Université du Québec

Institut national de la recherche scientifique

Énergie Matériaux Télécommunications Research Centre

DEVELOPMENT OF GENERALIZABLE, INTERPRETABLE, AND PRIVACY-PRESERVING HUMAN-CENTRIC AUDIO APPLICATIONS

By

Yi Zhu

A thesis submitted in fulfillment of the requirements for the degree of
Doctorate of Sciences, Ph.D
in Telecommunications

Evaluation Committee

Committee president and internal evaluator	Prof. Douglas O'Shaughnessy INRS-EMT
External evaluator 1	Prof. Ina Kodrasi Idiap Research Institute
External evaluator 2	Prof. Ali Etemad Queen's University
Research Advisor	Prof. Tiago Falk INRS-EMT

ACKNOWLEDGEMENTS

I would like to acknowledge the support and supervision of Prof. Tiago Falk, and colleagues and collaborations from MuSAE lab as well as research groups from different institutes : Abhishek Tiwari, Shruti Kshirsagar, João Monteiro, Mahsa Abdollahi, Heitor R. Guimarães, Ségolène Maucourt, Nico Coallier, Bennet Fischer, Mario Chemnitz, Prof. Pierre Giovenazzo, Prof. Roberto Morandotti, Prof. Alex Mariakakis, and Prof. Eyal De Lara.

While working towards the Ph.D., I was fortunate to take on an internship role during my last year at Reality Defender. I would like to thank my colleagues for their collaboration and guidance : Surya Koppisetti, Trang Tran, and Gaurav Bharaj.

I would like to thank my family for their unconditional love and support, none of my achievements today would become true without them : my mom Wenyu Hu, my dad Shuhan Zhu, my grandparents Maosen Hu and Ziwen Wang, and Zixia Wang, who raised me and inspired me to pursue this field. Her memory has stayed with me throughout my journey. Lastly, I would like to largely thank my partner Nan Ma for her encouragement and care along this road.

RÉSUMÉ

Les signaux audio centrés sur l'humain, tels que la parole, la toux et la respiration, sont utilisés dans diverses applications, telles que la reconnaissance automatique de la parole, la vérification du locuteur et le diagnostic de santé. Dans des scénarios réels, pour garantir une performance robuste, il est essentiel de se concentrer sur trois propriétés clés : la généralisabilité, l'interprétabilité et la sécurité. La généralisabilité assure que les modèles maintiennent une précision prédictive élevée lorsqu'ils sont appliqués à des données non vues et à des distributions inconnues, ce qui est fréquent dans les données collectées en conditions réelles, souvent influencées par des facteurs tels que le bruit de fond. Par ailleurs, dans des domaines nécessitant une grande confiance (par exemple, la santé), la précision seule n'est pas suffisante ; le modèle doit être interprétable afin que des experts humains puissent évaluer la fiabilité de ses résultats. Enfin, les signaux vocaux et respiratoires étant issus de la coordination des systèmes respiratoire et articulatoire, ils sont considérés comme des signaux biométriques, pouvant identifier un individu. À mesure que les applications audio se démocratisent, le partage de données vocales présente un risque de fuite d'identité. De plus, les modèles génératifs récents permettent à des attaquants malveillants de cloner la voix d'un individu, facilitant ainsi l'usurpation d'identité et la fraude. Il est donc crucial de mettre en place des mesures de sécurité pour protéger la vie privée des utilisateurs.

Dans cette thèse, nous proposons plusieurs innovations pour améliorer ces propriétés dans les applications audio centrées sur l'humain. Nous nous intéressons tout particulièrement au diagnostic de santé, un domaine émergent qui soulève des enjeux relatifs aux trois aspects mentionnés. Nous proposons tout d'abord deux ensembles de caractéristiques novateurs, fondés sur la connaissance, pour caractériser la santé à partir de la parole et de la toux pathologiques. Nous avons créé et rendu public un ensemble de données de toux annotées manuellement, comprenant plus de 1000 enregistrements de toux liés au COVID-19, avec des annotations fines des phases de toux. Nous démontrons également que les caractéristiques proposées, associées à des modèles d'apprentissage automatique inspirés de la physiologie, se généralisent bien aux ensembles de données non vus et surpassent plusieurs réseaux de neurones profonds complexes.

Ensuite, nous proposons deux nouvelles stratégies d'apprentissage, l'une supervisée et l'autre auto-supervisée, pour obtenir des représentations profondes généralisables et interprétables dans le cadre du diagnostic de santé et de la détection de la parole synthétique frauduleuse. Nous montrons que ces représentations atteignent des performances de pointe (SOTA) tout en étant interprétables par des humains. En outre, nous réalisons une évaluation approfondie de plusieurs méthodes d'anonymisation de la voix afin d'examiner leur impact sur le diagnostic de santé. Cette analyse met en lumière les limitations et compromis associés à l'anonymisation pour la dissimulation de l'identité, et explore les causes des variations de performance. Nous montrons également que ces anonymiseurs, utilisés comme générateurs de fausses pathologies vocales, entraînent une baisse importante de la précision des détecteurs de fausses voix SOTA, suggérant que les fausses pathologies sont plus difficiles à détecter que les fausses voix classiques.

Enfin, nous proposons une architecture de modèle générique, nommée WavRx, qui intègre les trois propriétés mentionnées. Ce modèle peut être appliqué à la fois pour le diagnostic de santé et la détection de la parole synthétique frauduleuse, et génère une représentation dynamique de l'énoncé permettant d'obtenir des performances de pointe sur six ensembles de données pathologiques et deux ensembles de données de fausses voix. Il surpasse de manière significative les

représentations universelles existantes en termes de performance sans apprentissage préalable (zero-shot), démontrant ainsi sa généralisabilité. De plus, nous montrons que ces représentations peuvent être utilisées pour identifier des anomalies dans la production de la parole et des artefacts de la parole synthétique, renforçant ainsi l'explicabilité. Par ailleurs, la représentation dissocie les attributs du locuteur des attributs liés à la tâche, ce qui en fait un bon candidat pour des applications respectueuses de la vie privée.

Mots-clés Apprentissage des représentations, Généralisabilité, Interprétabilité, Respectueux de la vie privée, Audio, Diagnostic de santé, Anonymisation de la voix, Anonymisation de la voix.

ABSTRACT

Human-centric audio signals, including speech, cough, and breathing, have been explored for a variety of applications, such as automatic speech recognition, speaker verification, health diagnostics, just to name a few. When deployed in real-world scenarios, achieving robust application performance demands a focus on three essential properties : generalizability, interpretability, and security. Generalizability guarantees that models can maintain high predictive accuracy when tested on unseen data with unknown distributions, which is often the case for in-the-wild data with unwanted biasing factors (e.g., background noise). Meanwhile, for applications where trustworthiness is required (e.g., healthcare), accuracy alone is not sufficient as the model needs to be interpretable in order for human experts to assess the reliability of its output. Lastly, since voice and respiratory signals are generated by the coordination of human respiratory and articulatory system, they are regarded as 'biometric' signals, which can be used to represent the user identity. As audio applications became increasingly accessible, sharing voice data can lead to potential identity leakage. With the rapid advancement of generative models, one concerning consequence is that attackers with malicious intent can easily clone someone's voice, enabling impersonating of individuals and potentially commit fraud. As such, it is crucial to develop secure measures to protect user privacy.

In this dissertation, we propose several innovations to improve the aforementioned properties of human-centric audio applications. One task that we focus on is health diagnostics, since it is an emerging field with issues spanning all three aspects mentioned above. For this, we firstly propose two novel knowledge-based feature sets to characterize health from pathological speech and cough modalities. For the latter, we curated and open-sourced a human-labeled cough dataset comprising 1000+ COVID-19 cough recordings with fine-grained cough-phase annotations. We further show that the proposed features combined with physiology-inspired machine learning (ML) models can generalize well to unseen datasets, which outperform several complex deep neural networks (DNNs).

Secondly, we propose two novel learning strategies, one in the supervised learning paradigm and the other in self-supervised learning, to obtain generalizable and interpretable deep representations for health diagnostics and deepfake speech detection, respectively. We show that the resultant deep representations not only achieve state-of-the-art (SOTA) performance, but also can be interpreted and understood by humans. Thirdly, as voice privacy has become a concern for numerous speech applications, we perform a comprehensive evaluation of several voice anonymization methods to investigate their impact on the health diagnostics task. Our evaluation has revealed the limitations and trade-off of using voice anonymization for identity concealing, and provides a deep analysis of the causes behind changes in performance. On the other side, we show that when these anonymizers are used as pathological deepfake generators, a significant drop is seen in the accuracy obtained by SOTA deepfake detectors, suggesting that pathological deepfakes are more challenging to be detected than regular deepfakes.

Finally, we propose a generic model architecture 'WavRx' that incorporates all three properties raised above. The model can be applied to both health diagnostics and deepfake detection, which encodes an utterance-level dynamics representation that helps achieve SOTA performance on six different pathological datasets and two deepfake datasets. It significantly outperforms existing universal representations in terms of zero-shot performance, demonstrating its generalizability.

Meanwhile, we show that the representations can be used to pinpoint speech production abnormalities as well as deepfake speech artifacts for explainability. Furthermore, the representation disentangles speaker attributes from the task-related attributes, making it a good candidate for privacy-preserving applications.

Keywords Representation learning, generalizability, interpretability, privacy-preserving, audio, health diagnostics, voice anonymization, deepfake detection.

TABLE DES MATIÈRES

LIST OF FIGURES

LISTE DES TABLEAUX

SYNOPSIS

DÉVELOPPEMENT D'APPLICATIONS AUDIO CENTRÉES SUR L'HUMAIN :
GÉNÉRALISABLES, INTERPRÉTABLES ET ESPECTUEUSES DE LA VIE PRIVÉE

0.1 Introduction

Les applications audio centrées sur l'humain, telles que la reconnaissance automatique de la parole (ASR), la vérification du locuteur (SV) et les diagnostics de santé respiratoire, sont devenues essentielles dans divers domaines, tels que le service client (Wang et al., 2023b), la conduite autonome (Cui et al., 2024), les soins de santé (Lugović et al., 2016) et la biométrie (Markowitz, 2000), pour n'en citer que quelques-uns. À mesure que ces applications sont déployées dans des contextes réels, garantir des performances robustes nécessite de se concentrer sur trois propriétés essentielles : *la généralisabilité*, *l'interprétabilité* et *la sécurité*.

La généralisabilité désigne la capacité d'un modèle à maintenir une précision prédictive élevée lorsqu'il est testé sur des données présentant des distributions non vues, c'est-à-dire des distributions différentes de celles des données utilisées lors de l'entraînement (Wang et al., 2022). Les modèles d'apprentissage automatique (AA) étant généralement gourmands en données, l'une des manières d'améliorer la *généralisabilité* est de manipuler les données, par exemple par augmentation ou génération, afin d'accroître la diversité et la quantité des données d'entraînement (Adila et al., 2022). Une autre approche consiste à explorer de nouvelles architectures de modèles et stratégies d'apprentissage, visant à apprendre une représentation généralisable tout en conservant intactes les données d'entraînement (Wang et al., 2022). Une représentation généralisable est censée de ne contenir que des attributs liés à la tâche, tout en restant indépendante des facteurs indésirables. Dans le cadre des diagnostics de santé par la parole, par exemple, une représentation généralisable devrait ne comporter que des informations relatives à la santé, tout en étant robuste aux autres attributs acoustiques tels que le sexe, la langue et le contenu de la parole.

Pour certaines applications de la parole, telles que les soins de santé, une haute précision à elle seule ne suffit pas, car le processus décisionnel des modèles AA/AP doit être compréhensible par des experts humains pour établir la confiance. Dans ces cas, les modèles doivent être interprétables afin que les décisions, qu'elles soient correctes ou incorrectes, puissent être expliquées et validées par des humains. Prenons l'exemple du diagnostic de santé, l' *interprétabilité* permet aux cliniciens de comprendre les fondements des prédictions d'un modèle, telles que des biomarqueurs vocaux spécifiques ou des motifs acoustiques indicatifs de certaines pathologies (Doshi-Velez et al., 2017). Un modèle interprétable aide les experts à évaluer la fiabilité de ses résultats, facilitant ainsi des décisions cliniques éclairées, en adéquation avec les connaissances

du domaine. Cependant, l'un des principaux défis des modèles AA/AP réside dans le compromis entre la performance et *l'interprétabilité*. Tandis que les modèles purement basés sur les données offrent généralement de meilleures performances, le manque d' *interprétabilité* limite leur utilisation dans des scénarios réels.

Enfin, étant donné que la voix contient de nombreuses informations spécifiques au locuteur, telles que le sexe et l'âge, elle est utilisée comme modalité biométrique dans diverses applications de vérification d'identité, telles que l'accès aux comptes bancaires personnels (Markowitz, 2000). Parallèlement, avec le nombre croissant d'outils de clonage vocal, il est désormais possible de reproduire la voix de quelqu'un à partir d'un enregistrement vocal de seulement 3 secondes (Wang et al., 2023a). Ainsi, des régulations ont été instaurées à l'échelle mondiale pour protéger la biométrie vocale des individus et limiter l'utilisation des outils de clonage vocal. Parmi ces régulations, on trouve le Règlement général sur la protection des données (RGPD) en Europe (Zarsky, 2016) et la Loi sur la protection des informations personnelles (PIPL) en Chine (Calzada, 2022). Dans cette thèse, nous proposons de traiter la question de *la sécurité* et de *la protection de la vie privée* par l'anonymisation de la voix, qui permet de masquer l'identité de l'utilisateur lors du partage de données, ainsi que par la détection de deepfake pour identifier les voix synthétiques.

0.1.1 Objectifs

Les recherches présentées dans cette dissertation ont pour objectif d'améliorer la généralisabilité, l'explicabilité et la sécurité des applications audio. Les objectifs spécifiques sont les suivants :

1. Développer de nouvelles caractéristiques et représentations visant à améliorer la généralisabilité et l'interprétabilité des modèles de diagnostic de santé et de détection de deepfake;
2. Explorer de nouvelles architectures de modèles adaptables à différents ensembles de données et tâches;
3. Élaborer des stratégies d'entraînement de modèles pour apprendre des représentations généralisables et renforcer l'explicabilité du processus décisionnel des modèles de type boîte noire;
4. Concevoir des techniques de préservation de la vie privée garantissant la sécurité de l'identité de l'utilisateur, tout en optimisant la performance des tâches associées.

0.1.2 Champ d'application

Dans ce manuscrit, nous nous concentrons sur trois applications distinctes de la parole qui nécessitent des modèles généralisables, interprétables et sécurisés. Les sous-sections suivantes présentent chacune de ces applications ainsi que les défis principaux auxquels elles sont confrontées.

Diagnostics de santé basés sur l'audio Ces dernières années, les signaux vocaux et respiratoires ont émergé comme des modalités prometteuses pour le diagnostic des maladies et la surveillance à distance de la santé. En comparaison avec les méthodes de tests classiques, telles que les radiographies, les tomodensitogrammes (CT-scans) ou les analyses de sang, les systèmes de diagnostic audio présentent l'avantage d'un temps de traitement réduit et de coûts de test faibles. Le diagnostic audio repose sur l'idée que les maladies causant des anomalies dans les systèmes articulatoire et/ou respiratoire entraînent des modèles atypiques dans la voix humaine et la respiration (Fagherazzi et al., 2021). Ces anomalies peuvent avoir diverses causes, telles qu'un contrôle neuromusculaire altéré ou une inflammation du tractus vocal et des poumons (Fagherazzi et al., 2021).

Bien que l'impact sur les signaux acoustiques puisse parfois être imperceptible pour l'humain, un modèle d'apprentissage automatique (ML) peut être utilisé pour détecter certains biomarqueurs numériques liés à des maladies. Dans un système typique de diagnostic de santé basé sur l'audio, les utilisateurs sont invités à lire plusieurs phrases ou à tousser quelques fois, tandis que les signaux acoustiques sont captés à distance via un microphone. Ces signaux sont ensuite envoyés au modèle diagnostique pour extraire les biomarqueurs numériques et déterminer si l'utilisateur est positif ou négatif à une maladie. Les modèles existants de diagnostic audio rencontrent des problèmes de généralisabilité et d'interprétabilité (Naudé, 2020). Le premier désigne les modèles formés sur un ensemble de données et qui montrent une mauvaise performance lorsqu'ils sont confrontés à des ensembles de données non vus ; le second fait référence à la nature opaque de la plupart des modèles diagnostiques, rendant le processus décisionnel incompréhensible pour un humain. Ces problèmes rendent difficile le déploiement des modèles existants dans des contextes réels. Ce manuscrit présente de nouvelles fonctionnalités et techniques d'apprentissage automatique visant à améliorer la généralisabilité et l'interprétabilité des modèles.

Détection de la parole synthétisée L'intérêt croissant pour les modèles génératifs a permis le développement d'outils permettant d'imiter de manière réaliste la voix d'une personne (Tan et al., 2021). Les systèmes de synthèse vocale (TTS) et de conversion de la voix (VC) peuvent désormais créer une voix artificielle à partir de seulement quelques secondes d'enregistrements vocaux réels (Tan, 2023). Lorsque ces outils sont utilisés à des fins malveillantes, les résultats, appelés 'deepfakes audio'(Khanjani et al., 2023), peuvent causer de graves problèmes. Parmi les exemples, citons l'usurpation d'identité de célébrités/famille pour des appels automatisés(Knibbs, 2024), l'accès illégal à des comptes bancaires protégés par la voix (Cox, 2023), ou la falsification de preuves en justice (Pender, 2023). Par conséquent, des modèles de détection de la parole synthétisée ont été explorés pour contrer les deepfakes audio. Toutefois, tout comme les systèmes de diagnostic audio, les modèles de détection des deepfakes souffrent de problèmes de généralisabilité face à des modèles génératifs inconnus (attacks non observées) et de manque d'explicabilité des décisions du modèle. Ce manuscrit présente de nouveaux modèles et stratégies d'apprentissage pour combler cette lacune.

Anonymisation de la voix Aujourd'hui, la majorité des applications de la parole reposent sur des architectures de réseaux neuronaux profonds (DNN), avec des modèles contenant des centaines de millions de paramètres, un nombre qui continue d'augmenter. Ces paramètres ne sont généralement pas stockés localement sur les appareils mobiles (Wang et al., 2018a), et les données vocales sont envoyées et traitées dans le cloud, les décisions étant ensuite renvoyées vers l'appareil de l'utilisateur. Avec l'augmentation des cyberattaques signalées (Stupp, 2019; Kaloudi et al., 2020; Yamin et al., 2021), cette transmission des données vocales via le cloud pourrait représenter de sérieuses menaces pour la vie privée et la sécurité des utilisateurs. Il a été rapporté que les assistants vocaux et de nombreuses applications tierces collectent la voix des utilisateurs à leur insu et la partagent avec des partenaires publicitaires (Iqbal et al., 2022). Par exemple, Amazon a breveté une méthode permettant de reconnaître l'état de santé d'un utilisateur via des conversations, puis de lui proposer des médicaments (Jin et al., 2018). Ce risque est particulièrement élevé pour les applications de diagnostic vocal, car la voix d'un utilisateur peut être liée à des informations médicales sensibles, comme son état de santé (Latif et al., 2020a), la progression d'une maladie (Harel et al., 2004) ou son état mental (Low et al., 2020). De ce fait, la préservation de la vie privée dans la parole est devenue un sujet d'intérêt mondial, notamment avec la mise en place de réglementations telles que le Règlement général sur la protection des données (RGPD) en Europe (Zarsky, 2016) et la loi sur la protection des informations personnelles (PIPL) en Chine (Calzada, 2022), qui se concentrent particulièrement sur les biométries personnelles (telles que la voix, l'image faciale et les empreintes digitales).

