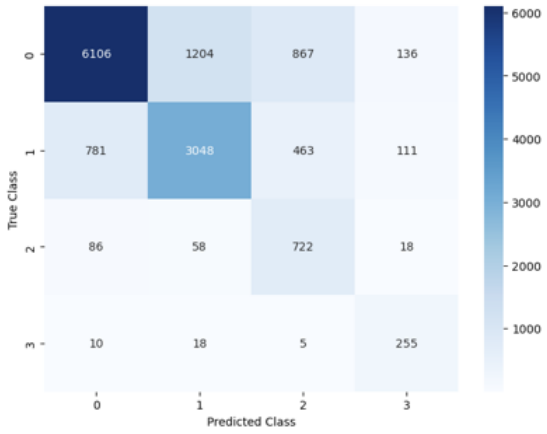
Features	Accuracy	F1-score	Precision	Recall	Nb of encoder layers	Class	Per-class accuracy
eGemaps	44	0.48	0.53	0.44	2	0 1 2 3	50 38 31 20
MFCCs	73	0.74	0.78	0.73	2	0 1 2 3	74 69 82 86
W2v2-CNN	71	0.72	0.75	0.71	3	0 1 2 3	70 69 76 95

Table 10.1: Results for the speaker-dependent scenario, using the transformer-based model. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

sented in table 10.2. In addition, we visualize the corresponding confusion matrices in figure 10.2.
Three feature configurations, eGeMAPS, MFCCs, and W2v2-CNN, are explored with either one or two encoder layers. The table provides each setup's overall accuracy, F1-score, precision, recall, and per-class accuracy.
eGeMAPS (1 encoder layer) yields the highest overall accuracy of 38% and an F1-score of 0.41. Although class 0 shows a decent accuracy of 49%, performance drops sharply for the other classes, with class 3 reaching only 4%. This wide disparity highlights the difficulty of capturing consistent speaker-independent cues using mostly prosodic features.
By contrast, MFCCs (2 encoder layers) lead to a lower overall accuracy of 31%, with an F1-score of 0.36. The class-wise results remain relatively low across all severity levels, ranging from 6% for class 3 to 36% for class 0. Despite offering more spectral detail than eGeMAPS, MFCCs appear to struggle with the additional variability introduced by new speakers, particularly for more severe dysarthria classes.
Finally, W2v2-CNN (1 encoder layer) falls in between these two methods with 32% overall accuracy and an F1-score of 0.37. Classes 0, 1, and 2 reach between 30% and 35% accuracy, and class 3 improves to 19%, outperforming the other two setups for the most severe dysarthria level, yet remaining under the overall accuracy recorded by eGeMAPS.
Overall, these results confirm the elevated difficulty of speaker-independent dysarthria severity classification, as all three feature sets show substantially lower performance compared to speaker-dependent experiments. While eGeMAPS achieves the best overall accuracy in this setting, its highly imbalanced class-wise performance is highly imbalanced. W2v2-CNN provides a bit more balanced performance (especially better for class 3), but



(a) eGemaps.



(b) MFCCs.



(c) Wav2vec2.0 embeddings.

Figure 10.1: The confusions matrices of the predictions using the transformer-based models, using (a) eGemaps, (b) MFCCs, and (c) wav2vec2.0 embeddings as inputs. Following the speaker-dependent scenario.

ends up with a lower overall accuracy.

10.2.2 Improving Transformer Model Via Contrastive Pre-Training

Contrastive learning is a self-supervised technique in which a model learns to distinguish between similar (positive) and dissimilar (negative) pairs of data. In our case, audios from participants with same dysarthria severity level are treated as positives, while audios from participants with different dysarthria severity level act as negatives. Through this process, the model forms more robust and speaker-invariant representations by clustering similar inputs and separating dissimilar ones in the embedding space, an especially valuable property for speaker-independent dysarthria severity classification.
In this section, we aim at improving the results obtained using the transformer-based model in combination with W2v2-CNN. This is done by the use of contrastive learning for better separation between different classes. By doing so, we aim to pull closer embeddings of audios from patients from same severity levels, while at the same time pulling

Features	Accuracy	F1-score	Precision	Recall	Nb of encoder layers	Class	Per-class accuracy
eGemaps	38	0.41	0.47	0.38	1	0 1 2 3	49 20 22 4
MFCCs	31	0.36	0.47	0.31	2	0 1 2 3	36 22 24 6
W2v2-CNN	32	0.37	0.49	0.32	1	0 1 2 3	32 30 35 19

Table 10.2: Results for the speaker-independent scenario, using the transformer-based model. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

away embeddings of audios from patients of different severity levels.

10.2.2.1 Data Partition

We perform only a speaker-independent data partition. The exact data distribution is shown in table 9.1.

10.2.2.2 Feature and Classifier

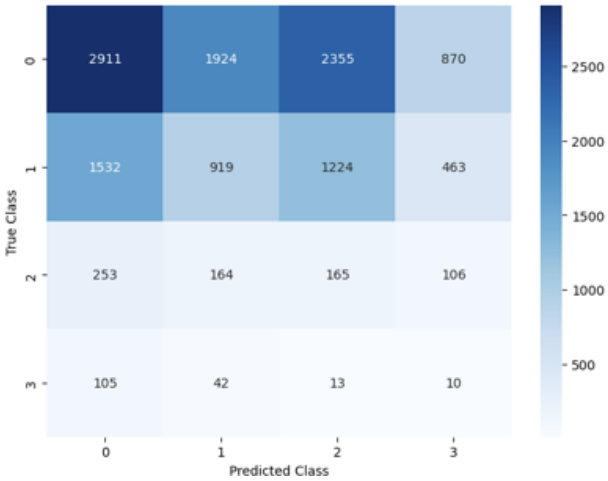
The same classifier as chapter 6 is used, while only the embeddings of wav2vec2.0 are used as input to the classifier (cf. section 6.2).

10.2.2.3 Training Settings

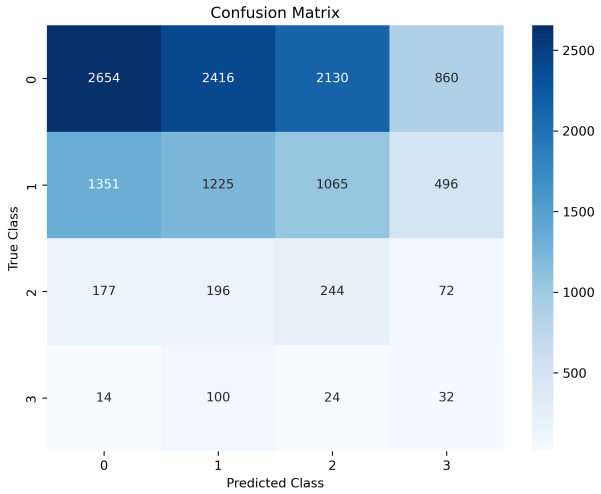
The same training settings presented in chapter 6 were used. Details could be found in section 6.2.
The training follows two stages: contrastive pre-training and classification training.
During the pre-training stage, the transformer-based model (without the classification head, see Section 6.2) is pre-trained with a contrastive loss. This loss pulls together embeddings of audio samples from the same class while pushing apart embeddings from different classes. A block diagram of this procedure is shown in figure 10.3.
After completing pre-training, we add the classification head and proceed with classification training. Here, the transformer encoder layers are initialized with the weights



(a) eGemaps.



(b) MFCCs.



(c) Wav2vec2.0 embeddings.

Figure 10.2: The confusions matrices of the predictions using the transformer-based models, using egemaps, MFCCs, and wav2vec2.0 embeddings as inputs. Following the speaker-independent scenario.

learned during contrastive pre-training, allowing the model to start with strong, class-discriminative features.

10.2.2.4 Results and Discussion

Table 10.3 shows the speaker-independent classification results for the W2v2-CNN configuration, comparing the baseline with no contrastive pre-training to a model that undergoes contrastive pre-training. In addition, we visualize the corresponding confusion matrices in figure 10.4.
As seen in the previous subsection, without pre-training, the model achieves an accuracy of 32%, with an F1-score of 0.37, a precision of 0.49, and a recall of 0.32. Class-wise performance remains moderate for classes 0 and 2 at 32% and 35%, respectively, while class 3 reaches 19%.

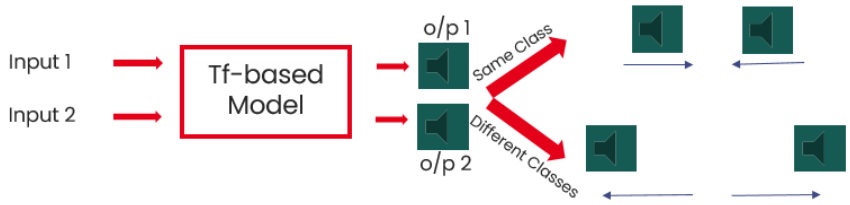


Figure 10.3: Block diagram of the contrastive learning. Tf refers to transformer, and o/p refers to output.

Introducing contrastive pre-training increases overall accuracy to 39%, and raises the F1-score to 0.43, precision to 0.53, and recall to 0.39. Notably, class 0 accuracy jumps from 32% to 41%, and class 1 from 28% to 36%, underscoring the model's improved ability to handle different speaker characteristics. Although the improvement for class 3 is more modest (19% to 20%), the overall trend indicates that leveraging unlabeled data enhances the model's generalization to unseen speakers and severity levels.
These findings confirm that contrastive pre-training can significantly boost performance in a speaker-independent setting by reinforcing speaker-invariant representations prior to the final dysarthria classification task.

Pre-training status	**Accuracy**	**F1-score**	**Precision**	**Recall**	**Class**	**Per-class accuracy**
Without	32	0.37	0.49	0.32	0 1 2 3	32 28 35 19
With	39	0.43	0.53	0.39	0 1 2 3	41 36 35 20

Table 10.3: Results for the speaker-independent scenario, using the contrastive pre-trained transformer-based model, with w2v2-CNN embeddings. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

10.3 Auto-Encoder

In this section, we investigate the use of auto-encoders for dysarthria severity classification. An auto-encoder is a type of neural network designed to learn a compressed, low-dimensional

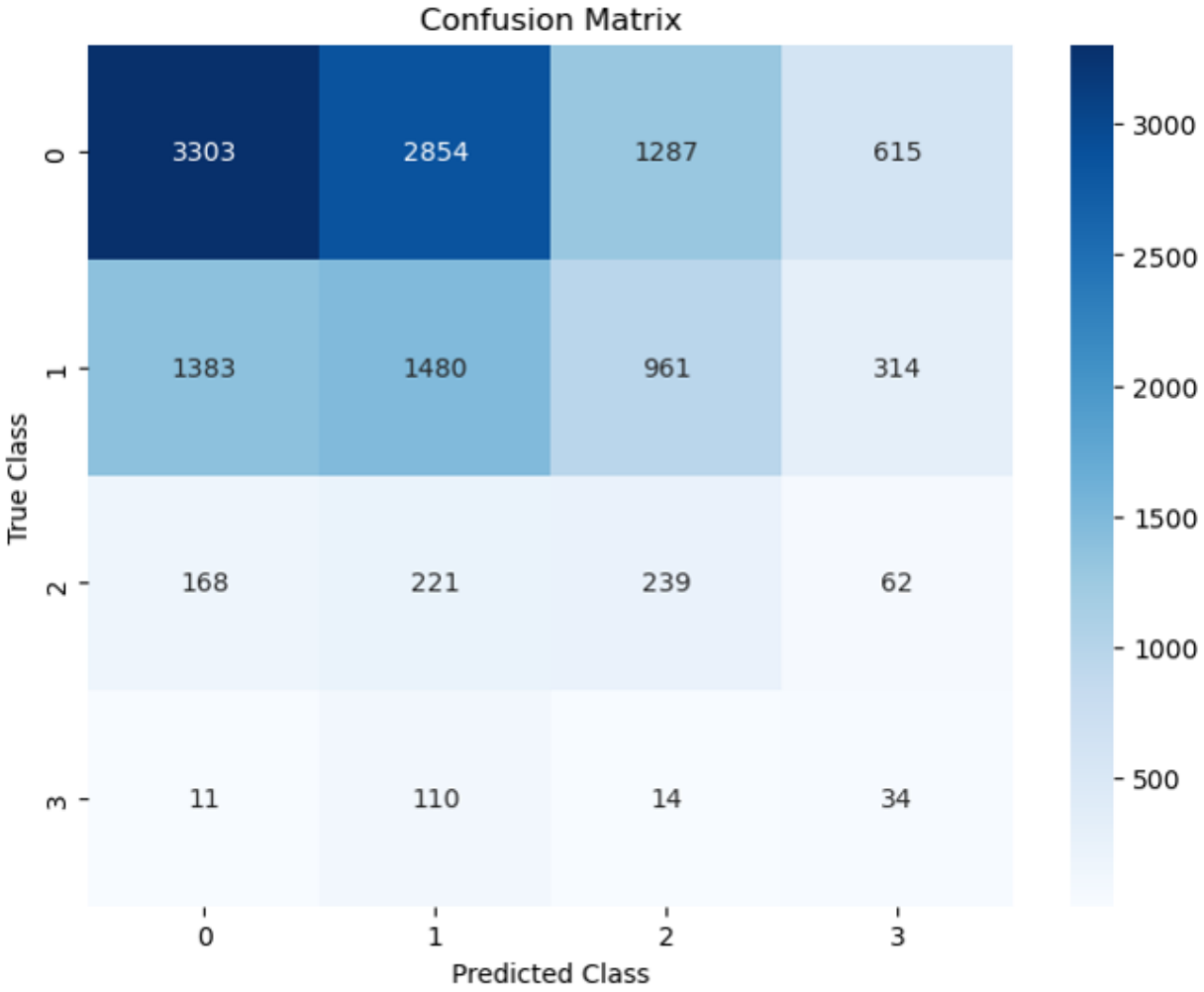


Figure 10.4: The confusions matrix of the predictions using the transformer model pretrained using contrastive learning, wav2vec2.0 embeddings as input.

representation of data and then reconstruct the original input from that compact representation. The encoder maps the input data into a smaller latent vector, and the decoder attempts to expand that vector back to the original dimension. By training on reconstruction tasks, the auto-encoder can learn meaningful features or patterns from the data in an unsupervised way, which makes it useful for dimensionality reduction, denoising, anomaly detection, and **feature learning**.

10.3.1 Proposed model

We propose a convolutional auto-encoder with two main parts: the encoder and the decoder. The encoder uses two convolution layers (each followed by ReLU) with strides of 2, effectively halving the dimensions at each step, and then flattens the result into a vector for a fully connected layer that produces the latent code. The decoder does the reverse: it takes the latent code, expands it back through another fully connected layer, and then

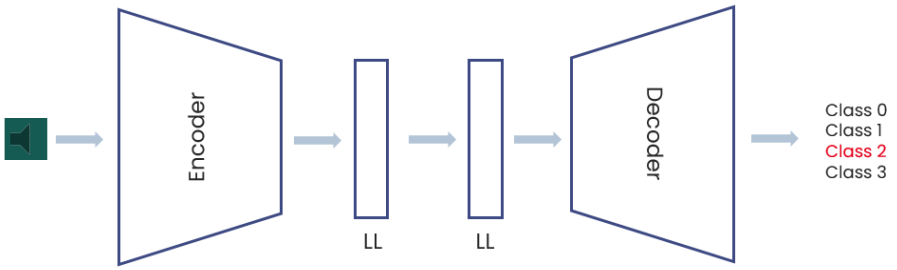


Figure 10.5: Block diagram of the auto-encoder model. LL refers to linear layer.

uses two transposed convolution layers (again with strides of 2) to rebuild the original shape.
A block diagram of the auto-encoder model is shown in figure 10.5.

As input to the model, we use MFCCs and wav2vec2.0 embeddings.
As for dysarthria severity classification, we build a classifier consisting of the encoder part of the aforementioned model, followed by two fully connected layers, that classifies the output into one of four possible classes.