Les méthodes d'anonymisation de la voix visent à modifier le signal vocal de manière à masquer l'identité du locuteur, tout en préservant les informations linguistiques et d'autres caractéristiques para-linguistiques (comme le timbre ou la naturalité). Étant donné l'essor de ce domaine, les séries de défis Voice Privacy Challenge (VPC) ont été organisées en 2020 et 2022 pour promouvoir l'avancée des techniques d'anonymisation vocale (Tomashenko et al., 2022b,a). Cependant, ces défis se sont concentrés sur le développement de méthodes d'anonymisation pour les tâches de reconnaissance automatique de la parole en aval (Tomashenko et al., 2022b,a; Fang et al., 2019; Meyer et al., 2022a), où le contenu linguistique est préservé mais pas les informations paralinguistiques. Une partie de cette thèse explore l'impact des techniques d'anonymisation vocale sur les tâches de diagnostic de santé, ainsi que de nouvelles approches permettant de préserver à la fois la confidentialité des utilisateurs et les informations sur leur santé.

0.2 Contexte et travaux connexes

0.2.1 Production et application des signaux de la parole et de la toux

Signaux de la parole La production des signaux de la parole constitue un processus physiologique complexe impliquant la coordination de multiples structures anatomiques, telles que les pou-

mons, la trachée, le larynx, les cordes vocales et les articulations (Honda, 2008). La parole véhicule différentes couches d'information, telles que le contenu verbal, l'identité du locuteur, les émotions, la santé, etc. Une manière courante de les classer est de distinguer les aspects linguistiques des aspects paralinguistiques (Schuller et al., 2013a), les premiers se rapportant au contenu verbal et les seconds aux attributs tels que la prosodie, la hauteur, le ton, l'intensité et le rythme (Scherer, 2003; Schuller et al., 2013a). Parmi les nombreuses applications paralinguistiques de la parole, le diagnostic de santé basé sur la parole est un domaine émergent (Ramanarayanan et al., 2022). Étant donné que les systèmes respiratoire et articulatoire peuvent être affectés par de nombreuses pathologies, des modifications dans les modèles de parole peuvent survenir en raison de troubles dans ces régions, ce qui fait de la parole un outil potentiel pour le diagnostic et le suivi de la santé. Par exemple, les patients atteints de maladies neurodégénératives comme la maladie d'Alzheimer ou de Parkinson peuvent présenter des altérations spécifiques de leur parole, telles que des changements de volume, de clarté et de rythme, qui peuvent être analysées pour suivre l'évolution de la maladie (Brabenec et al., 2017). Récemment, un nombre considérable d'études a mis en évidence l'utilisation de la parole pour détecter le COVID-19 (Deshpande et al., 2020a).

Signaux de la toux Une autre modalité acoustique importante pour évaluer l'état de santé respiratoire est la toux, qui est un symptôme associé à plus de 100 maladies (Korpáš et al., 1996), telles que la BPCO, l'asthme et le COVID-19 (Ciotti et al., 2020), entre autres. Bien que la toux provienne du flux d'air des poumons, comme la parole, il s'agit d'une action réflexe ne nécessitant pas un contrôle moteur précis. Le processus de toux inclut une inhalation rapide suivie d'une expiration forcée contre une glotte fermée, qui s'ouvre soudainement pour libérer un jet d'air (Chung et al., 2008; Piirila et al., 1995). Le son de la toux peut être divisé en trois phases : (1) l'inhalation, où la glotte reste ouverte pour amener de l'air dans les poumons ; (2) la compression, où une expiration forcée se produit contre la glotte fermée ; et (3) l'expulsion, où la glotte s'ouvre pour produire le son explosif (Chung et al., 2008). Des études cliniques ont montré que les modèles de toux varient selon les pathologies respiratoires (Piirila et al., 1995) et que des anomalies dans les phases de toux peuvent être liées à diverses pathologies (Korpáš et al., 1996). Étant donné que la toux peut être collectée de manière similaire à la parole tout en présentant moins de biais (comme le contenu verbal), elle a été utilisée dans plusieurs études récentes pour des tâches diagnostiques de santé (Infante et al., 2017; Alqudaihi et al., 2021; Tena et al., 2022).

0.2.2 Représentations audio

Caractéristiques basées sur les connaissances Les recherches antérieures ont souligné l'importance des caractéristiques basées sur les connaissances pour caractériser les traits sous-jacents de la parole. Au-delà des caractéristiques classiques, telles que les mel-spectrogrammes ou les MFCC, de nombreuses études ont exploré des caractéristiques adaptées à des applications spécifiques. Par exemple, des caractéristiques de phonation et d'articulation pour les tâches

de contrôle moteur de la parole (Arias-Vergara et al., 2017; Vásquez-Correa et al., 2018), des caractéristiques basées sur la prédiction linéaire pour séparer les informations sur la source vocale et le tractus vocal (Zhu et al., 2023f), des caractéristiques prosodiques pour représenter la hauteur, l'énergie et le rythme (Schuller et al., 2013b), ainsi que des caractéristiques de flux glottal pour modéliser les cordes vocales (Drugman et al., 2012). De plus, des ensembles de caractéristiques plus vastes ont été proposés, agrégant des descripteurs de bas niveau de la parole et leurs statistiques. Par exemple, l'ensemble openSMILE ComParE (Eyben et al., 2010) est largement utilisé comme base de référence dans plusieurs tâches de détection d'attributs de la parole (Schuller et al., 2021, 2017). D'autres ensembles, tels que GeMAPS (Eyben et al., 2015), se concentrent sur la capture de la qualité vocale et l'expression émotionnelle, tandis que l'extension GeMAPS (eGeMAPS) vise à allier exhaustivité et efficacité informatique pour les tâches paralinguistiques (Eyben et al., 2015).

Représentations profondes Les progrès récents dans l'apprentissage profond ont considérablement amélioré les performances des modèles pour diverses tâches de parole grâce à l'apprentissage de représentations à partir de réseaux neuronaux profonds. Une méthode pour obtenir des représentations au niveau des énoncés repose sur l'apprentissage supervisé sur de grandes quantités de données, comme le modèle x-vector (Snyder et al., 2018) et ECAPA-TDNN (Desplanques et al., 2020), qui ont montré de bonnes performances sur des tâches paralinguistiques (Morais et al., 2022). Parallèlement, l'apprentissage auto-supervisé (SSL) a attiré l'attention en raison de sa capacité à apprendre des représentations riches à partir de données non étiquetées. Des modèles tels que Wav2Vec2 (Baevski et al., 2020), WavLM (Chen et al., 2022), et HuBERT (Hsu et al., 2021) illustrent ce paradigme. Wav2Vec2 utilise l'apprentissage contrastif pour modéliser des représentations de la parole en prédisant des portions masquées de l'audio (Baevski et al., 2020). Les embeddings des différentes couches du transformeur ont prouvé leur utilité pour plusieurs tâches en aval, telles que la reconnaissance de la parole, la vérification du locuteur et la reconnaissance des émotions. WavLM s'appuie sur Wav2Vec2 et améliore les performances dans des tâches paralinguistiques comme la reconnaissance des émotions et l'identification du locuteur (Chen et al., 2022). HuBERT utilise un cadre hiérarchique pour apprendre des représentations contextualisées de la parole (Hsu et al., 2021). Ces modèles SSL se sont révélés efficaces pour une large gamme de tâches liées à la parole (Yang et al., 2021a), prouvant l'efficacité du SSL pour apprendre des représentations généralisables. En comparaison avec les caractéristiques basées sur les connaissances et les représentations supervisées, les représentations SSL ont montré une amélioration marquée de la généralisation aux données non vues sans avoir besoin d'un apprentissage supervisé à grande échelle.

0.3 Jeux de données et métriques d'évaluation

Étant donné que certains jeux de données ont été réutilisés dans plusieurs projets, nous listons ici tous les jeux de données audio pathologiques et synthétiques utilisés par les articles inclus

dans cette thèse, afin d'éviter des descriptions répétitives. Comme les séparations entrainement-validation-test varient d'un projet à l'autre, ces informations sont détaillées dans la configuration expérimentale des chapitres correspondants.

0.3.1 Jeux de données audio pathologiques

Au total, cette dissertation utilise sept jeux de données couvrant quatre types de pathologies liées à la parole. Deux modalités sont explorées, à savoir les signaux de parole et de toux. Les détails concernant ces jeux de données sont décrits dans les paragraphes suivants.

Cambridge COVID-19 Sounds Dataset Au moment de la rédaction, le Cambridge COVID-19 Sounds Dataset constitue la plus grande base de données audio publique disponible, avec divers symptômes respiratoires et des informations sur le statut COVID-19 auto-déclaré (Xia et al., 2021). Il contient un total de $552\,\mathrm{h}$ de données audio enregistrées à distance de 36 116 individus à travers le monde via une interface d'application. Pendant la collecte des données, les volontaires ont été invités à effectuer trois tâches : (1) discours scénarisé, où tous les participants ont prononcé la même phrase — 'J'espère que mes données peuvent aider à gérer la pandémie de virus' — trois fois dans leur langue maternelle ; (2) toux volontaire trois fois ; et (3) respiration profonde par la bouche pendant trois à cinq minutes. De plus, ils ont également auto-déclaré leur statut COVID ainsi que certaines informations de métadonnées (par exemple, sexe, âge, état médical préexistant, symptômes respiratoires).

Bien que la base de données COVID-19 Sounds soit avantageuse en raison de sa taille, elle peut ne pas être la version optimale pour entraîner un modèle diagnostique, car plusieurs facteurs n'ont pas été contrôlés, tels que la langue, la fréquence d'échantillonnage ou l'environnement acoustique. Concernant la détection des symptômes respiratoires, nous avons mis en place deux sous-ensembles de discours à partir de la base de données d'origine en éliminant plusieurs facteurs de confusion potentiels. Le premier sous-ensemble a été publié avec la base de données d'origine et a été utilisé comme référence pour la tâche de prédiction des symptômes respiratoires dans l'article COVID-19 Sounds (Xia et al., 2021). Ce sous-ensemble est désormais appelé CS-Res. CS-Res contient des échantillons en anglais de 6 623 individus avec des symptômes respiratoires (par exemple, mal de gorge, toux, etc.), ce qui donne un total de $31.3\,\mathrm{h}$ de données de parole. Les taux d'échantillonnage variaient en fonction des appareils utilisés, la majorité étant échantillonnée à $44.1\,\mathrm{kHz}$ (67,4%) et $16\,\mathrm{kHz}$ (29,8%). CS-Res a été soigneusement sélectionné afin que la qualité des enregistrements et l'équilibre des classes soient contrôlés.

Le deuxième sous-ensemble est similaire à CS-Res (en ce sens que seuls les échantillons en anglais sont utilisés), mais sans contrôler les autres facteurs. Ce sous-ensemble est appelé CS-Res-L, avec un total de $123.1\,\mathrm{h}$ de discours, dont 57,1% ont été échantillonnés à $16\,\mathrm{kHz}$ et 40,4% à $44.1\,\mathrm{kHz}$, le reste (2,5%) étant échantillonné à $8\,\mathrm{kHz}$ et $12\,\mathrm{kHz}$. Pour les deux sous-ensembles, les

participants ont été classés en deux classes : les positifs, ayant déclaré au moins un symptôme respiratoire, et les négatifs, n'ayant déclaré aucun symptôme. Avec CS-Res, nous avons suivi les partitions officielles décrites dans (Xia et al., 2021). Avec CS-Res-L, une séparation personnalisée indépendante des locuteurs a été effectuée avec un ratio de 7 :1 :2 (entrainement :validation). Par ailleurs, nous avons veillé à ce que la répartition des étiquettes de symptômes, du sexe et de l'âge soit similaire dans les trois divisions.

De plus, nous avons utilisé un autre sous-ensemble, appelé CS-Task2, qui a servi de référence pour la tâche de détection COVID-19 dans l'article COVID-19 Sounds (Xia et al., 2021). Deux modalités étaient incluses, à savoir la toux et la parole. Le CS-Task2 contenait à l'origine 1 486 échantillons provenant de 1 000 sujets. Comme il a été utilisé uniquement comme jeu de test à l'aveugle dans nos expériences, nous avons supprimé les utilisateurs dupliqués pour simuler un environnement réel, ce qui a permis d'utiliser un total de 1 000 échantillons de discours dans nos expériences. Parmi ces échantillons, 88% des sujets COVID-positifs sont symptomatiques et 41% des sujets COVID-négatifs présentent des symptômes ressemblant à ceux du COVID.

INTERSPEECH 2021 ComParE COVID-19 Dataset Le INTERSPEECH 2021 ComParE COVID-19 Dataset (CSS) est l'un des premiers jeux de données COVID-19 publics disponibles, publié avec le défi INTERSPEECH 2021 ComParE (Schuller et al., 2021). Comme le Cambridge COVID-19 Sounds Dataset, le CSS contient à la fois des signaux de toux et de parole, où le contenu des énoncés de parole est identique à celui de CS. Les enregistrements des ensembles de toux et de parole sont échantillonnés à $16\,\mathrm{kHz}$.

DiCOVA2 Dataset Ce jeu de données contient des données de parole utilisées dans le défi Second Diagnosing COVID-19 using Acoustics organisé en Inde (Sharma et al., 2022). DiCOVA2 a collecté des données acoustiques multimodales (c'est-à-dire de la parole, de la toux et de la respiration) à distance de 965 participants via des applications Android et Web. Les participants ont été invités à garder l'appareil à $10\,\mathrm{cm}$ de leur bouche pendant l'enregistrement. Pour la piste de parole, les participants ont compté de 1 à 20 à un rythme normal en anglais. Les enregistrements étaient échantillonnés à $48\,\mathrm{kHz}$. De plus, les participants ont auto-déclaré leurs métadonnées, telles que le sexe, les symptômes rencontrés et leur statut COVID-19, qui a été regroupé en labels binaires (positif ou négatif).

TORGO Dataset Ce jeu de données comprend des enregistrements de parole et des caractéristiques articulatoires synchronisées en 3D collectées auprès de témoins sains et de locuteurs atteints de paralysie cérébrale (CP) ou de sclérose latérale amyotrophique (SLA), les deux causes les plus fréquentes de dysarthrie (Rudzicz et al., 2012). La version publique du TORGO comprend 8 individus atteints de dysarthrie et 7 témoins sains. Lors de la collecte des données, tous les sujets ont été invités à lire des textes en anglais affichés sur un écran. Les données de parole ont

été enregistrées à un taux d'échantillonnage de $22.1\,\mathrm{kHz}$, tandis que les autres caractéristiques ont été capturées à $44.1\,\mathrm{kHz}$. Tous les sujets ont effectué quatre tâches de lecture différentes : (1) mots sans signification (ex. voyelles à haute et basse fréquence) ; (2) mots courts (ex. "oui", "non", "dos", etc.) ; (3) phrases restreintes (ex. "Le rapide renard brun saute par-dessus le chien paresseux") ; (4) phrases non restreintes (ex. décrire spontanément 30 images provenant des cartes photo Webber). Nous avons inclus des données des quatre tâches dans notre analyse.

Nemours Dataset Il s'agit d'une collection d'enregistrements de parole provenant de 12 hommes, dont 11 avec différents niveaux de dysarthrie et 1 témoin sain (Menendez-Pidal et al., 1996). Chaque participant a été invité à enregistrer 74 phrases sans signification du type "Le X est en train de Y le Z." ($X \neq Z$). Les phrases ont été générées en sélectionnant au hasard X et Z dans un ensemble de 74 noms monosyllabiques et en sélectionnant Y dans un ensemble de 37 verbes dissyllabiques. Tous les enregistrements ont été effectués dans une petite salle insonorisée avec un microphone monté sur une table et numérisés par la suite à un taux d'échantillonnage de $16\,\mathrm{kHz}$. Nous avons étiqueté tous les locuteurs en deux classes : ceux ayant des symptômes relativement sévères (scores inférieurs à 74.68, soit 6 locuteurs dysarthriques) et ceux ayant des symptômes plus légers (scores supérieurs à 74.68, soit 5 locuteurs dysarthriques et 1 témoin sain).

NCSC Dataset Il s'agit du NKI CCRT Speech Corpus(Clapham et al., 2012). NCSC contient des enregistrements de parole et des évaluations perceptuelles de 55 locuteurs (10 femmes et 45 hommes), qui ont subi un traitement de chimiothérapie et de radiothérapie concomitants (CCRT) pour un cancer de la région tête et cou. Les enregistrements et évaluations ont été réalisés à trois moments : (1) avant le CCRT ; (2) 10 semaines après le CCRT ; (3) 12 mois après le CCRT. Tous les sujets ont lu un passage de 189 mots d'un conte de fées néerlandais dans une salle traitée acoustiquement avec un taux d'échantillonnage de $44.1\,\mathrm{kHz}$. Treize pathologistes de la parole ont évalué l'intelligibilité de ces enregistrements de parole sur une échelle de 1 à 7. Nous avons utilisé les données de NCSC publiées dans le cadre du sous-défi de pathologie INTERSPEECH 2012(Schuller et al., 2012), où tous les enregistrements ont été étiquetés comme étant soit "intelligibles", soit "non intelligibles", et ont été divisés en trois ensembles indépendants pour l'entraînement et l'évaluation des modèles.

0.3.2 Jeux de données de parole synthétisée

De manière similaire aux autres sections de ce manuscrit, nous présentons ici un résumé des jeux de données de parole synthétisée. Un total de six jeux de données sont inclus, chacun étiquetant les échantillons comme réels (authentiques) ou faux (spoof). Les détails de ces jeux de données sont présentés dans les paragraphes suivants.

ASVspoof2019 La piste d'accès logique (LA) du défi ASVspoof 2019 est tirée du corpus multi-locuteurs VCTK (Todisco et al., 2019). Les énoncés de deepfake (DF) dans l'ensemble d'évaluation ont été générés à l'aide de 17 algorithmes différents de synthèse vocale (TTS) et de conversion vocale (VC), dont six ont été inclus dans les ensembles d'entraînement et de validation, tandis que les 11 autres étaient des attaques non vues. Tant les échantillons de parole réelle que générée sont échantillonnés à $16\,\mathrm{kHz}$.

ASVspoof2021 Un défi de détection de deepfakes (DF) a été ajouté au ASVspoof 2021 (Yamagishi et al., 2021). Les ensembles d'entraînement et de validation demeurent identiques à ceux de ASVspoof 2019, mais l'ensemble d'évaluation a été considérablement élargi, comportant 600 000 énoncés générés par plus de 100 modèles de TTS ou VC. De plus, les conditions de données et les techniques de compression audio utilisées dans l'ensemble d'évaluation diffèrent de celles des ensembles d'entraînement et de validation, créant ainsi un cadre plus exigeant pour évaluer la capacité de généralisation des détecteurs de deepfakes. Comme dans ASVspoof2019, tous les fichiers audio sont échantillonnés à $16\,\mathrm{kHz}$.