10.3.1.1 Data Partition

We perform only a speaker-independent data partition. The exact data distribution is shown in table 9.1.

10.3.2 Input Features

Two input features to the model were used: MFCCs and wav2vec2.0 embeddings.

10.3.3 Training settings

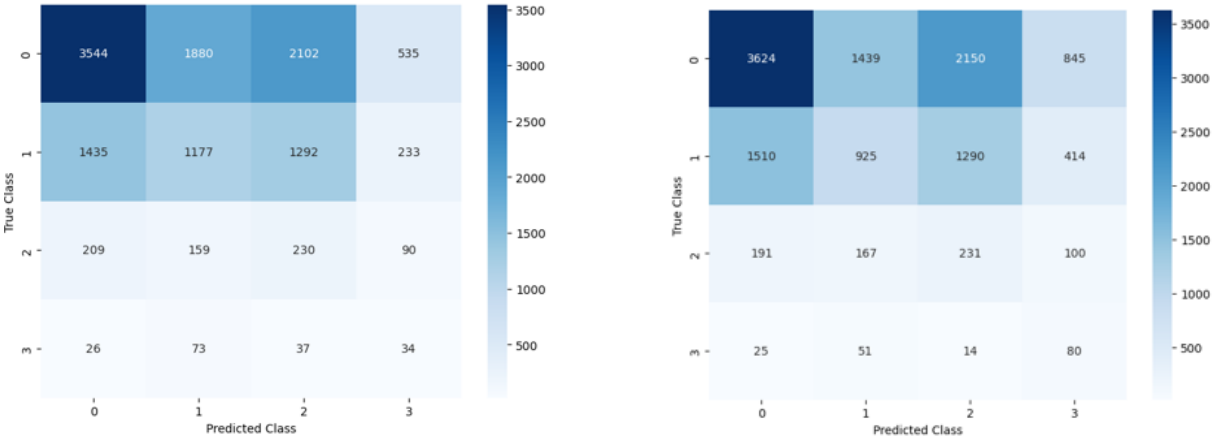
We start by training the auto-encoder to decrease the reconstruction error.
The model is trained using the Adam optimizer, and mean-squared error loss. We train using different learning rates (0.00001, 0.0001, 0.001, 0.01). We train the model for 50 epochs using a batch size of 64. Early stopping is applied when the loss stops decreasing for 5 consecutive epochs.
Once the AE is trained, we take the encoder part, freeze its weights, and use it for the dysarthria classifier.
The classifier is trained using the Adam optimizer, cross-entropy loss, with different learning rates (0.00001, 0.0001, 0.001, 0.01).
We train the models for 50 epochs using a batch size of 64. Early stopping is applied to prevent overfitting, with training terminating if validation loss does not improve for 5 consecutive epochs. The validation was performed using 10% of the training data.
Experiments were conducted following only the speaker-independent data partition.

10.3.4 Results and Discussion

The results of our experiments are presented in table 10.4. In addition, we visualize the corresponding confusion matrices in figure 10.6.
The table compares the two feature configurations in terms of overall accuracy, F1-score, precision, recall, and per-class accuracies.
When MFCCs are used as input, the model attains an overall accuracy of 38%, with an F1-score of 0.41, a precision of 0.51, and a recall of 0.38. Class 0 reaches 44%, while classes 1 and 2 hover around the high 20s and low 30s, and class 3 stands at 20%. Although these results show moderate performance for the two milder classes (0 and 2), the reduced scores for classes 1 and 3 suggest difficulties in capturing more pronounced articulation differences without additional refinement or adaptation.
Switching to the W2v2-CNN configuration yields an accuracy of 37%, which is slightly lower than MFCCs overall, but the F1-score improves to 0.43, and precision remains at 0.54. Notably, class 3 accuracy increases to 47%, more than double that achieved with MFCCs, indicating that the W2v2-CNN embeddings may be capturing severity-related cues more effectively for the most challenging class. However, class 1 drops to 22%, reflecting a potential trade-off in how well the learned representation generalizes across different severity levels.
In summary, these results indicate that while the auto-encoder's frozen encoder can provide a solid foundation for speaker-independent dysarthria classification, the choice of input features significantly affects both overall performance and per-class balance. MFCCs lead to higher aggregate accuracy, whereas W2v2-CNN better captures extreme severity levels (particularly class 3). Further exploration, such as feature fusion, more advanced pre-training strategies, or targeted data augmentation, may help balance performance across all dysarthria classes in this speaker-independent setting.

Features	**Accuracy**	**F1-score**	**Precision**	**Recall**	**Class**	**Per-class accuracy**
MFCCs	38	0.44	0.54	0.38	0	44
					1	28
					2	33
					3	20
W2v2-CNN	37	0.43	0.54	0.37	0	45
					1	22
					2	34
					3	47

Table 10.4: Results for the speaker-independent scenario, using the AE model. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.



(a) MFCCs embeddings.

(b) Wav2vec2.0 embeddings.

Figure 10.6: The confusions matrices of the predictions using the autoencoder models, (a) MFCCs and (b) wav2vec2.0 embeddings as inputs. Classes 0, 1, 2, 3 correspond to no dysarthria, low, medium, and high severity.

10.4 Results Summary and Comparison

A summary, comparative table of the best results from each of the presented methods are presented in tables 10.5 and 10.6, for speaker-dependent and speaker-independent scenarios, respectively.

Speaker-Dependent Scenario
The baseline model obtains an accuracy of 49%, an F1-score of 0.50, and a fairly uneven per-class distribution ranging from 35% to 63%. By contrast, wav2vec2.0 raises the accuracy to 72%, with a 0.74 F1-score and class-wise performance peaking at 89% for class 3. The AST model follows close behind at 71% accuracy and a 0.72 F1-score, but achieves even higher class-wise accuracies for classes 2 and 3 (87% and 91%, respectively). Finally, the TF-based approach attains the highest accuracy of 73% and ties wav2vec2.0 in F1-score at 0.74, balancing performance across classes in the range of 69% to 86%. These findings confirm that advanced transformer-based techniques substantially outperform the baseline in speaker-dependent dysarthria severity classification, highlighting the importance of robust feature extraction and architectural design for maximizing both overall and class-specific performance.

Speaker-Independent Scenario
The baseline method achieves the highest overall accuracy at 53%, but its extreme imbalance in per-class performance, most notably near-zero accuracies for classes 1 and 2, limits its practical utility. In contrast, wav2vec2.0 exhibits a more even distribution across severity levels, with an accuracy of 47% and per-class scores ranging from 30% to 54%. The AST model, at 34% accuracy, displays relatively higher accuracy for class 2 (42%) but struggles with class 3 (15%). Both the TF-based and TF+contrastive models reach lower overall accuracies (32% and 39%, respectively) but reduce the severe class imbalances observed in the baseline. Notably, TF+contrastive surpasses TF-based by seven

Model	Accuracy	F1-score	Precision	Recall	Class	Per-class accuracy
Baseline	49	0.5	0.6	0.4	0 1 2 3	0.35 0.45 0.56 0.63
Wav2vec2.0	72	0.74	0.76	0.72	0 1 2 3	72 73 68 89
AST	71	0.72	0.75	0.71	0 1 2 3	70 67 87 91
Tf-based	73	0.74	0.78	0.73	0 1 2 3	74 69 82 86

Table 10.5: A summary table of the best performing results of each conducted experiment, following the speaker-dependent data partition.

percentage points, suggesting that self-supervised feature learning can improve generalization. Finally, the auto-encoder method obtains 38% accuracy, matching the midrange performance of TF+contrastive while pushing class 3 accuracy to 20%. Collectively, these findings underscore the inherent difficulty of speaker-independent dysarthria severity classification, highlighting a trade-off between raw accuracy and balanced class-wise performance across the different models.

10.5 Conclusion

In this chapter, we revisited the transformer-based model proposed in chapter 6, and trained in on the newly collected french dataset. The model was trained following both the speaker-dependent and the speaker-independent data splits.
In order to improve the results for the speaker-independent scenario, we introduced a contrastive pre-trainig of the transformer-based model. This step resulted in a performance increase of 7% in terms of overall accuracy, and in general an increase in per-class performance.
A final experiment was carried out using an auto-encoder, for the speaker-independent scenario. This resulted in an overall accuracy of 38%.

Model	**Accuracy**	**F1-score**	**Precision**	**Recall**	**Class**	**Per-class accuracy**
Baseline	53	0.47	0.5	0.53	0 1 2 3	0.84 0.03 0.01 0.27
Wav2vec2.0	47	0.51	0.59	0.47	0 1 2 3	52 38 30 54
AST	34	0.39	0.52	0.34	0 1 2 3	36 28 42 15
Tf-based	32	0.37	0.49	0.32	0 1 2 3	32 30 35 19
Tf+contrastive	39	0.43	0.53	0.39	0 1 2 3	41 36 35 20
Auto-encoder	38	0.44	0.54	0.38	0 1 2 3	44 28 33 20

Table 10.6: A summary table of the best performing results of each conducted experiment, following the speaker-independent data partition.

Overall, these findings confirm the potential of leveraging robust feature extraction and pre-training strategies for dysarthria classification. While speaker-dependent setups achieve significantly higher accuracies, the speaker-independent scenario remains more challenging due to greater variability across speakers. Nevertheless, the observed gains from contrastive pre-training and pre-trained models like wav2vec2.0 indicate a promising path toward more accurate and generalizable dysarthria severity assessment.

Chapter 11

Conclusion and Perspectives

11.1 Conclusion

The objective of this dissertation was to design, implement, and rigorously evaluate an automated pipeline for the detection of dysarthria and the classification of its severity directly from running-speech audio, with the long-term goal of embedding such technology into day-to-day clinical practice and telehealth services. By focusing on both binary detection and severity estimation, the work set out to bridge the gap between research prototypes and tools that SLPs can reliably use to support diagnosis, therapy planning, and outcome tracking. To this end, several methods for automating speech pathology detection and classification have been proposed. Our work contributes to the state-of-the-art in dysarthria severity classification through a series of innovative experiments and methodical investigations, bridging the gap between traditional signal processing techniques and modern deep learning approaches. In particular, the work has centered on developing and rigorously evaluating novel models and strategies that accommodate the complexities inherent in dysarthric speech data.

In Chapter 6, we introduced a transformer-based model for dysarthria classification. Transformers are deep learning architectures built entirely on stacked self-attention and feed-forward layers, without any recurrent or convolutional operations, allowing every position in a sequence to model dependencies with every other position in a single computational step. To assess the applicability of this architecture to dysarthric speech, we implemented a model using existing transformer encoder blocks and trained it with different input features: MFCC, eGeMAPS, and wav2vec2 embeddings. For comparison, we implemented a baseline model using LSTM layers, trained on the same feature sets. Our experiments, conducted under both SD and SI scenarios, demo,started that the transformer-based model outperformed the LSTM baseline. The most notable results were obtained when using frozen wav2vec2-CNN embeddings as input: in this configuration, the model achieved classification accuracies of 92% in the SD scenario and 46% in the SI scenario. These findings suggest that attention mechanisms in transformers enhance dysarthria classification performance by effectively focusing on relevant temporal and spectral speech features. However, despite this improvement, the relatively low SI accuracy indicates that further generalization strategies are needed.

Building on these insights, Chapters 7 and 8 focused on enhancing model performance

and generalization, particularly under SI and cross-dataset conditions—by leveraging pre-trained models. These chapters provided a comprehensive comparative analysis of models operating on raw waveforms versus spectrograms for both dysarthria detection (Chapter 7) and severity classification (Chapter 8). The key idea was to address the performance drop observed in the SI scenario by using models that were already exposed to large-scale speech data during pre-training. In these Chapters, we conducted a detailed comparison between models trained on raw audio waveforms and those trained on spectrograms, evaluating their performance for both dysarthria detection and severity classification. For this, we used a consistent SI setup, in-database and cross-database testing, and conducted a detailed comparison between four SOTA models (wav2vec2, HuBERT, AST and whisper), under various configurations: using the model architectures without pre-trained weights as feature extractors, and under partial or full fine-tuning setups. The experiments showed that leveraging pre-trained architectures not only improves classification accuracy but also addresses challenges associated with limited annotated data. In particular, the best accuracies achieved for dysarthria detection were 96% (in-database) and 70% (cross-database), compared to 56% and 67% for dysarthria severity classification, respectively. These findings confirmed that using pre-trained models significantly improves robustness across different speaker populations and recording conditions.

Another contribution of this thesis was the data collection from French Parkinson's patients. This dataset was gathered in collaboration with Fondation Hopale and includes recordings from 114 participants. Each participant contributed samples across various categories: non-words, words, sentences, and short free speech segments. This data adds to the bank of data for speech pathology research in general, and for french language related research in particular. The data collection process was demonstrated in Chapter 5. Chapter 9 extended the investigation into the newly collected dataset. First, baseline experiments were conducted, before reapplying the experiments conducted in chapters 7, and 8, yielding an accuracy of 72% following the SD data split, vs an accuracy of 47% following the SI data split. Additionally, in Chapter 10, we explored the use of the transformer-based model, proposed in Chapter 6, and its improvement through contrastive pre-training. These experiments demonstrated that attention mechanisms inherent in transformers, when coupled with contrastive strategies, can better distinguish nuanced dysarthric patterns, yielding notable improvements in classification performance, in particular, following the SI data split, the accuracy of the classification increased from 32% to 39% when the contrastive pre-training was introduced. Finally, we explored an auto-encoder based approach. The auto-encoder model, equipped with carefully selected input features and optimized training settings, provided an alternative representation learning strategy for our classification framework. It achieved an accuracy of 38% following the SI data split.

Overall, the experiments detailed across Chapters 6 through 10 collectively contribute to a more adaptive, data-driven methodology for assessing dysarthria severity. The incorporation of attention mechanisms, pre-trained models, transformer-based learning with contrastive pre-training, and auto-encoder techniques has paved the way for upcoming efforts to improve the generality and interpretability of the model. Thus, this study provides a solid basis for creating useful, high-performing diagnostic instruments in clinical settings, which will ultimately lead to quicker and more efficient therapy interventions for dysarthria sufferers.

11.2 Perspectives

Despite the advances reported in the literature and in this thesis, there is still a long way to go before AI-driven methods can be efficiently deployed for diagnosis and rehabilitation. In this section, we propose potential directions for future research, aiming to address current limitations and pave the way for more comprehensive and adaptable dysarthria classification systems.

Expanding Transformer-Based Models
Building on the promising results observed when incorporating contrastive pre-training into transformer-based models, future work could focus on expanding this approach by integrating a larger and more diverse dataset. Increasing the volume and variety of training data, by encompassing a wider spectrum of languages, accents, recording devices, background-noise conditions, and dysarthria severities, would enable the self-attention layers to disentangle subtler timing, prosodic, and spectral cues. That, in turn, should sharpen diagnostic specificity, improve the detection of small yet clinically meaningful changes during therapy or disease progression, and make the system more tolerant of the acoustic quirks that inevitably crop up in a kitchen or living-room telehealth session. At the same time, a richer input space risks overwhelming the capacity of a shallow transformer, therefore, some upgrades, such as widening the feed-forward bottleneck, adding grouped convolutions, or the number of encoder layers, will be needed to keep accuracy rising without inflating latency, or memory usage. At the same time, scaling must be done carefully to avoid amplifying class imbalance and demographic bias, and any expanded model will still need clinician-friendly explanation modules to earn SLPs' trust. Finally, because much of the new data will arrive incrementally through telepractice platforms, privacy-aware continual-learning strategies are essential for seamless, secure integration into daily dysarthria care.