ASVspoof5 (2024) Tout en suivant une configuration similaire à celle de ASVspoof2021, ASVspoof5 inclut un nombre substantiellement plus élevé d'échantillons de parole dans les ensembles d'entraînement et de développement. L'ensemble d'évaluation contient des codecs et attaques inédits. Contrairement aux années précédentes, ASVspoof5 introduit également des échantillons d'attaques adversariales, ajoutant un défi supplémentaire à la généralisation des modèles.

In-the-wild Contrairement aux jeux de données des séries de défis ASVspoof, où la majorité des échantillons réels sont collectés en laboratoire, le jeu de données In-the-wild se distingue par l'inclusion de deepfakes audio (et de leurs contreparties réelles) pour un ensemble de politiciens et d'autres personnalités publiques, collectés à partir de sources accessibles au public telles que les réseaux sociaux et les plateformes de vidéo en streaming. Ce jeu de données comprend un total de 58 célébrités et politiciens, offrant ainsi $20.8\,\mathrm{h}$ de parole authentique et $17.2\,\mathrm{h}$ de parole usurpée, le tout échantillonné à $16\,\mathrm{kHz}$.

MLAAD Basé sur des échantillons de voix réelles issus du jeu de données M-AILABS, MLAAD constitue un jeu de données audio deepfake multilingue comprenant $160.2\,\mathrm{h}$ de données vocales synthétiques dans 23 langues différentes. Un total de 52 modèles génératifs ont été utilisés, couvrant 22 architectures différentes. En comparaison avec les données des séries de défis ASVspoof, MLAAD intègre des modèles génératifs plus récents, tels que VITS (Kim et al., 2021). Les audios générés sont échantillonnés à $16\,\mathrm{kHz}$.

WaveFake WaveFake contient de la parole resynthétisée à partir de 9 vocodeurs différents, utilisant les données de référence du jeu de données LJspeech (Frank et al., 2021). Contrairement aux jeux de données ASVspoof, les échantillons authentiques et les deepfakes dans WaveFake sont appariés, c'est-à-dire qu'ils possèdent le même contenu parlé et le même locuteur. Cela en fait un jeu de données pertinent pour analyser les artefacts liés aux vocodeurs. À l'instar de MLAAD, les audios générés sont échantillonnés à $16\,\mathrm{kHz}$.

0.3.3 Métriques d'évaluation

La performance diagnostique est généralement mesurée à l'aide de deux métriques principales, à savoir l'aire sous la courbe caractéristique de fonctionnement du récepteur (AUC-ROC) et le score F1. La première a été largement utilisée dans les tâches de détection de maladies comme métrique de base (Sharma et al., 2022; Xia et al., 2021). Cependant, il a été démontré que l'AUC-ROC est trop optimiste lorsqu'elle est appliquée à des ensembles de données fortement déséquilibrés (Fernández et al., 2018). En revanche, le score F1 s'avère être plus robuste dans des situations de déséquilibre. De plus, des métriques telles que la sensibilité (ou taux de vrais positifs, TPR), la spécificité (ou taux de vrais négatifs, TNR), et le rappel moyen non pondéré (UAR) sont souvent utilisées pour évaluer la performance du modèle sur une classe spécifique.

La détection de deepfakes s'apparente aux tâches diagnostiques où une décision binaire est obtenue, ce qui permet d'utiliser l'AUC-ROC et le score F1 pour évaluer la précision de la classification. De plus, le taux d'erreur égal (EER) est fréquemment utilisé comme métrique d'évaluation standard dans les défis ASVspoof passés (Todisco et al., 2019; Liu et al., 2023). L'EER est défini comme la moyenne arithmétique entre le TPR et le taux de faux positifs (FPR). Le seuil de l'EER représente le point où le taux de faux positifs (FPR) est égal au taux de faux négatifs (FNR).

Dans le cadre de l'anonymisation vocale, l'objectif est de tromper un modèle de reconnaissance de locuteur pré-entraîné afin d'obscurcir l'identité du locuteur. Dans un cas idéal, le modèle de reconnaissance de locuteur devrait obtenir une précision de 0% lorsqu'on lui demande de reconnaître l'utilisateur. Ainsi, nous utilisons le taux de mauvaise classification d'un modèle de reconnaissance de locuteur pour évaluer l'efficacité de l'anonymisation, un taux plus élevé étant synonyme d'une meilleure performance d'anonymisation.

0.4 Organisation de la thèse

Le reste de ce document est organisé comme suit : le Chapitre 2 présente un aperçu des signaux acoustiques humains ainsi que leurs représentations couramment utilisées. Il décrit également l'ensemble des jeux de données utilisés pour les études incluses dans cette thèse. Dans le Chapitre 3, nous introduisons les caractéristiques basées sur les connaissances extraites des si-

gnaux de toux et de parole, ainsi que leurs performances dans des tâches de diagnostic de santé, avec une interprétation des caractéristiques. Le Chapitre 4 présente deux cadres d'apprentissage de représentations pour des tâches de classification généralisables et interprétables. L'un repose sur un entraînement supervisé guidé par une carte de saillance pour le diagnostic de santé, et l'autre exploite l'apprentissage contrastif auto-supervisé pour la détection des deepfakes. Dans le Chapitre 5, nous explorons la faisabilité de l'utilisation de l'anonymisation vocale pour la protection de l'identité des utilisateurs dans des tâches de diagnostic de santé. Nous montrons également que les modèles génératifs utilisés pour l'anonymisation peuvent être modifiés pour générer des échantillons de deepfake afin d'attaquer les systèmes de détection. Dans le Chapitre 6, nous introduisons une architecture de modèle générique avec des performances de pointe dans les tâches de diagnostic de santé et de détection de deepfake. Elle démontre une bonne généralisabilité en zéro coup et une grande interprétabilité, tout en séparant les attributs liés au locuteur. Enfin, les conclusions sont résumées dans le Chapitre 7, accompagnées de discussions sur les perspectives de travaux futurs.

0.5 Chapitre 3 : Conception de caractéristiques basées sur les connaissances et apprentissage automatique pour le diagnostic de santé

Les caractéristiques basées sur les connaissances et les systèmes d'apprentissage automatique sont souvent privilégiées dans les scénarios où les données sont limitées et où l'interprétabilité est essentielle. Dans le cadre du diagnostic de santé, les ensembles de données vocales pathologiques présentent souvent un volume de données relativement plus faible par rapport à d'autres ensembles de données de parole, limitant ainsi l'efficacité des modèles d'apprentissage profond. De plus, le diagnostic de maladies nécessite fréquemment un processus décisionnel explicable, où une haute précision seule ne suffit pas pour une application dans le monde réel. Les caractéristiques acoustiques basées sur les connaissances, fondées sur des connaissances acoustiques et physiologiques, fournissent souvent des marqueurs fiables et interprétables, tels que les tremblements vocaux ou les variations irrégulières de la hauteur, qui sont corrélés avec des pathologies respiratoires et neuromusculaires (Schuller et al., 2021; Little et al., 2008). Dans des régimes de faible volume de données, des caractéristiques bien conçues, combinées à des modèles ML appropriés, peuvent surpasser les modèles DL (Zhu et al., ress), offrant ainsi une classification précise et une interprétabilité accrue. Dans cette section, nous présentons plusieurs ensembles de caractéristiques basées sur les connaissances pour la parole et la toux, ainsi que des systèmes ML inspirés de la physiologie pour aborder la détection du COVID-19.

Les principales contributions de ce chapitre sont résumées comme suit : Nous proposons d'abord deux ensembles de caractéristiques faites à la main et des systèmes ML pour les signaux de parole et de toux, respectivement, afin de capturer les anomalies liées à la respiration et à l'articulation. Nous montrons ensuite de manière empirique que les caractéristiques proposées

surpassent les caractéristiques faites à la main existantes et les caractéristiques profondes dans les tâches de détection du COVID-19 dans un cadre de croisement de jeux de données. Enfin, nous montrons que les caractéristiques proposées peuvent être utilisées pour identifier les facteurs affectant à la fois les systèmes d'articulation et respiratoire, fournissant ainsi des explications cliniques plus détaillées des résultats obtenus.

0.5.1 Méthodes proposées

Caractéristiques de la parole Nous proposons d'utiliser des caractéristiques basées sur le spectre de modulation et la prédiction linéaire (LP) pour caractériser les anomalies des articulations et du système respiratoire. Le spectrogramme couramment utilisé fournit des informations sur l'évolution des composantes fréquentielles en fonction du temps. Cependant, les sources de bruit communes se chevauchent à la fois dans le temps et dans la fréquence, ce qui rend cette représentation sous-optimale pour les tâches de classification reposant sur des données bruyantes. Le spectre de modulation est connu pour capturer des périodicités d'ordre supérieur du signal, qui ne sont pas évidentes dans les domaines temporel et temps-fréquence (Falk et al., 2010b; Avila et al., 2018). Cette propriété s'est révélée utile pour la caractérisation des maladies, telles que la détection des troubles du spectre de l'autisme à partir des pleurs de nourrissons et des vocalisations non verbales (Bedoya et al., 2020), ainsi que pour la surveillance automatisée de l'intelligibilité de la parole dysarthrique (Falk et al., 2012). Nous l'explorons ici comme une représentation utile et robuste pour la détection du COVID-19. De plus, nous partons de l'hypothèse que le COVID-19 peut affecter à la fois les propriétés du tractus vocal (par exemple, en raison de la fatigue musculaire accrue (Solomon, 2006; Helms et al., 2020)) et le signal d'excitation (par exemple, une phonation altérée (Quatieri et al., 2020)). Alors que les caractéristiques du spectre de modulation peuvent fournir des informations sur les premiers, nous proposons également d'utiliser l'analyse de prédiction linéaire (LP) pour décomposer le signal de parole en paramètres du tractus vocal (c'est-à-dire, les coefficients LP) et en source d'excitation (c'est-à-dire, le résiduel LP) (Makhoul, 1975). En général, nous extrayons plusieurs descripteurs et statistiques à partir des deux représentations de signal susmentionnées et combinons les caractéristiques extraites dans un seul ensemble pour la détection du COVID-19.

Caractéristiques de la toux L'information sur la phase de la toux s'est révélée cruciale pour le diagnostic de maladies, bien qu'elle ait été négligée par les modèles existants de traitement de la toux. Par conséquent, nous proposons de segmenter les enregistrements de toux en trois phases différentes (c'est-à-dire, inhalation, compression et expulsion) et d'extraire des caractéristiques de chaque phase séparément pour explorer leur utilité dans la détection du COVID-19. Cependant, comme l'information de phase n'est pas étiquetée dans les ensembles de données existants, nous avons d'abord effectué une segmentation manuelle des enregistrements de toux pour obtenir une segmentation précise des phases. L'une des six étiquettes suivantes est attribuée à chaque évé-

nement audio : (1) inhalation, (2) compression, (3) expulsion (toux), (4) bruit, (5) silence, et (6) autre (qui inclut tous les types de sons articulatoires autres que les sons de toux). Nous avons ensuite extrait huit caractéristiques temporelles des toux annotées par phase, telles que la durée de chaque phase et leurs relations, qui complètent l'information manquante dans les caractéristiques acoustiques.

0.5.2 Résultats et discussions

Nous avons évalué les deux ensembles de caractéristiques sur les ensembles de données ComParE, DiCOVA2 et Cambridge pour la parole et la toux. Pour les tâches liées à la parole, la fusion des caractéristiques basées sur le spectre de modulation et la prédiction linéaire (LP) atteint des performances systématiquement supérieures à celles des systèmes de référence (c'est-à-dire, basés sur openSMILE, les réseaux BiLSTM et VGGish) dans les tests intra- et inter-ensembles de données. Ces résultats démontrent que les caractéristiques proposées pour la parole sont plus généralisables à travers différentes distributions de données. Une analyse approfondie des caractéristiques les mieux classées a mis en évidence des corrélations avec des phénomènes tels que la souffle dans la parole, l'enrouement vocal et l'inflammation du système articulatoire, qui sont identifiés comme des symptômes majeurs du COVID-19.

Concernant les tâches liées à la toux, nous avons suivi une configuration de tests similaire à celle des tâches de parole, où différentes caractéristiques ont été évaluées dans des contextes intra- et inter-ensembles de données. Nous avons constaté qu'un petit ensemble de 11 caractéristiques temporelles liées aux différentes phases de la toux permet déjà d'obtenir des performances comparables à celles de plus de 6000 caractéristiques acoustiques sur DiCOVA2. Lorsque ces caractéristiques temporelles sont fusionnées avec des caractéristiques acoustiques segmentées par phase, une amélioration significative est observée dans les tests intra- et inter-ensembles de données. Ces résultats confirment notre hypothèse selon laquelle les caractéristiques des phases de la toux peuvent compléter les informations manquantes dans les caractéristiques acoustiques, et que la segmentation en phases est essentielle pour capturer des biomarqueurs liés aux maladies pouvant se généraliser à travers différents ensembles de données.

0.6 Chapitre 4 : Apprentissage profond de représentations généralisables et interprétables

Dans le chapitre précédent, nous avons démontré que les caractéristiques basées sur les connaissances peuvent favoriser l'interprétation des décisions, tout en surpassant certains modèles d'apprentissage profond (DL) dans des contextes de données limitées, comme la détection de voix pathologiques. Plus récemment, des ensembles de données plus volumineux ont été créés (Xia et al., 2021; Coppock et al., 2022b), ce qui motive l'apprentissage de représentations gé-

néralisables dans une approche de bout en bout en utilisant des modèles plus grands. Cependant, des études ont montré que les améliorations sont plus significatives pour les tests intra-domaine (c'est-à-dire entraînés et testés sur le même jeu de données), tandis que la généralisation à des données non vues reste une tâche difficile.

Une manière d'améliorer la généralisabilité et l'interprétabilité des schémas d'apprentissage des représentations profondes consiste à injecter des connaissances spécifiques au domaine dans la conception de la stratégie d'apprentissage (Dash et al., 2022). En appliquant certaines contraintes lors du processus d'apprentissage de bout en bout, on peut forcer les modèles à accorder moins d'importance aux biais non pertinents pour la tâche et à se concentrer sur les attributs liés à la tâche. De cette manière, les modèles peuvent encore apprendre des représentations qui contiennent plus d'informations que les caractéristiques conçues manuellement, tout en respectant les attentes spécifiques à la tâche. Dans ce chapitre, nous introduisons deux approches d'apprentissage profond des représentations, respectivement pour les diagnostics de santé et la détection de deepfakes.

Les principales contributions de ce chapitre sont résumées comme suit. Nous montrons comment les connaissances spécifiques au domaine peuvent être intégrées dans la conception de différents cadres d'apprentissage des représentations pour améliorer la généralisabilité et l'interprétabilité. Nous nous concentrons sur deux schémas d'apprentissage : l'apprentissage supervisé pour les diagnostics de santé et l'apprentissage auto-supervisé pour la détection de deepfakes. Dans les deux scénarios, les systèmes proposés atteignent des performances à l'état de l'art dans les tests de généralisation avec des résultats interprétables.

0.6.1 Méthodes proposées

MTR-CRNN. Dans le chapitre précédent, nous avons démontré que le spectrogramme de modulation contient des informations cruciales pour les diagnostics médicaux. Cependant, ce spectrogramme de modulation était moyenné temporellement, ce qui signifie que la représentation ne captait que les variations de fréquence et de fréquence de modulation, tout en négligeant les dynamiques temporelles. De plus, les informations spectrales étaient décrites par des caractéristiques conçues manuellement, ce qui pouvait limiter leur pouvoir discriminatif. Pour surmonter ces limitations, nous proposons le modèle MTR-CRNN, qui utilise une version tridimensionnelle du spectrogramme de modulation (c'est-à-dire le modulation tensorgram, ou MTR) en entrée, et l'analyse à l'aide d'un réseau neuronal convolutif récurrent (CRNN) pour explorer les informations spectrales et temporelles.

Une autre innovation du modèle proposé est la carte de saillance spectro-temporelle, qui agrège les informations spectrales et temporelles à partir du MTR et identifie les régions les plus discriminatives. Des études antérieures ont suggéré que différentes régions du spectre de modulation correspondent à des propriétés spécifiques du signal vocal (Greenberg et al., 1997;

Sarria-Paja et al., 2013; Zhu et al., 2022). Une meilleure compréhension des régions du spectre de modulation utilisées par le modèle permettrait une interprétation plus approfondie des résultats et pourrait fournir des informations sur les propriétés acoustiques de la parole affectée par la COVID-19. Cependant, les cartes de saillance par gradient existantes sont principalement conçues pour des images 2D, où seules les informations spatiales sont prises en compte. Dans notre cas, les entrées sont des tensorgrammes 3D où les motifs spectraux 2D évoluent au fil du temps. Par conséquent, nous avons conçu une méthode de carte de saillance spectro-temporelle qui intègre l'importance temporelle lors du calcul de la saillance pour chaque pixel du motif spectral.

SLIM. Pour la majorité des modèles de génération vocale, les sous-espaces du style et des informations linguistiques sont supposés indépendants l'un de l'autre (Tan et al., 2021; Kaur et al., 2023; Triantafyllopoulos et al., 2023; Mohammadi et al., 2017). Par exemple, les systèmes de conversion vocale (VC) modifient la voix d'un énoncé en remplaçant les embeddings du locuteur source par ceux du locuteur cible (Mohammadi et al., 2017; Triantafyllopoulos et al., 2023), en supposant que ces embeddings ne contiennent aucune information linguistique. De même, les systèmes modernes de synthèse vocale (TTS) reposent sur des représentations apprises de manière indépendante pour modéliser différents aspects de la parole (par exemple, le texte, le locuteur, l'émotion) afin de produire une parole expressive (Baevski et al., 2022; Desplanques et al., 2020; Triantafyllopoulos et al., 2024).

En raison de cette hypothèse de désentrelacement, un décalage est susceptible d'exister entre les informations de style et linguistiques dans les discours générés par TTS/VC, ce qui les différencie des discours réels. Pour capturer ce décalage, nous proposons un modèle de désalignement style-linguistique (style-linguistics mismatch), ou SLIM. SLIM est entraîné en deux étapes. La première étape fonctionne uniquement sur les données réelles et utilise un apprentissage auto-supervisé pour construire des représentations de style et linguistiques ainsi que leurs dépendances pour la parole réelle. Lors de la deuxième étape, un classifieur est ajusté sur les représentations apprises via un entraînement supervisé sur des ensembles de données deepfake avec des étiquettes binaires (réel/faux).