Integrating Pathological Speech in Pre-trained Models
Pre-trained models have demonstrated potential for feature extraction in speech pathology tasks; however, their performance may be constrained by the nature of the training data. A promising direction for future research is to incorporate pathological speech directly into the training of these models. Incorporating a substantial volume of pathological speech during pre-training promises several benefits: the network can internalize disease-specific cues instead of treating them as outliers, deliver sharper severity estimates across a wider range of impairments, and transfer more readily to downstream applications such as automatic intelligibility scoring or remote therapy progress tracking. At the same time, this strategy introduces challenges that must be tackled from both a technical and clinical perspective. Pathological corpora remain scarce, fragmented across hospitals, and often protected by strict privacy regulations, curating a balanced, demographically diverse dataset therefore requires multi-institutional collaboration, sophisticated identification pipelines, and federated training protocols that keep raw audio behind each hospital's firewall. Even after data sharing obstacles are cleared, severe cases frequently contain long stretches of unintelligible or phonation-less segments that can confuse contrastive objectives and lead to unstable training unless specialized masking or loss-weighting schemes are introduced. Finally, SLPs will require transparent feedback,

confidence scores, saliency overlays, and alerts when predictions fall outside the model's validated operating range, to trust that a system fine-tuned on pathological speech will behave reliably during telehealth sessions or bedside evaluations.

Multi-Modality and Model Fusion

Our findings suggest that different models may excel in classifying distinct classes of dysarthria. An interesting avenue for future work is to explore the integration of several modalities or the fusion of multiple models into a single framework. The core acoustic waveform remains indispensable because it encodes spectral detail and voice-quality cues, but it can be augmented with: 1- low-cost facial video from a tablet or laptop camera, which reveals lip-rounding, jaw displacement, and timing of bilabial closures, 2- smartphone-based inertial signals or respiratory belts that track breath-group duration and chest-wall movement, useful for hypokinetic speakers who run out of air mid-utterance, and 3- lexical-linguistic features automatically derived from real-time ASR transcripts, such as substitution or pause patterns that flag ataxic scanning speech. These modalities are already available on devices patients routinely use for home practice, keeping deployment friction low. By exploiting such multimodal or ensemble frameworks, we can build a dysarthria-classification tool that not only boosts accuracy across diverse impairment profiles but also remains robust when used in everyday telehealth sessions or during unsupervised home-exercise monitoring.

Exploring the Use of Prompts and Large Language Models

The increasing success of prompt-based approaches in natural language processing opens new possibilities for enhancing speech pathology classification systems. Future research should explore the integration of prompts and the utilization of LLMs as supplementary components. Prompt-based LLMs can be folded into our dysarthria pipeline by exploiting the fact that patients read pre-scripted passages whose canonical text is already known to the system. The LLM therefore focuses on the mismatch between the expected script and what the patient actually produced. After automatic speech recognition aligns the utterance to the script, the deviation pattern, insertions, deletions, substitutions, stutters, unusually long pauses, and their precise time stamps, together with the acoustics could be fed to a model that learns to classify the audio according to dysarthria severity level.

Conducting Multilingual an Multi-Pathology Experiments

Conducting experiments across multiple languages, as well as across the full spectrum of motor-speech pathologies (ataxic, spastic, flaccid, mixed, hypokinetic Parkinsonian speech, and the dysarthrias that accompany ALS or cerebral palsy) represents a valuable perspective for advancing dysarthria classification. Given the variability in phonetic and prosodic characteristics across languages and diseases, future studies should focus on assessing model generalization and adaptability in multilingual and multi-pathology contexts. One promising architecture is a hierarchical mixture-of-experts, where a shared base encoder learns universal speech representations, and smaller "experts" (specialist sub-models) handle specific languages or pathologies. This design keeps each expert small enough for real-time deployment on a tablet while allowing new languages or pathologies to be added incrementally without retraining the entire stack. Pursuing such multilingual,

multi-pathology experiments will therefore move the field from laboratory prototypes toward practical, equitable tools that can serve diverse patient populations in clinics and telehealth platforms worldwide.

Personalized Patient Models for Continuous Monitoring
An alternative to large generalist systems is the development of personalized models tailored to individual speakers. These models could adapt over time, track speech evolution, and offer highly accurate, low-latency feedback on personal devices. Compared to large-scale architectures, such models are compact, interpretable, and privacy-friendly—well suited for continuous monitoring and self-guided rehabilitation. This direction emphasizes flexibility and personal relevance, with potential to empower patients in managing their own therapy between clinical sessions.

Adaptive Treatment Planning and Tailored Therapeutic Interventions
Another promising future direction involves adapting the developed models to design individualized treatment exercises and adjust rehabilitation strategies according to each patient's specific needs. While the personalized patient models focus primarily on continuous assessment and monitoring, this perspective emphasizes the dynamic customization of therapy. By integrating patient-specific data into the model, it becomes possible to formulate targeted treatment protocols that evolve over time in response to the patient's progress, ensuring that interventions are both relevant and effective. This could be achieved through rule-based systems linked to model outputs, or through LLMs capable of mapping patterns in speech errors to evidence-based intervention strategies. By closing the loop between assessment and treatment, such tools would support clinicians in delivering tailored, responsive care, and offer patients more engaging and effective therapy.

Taken together, these directions outline a roadmap toward more personalized, explainable, and equitable AI systems for dysarthria care, moving beyond lab settings into real-world clinical impact.

Bibliography

[1] Jordan R Green, Yana Yunusova, Mili S Kuruvilla, Jun Wang, Gary L Pattee, Lori Synhorst, Lorne Zinman, and James D Berry. Bulbar and speech motor assessment in als: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14(7-8):494–500, 2013.

[2] Adina T Michael-Titus and Peter Shortland. *The Nervous System, E-Book: The Nervous System, E-Book*. Elsevier Health Sciences, 2022.

[3] Mohammad Jahanaray, Ali Jahanaray, and Zahra Zohoorian. Brain regions involved in speech production, mechanism and development. *Neuroscience Research Notes*, 5(4):178–178, 2022.

[4] Envelope and ADSR anything | Reason Experts. https://www.reasonexperts.com/envelope-and-adsr-anything.html.

[5] Modulation Definition. https://techterms.com/definition/modulation, urldate = 2024-11-12.

[6] Svm - support vector machines.

[7] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[10] Jonathan Boigne. HuBERT: How to Apply BERT to Speech, Visually Explained. https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html, October 2021.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[12] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[13] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R Gunderson, Thomas S Huang, Kenneth L Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Interspeech*, volume 2008, pages 1741–1744, 2008.

[14] Sony Enceinte tour de cou sans fil SRS-NS7.

[15] Dorothy VM Bishop, Margaret J Snowling, Paul A Thompson, Trisha Greenhalgh, and Catalise Consortium. Catalise: A multinational and multidisciplinary delphi consensus study. identifying language impairments in children. *PLOS one*, 11(7):e0158753, 2016.

[16] Echo. https://research.ncl.ac.uk/echo/. Accessed 2025-04-12.

[17] Hanne B Søndergaard Knudsen, Niloufar Jalali-Moghadam, Silvia Nieva, Ewa Czaplewska, Marja Laasonen, Ellen Gerrits, Cristina McKean, and James Law. Allocation and funding of speech and language therapy for children with developmental language disorders across europe and beyond. *Research in Developmental Disabilities*, 121:104139, 2022.

[18] Statistics on speech language disorders.

[19] S Jothilakshmi. Automatic system to detect the type of voice pathology. *Applied Soft Computing*, 21:244–249, 2014.

[20] George S Liu, Nedeljko Jovanovic, C Kwang Sung, and Philip C Doyle. A scoping review of artificial intelligence detection of voice pathology: challenges and opportunities. *Otolaryngology–Head and Neck Surgery*, 171(3):658–666, 2024.

[21] Nicolas Saenz-Lechon, Juan I Godino-Llorente, Víctor Osma-Ruiz, and Pedro Gómez-Vilda. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*, 1(2):120–128, 2006.

[22] Constance Holden. The origin of speech, 2004.

[23] Michael Balter. 'speech gene'tied to modern humans, 2002.

[24] Lindsey I Black, Anjel Vahratian, and Howard J Hoffman. Communication disorders and use of intervention services among children aged 3-17 years: United states, 2012. nchs data brief. number 205. *Centers for Disease Control and Prevention*, 2015.

[25] Neil Bhattacharyya. The prevalence of voice problems among adults in the united states. *The Laryngoscope*, 124(10):2359–2362, 2014.

[26] Dorothy VM Bishop, Margaret J Snowling, Paul A Thompson, Trisha Greenhalgh, and Catalise Consortium. Catalise: A multinational and multidisciplinary delphi consensus study. identifying language impairments in children. *PLOS one*, 11(7):e0158753, 2016.

[27] Hanne B Søndergaard Knudsen, Niloufar Jalali-Moghadam, Silvia Nieva, Ewa Czaplewska, Marja Laasonen, Ellen Gerrits, Cristina McKean, and James Law. Allocation and funding of speech and language therapy for children with developmental language disorders across europe and beyond. *Research in Developmental Disabilities*, 121:104139, 2022.

[28] Nancy Pearl Solomon. Evaluation of speech. *Neurologic and Neurodegenerative Diseases of the Larynx*, pages 67–77, 2020.

[29] Thomas P Marquardt and Christine L Matyear. Speech science. *Communication Sciences and Disorders: From Science to Clinical Practice*, page 91, 2015.

[30] Nancy Pearl Solomon. Evaluation of speech. *Neurologic and Neurodegenerative Diseases of the Larynx*, pages 67–77, 2020.

[31] Frank H. Guenther and Gregory Hickok. Chapter 58 - neural models of motor speech control. In Gregory Hickok and Steven L. Small, editors, *Neurobiology of Language*, pages 725–740. Academic Press.

[32] Types of Neurons: Parts, Structure, and Function. https://www.verywellhealth.com/types-of-neurons-5201172. Section: Verywell.

[33] Nicola J Allen and Ben A Barres. Glia—more than just brain glue. *Nature*, 457(7230):675–677, 2009.

[34] Carole R. Roth and Ignatius Nip. *Dysarthria*, pages 1–5. Springer International Publishing, Cham, 2017.

[35] Dysarthria - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/dysarthria/symptoms-causes/syc-20371994.

[36] Dysarthria (Slurred Speech): Symptoms, Causes, and Treatment. https://www.healthline.com/health/dysarthria, September 2013.

[37] Dysarthria. https://my.clevelandclinic.org/health/diseases/17653-dysarthria.

[38] Karen W Krigger. Cerebral palsy: an overview. *American family physician*, 73(1):91–100, 2006.

[39] Małgorzata Sadowska, Beata Sarecka-Hujar, and Ilona Kopyta. Cerebral palsy: current opinions on definition, epidemiology, risk factors, classification and treatment options. *Neuropsychiatric disease and treatment*, pages 1505–1518, 2020.

[40] CL Balf and TTS Ingram. Problems in the classification of cerebral palsy in childhood. *British Medical Journal*, 2(4932):163, 1955.

[41] B Hagberg, G Hagberg, and I Olow. The changing panorama of cerebral palsy in sweden 1954–1970. *Acta Pædiatrica*, 64(2):193–200, 1975.

[42] WORKINGER MS. Perceptual analysis of the dysarthrias in children with athetoid and spastic cerebral palsy. *Dysarthria and apraxia of speech: Perspectives on management*, 1991.

[43] Lindsay Pennington. Speech and communication in cerebral palsy. *Eastern journal of medicine*, 17(4):171, 2012.

[44] Nancy Pearl Solomon and Sandy Charron. Speech breathing in able-bodied children and children with cerebral palsy: A review of the literature and implications for clinical intervention. *American Journal of Speech-Language Pathology*, 7(2):61–78, 1998.

[45] Rupal Patel. Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy. 2003.

[46] Maarten C De Rijk, WA Rocca, DW Anderson, MO Melcon, MMB Breteler, and DM Maraganore. A population perspective on diagnostic criteria for parkinson's disease. *Neurology*, 48(5):1277–1281, 1997.

[47] Parkinson's disease - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055.

[48] Serge Pinto, Canan Ozsancak, Elina Tripoliti, Stéphane Thobois, Patricia Limousin-Dowsey, and Pascal Auzou. Treatments for dysarthria in parkinson's disease. *The Lancet Neurology*, 3(9):547–556, 2004.

[49] Pamela Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980.

[50] Annual workforce data: 2023 asha-certified audiologist- and speech-language pathologist-to-population ratios.

[51] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Transactions on Graphics (ToG)*, 23(3):559–568, 2004.

[52] Laurent Demany, Guilherme Monteiro, Catherine Semal, Shihab Shamma, and Robert P Carlyon. The perception of octave pitch affinity and harmonic fusion have a common origin. *Hearing research*, 404:108213, 2021.

[53] Sound intensity.

[54] Loudness and sound intensity.

[55] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[56] Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. A comprehensive survey on contrastive learning. *Neurocomputing*, page 128645, 2024.

[57] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

[58] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

[59] Zhihao Fan, Yeyun Gong, Dayiheng Liu, Zhongyu Wei, Siyuan Wang, Jian Jiao, Nan Duan, Ruofei Zhang, and Xuanjing Huang. Mask attention networks: Rethinking and strengthen transformer. *arXiv preprint arXiv:2103.13597*, 2021.

[60] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[61] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[62] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[63] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[64] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[65] Ying Qin, Yuzhong Wu, Tan Lee, and Anthony Pak Hin Kong. An end-to-end approach to automatic speech assessment for cantonese-speaking people with aphasia. *Journal of Signal Processing Systems*, 92(8):819–830, 2020.

[66] Tara N Sainath, Ron J Weiss, Andrew W Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Interspeech*, pages 1–5. Dresden, Germany, 2015.

[67] Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022.

[68] Zeinab Noroozi, Azam Orooji, and Leila Erfannia. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific reports*, 13(1):22588, 2023.

[69] Harsha Jakkanahalli Vishnukumar, Björn Butting, Christian Müller, and Eric Sax. Machine learning and deep neural network — artificial intelligence core for lab and real-world test and validation for adas and autonomous vehicles: Ai for efficient and quality test and validation. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 714–721, 2017.

[70] End-to-end models for complex ai tasks. https://www.capitalone.com/tech/machine-learning/pros-and-cons-of-end-to-end-models/. Accessed 2025-05-15.

[71] Overfitting, variance, bias and model complexity in machine learning. https://www.pico.net/kb/overfitting-variance-bias-and-model-complexity-in-machine-learning/. Accessed 2025-05-15.

[72] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022.

[73] Federico Calà, Lorenzo Frassineti, Giovanna Cantarella, Giulia Buccichini, Ludovica Battilocchi, Claudia Manfredi, and Antonio Lanatà. Towards an explainable artificial intelligence system for voice pathology identification and post-treatment characterisation. *Biomedical Signal Processing and Control*, 104:107530, 2025.

[74] Lauren Stumpf, Balasundaram Kadirvelu, Sigourney Waibel, and A Aldo Faisal. Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning. *arXiv preprint arXiv:2403.00854*, 2024.

[75] Richard Palmer, Roslyn Ward, Petra Helmholz, Geoffrey R Strauss, Paul Davey, Neville Hennessey, Linda Orton, and Aravind Namasivayam. Facial movements extracted from video for the kinematic classification of speech. *Sensors*, 24(22):7235, 2024.

[76] Ghadeer Alharbi, Najwa Alamri, and Sahar Sabbeh. Automatic classification of speech dysarthric intelligibility levels using textual feature. *IEEE Access*, 2025.

[77] Sari Masri, Ahmad Hasasneh, Mohammad Tami, and Chakib Tadj. Exploring the impact of image-based audio representations in classification tasks using vision transformers and explainable ai techniques. *Information*, 15(12):751, 2024.

[78] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[79] Amlu Anna Joshy and Rajeev Rajan. Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157, 2022.

[80] Veena Karjigi, N Sreedevi, et al. Speech intelligibility assessment of dysarthria using fisher vector encoding. *Computer Speech & Language*, 77:101411, 2023.

[81] Aastha Kachhi, Anand Therattil, Ankur T Patil, Hardik B Sailor, and Hemant A Patil. Teager energy cepstral coefficients for classification of dysarthric speech severity-level. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1462–1468. IEEE, 2022.

[82] Krishna Gurugubelli and Anil Kumar Vuppala. Analytic phase features for dysarthric speech detection and intelligibility assessment. *Speech Communication*, 121:1–15, 2020.

[83] Garima Vyas, Malay Kishore Dutta, Jiri Prinosil, and Pavol Harár. An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 515–518. IEEE, 2016.

[84] N P Narendra and Paavo Alku. Automatic intelligibility assessment of dysarthric speech using glottal parameters. *Speech Communication*, 123:1–9, 2020.

[85] Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, 158:103047, 2024.

[86] Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech. *IEEE journal of biomedical and health informatics*, 2024.

[87] Lauren Stumpf, Balasundaram Kadirvelu, Sigourney Waibel, and A Aldo Faisal. Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning. *arXiv preprint arXiv:2403.00854*, 2024.

[88] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 4040–4041, 2021.

[89] The Challenges of Data in Machine Learning. https://www.ailabs.global/blog/the-challenges-of-data-in-machine-learning.

[90] IABAC®. Issues in Machine Learning. https://iabac.org/blog/issues-in-machine-learning, January 2024.

[91] (7) 10 Data Challenges when considering a Data-Driven (ML) Solution | LinkedIn. https://www.linkedin.com/pulse/10-data-challenges-when-considering-data-driven-ml-steven-reece-phd-r8rle/.

[92] Universal Access (UA) Speech Corpus Development – Heejin Kim. https://publish.illinois.edu/hkim17/universal-access-ua-speech-corpus-development-universal-access-automatic-speech-

[93] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46:523–541, 2012.

[94] 3D Electromagnetic Articulograph | Carstens Medizinelektronik GmbH. ://www.articulograph.de/.

[95] The TORGO database. https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html.

[96] Kathryn M Yorkston, David R Beukelman, and Charles Traynor. *Assessment of intelligibility of dysarthric speech*. Pro-ed Austin, TX, 1984.

[97] Jeremy H Clear. The british national corpus. *the digital world*, pages 163–187, 1993.

[98] Ray D Kent, Gary Weismer, Jane F Kent, and John C Rosenbek. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54(4):482–499, 1989.

[99] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[100] Alan Wrench. The mocha-timit articulatory database, 1999.

[101] Lori F Lamel, Robert H Kassel, and Stephanie Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proc. SIOA 1989*, pages Vol–2, 1989.

[102] SG Webber. Webber photo cards: Story starters, 2005.

[103] Jan G Svec and Svante Granqvist. Guidelines for selecting microphones for human voice production research. 2010.

[104] Rita R Patel, Shaheen N Awan, Julie Barkmeier-Kraemer, Mark Courey, Dimitar Deliyski, Tanya Eadie, Diane Paul, Jan G Švec, and Robert Hillman. Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function. *American journal of speech-language pathology*, 27(3):887–905, 2018.

[105] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[106] Gokul Yenduri, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Rutvij H Jhaveri, Weizheng Wang, Athanasios V Vasilakos, Thippa Reddy Gadekallu, et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*, 2023.

[107] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[108] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.

[109] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[110] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[111] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[112] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[113] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[114] MP Lewis, GF Simon, and CD Fennig. Ethnologue: Languages of the world nineteenth edition. *SIL International. http://www. ethnologue. com*, 2016.

[115]

[116] Ons Aouedi, Alessio Sacco, Kandaraj Piamrat, and Guido Marchetto. Handling privacy-sensitive medical data with federated learning: challenges and future directions. *IEEE journal of biomedical and health informatics*, 27(2):790–803, 2022.

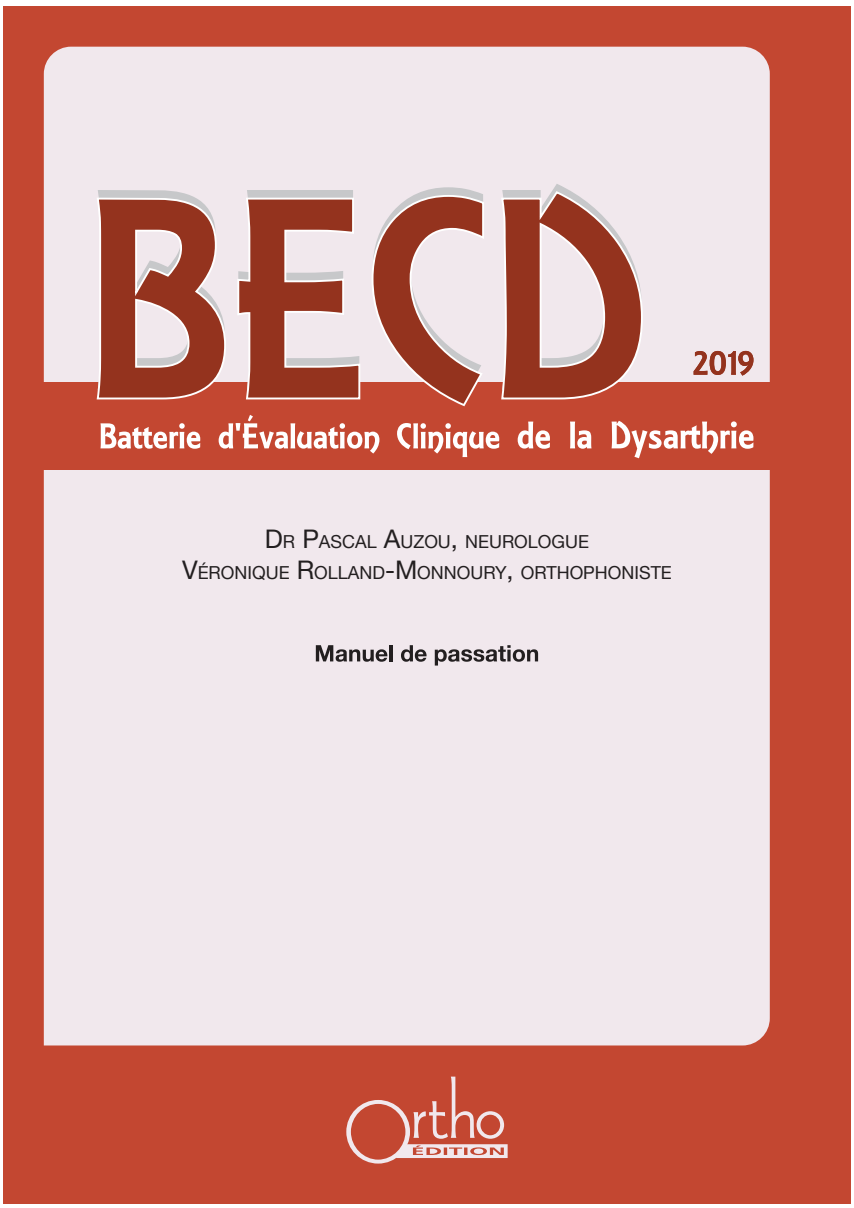
[117] Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, and Kristin Mcleod. Generating synthetic labeled data from existing anatomical models: an example with echocardiography segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2783–2794, 2021.

[118] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.

[119] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[120] Helin Wang, Yuexian Zou, Dading Chong, and Wenwu Wang. Environmental sound classification with parallel temporal-spectral attention. *arXiv preprint arXiv:1912.06808*, 2019.

Appendix I: BECD Manual



I. Contenu de la batterie

La Batterie d'Évaluation Clinique de la Dysarthrie (BECD) 2019 comprend :

- Un ouvrage introductif abordant la physiologie de la parole ; la définition, la classification et les descriptions des dysarthries ; les principes de l'évaluation des dysarthries ; la présentation des outils constitutifs de la batterie ; la validation des outils.

- Un manuel de passation.

- L'accès à une plateforme informatique pour :
  - la cotation des épreuves en ligne ;
  - le téléchargement de documents ;
  - l'utilisation de fichiers audio pour la passation du Score d'Intelligibilité auprès des enfants non-lecteurs.

- Le matériel nécessaire à la passation du Score d'Intelligibilité (SI)
  - Deux jeux de 50 cartes pour l'épreuve de compréhension de mots (série A et série B pour les enfants).
  - Deux jeux de 50 cartes pour l'épreuve de compréhension de phrases (série A et série B pour les enfants).

- Le matériel nécessaire à la passation du Test Phonétique d'Intelligibilité (TPI) : 5 carnets numérotés de 1 à 5, présentant chacun les 52 mots dans un ordre différent.

- Un texte extrait de « Pierrot » de Guy de Maupassant pour le support de lecture.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

II. Présentation générale

La BECD a plusieurs objectifs :

- permettre aux orthophonistes de réaliser un bilan de dysarthrie quels que soient le degré de sévérité et la pathologie à l'origine des troubles ;
- recenser et analyser précisément les troubles afin de pouvoir établir le projet thérapeutique individuel le mieux adapté ;
- recueillir des données quantitatives nécessaires au suivi thérapeutique pour juger de la pertinence et de l'efficacité de la prise en charge ;
- favoriser et harmoniser les échanges entre les cliniciens.

La BECD est un outil élaboré pour être aussi complet que possible tout en étant facilement utilisable dans la pratique clinique. Elle comporte plusieurs parties permettant de regrouper les différentes modalités d'approches disponibles pour évaluer cliniquement une dysarthrie :

- évaluation de la sévérité ;
- analyse perceptive ;
- analyse phonétique ;
- examen de la motricité ;
- auto-évaluation ;
- relevés acoustiques instrumentaux.

Pour chacun de ces domaines, différents outils sont proposés. Ils sont rapportés dans le tableau 1.

Domaine étudié	Outils disponibles
Sévérité	Score Perceptif (SP) Score d'Intelligibilité (SI) Test Phonétique d'Intelligibilité (TPI)
Analyse perceptive	Analyse globale Grille perceptive
Analyse phonétique	Répétitions de phonèmes, de mots simples, de mots complexes Test Phonétique d'Intelligibilité (TPI)
Examen de la motricité	Grille motrice
Auto-évaluation	Speech Handicap Index (SHI)
Analyse acoustique	Relevé de données

Tableau 1 : liste des approches permettant l'analyse de la dysarthrie avec les outils correspondants proposés dans la BECD 2019.

L'examinateur dispose ainsi de plusieurs épreuves qui lui permettront de tester les différents domaines successivement pour préciser la nature des troubles dans chacun d'eux. Il nous semble important lors d'un bilan de dysarthrie d'analyser les différents domaines distinctement. Dans cette démarche, on pourra constater qu'une parole très altérée peut avoir conservé toute son intelligibilité.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

III. Utiliser la BECD

Dès le premier contact avec un patient dysarthrique, de nombreuses informations sont transmises à l'examinateur. Un ou plusieurs traits prédominants peuvent s'imposer, tels qu'une hypernasalité, une hypophonie, la lenteur d'un débit, l'imprécision articulatoire, etc. Par les différents outils qu'elle propose, la BECD a pour objectif de permettre la réalisation du bilan le mieux adapté en fonction des premières informations reçues.

Selon la sévérité de la dysarthrie, la pathologie concernée, les troubles associés, l'état général du patient, l'examinateur pourra décider de faire passer toutes les épreuves ou choisira celles qui lui paraissent les plus pertinentes et les mieux adaptées.

Ainsi, pour estimer la sévérité de la dysarthrie, si celle-ci apparaît d'emblée majeure, on choisira de faire passer le Test Phonétique d'Intelligibilité plutôt que le Score d'Intelligibilité. Le Score Perceptif sera, quant à lui, toujours renseigné.

En ce qui concerne l'analyse phonétique, seule l'épreuve de réalisation phonétique isolée pourra être passée en cas de dysarthrie majeure, alors que l'épreuve des mots complexes sera proposée en cas de dysarthrie légère et/ou modérée.

L'observation du patient dans la première partie du bilan et les données recueillies par l'analyse perceptive permettront d'orienter un choix éventuel des items à faire passer pour l'examen de la motricité.

Lorsqu'il sera impossible d'envisager un bilan complet de la dysarthrie, du fait par exemple de la présence d'autres déficits prédominants (aphasie, troubles cognitifs…), ou lors de la phase aiguë d'une pathologie, il pourra être intéressant de remplir la grille perceptive et de renseigner le score perceptif, afin de disposer de données pouvant être utiles pour le suivi du patient.

De la passation de la BECD aux conclusions

Les différents domaines de l'évaluation sont distincts mais complémentaires. Les données qu'ils permettent de recueillir peuvent être reliées entre elles lors de l'analyse des résultats. Cela conduit à émettre des hypothèses sur le fonctionnement physiopathologique sous-jacent, à comprendre le trouble, et donc à déterminer les axes thérapeutiques.

Par exemple, une perturbation de l'articulation des consonnes, notée lors de l'analyse perceptive (réalisée en approche globale ou systématique par la passation de la grille perceptive), pourra être précisée par les analyses phonétiques (fournies par le TPI et les épreuves de répétition). Puis elle pourra être reliée à un dysfonctionnement des organes remarqué lors de la passation de la grille motrice. Ainsi, le *critère perceptif* : imprécision de l'articulation, pourra être précisé par le *critère phonétique* : production altérée d'une occlusive évoquant une fricative (/p/ => /f/), lui-même relié à un *dysfonctionnement moteur sous-jacent* : déficit de tonus labial ou déficit du souffle respiratoire ne permettant pas l'installation d'une pression intra-orale suffisante ou incompétence vélaire.

Divers dysfonctionnements pouvant conduire à une même perturbation, il est essentiel d'établir des liens entre les données recueillies lors de l'analyse perceptive, l'analyse phonétique et l'examen moteur pour définir le travail rééducatif adapté.

À l'issue du bilan, l'examinateur pourra aussi déterminer le degré de sévérité global de la dysarthrie, et le degré de sévérité de chaque trouble. Il pourra ainsi hiérarchiser les troubles afin de fixer les objectifs principaux et l'objectif prioritaire de la rééducation. Ce dernier étant l'amélioration du paramètre qui apportera le plus de bénéfice sur l'intelligibilité, le caractère naturel ou l'efficacité de la parole, selon le degré de sévérité de la dysarthrie (Duffy, 2005).

Les renseignements recueillis lors de l'interrogatoire et l'auto-évaluation faite par le Speech Handicap Index, permettront de connaître les besoins, les objectifs et la motivation du patient qui détermineront le rythme et la durée de la prise en charge. Dans le domaine de la dysarthrie, celle-ci sera de préférence précoce, intensive et limitée dans le temps.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Compte-rendu de bilan et projet thérapeutique

Le plan du compte-rendu de bilan d'une dysarthrie pourra correspondre aux différentes étapes de l'évaluation et s'intégrer dans l'architecture rédactionnelle de bilan défini par la nomenclature. Il comprendra ainsi plusieurs paragraphes :

- Renseignements administratifs.
- Anamnèse.
- Objectifs et besoins de communication.
- Compte-rendu des différentes étapes du bilan :
    - sévérité de la dysarthrie ;
    - analyse perceptive ;
    - réalisation phonétique ;
    - examen moteur ;
    - auto-évaluation ;
    - relevés instrumentaux.
- Diagnostic orthophonique.
- Projet rééducatif.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

IV. Présentation détaillée et consignes de passation de la batterie

1. Renseignements généraux, médicaux et contextuels

La première page du fichier de cotation permet de noter les renseignements administratifs, médicaux, personnels et contextuels. Ceux-ci seront recueillis par l'interrogatoire du patient et/ou de son entourage et à travers les éléments transmis par le dossier médical.

Les renseignements médicaux concernent le diagnostic médical, la date d'apparition des troubles, éventuellement l'évolution. Ils sont complétés par la recherche des troubles associés (comportement, cognition, langage, déglutition), et des facteurs associés (audition, denture, vision, posture) pouvant avoir une influence sur la dysarthrie ou la prise en charge. Ces troubles et facteurs associés peuvent être avérés ou suspectés lors du bilan.

Les renseignements personnels et contextuels sont obtenus par des questions permettant au patient d'exprimer sa plainte, ainsi que ses objectifs concernant la prise en charge. On notera les besoins en communication et la motivation du patient. Il conviendra de garder à l'esprit que ces éléments sont des variables individuelles qui ne sont pas obligatoirement corrélées à la sévérité du trouble et dont il faudra tenir compte dans le projet thérapeutique. Par exemple, un trouble isolé de la résonance nasale peut être insupportable pour un enseignant alors qu'une dysarthrie sévère sera peu handicapante pour une personne à la retraite, vivant seule.