0.6.2 Résultats et discussions

Pour évaluer les performances du modèle MTR-CRNN proposé, nous le comparons aux modèles les plus performants sur CSS et DiCOVA2 dans des configurations intra-dataset et inter-dataset. Sans l'utilisation des cartes de saillance spectro-temporelles, MTR-CRNN obtient de meilleures performances sur CSS et des performances légèrement inférieures sur DiCOVA2 par rapport aux modèles de référence. Dans le contexte inter-dataset, bien que MTR-CRNN surpasse les modèles de référence, les scores AUC-ROC globaux sont significativement inférieurs à ceux obtenus en intra-dataset. Cette tendance correspond à celle observée avec les caractéristiques

basées sur les connaissances dans le chapitre précédent, où tous les ensembles de caractéristiques ont montré des difficultés lorsqu'ils ont été testés sur des données non vues. En filtrant les régions moins discriminatives de MTR à l'aide des cartes de saillance spectro-temporelles, une amélioration notable de 10% des scores AUC-ROC peut être observée en termes de généralisabilité inter-dataset (.600 à .705 pour CSS et .509 à .651 pour DiCOVA2). À notre connaissance, ce sont les meilleures performances inter-dataset rapportées jusqu'à présent, ce qui démontre que le modèle MTR-CRNN proposé, associé aux cartes de saillance spectro-temporelles, peut efficacement capturer les biomarqueurs liés aux maladies.

En ce qui concerne les tâches de détection de deepfakes, nous comparons SLIM à plusieurs modèles parmi les plus performants sur différents ensembles de données de deepfakes, dont la plupart sont construits sur des représentations SSL. Avec seulement 5 millions de paramètres entraînables, SLIM surpasse de loin toutes les représentations SSL gelées sur des ensembles de données non vus (réduction de 12 à 20% de l'EER sur MLAAD et In-the-wild). En mode de fine-tuning complet, des améliorations supplémentaires sont obtenues, avec une diminution de 5% de l'EER par rapport aux SOTA. Ces résultats démontrent que la stratégie proposée d'apprentissage auto-supervisé sur des discours réels permet d'apprendre une meilleure représentation pour les tâches de détection de deepfakes. De plus, nous quantifions le désalignement style-linguistique dans les discours réels et artificiels et observons une distance cosinus significativement plus grande entre les incorporations de dépendance style-linguistique pour les discours artificiels, ce qui corrobore notre hypothèse initiale sur le désalignement style-linguistique dans les discours artificiels.

À partir des deux tâches introduites ci-dessus, nous avons montré que les connaissances du domaine peuvent être intégrées dans les approches d'apprentissage supervisé et auto-supervisé, et leur apporter des bénéfices. Comparées aux caractéristiques conçues manuellement et aux représentations de type boîte noire, ces représentations profondes basées sur les connaissances offrent l'avantage d'une performance élevée en détection tout en étant interprétables.

0.7 Chapter 5 : Privacy-preserving speech applications via voice anonymization

Dans le chapitre précédent, nous avons démontré que les représentations encodées par de grands modèles peuvent considérablement améliorer la précision par rapport aux caractéristiques basées sur les connaissances. Ces grands modèles contiennent généralement des centaines de millions à des milliards de paramètres, nécessitant des ressources informatiques importantes tant pour l'entraînement que pour l'inférence. Prenons l'exemple des diagnostics de santé à distance : les poids des modèles ne sont généralement pas stockés localement sur les appareils mobiles (Wang et al., 2018a), et les données vocales sont envoyées et traitées dans le cloud. Les décisions sont ensuite transmises à l'appareil de l'utilisateur. Cependant, cette transmission des données vocales via le cloud peut poser de graves menaces pour la vie privée des utilisateurs, puisque leur

voix peut être associée à des informations médicales sensibles telles que leur état de santé (Latif et al., 2020a), l'évolution d'une maladie (Harel et al., 2004), ou leur état mental (Low et al., 2020), pour ne citer que quelques exemples. Une technique visant à préserver l'identité de l'utilisateur est l'anonymisation de la voix (Tomashenko et al., 2022b,a). Traditionnellement, l'anonymisation de la voix est utilisée dans les tâches de reconnaissance automatique de la parole (ASR) pour dissimuler l'identité de l'utilisateur tout en conservant le contenu du discours. Dans ce chapitre, nous explorons si l'anonymisation de la voix peut être utilisée pour protéger la vie privée dans les applications de diagnostic vocal. Notre objectif final est de développer un modèle d'anonymisation capable de masquer l'identité du locuteur tout en préservant les informations liées à la santé. En parallèle, étant donné que certains systèmes d'anonymisation reposent sur des modèles génératifs vocaux, les discours anonymisés produits peuvent être considérés comme des "deepfakes pathologiques". Lorsqu'un locuteur cible est déterminé, le modèle d'anonymisation peut être utilisé pour convertir l'identité du locuteur tout en conservant l'état de santé du locuteur source. Cela représente une nouvelle menace pour les modèles de détection de deepfakes, car ces derniers n'ont pas été entraînés ni évalués avec des discours pathologiques. Ainsi, un second objectif de ce chapitre est d'examiner la vulnérabilité des modèles existants de détection de deepfakes face aux discours pathologiques anonymisés.

0.7.1 Méthodes proposées

Dans la première partie, nous avons réalisé une évaluation approfondie des effets de différents anonymisateurs sur plusieurs modèles de diagnostic. Trois types d'anonymisateurs ont été utilisés. Le premier repose sur une technique classique de traitement du signal et ne nécessite pas d'entraînement de modèle. Il utilise la méthode dite du coefficient de McAdams (McAdams, 1984) pour décaler la position des formants mesurés à l'aide de la codification linéaire prédictive (O'Shaughnessy, 1988). Par souci de simplicité, nous l'appelons méthode McAdams dans le texte suivant. Les deuxième et troisième anonymisateurs s'appuient sur un pipeline similaire à celui des modèles TTS/VC. Le deuxième anonymisateur, que nous nommons Ling-GAN, transcrit la parole en séquences de phonèmes et sépare les embeddings du locuteur à l'aide de modèles pré-entraînés d'ASR et de SV, puis remplace les embeddings du locuteur d'origine par des embeddings générés par GAN pour masquer l'identité du locuteur. Ces embeddings générés par GAN, ainsi que la séquence de phonèmes, sont ensuite envoyés dans un modèle acoustique pour la synthèse vocale. Il est important de souligner que ce GAN préexistant n'a pas été entraîné sur des données de parole pathologique (Meyer et al., 2022a). En conséquence, les embeddings générés peuvent ne pas refléter les attributs liés à la santé, ce qui peut affecter la précision des diagnostics. Le troisième anonymisateur peut être considéré comme une version améliorée du deuxième, isolant et préservant davantage les informations prosodiques tout en affinant le GAN sur des données de parole pathologique afin d'intégrer des informations de santé dans les embeddings du locuteur. À l'instar de Ling-GAN, Ling-Pros-GAN décompose d'abord la parole en séquence de phonèmes,

prosodie et embeddings du locuteur, puis remplace les embeddings d'origine par ceux générés par le GAN. Nous utilisons un seuil de 0.3 lors du remplacement des embeddings du locuteur, seuil qui s'est avéré efficace pour supprimer l'identité du locuteur tout en préservant au maximum les informations liées à la santé. Concernant les modèles de diagnostic, nous nous appuyons sur les modèles de référence et sur nos modèles basés sur les caractéristiques de modulation présentés dans les chapitres précédents. Nous explorons une liste exhaustive de différentes conditions d'anonymisation, où les données d'entraînement et de test peuvent être anonymisées par différents anonymisateurs (ou non). Au total, nous incluons cinq modèles de diagnostic et les évaluons dans treize conditions différentes. Pour évaluer l'efficacité de l'anonymisation, nous utilisons également un modèle de vérification de locuteur pré-entraîné afin de vérifier si les paroles originales et anonymisées proviennent du même locuteur. Un anonymisateur idéal devrait tromper le modèle de vérification tout en maintenant la même décision diagnostique.

Dans la deuxième partie, nous exploitons les paroles anonymisées générées précédemment, que nous introduisons avec des échantillons de paroles originales (pathologiques et saines) dans des modèles de détection de deepfake pré-entraînés. Ces modèles de détection ont été entraînés uniquement sur des paroles saines, rendant leur robustesse face aux paroles pathologiques inconnue. Dans un cas idéal, un modèle robuste de détection de deepfake devrait pouvoir classer les paroles originales (qu'elles soient saines ou pathologiques) comme réelles, et les paroles anonymisées comme fausses.

0.7.2 Résultats et discussions

Concernant les tâches de diagnostic sur les paroles anonymisées, toutes les méthodes d'anonymisation montrent une dégradation de la précision des diagnostics, la dégradation la plus importante étant observée avec les systèmes qui modifient directement les embeddings du locuteur, à savoir Ling-GAN et Ling-Pros-GAN. Ces deux systèmes atteignent cependant des performances quasi parfaites en termes d'obfuscation de l'identité du locuteur. La méthode McAdams, bien qu'elle préserve les informations liées à la santé, s'est révélée moins efficace pour l'anonymisation. Entre les deux méthodes basées sur les GAN, Ling-Pros-GAN engendre une moindre dégradation des performances de diagnostic. Notre analyse dans différents espaces de caractéristiques montre que le plus grand décalage dans les distributions est observé avec les paroles anonymisées par Ling-GAN, tandis que les deux autres méthodes montrent des décalages moindres. Ces résultats corroborent avec le classement des dégradations des performances de diagnostic observées avec les trois anonymisateurs. Globalement, nos résultats suggèrent que les méthodes existantes ne parviennent pas à préserver efficacement les informations diagnostiques tout en obfusquant les identifiants du locuteur.

Lorsqu'ils sont utilisés comme générateurs de deepfakes pathologiques, ces anonymisateurs entraînent une baisse significative des performances de détection, particulièrement lorsque les

informations de santé et de prosodie sont préservées. Cela suggère que les modèles actuels de détection de deepfakes ne sont pas encore robustes face à ce type unique de deepfake. Bien que les informations liées à la santé soient souvent négligées dans les modèles génératifs existants, les recherches futures devraient prendre en compte les deepfakes pathologiques pour améliorer la généralisabilité des modèles.

0.8 Chapitre 6 : Un modèle indépendant des tâches, explicable et respectueux de la vie privée

Dans les chapitres précédents, nous avons présenté différentes méthodes visant à améliorer la généralisabilité, l'interprétabilité et la préservation de la vie privée dans les applications vocales. Bien que ces techniques aient montré des résultats prometteurs, chacune aborde seulement un aspect du problème. Par exemple, les modèles proposés au chapitre 4 visent à améliorer la généralisation mais ne prennent pas en compte la confidentialité des utilisateurs, nécessitant ainsi le recours aux méthodes d'anonymisation vocale introduites au chapitre 5 pour combler cette lacune. Cependant, cela complique la conception globale du système et exige un réglage minutieux de chaque composant pour éviter une dégradation des performances. Des problèmes similaires se retrouvent dans de nombreuses applications vocales en conditions réelles, où l'agrégation de plusieurs modèles spécifiques à des tâches distinctes est nécessaire pour répondre à divers besoins.

Une façon de résoudre ce problème consiste à explorer des modèles capables d'être indépendants des tâches. Ainsi, des représentations universelles pour la parole, telles que Wav2vec2 (Baevski et al., 2020), WavLM (Chen et al., 2022) et HuBERT (Hsu et al., 2021), ont été proposées et ont suscité un intérêt croissant en raison de leur généralisabilité à différentes tâches en aval (Yang et al., 2021a). Toutefois, elles ont principalement été évaluées sur des tâches vocales conventionnelles, telles que la reconnaissance vocale, l'identification du locuteur et la reconnaissance des émotions. Avec l'intérêt croissant pour le diagnostic de santé et la détection de deepfakes, il n'est pas évident que ces représentations universelles soient optimales pour ces tâches. De plus, étant donné que ces représentations sont conçues pour inclure à la fois des informations linguistiques et paralinguistiques, il est difficile de déterminer les attributs acoustiques utilisés pour chaque tâche spécifique. Enfin, puisque ces représentations sont connues pour contenir des informations relatives à l'identité de l'utilisateur, leur utilisation dans des applications sensibles à la vie privée peut être limitée. Dans cette section, nous explorons la possibilité de concevoir une architecture de modèle applicable à une variété de tâches vocales (c'est-à-dire indépendante des tâches), capable d'apprendre une représentation explicable et dissociée de l'identité du locuteur.

0.8.1 Méthodes proposées

Les représentations universelles pour la parole existantes peuvent être considérées comme des représentations temporelles, où chaque cadre temporel (20-$25\,\mathrm{ms}$) est associé à un vecteur de caractéristiques de haute dimension. Ce design des encodeurs est principalement destiné à traiter des tâches nécessitant des décisions au niveau des cadres, comme la reconnaissance automatique de la parole (ASR) ou la détection d'événements acoustiques. Cependant, les tâches telles que le diagnostic de santé et la détection de deepfakes opèrent au niveau des énoncés et ne nécessitent pas de décisions à chaque cadre. Une solution couramment adoptée consiste à effectuer un regroupement temporel pour obtenir une décision globale par énoncé. Cette méthode repose sur l'hypothèse que les attributs au niveau des énoncés, tels que l'identité du locuteur et les informations de santé, peuvent également être encodés au niveau des cadres. Les statistiques temporelles peuvent alors être utilisées pour déduire le motif global. Cependant, nous soutenons que les statistiques temporelles (par exemple, la moyenne et l'écart type) ne peuvent pas capturer avec précision les attributs globaux. Une partie des informations au niveau des cadres (par exemple, les phonèmes) peut être redondante pour les tâches au niveau des énoncés et entraîner une dégradation des performances du modèle.

Par conséquent, nous proposons une nouvelle architecture de modèle, appelée WavRx, qui se concentre sur les dynamiques à long terme des représentations universelles. Nous nous inspirons du concept de spectre de modulation introduit dans les chapitres précédents, qui extrait les périodicités de chaque canal de fréquence dans le spectrogramme, et appliquons une opération similaire à chaque canal de caractéristiques des représentations universelles. Étant donné que les représentations universelles sont généralement encodées avec une résolution temporelle de 20-$25\,\mathrm{ms}$, nous appliquons une transformée de Fourier à court terme (STFT) le long de l'axe temporel, avec une fenêtre de taille beaucoup plus grande. Cela permet de convertir la représentation temporelle en domaine fréquentiel, capturant ainsi les dynamiques des variations temporelles, ce qui constitue le cœur de WavRx. Intuitivement, nous formulons l'hypothèse que les attributs globaux peuvent être mieux encodés par les fréquences des représentations universelles, ce qui élimine la majorité des détails au niveau des cadres. Finalement, WavRx étend la représentation universelle en intégrant une branche dédiée aux dynamiques, qui encode les fréquences des embeddings temporels. Pour les tâches de diagnostic et de détection de deepfakes, les représentations dynamiques et temporelles peuvent ensuite être fusionnées afin d'obtenir les meilleures performances.

0.8.2 Résultats et discussions

Nous avons évalué WavRx sur deux types de tâches : le diagnostic de santé, où six ensembles de données pathologiques couvrant quatre maladies différentes sont utilisés, et la détection de deepfakes, avec trois ensembles de données dédiés. Lors de l'évaluation, nous avons suivi une

configuration similaire à celle des chapitres précédents, avec des entraînements et tests réalisés dans des contextes intra-dataset et inter-datasets. Concernant le diagnostic de santé, WavRx dépasse les autres grands modèles de la parole (par exemple, Wav2vec2, WavLM, HuBERT, AST, ECAPA-TDNN) sur cinq des six ensembles de données évalués. Il démontre également une capacité de diagnostic en zéro-shot lorsqu'il est testé sur des données inconnues présentant des pathologies similaires à celles des données d'entraînement (par exemple, la dysarthrie). En outre, lorsqu'un classificateur est entraîné sur les embeddings de WavRx pour l'identification de locuteurs, la précision obtenue est nettement inférieure à celle obtenue avec des représentations universelles. En visualisant les embeddings projetés de WavRx, ceux provenant de différentes maladies sont bien discriminés, tandis que ceux provenant de différents locuteurs sont regroupés. Ensemble, ces résultats suggèrent que l'identité du locuteur est mieux dissociée des attributs pertinents pour la tâche dans les embeddings de WavRx, sans aucune dégradation des performances pour la tâche aval. De manière similaire, WavRx surpasse également les représentations universelles pour les tâches de détection de deepfakes. Lorsque les modèles sont entraînés avec ASVspoof2019 et testés sur ASVspoof2021, une dégradation moindre est observée lorsque les embeddings dynamiques sont utilisés.

Par ailleurs, alors que les représentations universelles sont connues pour être difficiles à interpréter en raison des variations temporelles entre les échantillons, WavRx présente l'avantage de fournir une visualisation statique dans le domaine fréquentiel. Avec les tâches de diagnostic et de détection de deepfakes, des motifs similaires dans le domaine fréquentiel sont observés pour des classes similaires, par exemple, des maladies ayant des pathologies proches ou des deepfakes générés à l'aide des mêmes vocodeurs.

0.9 Conclusions

Dans cette thèse, nous nous sommes concentrés sur le développement d'applications acoustiques généralisables, interprétables et respectueuses de la vie privée, telles que les diagnostics de santé et la détection des deepfakes. Les principales contributions de cette thèse peuvent être résumées comme suit :

Jeu de données. Les jeux de données de toux existants manquent d'annotations sur les phases de la toux, qui sont fréquemment utilisées en évaluation clinique pour aider au diagnostic des maladies respiratoires. Pour combler cette lacune, nous avons ouvert un jeu de données annoté comprenant plus de 1000 enregistrements de toux liés à la COVID-19, annotés manuellement. Il s'agit du plus grand jeu de données de toux avec des annotations détaillées à ce jour. La mise à disposition de ce jeu de données encouragera le développement de nouveaux modèles de traitement de la toux et permettra une meilleure compréhension de la physiologie de la toux.

Ensembles de caractéristiques. Nous avons proposé deux ensembles de caractéristiques basées sur des connaissances pour les tâches de diagnostic de santé. Le premier combine des caractéristiques extraites du spectre de modulation et des caractéristiques de prédiction linéaire, visant à capturer les anomalies articulatoires et respiratoires dans la parole pathologique. Nous avons démontré leur efficacité dans les tâches de détection de la COVID-19, surpassant les caractéristiques largement utilisées d'openSMILE en termes de généralisation sur des données non vues. Le second ensemble de caractéristiques, basé sur notre jeu de données annoté, consiste en des descripteurs liés aux phases qui identifient des anomalies au niveau des poumons, de la trachée et des cordes vocales. Nous avons montré que les caractéristiques liées aux phases surpassent de manière significative les caractéristiques acoustiques extraites à l'aide d'outils standard.