La conversation semi-dirigée qui s'installe lors de ce temps initial du bilan permet à l'examinateur d'avoir une première approche perceptive globale des caractéristiques de la parole du patient. Si un enregistrement audio ou vidéo est effectué à ce moment, il pourra être utilisé ultérieurement pour remplir la grille perceptive.

La rubrique « divers » permet de noter les examens complémentaires souhaités et/ou réalisés.

Sur la plateforme, ces éléments peuvent être rentrés en créant la fiche du patient.

2. Le score perceptif

Le Score Perceptif (SP) permet de quantifier la sévérité de la dysarthrie, en prenant en compte toutes les dimensions de la parole. Il est calculé par la somme obtenue à 5 items permettant de décrire une parole pathologique : 4 items catégoriels concernent la qualité vocale, la réalisation phonétique, la prosodie, l'intelligibilité et 1 item global concerne le caractère naturel de la parole. Ces items font partie de la grille perceptive présentée plus loin.

Consignes

Le score perceptif peut être renseigné soit par une approche globale, soit en reportant les scores obtenus aux items 1 – 13 – 19 – 34 – 35 après la passation de la grille d'analyse perceptive.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Cotation

Chaque item est coté de 0 à 4 :

0 : pas d'anomalie.
1 : anomalie discrète, uniquement repérable par un examinateur averti, ou anomalie rare.
2 : anomalie modérée ou anomalie occasionnelle.
3 : anomalie marquée ou anomalie fréquente.
4 : anomalie sévère ou anomalie quasi permanente.

La somme des 5 items notés sur 4 donne le score perceptif noté sur 20. Plus le score est élevé, plus le degré de sévérité de la dysarthrie est important (0 = parole normale, 20 = dysarthrie très sévère).

Dans le cahier de cotation informatisé, des menus déroulants proposent le choix des notes de 0 à 4 pour chaque item et le calcul du score total est automatique.

3. Le Score d'Intelligibilité

Le Score d'Intelligibilité (SI) permet de coter l'intelligibilité sur des tâches de lecture de 10 mots, lecture de 10 phrases et sur un échantillon de parole spontanée. Chaque épreuve est notée sur 8 pour un score d'intelligibilité total sur 24.

3.1 Compréhension de mots

Matériel

L'examinateur dispose de deux jeux de 50 cartes sur lesquelles sont écrits les mots. La série A correspond aux noms utilisés dans la version initiale ; la série B correspond aux noms déterminés pour la passation auprès des enfants. (Voir listes de mots en annexe 1).

Consignes

Mélanger les cartes et les placer face écrite (non visible) sur la table. Choisir 12 cartes au hasard sans les regarder. Montrer chaque carte au patient sans la lire. Le patient lit la carte puis l'examinateur vérifie si le mot identifié correspond au mot lu. Les deux premières cartes sont des cartes d'entraînement, les 10 autres servent à la cotation. Lorsque le patient a lu l'ensemble des cartes, comptabiliser le nombre de mots correctement interprétés et établir le score selon le système de cotation ci-dessous.

Cotation

- **8 points** : les 10 mots sont correctement identifiés par l'examinateur et facilement intelligibles, la parole est normale.
- **6 points** : les 10 mots sont correctement identifiés par l'examinateur, mais il doit être particulièrement attentif pour écouter et interpréter ce qu'il entend.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

- **4 points** : 7 à 9 mots sont identifiés correctement.
- **2 points** : 5 mots sont identifiés correctement.
- **0 point** : 2 mots ou moins sont identifiés correctement.

Des scores intermédiaires sont possibles. On pourra ainsi attribuer 7 points lorsque les 10 mots sont compris, sans attention particulière de l'auditeur, mais la parole est anormale ; 3 points pour 6 mots interprétés correctement ; 1 point pour 3 et 4 mots interprétés correctement.

3.2 Compréhension de phrases

Matériel

L'examinateur dispose de deux jeux de 50 cartes sur lesquelles sont inscrites des phrases comportant un groupe sujet identique « L'enfant » suivi d'un verbe. Le contexte phonétique (co-articulation) est donc différent par rapport à l'épreuve de compréhension de mots et justifie les deux types d'épreuves. La série A correspond aux phrases utilisées dans la version initiale ; la série B correspond aux phrases déterminées pour la passation auprès des enfants. (Voir listes de phrases en annexe 1).

Les consignes de passation sont identiques à celles de l'épreuve de mots.

Cotation

- **8 points** : les 10 phrases sont correctement identifiées par l'examinateur et facilement intelligibles, la parole est normale.
- **6 points** : les 10 phrases sont correctement identifiées par l'examinateur, mais il doit être particulièrement attentif pour écouter et interpréter ce qu'il entend.
- **4 points** : 7 à 9 phrases sont identifiées correctement.
- **2 points** : 5 phrases sont identifiées correctement.
- **0 point** : 2 phrases ou moins sont identifiées correctement.

Des scores intermédiaires sont possibles. On pourra ainsi attribuer 7 points lorsque les 10 phrases sont comprises, l'intelligibilité est conservée mais la parole est anormale ; 3 points pour 6 phrases interprétées correctement ; 1 point pour 3 et 4 phrases interprétées correctement.

3.3 Conversation

Engager la conversation avec le patient pendant quelques minutes. Le questionner sur son travail, ses loisirs, sa famille, etc.

Cotation

- **8 points** : pas d'anomalie, la parole est normale.
- **6 points** : la parole est intelligible mais est anormale ; le patient doit occasionnellement répéter.
- **4 points** : la parole est sévèrement altérée, pouvant être comprise la moitié du temps. Le patient doit très fréquemment répéter.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

- **2 points** : quelques mots sont déchiffrables occasionnellement.
- **0 point** : la parole est totalement inintelligible.

Les scores intermédiaires sont possibles. On pourra attribuer 7 points quand la parole est intelligible mais a perdu son caractère naturel ; 5 points quand plusieurs mots échappent à la compréhension et sont devinés grâce au contexte, le patient doit répéter souvent ; 3 points quand des portions du discours sont comprises sans pour autant permettre une compréhension globale de l'énoncé.

Dans le cahier de cotation informatisé, des menus déroulants permettent de choisir le nombre d'items compris pour les mots et les phrases ainsi que la cotation de la conversation. Les notes et le score total du SI sur 24 sont affichés automatiquement.

Pour mémoire :

SI total de 23 à 18 : dysarthrie légère ;

SI total de 17 à 12 : dysarthrie modérée ;

SI total de 11 à 7 : dysarthrie sévère ;

SI total 6 à 0 : dysarthrie massive avec une parole inintelligible.

Le Score d'Intelligibilité pour les enfants non-lecteurs

Pour les enfants non-lecteurs, les épreuves d'intelligibilité de mots et de phrases seront passées en répétition à l'aide des fichiers audio disponibles sur la plateforme. L'enfant sera muni d'un casque ou d'écouteurs et l'examinateur veillera à ne pas entendre les items ce qui altérerait la fiabilité de l'épreuve.

Consignes

Sur la page Score d'Intelligibilité de la plateforme, l'examinateur choisit le bouton « version audio ». Un choix aléatoire de douze mots puis de douze phrases est proposé pour la répétition. Au clic sur l'icône du haut-parleur, l'enfant entend le mot (ou la phrase) qu'il doit répéter. Après sa répétition, l'examinateur clique sur l'icône de visualisation, le mot (ou la phrase) entendu(e) s'affiche. Si le mot (ou la phrase) affiché(e) correspond à ce qui a été identifié, l'examinateur clique sur [+], sinon clique sur [-]. Les scores sont calculés automatiquement.

L'épreuve de conversation pourra être remplacée par la restitution d'une histoire entendue avec support d'images séquentielles, afin de faciliter l'expression orale et obtenir un échantillon de parole suffisamment important.

Les cotations restent identiques.

4. Le Test Phonétique d'Intelligibilité (TPI)

Le Test Phonétique d'Intelligibilité (TPI) est fondé sur le repérage des altérations phonétiques conduisant à une identification erronée des phonèmes et donc à une réduction de l'intelligibilité. Cet outil permet d'une part une mesure quantitative de l'intelligibilité et donc de la sévérité de la dysarthrie ; d'autre part, une analyse qualitative des perturbations phonétiques. Sa passation est simple et rapide.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Le TPI est un test d'identification de 52 mots bi-syllabiques en choix multiple. Ces 52 mots représentent 13 séries de contrastes phonétiques. Chaque contraste est testé 8 fois. Pour chacun des mots lus, 4 mots sont proposés : le mot cible et trois distracteurs. Les mots sont alternativement en situation de cibles et de distracteurs ; chaque série est donc présentée 4 fois. Dans chacune des séries, les mots s'opposent deux à deux par un ou deux contrastes phonétiques. Par exemple, dans la série « donna-donnant-tonna-tonnant », le contraste consonne initiale sourde/sonore distingue « donna » et « tonna » ainsi que « donnant » et « tonnant » ; et le contraste voyelle orale/voyelle nasale distingue « donna » et « donnant » et « tonna » et « tonnant ».

Les contrastes phonétiques et les séries correspondantes sont présentés en annexe 2.

Matériel

Pour éviter l'effet d'apprentissage des listes par l'examinateur, le TPI est proposé en cinq versions différentes numérotées de 1 à 5 et qui se distinguent entre elles uniquement par l'ordre de passation des 52 items. Il convient donc de ne pas toujours utiliser la même version.

L'examinateur dispose de cinq carnets numérotés de 1 à 5, contenant les 52 mots (un mot par page) que le patient doit lire à haute voix, auxquels correspondent cinq fichiers de cotation numérotés de 1 à 5.

Consignes

Sur la plateforme, sélectionner une des cinq versions du TPI et choisir le carnet correspondant. Demander au patient de lire chaque mot une seule fois à voix haute. L'examinateur tourne les pages lui-même pour donner le rythme de lecture ou signale au patient quand il peut le faire lui-même. Nous recommandons de ne pas regarder le patient lorsqu'il lit pour privilégier l'écoute (sans lecture labiale par exemple).

À chaque mot lu, l'examinateur clique sur le mot compris parmi les 4 mots proposés à l'écran, ou à défaut sur celui qui se rapproche le plus de ce qui a été compris. La cotation s'effectuant en choix forcé, l'examinateur doit impérativement sélectionner un des quatre mots proposés pour chaque item. Au clic, le passage à l'item suivant est automatique.

Cotation

La passation du TPI permet de donner un score quantitatif et un score qualitatif en recensant les erreurs pour les contrastes phonétiques étudiés.

Le score quantitatif sur 52 est calculé en comptabilisant le nombre de mots compris correctement : parole normale = 52/52 ; plus le score est bas, plus l'intelligibilité est réduite.

En cas de mauvaise identification du mot, et en se référant à la liste des contrastes phonétiques, l'examinateur peut mener une analyse qualitative. Il obtient un score sur 8 pour chaque contraste phonétique : plus le score est élevé, plus les altérations sont nombreuses pour le contraste en question.

La cotation informatisée fournit automatiquement le score quantitatif sur 52 et les sous-scores d'erreurs sur 8 pour chaque contraste étudié.

Sans accès à la plateforme informatique, la passation du TPI peut être réalisée grâce aux feuilles de cotation et aux feuilles de correction disponibles au format PDF et téléchargeables. Les feuilles de correction devront être imprimées sur un support transparent, à la taille réelle.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

5. Analyse perceptive

5.1 Analyse perceptive globale

Dans cet espace, l'examinateur peut noter les caractéristiques perceptives prédominantes recueillies par une approche globale dès le premier contact avec le patient. Les différentes étapes du bilan, la conversation spontanée, la lecture du texte pourront servir de support pour noter les caractéristiques prédominantes et les hiérarchiser. Pour cela, on peut se poser la question « Quels sont les quatre mots que je choisirais pour qualifier cette dysarthrie ? »

Dans le cahier de cotation informatisé, un menu déroulant propose des choix pour guider l'analyse de l'examinateur.

5.2 La grille perceptive

La grille perceptive propose une analyse systématique portant sur une liste de 35 critères définis en quantifiant le degré de perturbation à l'aide d'une échelle de cotation en 5 points.

Les critères sont de trois niveaux de hiérarchie : global (1 item), catégoriel (4 items) et spécifique (30 items). Ils sont ainsi classés :

- qualité vocale : 1 item catégoriel (1) et 11 items spécifiques (2 à 12) ;
- réalisation phonétique : 1 item catégoriel (13) et 5 items spécifiques (14 à 18) ;
- prosodie : 1 item catégoriel (19) et 11 items spécifiques (20 à 30) ;
- respiration : 3 items spécifiques (31 à 33) ;
- intelligibilité : 1 item catégoriel (34) ;
- caractère naturel de la parole : 1 item global (35).

À titre d'exemple, l'item catégoriel « Prosodie » évaluera l'ensemble des caractéristiques supra-segmentales de la parole : organisation temporelle du discours (débit et rythme) ; modulations de hauteur et modulations d'intensité ; les 11 items spécifiques de cette catégorie évalueront les traits prosodiques de façon plus analytique.

La grille perceptive est présentée en annexe 3 et les définitions des critères sont précisées en annexe 3.1.

Consignes

La grille peut être complétée en utilisant comme support la conversation libre ou la lecture du texte extrait de « Pierrot » de Guy de Maupassant. La cotation est difficilement réalisable en direct. Il est conseillé de remplir la grille à partir d'un enregistrement audio ou vidéo qui permet plusieurs écoutes pour une analyse plus systématique, sans risque de perte d'informations. Ces enregistrements sont par ailleurs utiles pour une comparaison post-rééducation.

Cotation

La cotation de chacun des items se fait à l'aide de la définition des critères et avec une échelle de cotation en 5 points de 0 à 4 :

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

0 = pas d'anomalie
1 = anomalie discrète, uniquement repérable par un examinateur averti, ou anomalie rare
2 = anomalie modérée ou anomalie occasionnelle
3 = anomalie marquée ou anomalie fréquente
4 = anomalie sévère ou anomalie quasi permanente

Pour 3 critères (hauteur, intensité, débit), l'anomalie peut varier dans deux directions. En plus de la cotation chiffrée, un signe additionnel [+] ou [-] définit conventionnellement le sens de la variation. (Hauteur de la voix jugée trop aiguë [+], trop grave [-] ; intensité augmentée [+], diminuée [-] ; débit accéléré [+], ralenti [-]).

Certains critères s'opposant, ils ne peuvent être cotés chez un même patient (par exemple voix soufflée/voix étranglée).

Les critères spécifiques sont cotés en priorité et l'ensemble de leurs résultats permet de coter les critères de niveau catégoriel en calculant une moyenne (total des scores aux items spécifiques/nombre d'items cotés).

Dans le cahier de cotation informatisé, des menus déroulants proposent les notes pour chaque critère spécifique et la note du critère catégoriel est calculée automatiquement.

À l'issue de la passation de la grille, l'examinateur peut établir le score perceptif en reportant les notes des 4 critères catégoriels et du critère global (items 1 – 13 – 19 – 34 – 35).

Pour les enfants dysarthriques, la grille peut être utilisée et cotée sur des échantillons de parole spontanée ou à partir de la restitution d'une histoire entendue préalablement avec support imagé, afin de faciliter les productions orales et de recueillir un échantillon de parole suffisamment important.

6. Analyse phonétique

Trois épreuves de complexité croissante sont proposées pour recenser les perturbations phonétiques en utilisant comme support des répétitions de phonèmes, syllabes et mots. Selon la sévérité de la dysarthrie, la fatigabilité du patient, il sera opportun de ne pas faire passer les trois épreuves dans leur totalité, mais de sélectionner celles qui paraissent les plus pertinentes. Dans le cas d'une dysarthrie sévère, seule l'épreuve de réalisation phonétique isolée sera proposée. Dans le cas d'une dysarthrie légère, on pourra proposer uniquement l'épreuve de mots complexes.

Matériel

6.1 Réalisation phonétique isolée

Cette épreuve permet de tester le tableau articulatoire du patient. (Voir annexe 3). Elle comprend :

- la répétition des voyelles présentées de façon isolée,
- la répétition des semi-voyelles associées aux phonèmes /a/ ou /i/,
- la répétition des consonnes associées au phonème /a/.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

La production de 33 phonèmes est ainsi évaluée et peut donner un score d'erreurs sur 33. Certaines distinctions (celles existant entre le /a/ de « bas » et le /ɑ/ de « âne » ; le /ɛ̃/ de « brin » et le /œ̃/) de « brun » ; le /ə/ de « le » et le /ø/ de « feu ») n'ont pas été retenues car elles ne nous semblent pas pertinentes dans l'évaluation d'une dysarthrie et sont de moins en moins prégnantes dans le langage oral. Le /ŋ/ final de parking n'est pas coté.

6.2 Mots simples

Cette épreuve propose une liste de 37 mots sélectionnés pour leur structure phonétique simple (consonne-voyelle ou voyelle-consonne), leur fréquence (aucun mot rare), leur longueur (une, deux syllabes pour la quasi-totalité). Ils ont été choisis également pour permettre de présenter les 30 phonèmes vocaliques et consonantiques en position initiale, médiane et finale dans une liste aussi courte que possible. Les douze premiers mots comportent les 30 phonèmes dans au moins une position. Deux phonèmes ne sont pas retrouvés en position finale (/œ/ et /ɔ/), cela conduit à tester 88 possibilités et peut donner un score d'erreurs sur 88.

6.3 Mots complexes

Cette épreuve propose une liste de 25 mots qualifiés de complexes car ils contiennent des semi-voyelles et un ou plusieurs groupes consonantiques de deux, trois ou quatre consonnes. Ils sont proposés par ordre de complexité croissante. Au total, la réalisation de 30 semi-voyelles et groupes consonantiques est testée et peut donner un score d'erreurs sur 30.

Les phonèmes, mots simples et mots complexes sont présentés sur écran avec leur traduction en Alphabet Phonétique International (API), les phonèmes sur fond rouge sont les cibles testées.

Consignes et cotation

Demander au patient de répéter une seule fois les phonèmes, syllabes et mots.

Pour l'épreuve de réalisation phonétique isolée, cliquer sur les phonèmes altérés pour obtenir le nombre d'erreurs sur 33 et définir l'altération. Plus le score est élevé, plus les perturbations sont nombreuses.

Pour l'épreuve de mots simples, cliquer sur les phonèmes sur fond rouge altérés pour obtenir le nombre d'erreurs sur 88 et définir l'altération. Plus le score est élevé, plus les perturbations sont nombreuses.

Pour l'épreuve des mots complexes, cliquer sur les semi-voyelles ou groupes consonantiques sur fond rouge perturbés pour obtenir le nombre d'erreurs sur 30 et définir l'altération. Plus le score est élevé, plus les perturbations sont nombreuses.

Un tableau est proposé pour permettre de récapituler les altérations et de qualifier la nature globale des perturbations.

Dans le cahier de cotation informatisé, le nombre d'erreurs est calculé automatiquement ; en cas d'erreur, un choix s'affiche et permet de qualifier le type de perturbation.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

7. Examen de la motricité

La grille d'examen moteur permet d'évaluer le fonctionnement des organes concernés dans la production de la parole : respiration, larynx, vélo-pharynx, langue, lèvres, joues, mâchoires. Elle propose également une observation de la posture, de la face, ainsi que la réalisation des mouvements alternatifs. Elle est donc constituée de 10 parties. Entre parenthèses, un rappel indique les structures et nerfs crâniens assurant le contrôle neurologique de la fonction ou de l'organe observé.

Dans le cadre d'un bilan de dysarthrie, l'évaluation de la motricité lors de la parole est à privilégier. Selon la pathologie concernée, le type de dysarthrie, le degré de sévérité et les renseignements recueillis lors des épreuves précédentes, l'examinateur pourra éventuellement choisir dans cette grille motrice les organes à évaluer. De même, il pourra sélectionner les tâches à effectuer, ajouter d'autres tâches qui ne sont pas proposées ici mais qui lui semblent adaptées.

La grille motrice permet de déterminer un score pour chacun des domaines évalués, en modalités verbale et non verbale, coté de 0 (normal) à 4 (très altéré).

Matériel nécessaire (non fourni)

L'examinateur aura à sa disposition : des gants, un abaisse-langue, une lampe frontale ou une lampe de poche, un chronomètre, un verre de plus de 12 cm de hauteur, gradué, une paille.

Consignes

Pour chaque domaine, des tâches sont proposées. Pour chacune d'elles, l'examinateur effectue une démonstration. Le patient réalise les tâches une fois la démonstration terminée. Elles peuvent être réalisées deux fois, la cotation s'effectue alors sur le second essai. Certaines tâches non verbales nécessitent la production d'un phonème pour observer le mouvement, il ne s'agit cependant pas de situation de parole vraie.

Dans chaque partie, les observations et les altérations les plus fréquemment rencontrées sont proposées mais les listes ne sont pas exhaustives. La grille complète est présentée en annexe 5. Les consignes détaillées pour chaque domaine et dans les deux modalités sont exposées en annexe 5.1.

Cotation

Pour chaque tâche demandée, l'examinateur note [+], [-] ou [+/-]. Selon les résultats, il attribue ensuite un score global avec une échelle de cotation en 5 points de 0 à 4 pour le domaine évalué et dans la modalité concernée.

0 = pas d'anomalie
1 = anomalie discrète, uniquement repérable par un examinateur averti ou anomalie rare
2 = anomalie modérée ou anomalie occasionnelle
3 = anomalie marquée ou anomalie fréquente
4 = anomalie sévère ou anomalie quasi permanente

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Les items d'observation en conversation peuvent être cotés lors d'une discussion, avant ou après la passation des autres tâches.

8. Le Speech Handicap Index (SHI)

Ce test d'auto-évaluation est constitué d'une grille comportant 30 items répartis dans deux domaines : parole (P) et psychosocial (Ps), présentée en annexe 6.

Consignes

Le patient complète seul la grille après avoir reçu toutes les explications nécessaires sur la façon de procéder. Pour chacune des 30 affirmations proposées, cinq réponses sont proposées à son choix : Jamais - Presque jamais - Parfois - Presque toujours - Toujours.

Le patient choisit la réponse correspondant le mieux à son ressenti pour chaque affirmation. Il est indispensable de s'assurer que le patient a bien compris la consigne et de lui préciser que seules ses impressions personnelles sont à prendre en compte pour compléter cette grille.

Cotation

Chaque item est coté par l'attribution de points selon le choix :

- Jamais = 0 point
- Presque jamais = 1 point
- Parfois = 2 points
- Presque toujours = 3 points
- Toujours = 4 points

Le score total est obtenu en effectuant la somme des points. Le score maximum est de 120. Plus le score est élevé, plus le handicap est important.

Un sous-score sur 56 peut être déterminé pour le ressenti concernant la parole (items 1-2-3-5-6-9-10-13-15-18-20-21-26-28) ainsi qu'un sous-score sur 56 concernant le ressenti psycho-social (items 4-7-8-11-12-14-16-17-19-24-25-27-29-30). Les items 22 et 23 ne sont pas classés dans un des deux sous-domaines.

La grille peut être remplie sur écran directement, en s'assurant que les possibilités de manipulations de la souris sont adaptées. Si ce n'est pas le cas, la grille sera remplie sur papier puis reportée dans le cahier informatisée par l'examinateur. En cas de déficits moteurs, on peut éventuellement aider le patient en ayant soin de rester neutre à l'énoncé de ses choix.

Dans le cahier de cotation informatisée, le score total et les sous-scores parole et psycho-social sont calculés automatiquement.

L'examinateur posera ou fera lire ensuite la question : « Comment trouvez-vous votre parole aujourd'hui ? ». Le patient choisira sa réponse parmi quatre propositions : excellente, bonne, moyenne, mauvaise.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

9. Relevés instrumentaux

La micro-informatique et plusieurs logiciels ou applications permettent maintenant aux orthophonistes de disposer de mesures concernant l'analyse acoustique de la parole. Dans ce domaine, les pratiques sont variables et difficilement harmonisables. Les logiciels ne fournissent pas tous des données absolument exactes, mais ils sont suffisamment fiables pour que les enregistrements servent de référence pour estimer l'évolution de la parole, ou pour évaluer l'efficacité de la rééducation. Il faut veiller à effectuer les mesures dans des conditions identiques (réglage carte son, distance micro…).

Nous avons donc choisi de réserver un emplacement pour noter quelques relevés instrumentaux. Ces relevés peuvent également être réalisés avec un décibelmètre et un sonomètre.

Nous proposons de noter des relevés concernant l'intensité et la hauteur.

Consignes

- Mesurer l'intensité moyenne en voix conversationnelle sur une production libre.
- Demander au patient de répéter une phrase simple et courte avec une voix faible, puis avec une voix normale, puis avec une voix forte, et avec une voix très forte. La phrase utilisée peut être « *Nous sommes à…..* ». Si la dysarthrie est majeure, on pourra effectuer les relevés sur des tenues vocaliques.
- Mesurer la fréquence fondamentale (F0) sur une production libre ou au cours d'une lecture.
- L'étendue vocale pourra être mesurée avec un ou plusieurs des exercices suivants, selon les possibilités du patient :
  - Faire réaliser une sirène ascendante, puis une sirène descendante, après avoir donné un modèle.
  - Demander de produire la note la plus aiguë possible, puis la plus grave possible après avoir donné un modèle.
  - Demander de réaliser des montées et des descentes par degrés après avoir donné un modèle.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 1

Matériel Score d'Intelligibilité

Mots Série A

abri	chaîne	grain	nain	pont
appris	chèvre	grand	noix	porc
arbre	clenche	gras	pain	prend
boxeur	cloche	groin	paire	sport
brin	craie	gros	paon	tasse
caché	crin	grue	paquet	taxe
café	croche	herbe	parquet	thym
calé	docteur	main	peur	toi
carré	franc	moi	plein	toile
cassé	frein	mouche	pleur	train

Mots série B

bain	craie	gomme	pain	rue
bouche	cri	gris	peau	sang
boue	doigt	joue	père	sieste
bras	drap	jour	peur	tache
chant	faim	lion	pierre	tasse
chat	faux	long	pluie	temps
ciel	fée	lui	pont	tête
clé	feuille	nez	pré	toi
clown	fleur	nid	roue	vrai
coquin	froid	nuit	rouge	zoo

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Phrases série A

L'enfant abrite.	L'enfant dicte.	L'enfant fraude.	L'enfant accorde.
L'enfant dîne.	L'enfant frôle.	L'enfant accourt.	L'enfant dispense.
L'enfant habite.	L'enfant achète.	L'enfant disperse.	L'enfant paraît.
L'enfant achève.	L'enfant dispose.	L'enfant paresse.	L'enfant arbitre.
L'enfant dit.	L'enfant parle.	L'enfant asperge.	L'enfant doit.
L'enfant part.	L'enfant aspire.	L'enfant donne.	L'enfant passe.
L'enfant assiège.	L'enfant écarte.	L'enfant pense.	L'enfant baigne.
L'enfant éclate.	L'enfant permet.	L'enfant baille.	L'enfant écrase.
L'enfant plie.	L'enfant boit.	L'enfant écrit.	L'enfant plisse.
L'enfant comprend.	L'enfant empile.	L'enfant se couche.	L'enfant coupe.
L'enfant emporte.	L'enfant se douche.	L'enfant court.	L'enfant expire.
L'enfant se mouche.	L'enfant décrit.	L'enfant forge.	L'enfant voit.
L'enfant dépose.	L'enfant forme.		

Phrases série B

L'enfant aide.	L'enfant goûte.	L'enfant s'assoit.	L'enfant aime.
L'enfant jette.	L'enfant sert.	L'enfant boit.	L'enfant laisse.
L'enfant sort.	L'enfant bricole.	L'enfant lance.	L'enfant souffle.
L'enfant cache.	L'enfant lit.	L'enfant suit.	L'enfant casse.
L'enfant mange.	L'enfant tape.	L'enfant coupe.	L'enfant marche.
L'enfant touche.	L'enfant court.	L'enfant mord.	L'enfant tourne.
L'enfant croit.	L'enfant nettoie.	L'enfant vend.	L'enfant danse.
L'enfant ouvre.	L'enfant voit.	L'enfant déchire.	L'enfant parle.
L'enfant dessine.	L'enfant passe.	L'enfant dit.	L'enfant perd.
L'enfant doit.	L'enfant porte.	L'enfant donne.	L'enfant pleure.
L'enfant dort.	L'enfant prend.	L'enfant écoute.	L'enfant range.
L'enfant écrit.	L'enfant regarde.	L'enfant essuie.	L'enfant rend.
L'enfant ferme.	L'enfant rit.		

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 2

TPI : contrastes phonétiques étudiés

	Contrastes phonétiques	Séries
A	Voyelle antérieure/postérieure	• la rue - la roue - la russe - la rousse • début - débute - des bouts - déboute
B	Voyelle ouverte/fermée	• des bains - dix bains - des mains - dix mains • si oui - ses oui - scierie - série
C	Voyelle labiale/non labiale	• le fer - le verre - les fers - les verres • deux pots - des pots - deux faux - des faux
D	Voyelle orale/nasale	• donna - donnant - tonna - tonnant • des thons - des taux - des sons - des seaux
E	Consonne initiale sourde/sonore	• donna - tonna - donnant - tonnant • tes doigts - tes draps - des doigts - des draps
F	Consonne médiane sourde/sonore	• le fer - le verre - les fers - les verres • l'étal - les dalles - les cales - légal
G	Changement de lieu d'articulation des fricatives	• tracer - trachée - tasser - tâcher • masser - mâcher - basset - bâcher
H	Changement de lieu d'articulation des occlusives	• roter - roquet - crotter - croquer • l'étal - les cales - les dalles - légal
I	Occlusive/fricative	• des thons - des sons - des taux - des seaux • deux pots - deux faux - des pots - des faux
J	Occlusive/nasale	• basset - masser - bâcher - mâcher • des bains - des mains - dix bains - dix mains
K	/R/-/w/	• scierie - si oui - série - ses oui • des draps - des doigts - tes doigts - tes draps
L	Initiale complexe/simple	• crotter - roter - croquer - roquet • tracer - tasser - trachée - tâcher
M	Consonne finale/élision	• la rue - la roue - la russe - la rousse • début - débute - des bouts - déboute

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 3

Grille d'analyse perceptive

0 : pas d'anomalie.
1 : anomalie discrète, uniquement repérable par un examinateur averti, ou anomalie rare.
2 : anomalie modérée ou anomalie occasionnelle.
3 : anomalie marquée ou anomalie fréquente.
4 : anomalie sévère ou anomalie quasi permanente.
Le signe [-] signifie grave, faible ou lent ; le signe [+] aigu, fort ou rapide.

	Qualité vocale	
1	Qualité vocale	
2	Hauteur (+/-)	
3	Intensité (+/-)	
4	Voix rauque, enrouée	
5	Voix éraillée	
6	Voix soufflée, voilée	
7	Voix mouillée, humide, gargouillante	
8	Voix étranglée, hyperfonctionnelle	
9	Tremblements dans la voix	
10	Voix avec fry	
11	Dégradation de l'intensité vocale	
12	Périodes d'aphonie	
	Réalisation phonétique	
13	Réalisation phonétique	
14	Modification de la durée des phonèmes	
15	Imprécision des consonnes	
16	Imprécision des voyelles	
17	Trouble de la résonance nasale	
18	Souffle nasal	

	Prosodie	
19	Prosodie	
20	Perte des modulations de la hauteur	
21	Fluctuations excessives de la hauteur	
22	Perte des modulations de l'intensité vocale	
23	Fluctuations excessives de l'intensité vocale	
24	Débit (+/-)	
25	Fluctuations du débit	
26	Accélérations paroxystiques du débit	
27	Ruptures de la fluence	
28	Silences inappropriés	
29	Palilalies	
30	Difficultés d'initiation motrice	
	Respiration	
31	Inspirations-expirations forcées	
32	Bruit inspiratoire	
33	Contrôle respiratoire	
	Intelligibilité – Caractère naturel de la parole	
34	Intelligibilité	
35	Caractère naturel de la parole	

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 3

Grille d'analyse perceptive

0 : pas d'anomalie.
1 : anomalie discrète, uniquement repérable par un examinateur averti, ou anomalie rare.
2 : anomalie modérée ou anomalie occasionnelle.
3 : anomalie marquée ou anomalie fréquente.
4 : anomalie sévère ou anomalie quasi permanente.
Le signe [-] signifie grave, faible ou lent ; le signe [+] aigu, fort ou rapide.