Architectures de modèles et stratégies d'apprentissage. Nous avons proposé plusieurs nouvelles architectures de modèles et stratégies d'apprentissage avec une meilleure généralisabilité et interprétabilité. Cela inclut : (1) MTR-CRNN, qui apprend à capturer des biomarqueurs respiratoires en se concentrant sur des régions spécifiques dans la représentation du modulation tensorgram, identifiées grâce à une carte de saillance spectrale-temporelle. Sa généralisabilité inter-données pour la détection de la COVID-19 est nettement supérieure aux réseaux de référence comme VGGish, ainsi qu'à d'autres systèmes primés basés sur des connaissances. (2) SLIM, un cadre d'apprentissage innovant qui capture la dépendance entre le style et le contenu linguistique dans la parole réelle et l'utilise pour la détection des deepfakes. Contrairement aux modèles existants qui s'appuient uniquement sur l'apprentissage supervisé, SLIM intègre une phase d'apprentissage contrastif auto-supervisé pour capturer les modèles de parole réels absents des représentations SSL existantes. Sans fine-tuning complet, nous avons montré que SLIM surpasse de manière significative d'autres grands modèles de parole entièrement ajustés, en particulier pour la détection d'attaques inconnues. (3) Pour protéger la vie privée des utilisateurs dans les systèmes de diagnostic vocal, nous avons évalué les méthodes existantes d'anonymisation de la voix et proposé une méthode améliorée qui atténue la dégradation des précisions diagnostiques. Enfin, (4) nous avons introduit WavRx, un modèle indépendant des tâches, interprétable et préservant la vie privée. WavRx atteint non seulement l'état de l'art sur divers jeux de données de parole pathologique et synthétisée, mais produit également des représentations plus facilement interprétables et indépendantes des locuteurs. Ces caractéristiques démontrent son potentiel pour une utilisation dans des applications réelles.

1 INTRODUCTION

Human-centric audio applications, such as automatic speech recognition (ASR), speaker verification (SV), and respiratory health diagnostics, have become integral to diverse fields, including customer service (Wang et al., 2023b), autonomous driving (Cui et al., 2024), healthcare (Lugović et al., 2016), and biometrics (Markowitz, 2000), just to name a few. As these applications transition into real-world use, achieving robust performance demands a focus on three essential properties : *generalizability*, *interpretability*, and *security*.

Generalizability refers to a model's capability to maintain high predictive accuracy when tested on data with unseen distributions, i.e., distributions that differ from the data seen during training (Wang et al., 2022). While machine learning (ML) systems usually assume the similar data distributions for training and test sets, this is usually hard to achieve in real-world scenarios, where various factors, such as background noise, recording conditions, or different dialects, could mislead the model to make false predictions. As ML models are known to be data-hungry, one way to improve *generalizability* is through data manipulation, such as augmentation or generation, to expand the diversity and quantity of training data (Adila et al., 2022). While such methods are simple and cheap to implement, they can be ineffective when little knowledge is known about the testset distributions and may introduce unwanted model biases (Maharana et al., 2022). Another line of work focuses on investigating novel model architectures and learning strategies, which aim at learning a generalizable representation while keeping training data intact (Wang et al., 2022). A generalizable representation is expected to contain only task-related attributes while being independent of unwanted factors. In the case of speech health diagnostics, for example, a generalizable representation should contain only health-related information while being robust to other acoustic attributes, such as gender, language, and speech content.

For some speech applications, such as healthcare, high accuracy alone is not sufficient, since the decision-making process of ML/DL models needs to be understood by human experts in order to establish trust. In these cases, models need to be interpretable so that both correct and incorrect decisions can be explained and validated by humans. Take health diagnostics as an example, *interpretability* enables clinicians to understand the basis of a model's predictions, such as specific vocal biomarkers or acoustic patterns indicative of certain conditions (Doshi-Velez et al., 2017). An interpretable model helps experts assess the reliability of its outputs, facilitating informed clinical decisions that align with domain knowledge. However, one challenging issue faced by ML/DL models is the trade-off between performance and *interpretability*. While pure data-driven models usually provide better performance, the lack of *interpretability* limits their application in real-world scenarios.

Lastly, since voice carries a variety of speaker-specific information, such as gender and age, it has been used as a biometric modality in many user verification applications, such as gaining

access to personal banks (Markowitz, 2000). Meanwhile, with the rapidly growing number of voice cloning tools, it is now possible to replicate someone's voice using only a 3-second voice recording (Wang et al., 2023a). As such, regulations have been made worldwide to protect individuals' voice biometrics and to limit the use of voice cloning tools, some examples include the General Data Protection Regulation (GDPR) in Europe (Zarsky, 2016) and the Personal Information Protection Law (PIPL) in China (Calzada, 2022). In this dissertation, we propose to address the *security* and *privacy-preserving* issue via voice anonymization which obfuscates user identity when sharing the data, and deepfake detection to identify synthesized voice.

1.1 Objectives

The research work described in this dissertation aims at improving the generalizability, explainability, and security of audio applications. The detailed objectives are summarized as follows :

1. Design novel features and representations for improving the generalizability and interpretability of health diagnostics and deepfake detection models;
2. Explore novel model architectures that can be reused for different datasets and tasks;
3. Develop model training strategies to learn generalizable representations and improve the explainability of decision-making process of black-box models;
4. Design privacy-preserving techniques that secure user's identity while maximizing downstream task performance.

1.2 Summary of Contributions

The contribution of the research work reported in this dissertation can be summarized as follows :

1. We curated a phase-annotated cough dataset and developed several physiological-inspired systems for speech and cough based health diagnostics. Different from the black-box models, the proposed features and models are motivated from the nature of human articulation and respiration, hence can be better explained and generalized across different datasets.
2. We designed a novel end-to-end model with a learnable spectral-temporal saliency map for health diagnostics. With the guidance from the saliency map, the proposed model showed to generalize well to speech samples from unseen datasets with more interpretable decision-making processes.
3. We conducted a comprehensive evaluation of the feasibility of using voice anonymization to protect the user identity leakage for speech-based health diagnostic models. We further designed an improved anonymization approach to tackle limitations of the existing techniques.
4. We explored the robustness of current deepfake detection models to anonymized health and pathological speech, and further discussed the observed limitations and challenges.

5. We proposed a novel pre-training strategy for generalized audio deepfake detection, which focuses on the style-linguistic mismatch in generated speech. Existing models typically rely on supervised training with universal representations, our proposed representation, on the other hand, is learned in a self-supervised manner without the need for fake speech samples, and is shown to generalize better to unseen attacks than other widely used universal representations.

6. We proposed a generic model architecture that focuses on the temporal dynamics of universal speech representations. The model is shown to achieve state-of-the-art on multiple pathological speech datasets as well as deepfake speech datasets, along with good zero-shot generalizability. We also showed that the learned dynamics are independent of speaker identities, hence can be used for privacy-preserving applications.

1.3 Publications

Here, we list the published papers related to the projects conducted during the PhD program.

Publications included in the dissertation

Journal publications

Zhu, Yi, and Tiago Falk. "WavRx : a Disease-Agnostic, Generalizable, and Privacy-Preserving Speech Health Diagnostic Model." IEEE Journal of Biomedical Health Informatics, early access, 14 pages, Sept. 2024.

Zhu, Yi, and Tiago H. Falk. "Spectral–temporal saliency masks and modulation tensorgrams for generalizable COVID-19 detection." Computer Speech & Language, Vol. 86, June 2024.

Zhu, Yi, Mohamed Imoussaïne-Aïkous, Carolyn Côté-Lussier, and Tiago H. Falk. "On the Impact of Voice Anonymization on Speech Diagnostic Applications : a Case Study on COVID-19 Detection." IEEE Transactions on Information Forensics and Security, Vol. 19, pp.5151-5165, April 2024.

Zhu, Yi, Abhishek Tiwari, João Monteiro, Shruti Kshirsagar, and Tiago Henrique Falk. "COVID-19 detection via fusion of modulation spectrum and linear prediction speech features." IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 31, pp. 1536-1549, April 2023.

Conference proceedings

Zhu, Yi, Surya Koppisetti, Trang Tran, and Gaurav Bharaj. "SLIM : Style-Linguistics Mismatch Model for Generalized Audio Deepfake Detection." Advances in neural information processing systems (NeurIPS) 2024.

Zhu, Yi, Saurabh Powar, and Tiago H. Falk. "Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection." In 2024 IEEE International Confe-

rence on Acoustics, Speech, and Signal Processing (ICASSP) Workshops, pp. 139-143. IEEE, 2024.

Zhu, Yi, Mohamed Imoussaıne-Aıkous, Carolyn Côté-Lussier, and Tiago H. Falk. "Investigating biases in COVID-19 diagnostic systems processed with automated speech anonymization algorithms." In 3rd Symposium on Security and Privacy in Speech Communication, pp. 46-54. 2023.

Zhu, Yi, Mahil Hussain Shaik, and Tiago H. Falk. "On the importance of different cough phases for COVID-19 detection." International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

Zhu, Yi, and Tiago H. Falk. "Fusion of modulation spectral and spectral features with symptom metadata for improved speech-based COVID-19 detection." International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8997-9001. IEEE, 2022.

Zhu, Yi, Alex Mariakakis, Eyal De Lara, and Tiago H. Falk. "How generalizable and interpretable are speech-based COVID-19 detection systems? : A comparative analysis and new system proposal." In 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1-5. IEEE, 2022.

Other publications

Journal publications

Zhu, Yi, Mahsa Abdollahi, Ségolène Maucourt, Nico Coallier, Heitor R. Guimarães, Pierre Giovenazzo, and Tiago H. Falk. "MSPB : a longitudinal multi-sensor dataset with phenotypic trait measurements from honey bees." Nature Scientific Data 11, no. 1 (2024) : 860.

Fischer, Bennet, Mario Chemnitz, **Yi Zhu**, Nicolas Perron, Piotr Roztocki, Benjamin MacLellan, Luigi Di Lauro et al. "Neuromorphic Computing via Fission-based Broadband Frequency Generation." Advanced Science 10, no. 35 (2023) : 2303835.

Tiwari, Abhishek, Raymundo Cassani, Shruti Kshirsagar, Diana P. Tobon, **Yi Zhu**, and Tiago H. Falk. "Modulation spectral signal representation for quality measurement and enhancement of wearable device data : A technical note." Sensors 22, no. 12 (2022) : 4579.

Conference proceedings

Zhu, Yi, Mahsa Abdollahi, Ségolène Maucourt, Nico Coallier, Heitor R. Guimarães, Pierre Giovenazzo, and Tiago H. Falk. "Early prediction of honeybee hive winter survivability using multimodal sensor data." In 2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), pp. 657-662. IEEE, 2023.

Guimarães, Heitor R., Mahsa Abdollahi, **Yi Zhu**, Ségolène Maucourt, Nico Coallier, Pierre Giovenazzo, and Tiago H. Falk. "Adapting Self-Supervised Features for Background Speech Detection

in Beehive Audio Recordings." In 2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), pp. 663-667. IEEE, 2023.

Guimarães, Heitor R., **Yi Zhu**, Orson Mengara, Anderson R. Avila, and Tiago H. Falk. "Assessing the vulnerability of self-supervised speech representations for keyword spotting under white-box adversarial attacks." In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1672-1677. IEEE, 2023.

1.4 Open-source code and model hyperparameters

Here, we list the links to open-source code produced within research projects reported in this document.

1. `https://github.com/zhu00121/WavRx`
2. `https://github.com/zhu00121/Anonymized-speech-diagnostics`
3. `https://github.com/zhu00121/Universal-representation-dynamics-of-deepfake-speech`
4. `https://github.com/MuSAELab/modulation_filterbanks`
5. `https://github.com/zhu00121/two_stage_fusion`
6. `https://github.com/MuSAELab/COVID_Cough_Phases`

The hyper-parameters of the open-sourced models are described in detail in the corresponding publications. To facilitate better readability of this dissertation, we leave out these details in the text, and encourage interested readers to refer to the publications listed in Section 1.3.

1.5 Thesis organization

The remainder of this document is organized as follows : Chapter 2 includes some background on the human acoustic signals, as well as their commonly-used representations. It also describes all the datasets used for the studies included in this dissertation. In Chapter 3, we introduce the knowledge-based features extracted from cough and speech signals, as well as their performance on health diagnostic tasks with feature interpretation. In Chapter 4, we show two different representation learning frameworks for generalizable and interpretable classification tasks. One relies on saliency-map guided supervised training for health diagnostics, and the other leverages self-supervised contrastive learning for deepfake detection. In Chapter 5, we explore the feasibility of using voice anonymization for user identity protection in health diagnostics tasks. We further show that the generative models used for anonymization can be modified to generate deepfake samples to attack detection systems. In Chapter 6, we introduce a generic model architecture with state-of-the-art performance on health diagnostics and deepfake detection tasks. It demonstrates good

zero-shot generalizability and interpretability, while disentangling the speaker-related attributes. Finally, conclusions are summarized in Chapter 7 along with discussions on future work directions.

2 BACKGROUND

In this Section, we firstly introduce the physiological nature behind the production of different sound modalities. We then describe the knowledge-based features as well as deep representations extracted from these sound modalities for different downstream tasks. Lastly, we dive into three major applications this dissertation focuses on and introduce their background and current challenges.

2.1 Production and Application of Speech and Cough Signals

2.1.1 Speech signals

The production of speech signals is a complex physiological process that involves the coordination of multiple anatomical structures. Speech production begins with the respiratory system, where airflow is generated from the lungs (Honda, 2008). The air is expelled through the trachea and into the larynx, where the vocal folds are located. When the vocal folds are closed and air passes through them, they vibrate to produce voiced sounds. Meanwhile, speech also includes unvoiced segments, where airflow is not interrupted by the vocal folds, resulting in sounds like fricatives (e.g., /s/ or /f/), plosives (e.g., /p/, /t/), and nasal stops (e.g., /m/, /n/). These sound waves are further shaped into distinct speech sounds through the actions of the articulatory system, which includes the tongue, lips, teeth, and palate (Stevens, 2000). The movement of these articulators modifies the airflow which leads to the speech that can be perceived by humans.

Speech carries layers of information, such as the speech content, speaker identity, emotion, health, etc. One common way to categorize them is to divide speech into linguistic and paralinguistic aspects (Schuller et al., 2013a). Linguistic aspects refer to the explicit verbal content, including vocabulary and syntax, while paralinguistic attributes encompass non-verbal elements that influence how the message is interpreted (Scherer, 2003; Schuller et al., 2013a). For example, paralinguistic attributes like prosody, pitch, tone, loudness, and speech rate, significantly impact the listener's perception and understanding of the spoken message.

Among many paralinguistic speech applications, one emerging topic is speech-based health diagnostics (Ramanarayanan et al., 2022). Since respiratory and articulatory systems can be affected by a large number of diseases, such as COVID-19, Parkinson's disease, Dysarthria, or Chronic Obstructive Pulmonary Disease (COPD), changes in speech patterns can be induced by abnormalities in these regions, which makes speech a potential candidate for health diagnostics and monitoring. For instance, individuals with neurodegenerative diseases, such as Alzheimer's or Parkinson's, may exhibit distinct alterations in their speech, including changes in volume, clarity,

and rhythm, which can be analyzed to monitor disease progression (Brabenec et al., 2017). More recently, a substantial body of work has shown the use of speech for COVID-19 detection (Deshpande et al., 2020a).

2.1.2 Cough signals

Another acoustic modality entailing information regarding the respiratory health condition is the cough. While coughs originate from the airflow from lungs, which is similar to speech, it is a reflex action and does not require fine motor control. The process of coughing involves a rapid inhalation followed by a forceful exhalation against a closed glottis, which then abruptly opens, releasing a burst of air (Chung et al., 2008; Piirila et al., 1995). For involuntary coughs, the vocal folds do not vibrate, as the primary goal is not sound production but the expulsion of substances from the respiratory tract (Fontana et al., 2006). For voluntary coughs (i.e., forced coughs), however, studies have shown more involvement of vocal folds in order to initiate the bursting sound production (Fontana et al., 2006; Lee et al., 2002).

Cough is an important symptom of over 100 diseases (Korpáš et al., 1996), such as COPD, asthma, and COVID-19 (Ciotti et al., 2020), to name just a few. Pathological abnormalities in the respiratory system (e.g., phlegm, inflammation in lung bifurcations) can be reflected by the changes in the characteristics of the coughs (Korpas et al., 1987). A cough sound can be generally divided into three phases : (1) inhalation (inspiration), where the glottis remains wide open to bring air into the lung area; (2) compression, where a forced expiratory effort is against the closed glottis; and (3) expulsion, where the glottis is open again in the moment of the transient explosive sound is generated (Chung et al., 2008). Clinical studies have shown that cough phase patterns vary across respiratory diseases (Piirila et al., 1995) and abnormalities in different cough phases can be linked to different pathological origins (Korpáš et al., 1996). For example, the length of the compression phase can be indicative of the location of secretions in lung airways (Piirila et al., 1995), while the first explosive sound reflects the condition of tracheal bifurcations (Korpáš et al., 1996). Given that cough can be collected in the same way as speech while carrying potentially fewer biasing factors (e.g., speech content), it has been used for health diagnostic tasks in several recent studies (Infante et al., 2017; Alqudaihi et al., 2021; Tena et al., 2022).

2.2 Audio Representations

2.2.1 Knowledge-based features

Earlier works have emphasized the use of knowledge-based features to characterize underlying speech characteristics. Beyond conventional speech features, such as mel-spectrograms or mel-frequency cepstral coefficients (MFCCs), various studies have explored a wide range of fea-

tures tailored to specific applications. Some examples include phonation and articulation features for speech motor control tasks (Arias-Vergara et al., 2017; Vásquez-Correa et al., 2018), linear prediction (LP)-based features for separating vocal source and vocal tract information (Zhu et al., 2023f), prosodic features to represent pitch, energy, and rhythm (Schuller et al., 2013b), as well as glottal flow features to model vocal fold behavior (Drugman et al., 2012). Furthermore, some larger feature sets have been proposed which aggregates different low-level descriptors of speech and their statistics. The openSMILE ComParE feature set (Eyben et al., 2010), for instance, has been widely adopted as a baseline across several tasks, including detection challenges involving different types of speech attributes (Schuller et al., 2021, 2017). Other widely used feature sets include GeMAPS (Eyben et al., 2015), which focuses on capturing voice quality and emotional expression, and the extended GeMAPS (eGeMAPS), designed to balance comprehensiveness and computational efficiency for paralinguistic tasks (Eyben et al., 2015).

Once extracted, these hand-crafted features are typically fed into classical ML classifiers, such as support vector machines (SVM) or random forests. The benefits of using hand-crafted features include their explainability, the suitability for small datasets, and their potential for better generalization across diverse datasets (Zhu et al., 2022). These features also facilitate the integration of domain-specific knowledge, offering insights into the underlying mechanisms of speech production and perception. These advantages make knowledge-based features great candidates for tasks with limited amount of data and varied data distributions, such as health diagnostics.

2.2.2 Deep representations

Recent advancements in deep learning have significantly enhanced model performance on various speech tasks through learning representations from deep neural networks (hereinafter referred to as deep representations). For utterance-level representations, one commonly-used deep representation is the x-vector (Snyder et al., 2018), which employs deep neural networks for speaker verification and identification tasks. Utilizing a time-delay neural network (TDNN), the x-vector model extracts speaker embeddings from variable-length speech segments, effectively capturing temporal dynamics within the audio signal (Snyder et al., 2018). The ECAPA-TDNN architecture further incorporates a context aggregation module and a lightweight convolutional design to enhance the representation of speaker characteristics (Desplanques et al., 2020). By leveraging attention mechanisms and multi-scale feature extraction, ECAPA-TDNN has been shown to outperform x-vector on speaker recognition benchmarks while minimizing the influence of noise and variability in recording conditions. While trained for speaker tasks, both representations have demonstrated good performance when transferred to other paralinguistic tasks (Morais et al., 2022).

In parallel to these supervised learning techniques, self-supervised learning (SSL) has gained attention for its ability to learn richer representations from unlabelled data. Models such as Wav2Vec2 (Baevski et al., 2020), WavLM (Chen et al., 2022), and HuBERT (Hsu et al., 2021)

exemplify this paradigm. Wav2Vec 2.0 employs contrastive learning to model speech representations by predicting masked portions of input audio based on contextual information (Baevski et al., 2020). The embeddings from different transformer layers have shown useful on several different downstream tasks, including automatic speech recognition (ASR), speaker verification, and emotion recognition. WavLM builds on Wav2Vec2 by performing denoising together with masked token prediction, and has shown better performance in paralinguistic tasks, such as emotion recognition and speaker identification (Chen et al., 2022). HuBERT utilizes a hierarchical framework to learn contextualized speech representations by masking input audio and predicting these masked segments (Hsu et al., 2021). These SSL models have proven effective for a wide range of speech-related tasks (Yang et al., 2021a), showcasing the efficacy of SSL for learning generalizable representations.

The success of self-supervised pre-training has extended to foundation models and universal representations for specific domains, where large transformer-based models are pre-trained with domain-specific data, such as respiratory sound (Baur et al., 2024; Zhang et al., 2024b,a). For example, Health-aware Audio Representation (HeAR) is a masked autoencoder (MAE) based model pre-trained on 313 million 2s-long respiratory audio clips (Baur et al., 2024), and achieves better performance than general SSL audio representations on health diagnostics tasks. Recent works have also explored other generative and contrastive methods for large-scale pre-training (Zhang et al., 2024a), as well as leveraging pre-trained large language models to improve generalizability (Zhang et al., 2024b). Compared to knowledge-based features and supervised trained representations, SSL representations have shown marked improvement on generalization to unseen data without the need for large-scale downstream supervised training.

2.3 Scope of Applications

In this manuscript, we focus on three distinct audio applications. The following subsections provide an introduction of each application and the key challenges they face.

2.3.1 Audio-based Health Diagnostics

During recent years, speech and respiratory signals have emerged as promising modalities for disease diagnosis and remote health monitoring. Compared to conventional testing methods, such as X-rays, CT-scans, and blood tests, audio diagnostic systems are advantageous in terms of short processing time and low testing cost. Audio diagnostics is based on the assumption that diseases causing abnormalities in articulatory and/or respiratory systems would lead to an atypical pattern in human voice and respiration (Fagherazzi et al., 2021). Such abnormality could be due to a variety of reasons, such as impaired neuromuscular control or an inflammation in the vocal tract and lungs (Fagherazzi et al., 2021).

While the impact on the acoustic signals may sometimes be imperceptible to humans, a ML model could be trained to detect certain disease-related digital biomarkers. In a typical audio-based health diagnostic system, users are prompted to read several sentences or cough a few times, while the acoustic signals are collected remotely via a microphone. The signals are then sent to the diagnostic model to extract digital biomarkers and decide if the user is disease-positive or negative. Existing audio diagnostic models are facing generalization and interpretability issues (Naudé, 2020). The former refers to the case where models trained on one dataset have been observed to perform poorly on unseen datasets; while the latter points to the black-box nature of most diagnostic models, where the decision-making process cannot be explained by a human. Together, these issues make it challenging for existing models to be deployed in real-world settings. In this manuscript, we introduce novel features and ML techniques that improve the model generalization and interpretability.

2.3.2 Synthesized Speech Detection

The growing interest in generative models has led to an expansion of publicly available tools that can closely mimic the voice of a real person (Tan et al., 2021). Text-to-speech (TTS) or voice conversion (VC) systems can now be used to synthesize a fake voice from only a few seconds of real speech recordings (Tan, 2023). When these generation tools are used by bad actors, their outputs, commonly referred to as 'audio deepfakes' (Khanjani et al., 2023), can pose serious dangers. Examples include impersonation of celebrities/family members for robocalls (Knibbs, 2024), illegal access to voice-guarded bank accounts (Cox, 2023), or forgery of evidence in court (Pender, 2023). Therefore, synthesized speech detection models have been explored to tackle audio deepfakes. However, similar to audio diagnostic systems, deepfake detection models lack generalization to unseen generative models (i.e., unseen attacks) as well as explainability of model decisions. In this manuscript, we introduce novel models and learning strategies that help bridge this gap.

2.3.3 Voice Anonymization

Today, the great majority of speech applications rely on deep neural network (DNN) architectures with models containing hundreds of millions of parameters with this number continuously rising. Commonly, these parameters are not stored locally on mobile devices (Wang et al., 2018a) and speech data are sent and processed in the cloud; decisions are then transmitted back to the user device. As more and more cases of cyberattacks are being reported (Stupp, 2019; Kaloudi et al., 2020; Yamin et al., 2021), this transmission of speech data over the cloud could pose serious threats to user privacy and security. It has been previously reported that voice assistants and many third-party applications collect users' voice without their knowledge and share it with advertising partners (Iqbal et al., 2022). For example, Amazon patented a technique which recognizes health

status via conversations with users and advertises the related medicines to them (Jin et al., 2018). This could be particularly risky for speech diagnostics applications, since the user's voice could be linked with sensitive medical information, such as health status (Latif et al., 2020a), disease progression (Harel et al., 2004), or mental state (Low et al., 2020), just to name a few. As such, speech privacy-preserving methods have gained increased attention globally, especially with the release of regulations, such as the General Data Protection Regulation (GDPR) in Europe (Zarsky, 2016) and the Personal Information Protection Law (PIPL) in China (Calzada, 2022). The latter is particularly aimed at personal biometrics (i.e., voice, facial image, and fingerprints).

Alternately, voice anonymization methods have emerged with the aim of manipulating the speech signal such that information about speaker identity is obfuscated, while the linguistic content and other para-linguistic attributes (e.g., timbre, naturalness) remain intact. Given the burgeoning interest in this domain, the Voice Privacy Challenge series (VPC) were held in 2020 and 2022 to foster development in speech anonymization techniques (Tomashenko et al., 2022b,a). However, these challenges were aimed at developing anonymization methods for downstream automatic speech recognition tasks (Tomashenko et al., 2022b,a; Fang et al., 2019; Meyer et al., 2022a), where linguistic content was preserved, but not para-linguistic information. Therefore, part of this dissertation is dedicated to exploring the impact of voice anonymization techniques on health diagnostics tasks, as well as new anonymization approaches that can simultaneously preserve health information and user privacy.

Moreover, as the field of voice-based remote diagnostics progresses, it is likely that in the future the first step in a diagnostic system will entail a deepfake detection step, to ensure that fraud is not being committed (e.g., see (Chen et al., 2021b)). Since voice anonymization is one application of voice conversion, the anonymized voice can be regarded as a special type of 'deepfake' and be blocked by future deepfake detection systems. As an exploratory goal, in this thesis we measure the robustness of existing deepfake detection models (trained on non-pathological speech) to anonymized pathological speech and discuss the limitations and challenges observed.

2.4 Datasets

Since some datasets have been reused across several projects, here we list all pathological and synthesized audio datasets that have been used by papers incorporated in this thesis to avoid repetitive descriptions in later chapters. Since the train-validation-test splits vary across projects, this information is detailed in the experimental setup in the corresponding chapters.

TABLEAU 2.1 : Summary of pathological audio datasets and their occurrences in different chapters in this dissertation.

Dataset	Modality	Pathology	Chapter
‡ CS-Res	Speech	Respiratory symptoms	6
‡ CS-Res-L	Speech	Respiratory symptoms	6
‡ CS-Task2	Cough	COVID-19	3
‡ CS-Task2	Speech	COVID-19	3 & 4 & 5
CSS	Speech & Cough	COVID-19	3 & 4 & 5
DiCOVA2	Speech & Cough	COVID-19	3 & 4 & 5
TORGO	Speech	Dysarthria	6
Nemours	Speech	Dysarthria	6
NCSC	Speech	Cervical cancer	6

‡ Subsets from the Cambridge COVID-19 Sound Dataset.

2.4.1 Pathological audio datasets

In total, in this dissertation we employed seven datasets covering four types of speech-related pathologies. Two modalities are explored, namely speech and cough signals. Details about these datasets are described in the following paragraphs. Table 2.1 outlines the datasets used in different chapters.

Cambridge COVID-19 Sounds Dataset At the time of writing, the Cambridge COVID-19 Sounds Dataset exemplifies the largest publicly available audio database with various respiratory symptoms and self-reported COVID-19 status (Xia et al., 2021). It contains a total of $552\,\mathrm{h}$ of audio data recorded remotely from 36,116 individuals around the globe via an app interface. During data collection, volunteers were prompted to conduct three tasks : (1) scripted speech, where all participants uttered the same sentence – 'I hope my data can help to manage the virus pandemic' – three times in their mother tongue; (2) voluntary cough for three times; and (3) deep breathing through the mouth for three to five minutes. In addition, they also self-reported their COVID-status along with certain metadata information (e.g., gender, age, pre-existing medication condition, respiratory symptoms). It should be emphasized that not all participants had conducted a PCR test before recording, hence the COVID-status was in the form of a subjective evaluation (e.g., 'I think never had COVID-19') rather than a binary label (i.e., positive vs negative). Such ambiguity in COVID-19 labels motivated us to use this database primarily for respiratory abnormality detection, and only as an additional evaluation set for COVID-19 detection.

Although the COVID-19 Sounds database is advantageous in its size, it may not be the optimal version to train a diagnostics model considering that multiple factors were not controlled, such as language, sampling rate, or acoustic environment. Regarding respiratory symptom detection, we set up two speech subsets from the original database by screening out several potential confounding factors. The first subset was released along with the original database, which was used as the

benchmark data for the respiratory symptom prediction task in the COVID-19 Sounds paper (Xia et al., 2021). This subset is henceforth referred to as CS-Res. CS-Res contains English samples from 6,623 individuals with respiratory symptoms (e.g., sore throat, cough, etc.), resulting in a total of $31.3\,\mathrm{h}$ speech data. The sampling rates varied upon different devices used, with the majority sampled at $44.1\,\mathrm{kHz}$ (67.4%) and $16\,\mathrm{kHz}$ (29.8%). CS-Res was carefully curated so that the recording quality and class balance were controlled.

The second subset is similar to CS-Res (in that only English samples are used) but without controlling for the other factors. This subset is referred to as CS-Res-L, with a total of $123.1\,\mathrm{h}$ of speech, of which 57.1% were sampled at $16\,\mathrm{kHz}$ and 40.4% at $44.1\,\mathrm{kHz}$ and the rest (2.5%) were sampled at $8\,\mathrm{kHz}$ and $12\,\mathrm{kHz}$. For both subsets, participants were labelled into two classes, namely the positive ones who reported at least one respiratory symptom, and the negative ones reporting no symptoms at all. With CS-Res, we followed the official partitions as described in (Xia et al., 2021). With CS-Res-L, a customized speaker-independent split was performed with a ratio of 7 :1 :2 (train :validation :test). Meanwhile, we ensured that the distribution of symptom labels, gender, and age were similar in all three splits.

Furthermore, we employed another subset, termed CS-Task2, which was used as the benchmark data for the COVID-19 detection task in the COVID-19 Sounds paper (Xia et al., 2021). Two modalities were included, namely cough and speech signals. The CS-Task2 originally contained 1,486 samples from 1,000 subjects. Since it was used only as a blind test set in our experiments, we removed the duplicated users to simulate a real-world setting, thus a total of 1,000 speech samples were used in our experiments. In these samples, 88% of the COVID-positive subjects are symptomatic and 41% of the COVID-negative subjects are with COVID-like symptoms.

INTERSPEECH 2021 ComParE COVID-19 Dataset The INTERSPEECH 2021 ComParE COVID-19 Dataset (CSS) is one of the earliest publicly available COVID-19 datasets, which was released together with the INTERSPEECH 2021 ComParE challenge (Schuller et al., 2021). Similar to the Cambridge COVID-19 Sounds Dataset, CSS contains both cough and speech signals, where the speech utterance content is the same as CS. Recordings in both cough and speech sets are sampled at $16\,\mathrm{kHz}$. With the speech set, a total of eight languages were included, with the majority of samples being uttered in English, Portuguese, Italian, and Spanish. The gender split was 45% female and 35% male (the remaining 20% chose *Prefer not to say* or *Other*). Close to 28% of the COVID-positive subjects were asymptomatic, while only 59% of the COVID-negative subjects were without respiratory symptoms.

DiCOVA2 Dataset This dataset contains speech data used in the Second Diagnosing COVID-19 using Acoustics challenge organized in India (Sharma et al., 2022). DiCOVA2 collected multimodal acoustic data (i.e., speech, cough, and breathing) remotely from a total of 965 participants via Android and Web apps. Participants were advised to keep the device $10\,\mathrm{cm}$ from their mouth

during recording. For the speech track, participants did number counting from 1 to 20 in a normal pace in English. The recordings were sampled at $48\,\mathrm{kHz}$. Furthermore, participants self-reported their metadata, such as gender, experienced symptoms, and COVID-19 status which was grouped into binary labels (either positive or negative). Since the test labels were not made accessible to the public, we used the validation data as the new test set, and partitioned the original training data into the new training set and validation set (80% :20%).

TORGO Dataset This dataset consists of speech recordings and synchronized 3D articulatory features collected from healthy controls and speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), the two most prevalent causes of dysarthria (Rudzicz et al., 2012). TORGO was originally designed to develop ASR models for dysarthric individuals. The publicly available version of TORGO includes 8 individuals with dysarthria and 7 healthy controls. During data collection, all subjects were asked to read English text from the screen. The speech data were recorded from two microphones, one facing the participant at a distance of $61\,\mathrm{cm}$ with a sampling rate of $22.1\,\mathrm{kHz}$ while the other is head-mounted with a sampling rate of $44.1\,\mathrm{kHz}$. Only the data from the front-facing microphone were employed herein. All subjects conducted four different reading tasks : (1) non-words (e.g., high- and low-pitch vowels); (2) short words (e.g. 'yes', 'no', 'back', etc.); (3) restricted sentence (e.g., "The quick brown fox jumps over the lazy dog"); (4) unrestricted sentence (e.g., spontaneously describe 30 images from the Webber Photo Cards). We included data from all four tasks in our analysis. As there were no official data partitions, we followed the speaker-independent principle to split all 15 subjects into three sets [1] : (1) training set ('FC02','F03','F01','MC04','MC03','M02'); (2) validation set ('MC02','FC01','M03','M01'); and (3) test set ('FC03','F04','MC01','M05','M04'). The average dysarthria severity was made similar for all three sets.

Nemours Dataset This is a collection of speech recordings from 12 males, 11 with different levels of dysarthria and 1 healthy control (Menendez-Pidal et al., 1996). Each participant was asked to record 74 nonsense sentences of the form "The X is Ying the Z." ($X \neq Z$). Sentences were generated by randomly selecting X and Z without replacement from a set of 74 monosyllabic nouns and selecting Y without replacement from a set of 37 disyllabic verbs. All recordings were collected in a small sound dampened room with one table-mounted microphone, and digitized subsequently using a $16\,\mathrm{kHz}$ sampling rate. Apart from recording sessions, Nemours also included a perception session where 5 listeners tried to identify the words of the nonsense sentences. The average number of correct identifications was calculated per speaker and the Frenchay speaker assessment scores were reported, which reflects the severity of dysarthria. The average assessment score of the dysarthric speakers is 74.68 with a standard deviation of 14.54. We labelled all speakers into two classes, namely the relatively severe individuals with scores lower than 74.68 (6 dysarthria

1. 'F' and 'M' stand for female and male; 'C' stands for healthy controls.

TABLEAU 2.2 : Summary of the synthesized datasets. The main characteristics are described in the notes to differentiate one from the others.

Dataset	Chapter	Notes
ASVspoof2019	4 & 6	Conventional generative models without codecs
ASVspoof2021	4 & 6	Different speech codecs; 100+ types of attacks
ASVspoof5 (2024)	4	Different codecs + more recent generative models
In-the-wild	4	Celebrity voice collected from in-the-wild conditions
MLAAD	4	Multilingual deepfakes
WaveFake	6	Vocoder-resynthesized speech

speakers), and the mild ones with scores higher than 74.68 (5 dysarthria speakers plus 1 healthy control).

NCSC Dataset This refers to the "NKI CCRT Speech Corpus" (Clapham et al., 2012). NCSC contains speech recordings and perceptual evaluations of 55 speakers (10 female and 45 male), who underwent concomitant chemo-radiation treatment (CCRT) for cancer of the head and neck region. Recordings and evaluations were made at three moments : (1) before CCRT; (2) 10-weeks after CCRT; and (3) 12-months after CCRT. All subjects read a 189-word passage from a Dutch fairy tale in a sound-treated room. Speech data were collected using a microphone with a $44.1\,\mathrm{kHz}$ sampling rate at a distance of $30\,\mathrm{cm}$ from mouth. Thirteen speech pathologists rated the intelligibility of these speech recordings on a scale of 1 to 7. We employed the NCSC data released by the *INTERSPEECH 2012 Pathology Sub-Challenge* (Schuller et al., 2012), where all recordings were labelled either as 'intelligible' or 'non-intelligible', and were split into three independent sets for model training and evaluation. However, since the test labels were not accessible to the public, we used the validation set as the new test set and split the original training set into the new training and validation set with a ratio of 80% :20%.

2.4.2 Synthesized speech datasets

Similarly, we present a summary of all the synthesized speech datasets included in this manuscript in Table 2.2. In total, there are six synthesized datasets where samples are labelled either as real (i.e., bonafide) or fake (i.e., spoof). Details of these datasets can be found in the following paragraphs.

ASVspoof2019 The Logical Access (LA) track of the ASVspoof 2019 challenge is derived from the multi-speaker VCTK corpus (Todisco et al., 2019). The DF utterances in the evaluation set were generated using 17 different TTS and VC algorithms, among which six were seen in the training and validation sets while the remaining 11 were unseen attacks. Both real and generated speech are sampled at $16\,\mathrm{kHz}$.

ASVspoof2021 A DF track was added to the 2021 ASVspoof challenge (Yamagishi et al., 2021). The training and validation sets remained the same as in ASVspoof 2019, but the evaluation set was increased to 600k utterances generated by more than 100 TTS or VC models. Furthermore, the data conditions and audio compression techniques of the evaluation data also differed from that of the training and validation set, hence posing a more stringent setting to evaluate the generalizability of DF detectors. Similar to ASVspoof2019, all speech files are sampled at $16\,\mathrm{kHz}$.

ASVspoof5 (2024) While sharing the similar setup as ASVspoof2021, ASVspoof5 includes a substantially large amount of speech samples in train and development sets, where the evaluation set contains unseen codecs and attacks. Different from previous years, ASVspoof5 includes adversarial attack samples to challenge the model generalization.

In-the-wild Different from the ASVspoof challenge series datasets where majority of the real samples are collected in lab, the In-the-wild dataset is unique for including audio deepfakes (and corresponding benign audio) for a set of politicians and other public figures, collected from publicly available sources such as social networks and video streaming platforms. A total of 58 celebrities and politicians are included, leading to $20.8\,\mathrm{h}$ of bona-fide and $17.2\,\mathrm{h}$ of spoofed audio, with a sampling rate at $16\,\mathrm{kHz}$.