	Qualité vocale	
1	Qualité vocale	
2	Hauteur (+/-)	
3	Intensité (+/-)	
4	Voix rauque, enrouée	
5	Voix éraillée	
6	Voix soufflée, voilée	
7	Voix mouillée, humide, gargouillante	
8	Voix étranglée, hyperfonctionnelle	
9	Tremblements dans la voix	
10	Voix avec fry	
11	Dégradation de l'intensité vocale	
12	Périodes d'aphonie	
	Réalisation phonétique	
13	Réalisation phonétique	
14	Modification de la durée des phonèmes	
15	Imprécision des consonnes	
16	Imprécision des voyelles	
17	Trouble de la résonance nasale	
18	Souffle nasal	

	Prosodie	
19	Prosodie	
20	Perte des modulations de la hauteur	
21	Fluctuations excessives de la hauteur	
22	Perte des modulations de l'intensité vocale	
23	Fluctuations excessives de l'intensité vocale	
24	Débit (+/-)	
25	Fluctuations du débit	
26	Accélérations paroxystiques du débit	
27	Ruptures de la fluence	
28	Silences inappropriés	
29	Palilalies	
30	Difficultés d'initiation motrice	
	Respiration	
31	Inspirations-expirations forcées	
32	Bruit inspiratoire	
33	Contrôle respiratoire	
	Intelligibilité – Caractère naturel de la parole	
34	Intelligibilité	
35	Caractère naturel de la parole	

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 3.1

Définitions des critères de la grille perceptive

Le critère de niveau global et les quatre critères de niveau catégoriel sont soulignés.

- *Critères relatifs à la qualité vocale*

1. <u>Qualité vocale</u> : la qualité de la voix est altérée. L'altération porte sur un ou plusieurs paramètres (hauteur, intensité, timbre) et est plus ou moins marquée, donnant une impression globale de sévérité de l'atteinte vocale.

2. Hauteur (+/-) : la hauteur de la voix est anormalement basse (voix grave) ou anormalement élevée (voix aiguë) pour l'âge et le sexe du sujet. Une voix anormalement basse est cotée par le signe - ; une voix anormalement haute par le signe +.

3. Intensité (+/-) : l'intensité de la voix est anormalement faible ou anormalement forte compte tenu de la situation. Une voix anormalement faible est cotée par le signe -, une voix anormalement forte par le signe +.

4. Voix rauque, enrouée : la voix peut être décrite comme rauque, rugueuse et enrouée. Bruit non périodique avec basse fréquence, probablement en rapport avec un certain type de vibrations irrégulières des cordes vocales.

5. Voix éraillée : la voix peut être décrite comme stridente, métallique ou grinçante. Bruit non périodique avec haute fréquence, probablement en rapport avec un certain type de vibrations irrégulières des cordes vocales.

6. Voix soufflée, voilée : la phonation est accompagnée par un souffle permanent ou intermittent traduisant une fuite d'air à travers la glotte.

7. Voix mouillée, humide, gargouillante : la voix est décrite comme gargarisée. Elle donne l'impression qu'un liquide vibre en même temps que les cordes vocales.

8. Voix étranglée, hyperfonctionnelle : les sons sont étouffés, atténués. La voix donne l'impression d'une glotte très étroite, de cordes vocales comprimées pendant la phonation, hypertendues et peu mobiles.

9. Tremblement de la voix : la voix présente un chevrotement ou des tremblements.

10. Voix avec fry : La voix peut être comparée aux craquements ou grincements d'une vieille porte rouillée que l'on ouvre. Ces grincements sont en rapport avec une vibration périodique de basse fréquence.

11. Dégradation de l'intensité vocale : le patient éprouve des difficultés à maintenir un volume sonore suffisant. Le volume sonore diminue progressivement au cours des segments de parole ou au cours du discours.

12. Périodes d'aphonie : la parole présente des épisodes sans phonation. Il y a des moments de chuchotements ou de perte de voix.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

- *Critères relatifs à la réalisation phonétique*

13. Réalisation phonétique : La production des consonnes (articulation) et/ou des voyelles (résonance) manque de netteté et de précision ; on note un trouble de la résonance nasale. Les altérations influent de façon plus ou moins marquée sur l'intelligibilité.

14. Modification de la durée des phonèmes : La durée des phonèmes est perturbée. Cette perturbation, consistant le plus souvent en un allongement du phonème, porte habituellement sur les voyelles.

15. Imprécision des consonnes : l'articulation des consonnes est imprécise. On note des distorsions, des simplifications et des omissions.

16. Imprécision des voyelles : les voyelles sont distordues. On note une réduction du triangle vocalique (par exemple mauvaise différenciation /i/ et /e/).

17. Trouble de la résonance nasale : la résonance peut être excessive, inappropriée, permanente (hypernasalité). Il peut y avoir un manque de résonance nasale (hyponasalité), ou une dénasalisation, par obstruction complète ou partielle de la cavité nasale. Les deux types d'anomalies peuvent s'associer sous la forme d'un trouble mixte de la résonance.

18. Souffle nasal : on note une fuite d'air audible par le nez lors de la parole.

- *Critères relatifs à la prosodie*

19. Prosodie : on note une altération des caractéristiques prosodiques ou supra-segmentales de la parole. Les modulations de la hauteur et/ou l'intensité sont inadéquates ; le débit est altéré. Le patient éprouve des difficultés à respecter les schémas prosodiques généraux (affirmation, exclamation, interrogation...).

20. Perte des modulations de la hauteur : la voix manque de modulations de hauteur adaptées au contexte et au sens du discours.

21. Fluctuations excessives de la hauteur de la voix : la hauteur de la voix varie brusquement. Elle est parfois trop haute, parfois trop basse compte tenu de l'âge et du sexe du sujet.

22. Perte des modulations de l'intensité vocale : la voix manque de modulations d'intensité adaptées au contexte et au sens du discours.

23. Fluctuations excessives de l'intensité vocale : l'intensité de la voix varie brusquement. Elle est parfois trop forte, parfois trop faible, compte tenu de la situation.

24. Débit (+/-) : la cotation porte sur les segments de parole (les périodes pendant lesquelles le patient parle) en faisant abstraction des pauses dans le discours. Le débit est anormalement lent ou anormalement rapide. L'intelligibilité du discours peut être perturbée. Un débit anormalement lent est coté par un signe -, un débit anormalement rapide par un signe +.

25. Fluctuations du débit : le débit est fluctuant, parfois rapide, parfois lent. Ces fluctuations ne sont pas en rapport avec le contenu du discours.

26. Accélérations paroxystiques du débit : le discours est produit par courtes phases d'accélération souvent accompagnées d'une dégradation de l'articulation.

27. Ruptures de la fluence : on note des ruptures dans la fluence du discours. Celui-ci est marqué par des pauses fréquentes réduisant la longueur des segments de parole.

28. Silences inappropriés : le discours est marqué par des pauses entre les mots, inadaptées au sens du discours et/ou des pauses anormales entre les syllabes ou les sons.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou

En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

29. Palilalie (répétition de sons, de mots, de phrases) : certains phonèmes, certains mots, certaines phrases sont répétés. Ces répétitions s'accompagnent fréquemment d'une accélération du débit, d'une diminution de l'intensité et d'une perte de la précision articulatoire. Les palilalies n'entraînent pas de rupture du discours.

30. Difficultés d'initiation motrice de la parole : on note des difficultés à initier la parole. Cette anomalie peut se traduire par un temps prolongé en position pré-articulatoire.

- *Critères relatifs à la respiration*

31. Inspirations-expirations forcées : le discours est marqué par des périodes d'inspirations et/ou d'expirations forcées.

32. Bruit inspiratoire : l'inspiration est audible (bruit, stridor).

33. Contrôle respiratoire : le contrôle respiratoire est défaillant. Le sujet peut être essoufflé ou donner l'impression de suffoquer. On note un non-respect des groupes de souffle.

- *Quatrième critère catégoriel*

34. Intelligibilité : il s'agit d'évaluer la perturbation de l'intelligibilité en tenant compte des efforts réalisés par l'auditeur pour comprendre le patient.

- *Critère global*

35. Caractère naturel de la parole : la parole du patient est perçue comme anormale. Elle a perdu son caractère naturel par une ou plusieurs perturbations altérant le plus souvent la prosodie, mais aussi la qualité vocale ou la réalisation phonétique. Les altérations peuvent être plus ou moins sévères. Un trouble discret et isolé a déjà des répercussions sur le caractère naturel de la parole.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 5

Grille d'examen moteur

0 : pas d'anomalie.
1 : anomalie discrète, uniquement repérable par un examinateur averti, ou anomalie rare.
2 : anomalie modérée ou anomalie occasionnelle.
3 : anomalie marquée ou anomalie fréquente.
4 : anomalie sévère ou anomalie quasi permanente.

POSTURE	
MODALITÉ VERBALE	*MODALITÉ NON VERBALE*
1. Observation en conversation	1. Observation au repos
o verticalité o antécolis o rétrocolis o mouvements anormaux o dystonies	o verticalité o antécolis o rétrocolis o mouvements anormaux o dystonies
Score /4	**Score /4**

RESPIRATION (tronc cérébral, moelle épinière)	
MODALITÉ VERBALE	*MODALITÉ NON VERBALE*
1. Observation en conversation	1. Observation au repos
2. Comptage jusqu'à 20	2. Souffle dans la paille T =.....
	3. Tenue du /S/
o voix sur inspiration o inspiration dans un mot o non-respect des groupes de souffle o variabilité o essoufflement o mouvements anormaux	o costo-abdominale o costale o abdominale o rythme accéléré o mobilisation des épaules o mouvements anormaux
Score /4	**Score /4**

LARYNX (X)	
MODALITÉ VERBALE	*MODALITÉ NON VERBALE*
1. Observation en conversation	1. Adduction des cordes vocales Coups de glotte - Raclements - Toux
2. Temps Maximum de Phonation sur /a/ (TMP)	2. Abduction des cordes vocales Inspiration par la bouche
3. Rapport S/Z	
4. Phonation interrompue sur /a/	
o altération de la voix o trouble de l'intensité o trouble de la hauteur o attaque brusque o présence d'un souffle	o inspiration bruyante o toux inefficace
Score /4	**Score /4**

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

VÉLOPHARYNX (IX-X-XI)			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Observation en conversation		1. Observation au repos	
2. /a/ tenu 4 secondes		2. Observation sur 5 /a/	
3. Série de /s/		o voile très bas o voile très haut o asymétrie o déviation ☐ D ☐ G	
4. Série de /p/			
5. Phrase avec phonèmes oraux			
6. Série de /a/ - /an/			
o trouble de la résonance o perception d'une fuite nasale o mouvements anormaux			
Score	**/4**	**Score**	**/4**

LANGUE (XII)			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Observation en conversation		1. Observation au repos	
2. Répétition de la syllabe /ta/		2. Protrusion	
3. Répétition de la syllabe /na/		3. Élévation de la pointe	
4. Répétition de la syllabe /la/		4. Claquements de langue (dos)	
5. Répétition de la syllabe /ka/		5. Clics de réprobation (apex)	
6. Répétition de la syllabe /ra/		6. Recul de la base de la langue	
o mouvements anormaux o interposition : ☐ dentale ☐ labiale o défaut de précision des consonnes o indifférenciation des voyelles		o atrophie o trouble du tonus ☐ D ☐ G o déviation ☐ D ☐ G o mouvements anormaux	
Score	**/4**	**Score**	**/4**

LÈVRES (VII)			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Observation en conversation		1. Observation au repos	
2. Répétition des syllabes /pa/ /ba/		2. Arrondissement	
3. Répétition des voyelles /i/ /u/ /y/		3. Pincement	
		4. Étirement	
o imprécision des voyelles o trouble de l'explosion (p, b) o incoordination des mouvements		o absence de mouvements o réduction des mouvements o déviation ☐ D ☐ G o asymétrie o trouble du tonus o mouvements anormaux	
Score	**/4**	**Score**	**/4**

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

JOUES (VII)			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Observation en conversation		1. Observation au repos	
o asymétrie : déformation ☐ D ☐ G o trouble du tonus		2. Gonflement des 2 joues	
		3. Gonflement à droite puis à gauche	
		o asymétrie : déformation ☐ D ☐ G o trouble du tonus	
Score /4		**Score /4**	

MÂCHOIRES (V)			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Observation en conversation		1. Observation au repos	
o dystonie oro-mandibulaire		2. Grande ouverture	
		o trouble de la position o déviation o tremblement	
Score /4		**Score /4**	

FACE (V - VII)			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Observation en conversation		1. Observation au repos	
o dysharmonie o asymétrie o expressivité réduite o expressivité non adaptée au discours o mouvements anormaux o regard o mouvements des paupières		o dysharmonie o déformation ☐ D ☐ G o asymétrie o expressivité réduite o mouvements anormaux o regard o mouvements des paupières	
Score /4		**Score /4**	

MOUVEMENTS ALTERNATIFS			
MODALITÉ VERBALE		*MODALITÉ NON VERBALE*	
1. Répétitions de la syllabe /pa/		1. Ouverture/fermeture de la bouche	
2. Répétition de /pata/		2. Protrusion/rétraction de la langue	
3. Répétition de /pataka/		3. Langue de gauche à droite	
o lenteur o ralentissement o accélération o irrégularité o dégradation de l'articulation o mouvements anormaux		o lenteur o ralentissement o accélération o irrégularité o déficit d'amplitude o dégradation de l'amplitude o mouvements anormaux	
Score /4		**Score /4**	

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 5.1

Examen moteur

Consignes détaillées pour chaque domaine

POSTURE

Observer la posture globale du patient au repos et en conversation. Noter la verticalité du tronc, la position de la tête (antécolis ou rétrocolis), la présence de mouvements anormaux, les dyskinésies, les dystonies cervico-faciale ou axiale. Remarquer d'éventuels changements lors de la parole.

Donner un score posture dans chaque modalité.

RESPIRATION (tronc cérébral, moelle épinière)

MODALITÉ VERBALE

1. Observation en conversation

Observer le patient au cours d'une conversation. Noter toute anomalie perçue : voix sur inspiration, inspiration dans un mot, non-respect des groupes de souffle, variabilité, essoufflement, mouvements anormaux.

2. Comptage jusqu'à 20

Demander au patient de compter rapidement jusqu'à 20 sur une seule expiration en précisant que la qualité de l'articulation n'est pas prise en compte. Normalement, cette tâche est effectuée sur une seule expiration.

Donner un score de 0 à 4 pour la respiration en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Observer le patient assis hors moment de parole. Ne pas hésiter à poser une main au niveau de la cage thoracique pour évaluer le type de respiration. Noter le type de respiration et toute information importante : rythme accéléré ou variable ; mobilisation des épaules, du cou ; posture anormale réduisant les mouvements de la cage thoracique, du diaphragme ; respiration bruyante ; mouvements anormaux…

2. Souffle dans la paille

Prendre un verre de plus de 12 cm de hauteur, le graduer en centimètres à partir du haut de 1 à 12, le remplir d'eau jusqu'à la graduation 12 cm et maintenir une paille enfoncée jusqu'à la graduation 5 cm à l'aide d'un trombone.