MLAAD Based on real voice samples from the M-AILABS dataset, MLAAD is a multilingual deepfake audio dataset comprising $160.2\,\mathrm{h}$ of synthetic voice data in 23 different languages. A total of 52 generative models are used, covering 22 different model architectures. Compared to the ASVspoof challenge series data, MLAAD includes more recent audio generative models, such as VITS (Kim et al., 2021). The generated audios are sampled at $16\,\mathrm{kHz}$.

WaveFake WaveFake contains speech resynthesized from 9 different vocoders using ground-truth from the LJspeech (Frank et al., 2021). Different from the ASVspoof datasets, genuine and DF samples in WaveFake are paired (i.e., same spoken content and speaker), thus serves as a good candidate to capture the vocoder-related artifacts. Similar to MLAAD, the generated audios are sampled at $16\,\mathrm{kHz}$.

2.5 Evaluation metrics

Diagnostics performance can usually be measured by two metrics, namely the area under the receiver operating curve (AUC-ROC) and the F1 score. The former has been used widely in disease detection tasks as a baseline metric (Sharma et al., 2022; Xia et al., 2021). However, AUC-ROC has been shown to be over-optimistic when evaluating on extremely imbalanced datasets (Fernández et al., 2018). The F1 metric, on the other side, is more robust in an imbalanced

setting. Furthermore, sensitivity (or true positive rate, TPR) and specificity (or true negative rate, TNR) are often used to gauge model performance on one specific class. For diagnostics tasks, we usually regard healthy as negative, and pathological as positive. Because positive samples are usually the minority class, while the missed prediction of a positive sample is more disastrous than that of a negative sample, the unweighted average recall (UAR) can then be used to obtain an average of TPR and TNR.

Deepfake detection is similar to diagnostics tasks where a binary decision is obtained, therefore AUC-ROC and F1 can be used for evaluating classification accuracy. Furthermore, the equal error rate (EER) is often used as a standard evaluation metric in the past ASVspoof challenges (Todisco et al., 2019; Liu et al., 2023). The EER value is defined as the arithmetic mean between TPR and FPR. The EER threshold represents the point where the false positive rate (FPR) equals the false negative rate (FNR).

For voice anonymization, the goal is to fool a pre-trained speaker recognition model so that the speaker identity can be obfuscated. In an ideal case, the speaker recognition model is expected to achieve 0% accuracy when asked to recognize the user. Therefore, we use misclassification rate of a speaker recognition model to reflect the anonymization efficacy, where higher rate corresponds to better anonymization performance.

3 KNOWLEDGE-BASED FEATURE ENGINEERING AND ML FOR HEALTH DIAGNOSTICS

3.1 Preamble

This Chapter is compiled from manuscripts published in the IEEE Transactions in Audio, Speech, and Language Processing (Zhu et al., 2023f), and the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022 & 2023 (Zhu et al., ress, 2023e).

3.2 Introduction

Knowledge-based features and ML systems are often favored in scenarios with limited data and a high demand for interpretability. In the case of health diagnostics, pathological voice datasets often come with a relatively smaller data volume compared to other speech datasets, therefore limiting the effectiveness of DL models. Furthermore, disease diagnosis often requires an explainable decision-making process, where high accuracy alone is not sufficient enough for the application in real-world. Knowledge-based acoustic features, rooted in acoustic and physiological insights, often provide reliable and interpretable markers, such as vocal tremor or irregular pitch, which correlate with respiratory and neuromuscular pathologies (Schuller et al., 2021; Little et al., 2008). In low-data regimes, well engineered features combined with properly designed ML models can outperform DL models (Zhu et al., ress), providing accurate classification and enhanced interpretability. In this section, we introduce several novel knowledge-based feature sets for speech and cough, together with physiology-inspired ML systems to tackle COVID-19 detection.

The main contributions of this chapter are summarized as follows :

1. We propose two hand-crafted feature sets and ML systems for speech and cough signals, respectively, to capture respiration and articulation related abnormalities;

2. We empirically show that the proposed features outperform existing hand-crafted and deep features on the COVID-19 detection tasks;

3. We evaluate the proposed features in a cross-dataset setting, where the proposed features demonstrate good generalizability to unseen datasets;

4. We show that the proposed features can be used to pinpoint factors affecting both the articulation and the respiratory systems, thus providing more detailed clinical explanations to the obtained findings.

3.3 Related work

During the past few years, the COVID-19 pandemic has affected hundreds of millions of lives (Heckman et al., 2020). Due to the scarcity of medical testing toolkits and the heavy reliance on medical professionals, researchers have explored using speech and cough signals as alternative methods to remotely diagnose the disease. It has been established that the COVID-19 virus can affect many of the systems involved with speech production (Quatieri et al., 2020). For example, the virus affects the respiratory system, targeting the lungs and other airway passages, thus resulting in shortness of breath, coughs, and atypical breathing modulations, to name a few symptoms (Vetter et al., 2020). These, in turn, can irritate the vocal cords, resulting in inflammation, sore throats, hoarseness, and breathiness. Moreover, temporary neuromuscular deficits have been reported (Helms et al., 2020), thus also affecting speech articulators, causing atypical changes in the acoustic properties of the produced speech signal (Quatieri et al., 2020). These factors suggest that automated analysis of speech and cough signals for COVID-19 detection can be possible.

Existing acoustics-based (i.e., speech and cough) COVID-19 detection systems have focused mainly on two aspects : i) the design of acoustic features and ii) data-driven ML algorithms that find non-linear (discriminatory) patterns in the acoustic signals. Regarding knowledge-based speech and cough features, one of the most widely used feature sets explored for COVID-19 monitoring has been the so-called ComParE acoustic feature set 2016 (Weninger et al., 2013). This ComParE set contains over 6,000 features that cover different speech signal representations such as mel-frequency cepstral coefficients (MFCC), pitch contour, voicing related information, as well as their low-level descriptors (LLDs). In the recent INTERPSEECH 2021 Computational Paralinguistics ChallengE COVID-19 speech and cough sub-tasks, the ComParE set 2016 was used as one of the benchmark features (Schuller et al., 2021).

With these features, Jing et al showed that it could be possible to predict the severity of the disease, patient anxiety, sleep quality, and fatigue (Han et al., 2020). Alternatively, as mel-spectrograms can be seen as an image representation of the acoustic signals, several attempts have been made to explore their use with convolutional neural networks (CNNs) (Deshpande et al., 2020a; Pinkas et al., 2020; Pahar et al., 2021; Nessiem et al., 2021). In fact, the majority of the existing speech-based systems have focused on the use of deep neural networks (DNNs) for the task at hand (e.g., (Pinkas et al., 2020; Pahar et al., 2021; Nessiem et al., 2021)). Notwithstanding, the COVID Challenge showed that support vector machines (SVM) could still outperform DNNs (Schuller et al., 2021). Furthermore, while larger transformer-based speech models have obtained promising results in some diagnostic tasks, it remains challenging to interpret the model decisions and pinpoint the symptom-related disease biomarkers.

Ultimately, we are interested in systems that are i) accessible to all, i.e., does not require very complex models that cannot be run on legacy devices; ii) interpretable, hence the input features, as well as the predictions made by the system, can be explained and understood; and iii) generalizable

to unseen conditions, such as speakers, languages, or microphones. To achieve these goals, we have proposed novel feature sets motivated by the physiology of speech and cough production, and developed ML-based COVID-19 diagnostic systems.

3.4 Modulation spectrum and linear prediction based COVID-19 speech detection system

In this section, we describe the proposed system and motivate the use of the modulation spectrum and linear prediction (LP) features. It should be noted that some datasets include samples with varying sampling rates, with the majority sampled at $16\,\mathrm{kHz}$. Hence, we resampled all recordings to $16\,\mathrm{kHz}$ prior to feature extraction.

3.4.1 Modulation spectral features

The commonly-used spectrogram provides information on how frequency components change as a function of time. However, common noise sources overlap in both time and frequency, thus making the representation sub-optimal for classification tasks relying on noisy data. The modulation spectrum, in turn, characterizes the rate-of-change of these frequency components, thus better separating signal and noise components, making it a better candidate for in-the-wild applications (Falk et al., 2010b; Avila et al., 2018). The modulation spectrum is also known to capture higher-order periodicities of the signal not otherwise obvious in time and time-frequency domains. This property has shown to be useful for disease characterization, such as autism spectrum disorder detection from toddler cries and non-verbal vocalizations (Bedoya et al., 2020), as well as automated intelligibility monitoring of dysarthric speech (Falk et al., 2012). We explore it here as a useful, robust representation for COVID-19 detection.

The signal processing steps involved in the computation of the modulation spectrogram are depicted in Fig. 3.1. First, the speech signal $x(t)$ is transformed to the time-frequency domain (spectrogram) via the short-time Fourier transform (STFT) (here, implemented using a 256-point FFT). A second transform is then applied across the time axis, for each frequency bin magnitude $|X(t, f)|$. This results in a frequency-frequency representation of the signal termed 'modulation spectrogram' ($X(f_m, f)$), which characterizes the rate-of-change of different spectral components. Here, f is used to characterize the conventional frequency (Hz) and f_m the modulation frequency. As the majority of the modulation spectral content of speech is known to lie below 20 Hz f_m (Falk et al., 2010a), parameters used in the computation of the modulation spectrogram have been chosen here such that $f_m = 0 - 20$ Hz and $f = 0 - 8$ kHz.

While one may choose to use the modulation spectrogram directly as input to ML algorithms (i.e., to be treated as an image, as most applications relying on spectrograms do (Laguarta et al.,

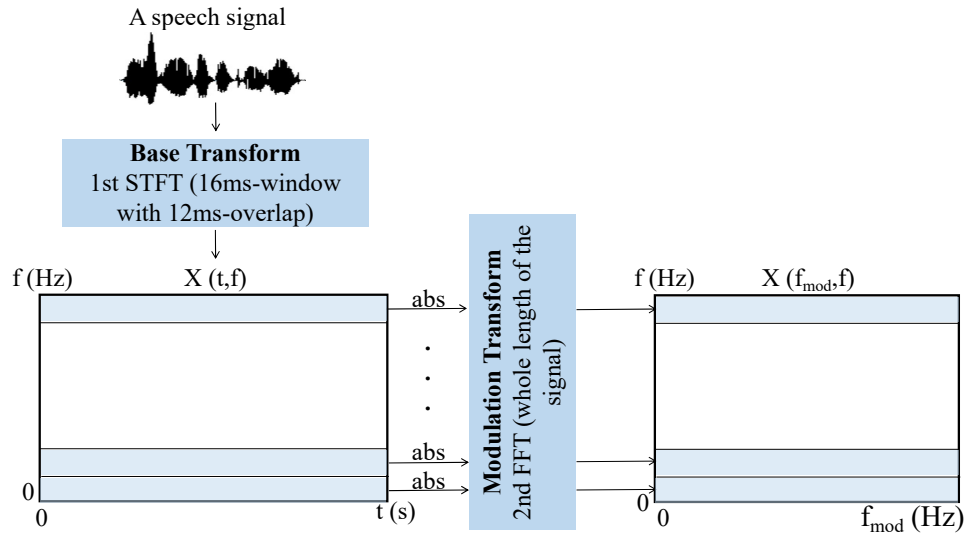


FIGURE 3.1 : Signal processing steps involved in the computation of the modulation spectrogram.

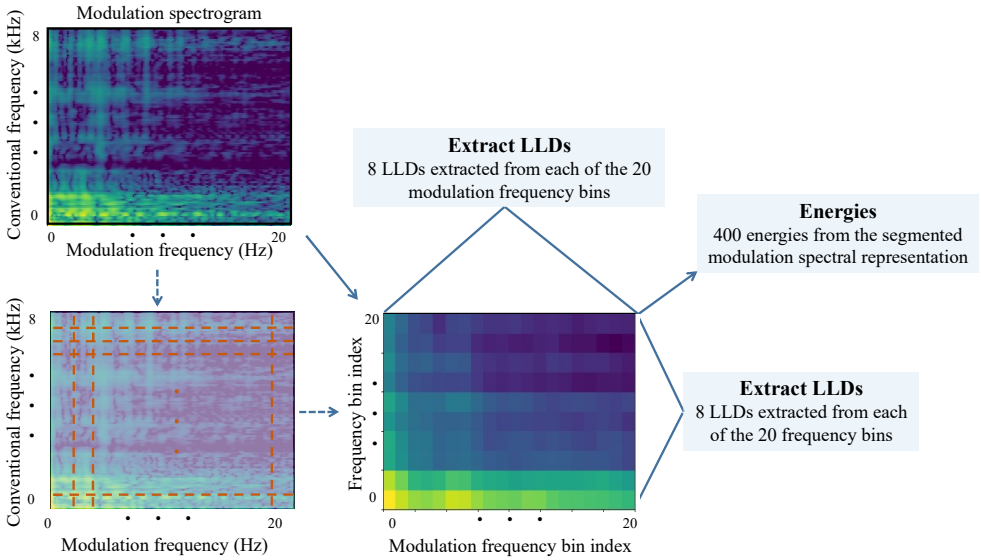


FIGURE 3.2 : Extraction of modulation spectrogram features.

2020)), it increases the model complexity and makes the system less interpretable. To improve the interpretability and simplicity of the proposed method, here we have decided to extract spectral power and descriptor features from the modulation spectrogram and use those instead to reduce the input dimensionality. Figure 3.2 depicts this feature extraction process.

We first quantize the modulation spectrogram into 400 bins by grouping the 0-20 Hz f_m axis into twenty 1-Hz bins. Similarly, the 0-8 kHz frequency axis is grouped into twenty 400-Hz bins. Next, the spectral power of each bin is then normalized by the total power of the modulation spectrogram, then used as modulation spectral energy features. Next, we compute eight spectral shape

descriptors for each of the 20 conventional frequency bins (i.e., descriptors computed across modulation frequency axis), as well as for the 20 modulation frequency bins (i.e., across conventional frequency). This results in an additional 320 features ($20 \times 8 \times 2$). The eight descriptors include : spectral centroid, entropy, spread, skewness, kurtosis, flatness, crest, and flux. A detailed description of these descriptors can be found in (Peeters, 2004; Wu et al., 2011). Overall, a total of 720 modulation spectral features (MSF) are computed and tested.

3.4.2 Vocal tract and excitation signal decomposition

In an effort to build COVID-19 diagnostic tools that have improved interpretability, we part from the hypothesis that COVID-19 can affect both the vocal tract properties (e.g., via increased muscle fatigue (Solomon, 2006; Helms et al., 2020)) and the excitation signal (e.g., impaired phonation (Quatieri et al., 2020)). While modulation spectral features may provide some insights from the former, we propose to also use linear prediction (LP) analysis to decompose the speech signal into vocal tract parameters (i.e., LP coefficients) and the excitation source (i.e., LP residual) (Makhoul, 1975). Linear prediction analysis assumes that speech is generated by the excitation of a linear time-varying filter (vocal tract) by impulses for voiced speech segments or random noise for unvoiced speech segments (Markel et al., 2013). The vocal tract can be modeled as an all-pole filter, of which the transfer function can be given by :

$$H(z) = \frac{1}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{3.1}$$

where G is the gain factor of the LP filter and is set to 1, p is the order of the LP filter and implies that the past p samples are used in the prediction of the current sample. More specifically, the speech signal $s(n)$ can be approximated by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n), \tag{3.2}$$

where the coefficients a_k are the linear predictive coefficients (LPCs) and $p = 18$ is used herein, as speech sampled at 16 KHz is used in our experiments.

The excitation signal, in turn, is represented by the prediction error signal $e(n)$, which is the difference between the estimated speech $\hat{s}(n)$ and the original speech $s(n)$, i.e.,

$$e(n) = s(n) - \hat{s}(n) = Gu(n), \tag{3.3}$$

where a_k in (4.8) can be estimated by minimizing the energy of $e(n)$ using Burg's method (Makhoul, 1977). Henceforth, we utilize features extracted from the LP residual as features characterizing the excitation and the $a_k, k = 1, \ldots 18$ as features characterizing the vocal tract.

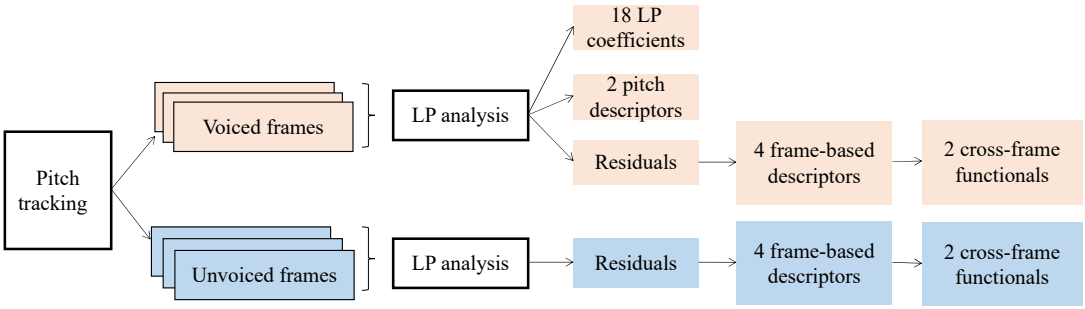


FIGURE 3.3 : Signal processing steps involved in the computation of the vocal tract and excitation signals from LP analysis.

The excitation signal, however, differs based on the periodicity of the speech segment. Voiced speech segments are produced when quasi-periodic pulses of air generated by the vibration of vocal folds resonate through the vocal tract, where the fundamental frequency of vibration of the vocal folds is usually interpreted as pitch (Zhang, 2016). As such, the residual signal has strong impulse-like peaks corresponding to the glottal pulses produced during voiced speech. Unvoiced segments, on the other hand, are produced without vibration of the vocal cords, and have residuals that are commonly modeled as noise (Zhang, 2016). In our analysis, we use the pYAAPT open source pitch tracker (Zahorian et al., 2008) in order to separate voiced and unvoiced segments prior to LP analysis. The following pitch tracking hyper-parameters were used in our analyses : 30 Hz minimum pitch searched, 400 Hz maximum pitch searched, 15 ms frame length, and 5 ms frame hops.

From the LP analysis, different features are then extracted from the vocal tract and excitation signals from the voiced and unvoiced segments. Figure 3.3 depicts the steps taken for the calculation of the proposed LP features. First, LP analysis is performed on *voiced* speech frames and the following features are extracted : (i) 18 LPCs, to indicate the shape and resonance characteristics of the vocal tract, (ii) mean and standard deviation of the pitch values computed over all voiced speech frames in a recording, and the (iii) mean, standard deviation, kurtosis, and skewness of the voiced LP residual, computed per frame. These latter four per-frame features are finally aggregated using the mean and standard deviation statistics, thus resulting in eight voiced LP residual temporal dynamics features. Lastly, the same eight residual temporal dynamics features are extracted, but for the *unvoiced* frames. A total of 36 LP features are computed ($18 + 2 + 8 + 8$).