Demander au patient de souffler dans la paille le plus longtemps possible. Si le souffle est maintenu durant 5 secondes pour produire un tourbillon de façon régulière, cela suggère que le fonctionnement respiratoire est suffisant pour obtenir la pression nécessaire à la production des énoncés les plus courants. (Hixon et al., 1982).

3. Souffle sur S

Demander au patient de prendre une inspiration ample et confortable puis de souffler sur /s/ le plus régulièrement et le plus longtemps possible. Observer le contrôle. Noter la durée.

Chez un homme, la tenue moyenne se situe entre 10 et 16 secondes ; elle est considérée comme insuffisante si elle est inférieure à 10 secondes. Chez une femme, la tenue moyenne se situe entre 9 et 15 secondes ; elle est considérée comme insuffisante si elle est inférieure à 9 secondes.

Donner un score de 0 à 4 pour la respiration en motricité non verbale.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

LARYNX (X)

MODALITÉ VERBALE

1. Observation en conversation

Écouter les productions vocales du patient. Noter si la voix est altérée, si le patient présente un trouble du volume sonore ou un trouble de la hauteur ; la présence d'attaques brusques, de souffle sur la voix.

2. Temps Maximum de Phonation sur /a/ (TMP)

Demander au patient de prendre une inspiration ample et d'émettre un /a/ tenu le plus longtemps possible sans arrêt, à hauteur et intensité confortables. Chronométrer et noter le temps réalisé. La norme, de 20 à 50 ans, (d'après Estienne, 1998) est de 22 s pour un homme et de 17,9 s pour une femme. Ne tenir compte que du temps où la qualité de la voix est bonne.

3. Rapport S/Z

Demander au patient de tenir un /s/ le plus long possible. Puis de faire de même sur un /z/. Chronométrer, noter les temps réalisés, calculer le rapport S/Z et le noter. Normalement, le rapport est égal à 1. Au-dessus, le score est pathologique et indique un trouble au niveau du fonctionnement laryngé (tenue du /z/ très inférieure à celle du /s/) ou une incoordination pneumo-phonique (tenue du /z/ inférieure à celle du /s/) (Mueller P.B. et al., 1992).

4. Phonation interrompue sur /a/

Demander au patient de réaliser un /a/ avec une intensité confortable sur un même souffle en s'interrompant plusieurs fois puis en reprenant l'émission. Observer la qualité de l'arrêt (précis) et de la reprise phonatoire (précise sans attaque brusque). Vérifier qu'il n'y a pas de reprise d'air pendant l'interruption de la phonation.

Donner un score de 0 à 4 pour le larynx en motricité verbale.

MODALITÉ NON VERBALE

1. Adduction des cordes vocales

- Demander au patient de réaliser 5 coups de glotte successifs, très scandés. Noter la qualité des coups de glotte et préciser s'il y a non-réalisation des derniers coups de glotte.

- Raclement de gorge. Demander au patient de se racler la gorge. Noter si le raclement est audible, s'il est faible.

- Toux volontaire. Demander au patient de tousser. Observer la qualité de la toux volontaire.

2. Abduction des cordes vocales

- Inspiration par la bouche. Demander au patient d'effectuer une inspiration rapide et profonde par la bouche. Noter si elle est bruyante, scandée, très brève.

Donner un score de 0 à 4 pour le larynx en motricité non verbale

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

VÉLOPHARYNX (IX, X, XI)

MODALITÉ VERBALE

1. Observation en conversation

Écouter les productions vocales du patient. Noter s'il y a un trouble de la résonance ou perception d'une fuite nasale.

2. /a/ tenu

Demander au patient de maintenir un /a/ pendant 4 secondes en plaçant un miroir sous son nez. Noter s'il y a une fuite d'air nasale et observer son importance.

3. Série de /s/

Demander au patient d'émettre 5 /s/ successifs en plaçant un miroir sous son nez. Noter s'il y a une fuite d'air nasale et observer son importance.

4. Série de /p/

Demander au patient d'émettre 5 /p/ successifs en plaçant un miroir sous son nez. Noter s'il y a une fuite d'air nasale et observer son importance.

5. Phrase avec phonèmes oraux

Demander au patient de répéter la phrase suivante contenant uniquement des phonèmes oraux : « Le pape salue la foule. » en plaçant un miroir sous son nez. Noter s'il y a une fuite d'air nasale et observer son importance.

6. Série de /a/ - /an/

Demander au patient de répéter 5 fois la série /a/ - /an/. Noter si la différenciation est nette, partielle ou inexistante, dégradée au fil des répétitions.

Donner un score de 0 à 4 pour le vélopharynx en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Demander au patient d'ouvrir très grand la bouche. Observer le voile à l'aide d'une source lumineuse et d'un abaisse-langue. Noter si le voile est très bas, très haut, asymétrique, s'il présente une déviation droite ou gauche ou s'il y a des mouvements anormaux.

2. Observation sur 5 /a/

Demander au patient d'émettre 5 /a/ successifs, bien différenciés, bouche ouverte. Observer l'amplitude, la régularité et la symétrie de l'élévation et de l'abaissement du voile.

Donner un score de 0 à 4 pour le vélopharynx en motricité non verbale.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

LANGUE (XII)

MODALITÉ VERBALE

1. Observation en conversation

Observer la mobilisation de la langue du patient au cours d'une conversation. Noter s'il y a des mouvements anormaux, une interposition dentale ou labiale, un défaut de précision des consonnes /l/, /t/, /d/, /n/ ou une indifférenciation des voyelles.

2. à 6. Répétitions successives de syllabes

Demander au patient de répéter successivement /ta/, /na/, /la/, /ka/, /ra/. Observer la précision d'articulation des consonnes.

Donner un score de 0 à 4 pour la langue en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Demander au patient d'ouvrir la bouche. Si le patient a des difficultés à maintenir cette position, utiliser un abaisse-langue posé latéralement sur les dents. Laisser un temps de latence de 10 secondes pour que la langue se mette au repos, (ce qui peut ne pas se faire immédiatement après l'ouverture de la bouche) et observer pendant au minimum 30 secondes. Noter s'il y a une atrophie, une hypo ou hypertonie droite ou gauche, une déviation droite ou gauche, des mouvements anormaux.

2. Protrusion

Demander au patient de tirer la langue le plus loin possible. Noter s'il y a une déviation droite ou gauche ou des mouvements anormaux. Observer l'amplitude du mouvement.

3. Élévation de la pointe

Demander au patient de toucher son palais en arrière des incisives supérieures avec la pointe de la langue, bouche ouverte. Observer l'amplitude et la qualité du mouvement.

4. Claquements de langue (dos)

Demander au patient de réaliser 5 claquements avec le dos de la langue comme pour imiter le bruit du pas du cheval. Observer l'accolement de la langue contre le palais.

5. Clics de réprobation (apex)

Demander au patient de réaliser des petits clics rapides en plaçant la langue juste derrière les incisives supérieures comme pour émettre des signes de réprobation (lieu articulatoire du /t/). Observer la qualité, l'aisance, la vitesse et la régularité de réalisation.

6. Recul de la base de la langue

Demander au patient de réaliser des claquements avec la base de la langue en émettant 5 /ka/ très explosifs. Observer le recul de la base de la langue, la qualité de l'explosion.

Donner un score de 0 à 4 pour la langue en motricité non verbale.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

LÈVRES (VII)

MODALITÉ VERBALE

1. Observation en conversation

Observer les lèvres du patient au cours d'une conversation. Noter s'il y a une absence, une réduction ou une incoordination des mouvements, un trouble de l'explosion, une imprécision des voyelles, une déviation droite ou gauche, des mouvements anormaux.

2. et 3. Répétitions successives

Demander au patient de répéter successivement /pa/, /ba/, /i/, /y/, / u/. Observer la qualité de l'explosion et la précision des voyelles.

Donner un score de 0 à 4 pour les lèvres en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Observer les lèvres au repos. Palper les lèvres du patient avec le bout du doigt puis pincer chaque lèvre de façon à l'étirer vers l'avant. Noter s'il y a une atrophie, une asymétrie, une hypo ou hypertonie droite, gauche, supérieure ou inférieure, des mouvements anormaux.

2. Arrondissement

Demander au patient de projeter les lèvres arrondies vers l'avant comme pour réaliser un /u/. Évaluer la qualité de l'arrondissement.

3. Pincement

Demander au patient de pincer ses lèvres l'une sur l'autre en maintenant quelques secondes. Évaluer la qualité du pincement, noter une asymétrie.

4. Étirement

Demander au patient de faire un sourire exagéré. Noter s'il y a une asymétrie, une déformation, un déficit d'amplitude.

Donner un score de 0 à 4 pour les lèvres en motricité non verbale.

JOUES (VII)

MODALITÉ VERBALE

1. Observation en conversation

Observer les joues en conversation. Noter les mouvements anormaux ou d'éventuelles vibrations.

Donner un score de 0 à 4 pour les joues en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Observer les joues au repos puis les palper. Noter s'il existe une hypo ou hypertonie droite ou gauche.

2. Gonflement des 2 joues

Demander au patient de gonfler les deux joues pendant 5 secondes. Observer l'amplitude du mouvement, la symétrie et l'éventuelle fuite d'air labiale.

3. Gonflement à droite puis à gauche

Demander au patient de gonfler la joue droite puis la joue gauche pendant 5 secondes. Observer l'amplitude du mouvement et l'éventuelle fuite d'air labiale.

Donner un score de 0 à 4 pour les joues en motricité non verbale.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

MÂCHOIRES (V)

MODALITÉ VERBALE

1. Observation en conversation

Observer les mouvements de la mandibule au cours d'une conversation. Noter s'il y a une réduction des mouvements, une déviation droite ou gauche, des secousses ou l'installation d'une dystonie oro-mandibulaire.

Donner un score de 0 à 4 pour les mâchoires en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Observer les mâchoires du patient au repos. Noter s'il y a un trouble de la position des mâchoires : prognatie, rétrognatie, déviation latérale droite ou gauche, un serrage excessif, un relâchement, des mouvements anormaux.

2. Grande ouverture

Demander au patient d'ouvrir la bouche le plus grand possible. Observer l'amplitude (norme à 4/5 cm) et noter une éventuelle déviation de la mandibule vers la droite ou la gauche.

Donner un score de 0 à 4 pour les mâchoires en motricité non verbale.

FACE (V, VII)

MODALITÉ VERBALE

1. Observation en conversation

Observer le visage du patient au cours d'une conversation. Noter s'il y a une dysharmonie, une déformation vers la droite ou la gauche, une asymétrie, un faciès figé, une adaptation de la mimique au discours, des mouvements anormaux.

Donner un score de 0 à 4 pour la face en motricité verbale.

MODALITÉ NON VERBALE

1. Observation au repos

Demander au patient de regarder vers l'avant et de se détendre. Si nécessaire, lui proposer de fermer les yeux un instant. Noter s'il y a une dysharmonie, une déformation de la face vers la droite ou la gauche, une asymétrie, une expressivité réduite, des mouvements anormaux.

Donner un score de 0 à 4 pour la face en motricité non verbale.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

MOUVEMENTS ALTERNATIFS

MODALITÉ VERBALE

1. Répétitions de la syllabe /pa/

Demander au patient de répéter rapidement et correctement la syllabe /pa/. Donner un modèle avec 7 répétitions en 1 seconde. Observer la vitesse de réalisation, le maintien de la vitesse (accélération, ralentissement) et la qualité de l'articulation au fil des répétitions.

2. Répétition de /pata/

Demander au patient de répéter rapidement et correctement la séquence /pata/. Faire une démonstration avec 5 répétitions de la séquence en 3 secondes. Observer la vitesse de réalisation, le maintien de la vitesse (accélération, ralentissement) et la qualité de l'articulation au fil des répétitions.

3. Répétition de /pataka/

Demander au patient de répéter rapidement et correctement la séquence /pataka/. Faire une démonstration avec 5 répétitions de la séquence en 5 secondes. Observer la vitesse de réalisation, le maintien de la vitesse (accélération, ralentissement) et la qualité de l'articulation au fil des répétitions.

Donner un score de 0 à 4 pour les mouvements alternatifs en motricité verbale.

MODALITÉ NON VERBALE

1. Ouverture/fermeture de la bouche

Demander au patient d'ouvrir et fermer la bouche rapidement. Faire une démonstration avec 5 mouvements. On doit entendre un léger bruit d'occlusion des lèvres. Observer la précision, la coordination, la vitesse, le respect du rythme et le maintien de l'amplitude des mouvements.

2. Protrusion/rétraction de la langue

Demander au patient de sortir la langue et de la rentrer en fermant la bouche rapidement. Faire une démonstration avec 5 mouvements. Observer la précision, la coordination, la vitesse, le respect du rythme et le maintien de l'amplitude des mouvements.

3. Langue de gauche à droite

Demander au patient de toucher les commissures des lèvres à droite et à gauche avec la langue rapidement. Faire une démonstration avec 5 mouvements. Observer la précision, la coordination, la vitesse, le respect du rythme et le maintien de l'amplitude des mouvements.

Donner un score de 0 à 4 pour les mouvements alternatifs en motricité non verbale.

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

Annexe 6

Speech Handicap Index (Degroote et al., 2012)

		Jamais	Presque jamais	Parfois	Presque toujours	Toujours
P 1	On me comprend difficilement à cause de ma parole.					
P 2	Je suis à bout de souffle quand je parle.					
P 3	L'intelligibilité de ma parole varie au cours de la journée.					
Ps 4	Je me sens incompétent à cause de ma parole.					
P 5	On me demande pourquoi je suis difficile à comprendre.					
P 6	Je suis ennuyé(e) quand les gens me demandent de répéter.					
Ps 7	J'évite de téléphoner.					
Ps 8	Je suis tendu(e) quand je parle avec d'autres personnes à cause de ma parole.					
P 9	Mon articulation n'est pas nette.					
P 10	On me comprend difficilement quand je parle dans une pièce bruyante.					
Ps 11	J'ai tendance à éviter les groupes de gens à cause de ma parole.					
Ps 12	Les gens semblent irrités par ma parole.					
P 13	Les gens me demandent de me répéter quand on parle face-à-face.					
Ps 14	Je parle moins souvent avec mes amis, mes voisins, ma famille à cause de ma parole.					
P 15	J'ai l'impression que je dois forcer pour parler.					
Ps 16	Je trouve que les gens ne comprennent pas mes problèmes pour parler.					
Ps 17	Mes difficultés à parler limitent ma vie personnelle et sociale.					

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.

P 18	L'intelligibilité de ma parole est imprévisible.					
Ps 19	Je me sens écarté(e) des conversations à cause de ma parole.					
P 20	Je fais beaucoup d'efforts pour parler.					
P 21	Ma parole est plus mauvaise le soir.					
22	Mes difficultés à parler entraînent des pertes de revenus.					
23	J'essaie de modifier ma parole afin qu'elle sonne différemment.					
Ps 24	Mes problèmes de parole me dérangent.					
Ps 25	Je suis moins extraverti(e) à cause de ma parole.					
P 26	Ma famille a du mal à me comprendre quand je les appelle dans la maison.					
Ps 27	Je me sens handicapé(e) à cause de ma parole.					
P 28	J'ai des difficultés à poursuivre une conversation à cause de ma parole.					
Ps 29	Je suis embarrassé(e) quand les gens me demandent de répéter.					
Ps 30	J'ai honte de mon problème de parole.					
	Total					

Score total SHI /120

Sous-score parole /56

Sous-score psychosocial /56

Seuil pathologique à 28/120

Comment trouvez-vous votre parole aujourd'hui ?

Excellente Bonne Moyenne Mauvaise

©Ortho Edition 2019 – BECD – V. Rolland-Monnoury & P. Auzou
En vertu du code régissant la propriété intellectuelle il est strictement interdit de diffuser cette page sur les réseaux sociaux, blogs ou sites internet et de la reproduire, dans n'importe quel format que ce soit, à des fins commerciales ou dans tout autre but.