3.4.3 Feature selection

In order to reduce the number of features to be input to ML algorithms and better understand the importance of the different extracted features, the Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithm was used. The MRMR algorithm finds the optimal feature subset, such that the top-ranked features are mutually dissimilar, while being maximally related to the

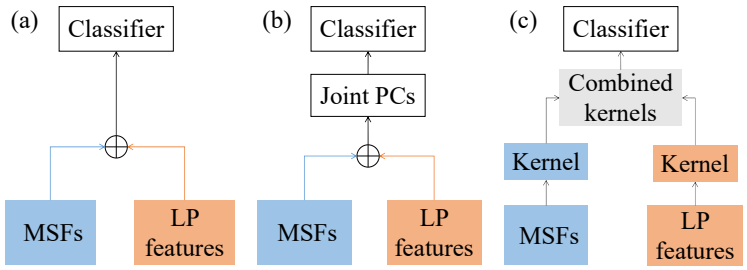


FIGURE 3.4 : Feature fusion schemes tested : (a) early-stage fusion, (b) intermediate-stage fusion via PCA, and (c) multiple kernel learning fusion.

outcome variable (i.e., in our case, the COVID-19 diagnostic) (Radovic et al., 2017). It is a filter-based feature selection method in the sense that feature ranking is performed independent of the downstream ML algorithm, hence allowing for experiments with different models to be performed using the selected top features. Moreover, it allows for assessment of the potential of new engineered features and not necessarily on the ML algorithms. As we are interested in gauging the benefits and interpretability of the proposed features, a filter method is preferred. In this study, the number of features to be selected was sequentially experimented from 5 to 100, where the optimal number was determined by the performance achieved during cross-validation trials.

3.4.4 Classifiers and fusion methods

As mentioned above, results from the 2021 ComParE COVID-19 Challenge have shown that conventional models, such as support vector machines (SVM), can achieve similar results relative to DNNs (Schuller et al., 2021). Meanwhile, as is pointed out in (Tu, 1996), overly complex models could easily overfit to small datasets, hence leading to lower generalizability across unseen data conditions. Therefore, we explore the use of conventional classifiers in order to achieve better interpretability as well as lower computational complexity. In particular, a linear SVM classifier is used, as well as a decision tree (DT) classifier, as these have shown increased interpretability in the past (Slack et al., 2019). Random forests (RF) are also explored, as they have shown high accuracy across different clinical applications and can be made interpretable (Hara et al., 2018). In all cases, the scikit-learn toolkit was utilized (Pedregosa et al., 2011). For the SVM, a linear kernel was used. For the decision tree classifiers, a maximum tree depth of 10 was selected since we observed larger values to result in over-fitting; other parameters were set to default values.

Next, we take the advantage of the complementarity of the proposed features and explore different feature fusion schemes. Since our focus is to investigate the importance of different feature sets and their complementarity, we explored rather conventional feature fusion methods and leave the optimal fusion strategies for future investigation. As shown in Fig. 3.4, three fusion methods are explored, namely : (a) early-stage fusion, where MSF and LP features are aggregated into one larger feature vector, (b) features from the two methods are aggregated via principal component

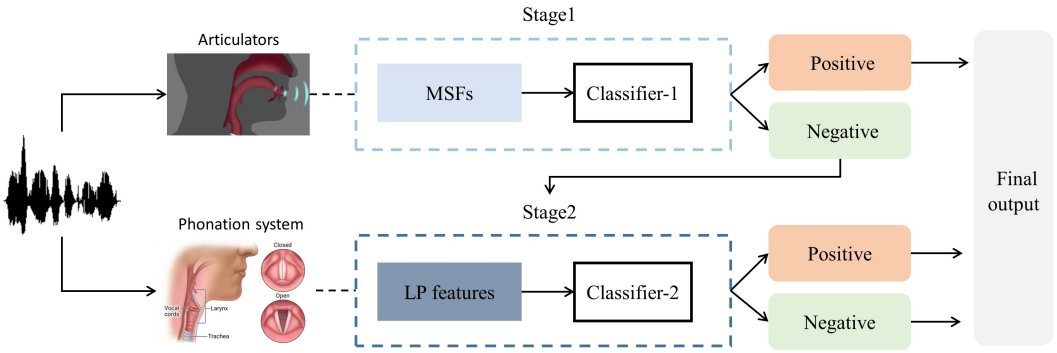


FIGURE 3.5 : Two-stage, decision-level fusion scheme proposed for improved COVID-19 detection.

analysis (PCA), and (c) in the case of the SVM classifier, a multiple kernel learning method is explored where different kernels are tested for each feature set, and the optimal kernel combination is found. Here, the open source MKLpy toolbox (Lauriola et al., 2020) was used for multi-kernel learning, where the regularization hyper-parameter λ was tuned empirically by experimenting values from 0 to 1.

In addition to feature fusion, decision level fusion combines decisions made by different systems. This architecture is favored when errors made by different systems are mutually exclusive (Ramachandram et al., 2017). Decision level fusion, however, requires additional data to train the meta-classifier. Motivated by the previous findings where a reduced coordination of speech subsystems was found in COVID-19 patients (Quatieri et al., 2020), we here propose a two-stage system to capture complications in the articulators or the phonation system (or both) (depicted by Fig. 3.5). While the first stage targets deficits in the articulators, samples deemed negative pass through a second stage to avoid potential abnormalities in the phonation system being overlooked. As such, the two-stage approach should lead to higher sensitivity levels. Our experiments showed that combining the top-40 MSFs and top-10 LP features, as ranked by MRMR, led to improved results. A linear SVM classifier was used for the MSFs, while a DT-based classifier was used for the top LP features. With this two-stage approach, recordings are first processed by the MSF-based system. Positive predictions are kept as is, whereas those deemed to be negative, are further processed by the LP based system to be fine-tuned, as shown in the figure.

3.5 Evaluation of speech features

3.5.1 Datasets and evaluation metrics

For the evaluation, we relied on the INTERSPEECH 2021 ComParE COVID- 19 Dataset (CSS), DiCOVA2 dataset, and the subsets of Cambridge COVID-19 Sounds Dataset. Regarding evaluation metrics, we used the metrics adopted by the COVID-19 detection challenges, namely AUC-

ROC, UAR, sensitivity, and specificity. Details of these three datasets and evaluation metrics can be found in Sections 2.4 and 2.5, respectively.

3.5.2 Benchmark systems

As two of the used datasets were part of international COVID-19 challenges, we utilize their benchmark systems as baselines in our experiments. As stated above, since data partitions are different from the original challenge ones due to limited data accessibility, we reproduced the baseline systems following the system architecture described in their corresponding papers (Schuller et al., 2021; Sharma et al., 2022; Xia et al., 2021). A brief description of the baseline systems is presented below.

ComParE + SVM : 6,373 acoustic features are extracted using the openSMILE toolbox (Eyben et al., 2010) and fed into a linear support vector machine (SVM) for final predictions. This system has been used as a benchmark on all three datasets. Note that the work in (Schuller et al., 2021) showed that using the bag-of-words (BoW) methodology on openSMILE features could lead to improved results. BoW on top of modulation features has also shown to be useful for speech emotion recognition (Kshirsagar et al., 2022). As we are interested in gauging the benefits of the proposed features, we compare them with the original openSMILE set and leave BoW extended features as future work.

Spectrogram + VGGish : Raw audio waveforms are first converted to 2-dimensional mel-spectrograms and then fed into a pre-trained VGGish network for classification. The VGGish backbone and the fully-connected layers are fine-tuned jointly on each dataset. This is a benchmark system used for the Cambridge subset.

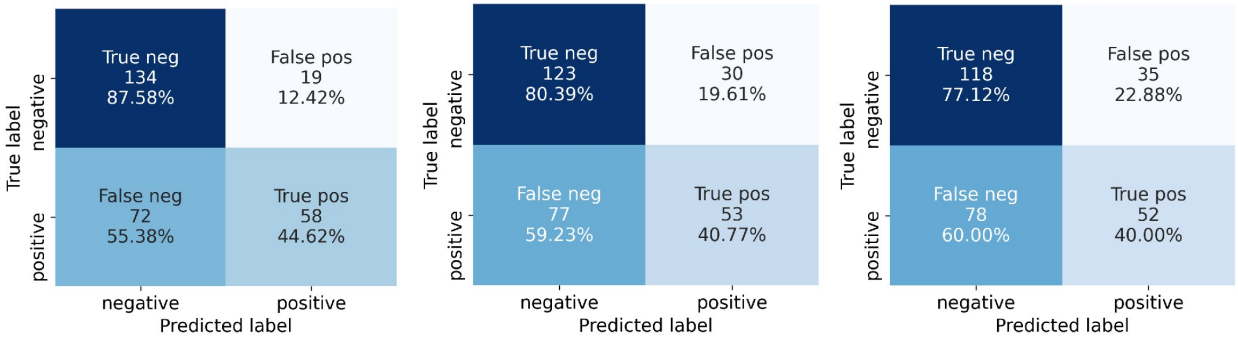
Spectrogram features + BiLSTM : 192 log mel-spectrogram features are extracted per speech frame, which are then fed into a bi-directional long short term memory (BiLSTM) recurrent neural network for classification. This is one of the benchmark systems for the DiCOVA2 dataset.

3.5.3 Two-stage fusion

To evaluate the efficacy of proposed features, Table 3.1 reports the UAR, sensitivity, and specificity levels achieved on the CSS dataset along with the top-performing classifier and the number of top-selected features used. As can be seen, the proposed MSF feature set achieves the highest UAR (0.661) and sensitivity (0.515). This is followed by the combined LP feature set, which achieves a UAR of 0.606 and sensitivity of 0.408. Both of the proposed feature sets outperform the benchmark system (ComParE+random forest) in terms of these two performance metrics, whilst requiring a significantly lower number of features (40). When comparing results achieved with the similar number of features (i.e., 40-50), the benchmark features, on the other hand, achieve the highest specificity value of 0.961. The marked increase in specificity achieved relative to sensiti-

TABLEAU 3.1 : Performance of proposed and benchmark features used individually with the CSS dataset. Bold values indicate the best system based on a given figure-of-merit. Statistically significant improvement relative to the highest benchmark UAR is highlighted with an asterisk.

Feature	#No.	Classifier	UAR	Sensitivity	Specificity
MSF	40	SVM	**0.661***	0.447	0.876
		DT	0.578	**0.515**	0.641
		RF	0.602	0.269	0.935
LP features	10	SVM	0.551	0.220	0.882
		DT	0.571	0.369	0.771
		RF	0.606	0.408	0.804
Benchmark	6373	SVM	0.537	0.139	0.935
		DT	0.566	0.426	0.706
		RF	0.586	0. 400	0.771
Benchmark	50	SVM	0.540	0.153	0.927
		DT	0.507	0.223	0.791
		RF	0.534	0.108	**0.961**



(a) **(b)** **(c)**

FIGURE 3.6 : Confusion matrices of systems tested on CSS (with best results chosen) for (a) MSFs based system, (b) LP features based system, and (c) Benchmark system.

vity suggests that all models are better at classifying non-COVID samples rather than COVID-19 samples. Confusion matrices achieved for the three systems are shown in Fig. 3.6. As can be seen, only 24.7% of positive predictions made by MSFs were false alarmed. The LP features, however, render a false alarming rate that is almost twice as high (40.2%). On the other hand, the correct rejection rate of MSFs and LP features are relatively close (34.9% and 39.7%). Based on the high precision achieved by MSFs, the development of a hierarchical system is motivated, where MSFs pre-screen COVID-19 samples at the first stage, with the negative predictions being then sent as inputs to a second stage classification using LP features (see Fig. 3.5).

Next, we explore the effects of fusion on system performance. Table 3.2 presents the results obtained with the four fusion methods described in Section 3.4.4. As can be seen, the proposed two-stage system achieves the highest UAR (0.682), hence a 3.2% increase relative to the best single feature reported in Table 3.1, and a 16.4% increase relative to the benchmark performance. The multi-kernel learning fusion method, in turn, results in the highest sensitivity (0.746), thus a 44.9% improvement relative to the sensitivity of the best single feature. Early-stage fusion achieves

TABLEAU 3.2 : Performance comparison of different fusion methods evaluated with CSS. Bold values indicate the best system based on a given figure-of-merit. Statistically significant improvement relative to the highest UAR obtained by single feature modality is highlighted with an asterisk.

Fusion type	Configuration	UAR	Sensitivity	Specificity
Early-stage	SVM	0.664	0.380	**0.948**
	DT	0.607	0.347	0.867
	RF	0.630	0.418	0.842
Intermediate-stage	PCA+SVM	0.641	0.493	0.789
	PCA+DT	0.593	0.311	0.895
	PCA+RF	0.610	0.346	0.874
MKL	Polynomial	0.652	**0.746**	0.558
Two-stage	SVM+DT	0.674	0.508	0.840
	SVM+RF	**0.682***	0.531	0.833

the highest specificity. Finally, the two-stage fusion method based on SVM and a random forest classifier achieves the best trade-off between sensitivity and specificity, hence could be a better candidate for use in the clinic. Overall, it is clear that fusion of both MSF and LP features results in the best accuracy, likely due to the fact that the different features are capturing different aspects of the disease in a complementary manner. Overall, the two-stage system is shown to statistically outperform the benchmark system in terms of UAR, whilst requiring a significantly lower number of features (i.e., 50 compared to over 6,000), thus is used throughout the remainder of this paper.

3.5.4 System generalizability

While the two-stage system performs significantly better than the benchmark system on the CSS dataset, it is crucial to examine whether such system architecture and the proposed features can also be useful on other datasets. Table 3.3 compares the performance of the proposed 2-stage system, as well as systems trained on single feature modalities and two benchmark systems, on the DiCOVA2 and Cambridge datasets. With an AUC-ROC of 0.612 achieved with the Cambridge set, the two-stage system outperforms both ComParE and VGGish-based benchmark systems. However, the improvement over the VGGish system is not statistically significant. Meanwhile, it can be noticed that using a small number of LP features alone achieves comparable performance relative to the more complex VGGish benchmark. Notwithstanding, the spectrogram with BiLSTM classifier remains the top-performer system on the DiCOVA2 dataset with an AUC-ROC of 0.770.

Moreover, it can be observed that the performance of all tested systems varies greatly across datasets. For example, the ComParE benchmark achieves an AUC-ROC of 0.751 on DiCOVA2 but only 0.521 on Cambridge and an UAR of 0.537 on the CSS dataset. A similar, but relatively smaller, variance can be observed with our proposed two-stage system, which achieves an AUC-ROC of 0.711 on DiCOVA2 and 0.612 on the Cambridge set. These findings suggest that the prediction tasks with CSS and Cambridge datasets may be more challenging relative to DiCOVA2, which may

TABLEAU 3.3 : Performance comparison on Cambridge and DiCOVA2 datasets. Scores reported are averaged over 10 different cross-validation runs. Bold values indicate the best system for a given metric. The statistically significant difference between the highest AUC-ROC obtained by benchmark systems and the two-stage system is highlighted with an asterisk.

Dataset	System	Evaluation metrics		
		AUC-ROC	Sensitivity	Specificity
Cambridge	MSF+SVM	0.543	**0.760**	0.366
	LP+RF	0.595	0.575	0.**625**
	Two-stage	**0.612**	0.722	0.413
	ComParE+SVM	0.521	0.431	0.619
	Spec+VGGish	0.608	0.582	0.615
DiCOVA2	MSF+SVM	0.693	0.729	0.567
	LP+RF	0.633	0.401	**0.806**
	Two-stage	0.711	0.708	0.689
	ComParE+SVM	0.751	**0.731**	0.671
	Spec+LSTM	**0.770***	0.653	0.791

TABLEAU 3.4 : Performance comparison in cross-dataset testing conditions. Scores reported are averaged over 10 different cross-validation runs. Bold values indicate the best system for a given condition. The system with a significantly improved average AUC-ROC relative to the other systems is highlighted with an asterisk. CA :Cambridge; DI :DiCOVA2.

System	CSS+CA→DI	CSS+DI→CA	CA+DI→CSS	Ave
Two-stage	**0.606**	**0.554**	**0.578**	**0.579***
ComParE+SVM	0.570	0.515	0.506	0.530
Spec+LSTM	0.489	0.484	0.477	0.483
Spec+VGGish	0.491	0.502	0.485	0.493

be due to the high percentage of asymptomatic COVID-19 samples and symptomatic non-COVID samples.

Finally, we test the generalizability of the investigated systems. Table 3.4 summarizes the AUC-ROC achieved with the different systems within cross-dataset testing conditions. In this experiment, training is performed on two combined datasets and tested on the third unseen set without any model fine-tuning. As can be seen, a marked drop is observed for all systems, which is in line with findings reported with other modalities (e.g., cough and image) (Roberts et al., 2021; Akman et al., 2021). Interestingly, for the deep learning based systems, all performances degrade to below chance levels, suggesting that the systems may be overfitting to specific database nuances and not necessarily COVID-19 information. In comparison, our two-stage system maintains an average AUC-ROC of 0.579 across the three cross-dataset conditions, significantly outperforming all other benchmarks. These findings suggest that the proposed two-stage system may indeed be capturing relevant phonation and articulators information important for COVID-19 detection.

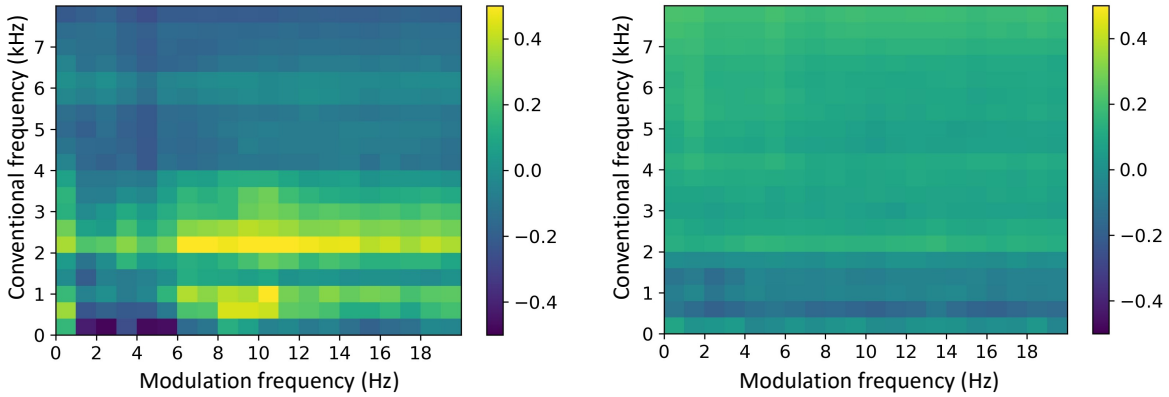


FIGURE 3.7 : Modulation spectrogram of COVID-19 speech (left) and non-COVID speech (right). Both are averaged across samples from the CSS training set.

3.5.5 Interpretation of speech features

As we are interested in developing interpretable models, we explore the discriminatory potential of the two proposed feature sets using the CSS dataset. Modulation spectrograms, normalized and averaged over the training set files, are depicted in Fig. 3.7 for the COVID-19 and non-COVID speech recordings. As can be seen, speech from COVID-19 patients shows some unique patterns, especially around $f_m = 6 - 10$ Hz and $f = 0.4 - 1.2$ kHz, and $f_m = 6 - 14$ Hz and $f = 2 - 2.8$ kHz.

To further explore the discriminatory potential of the modulation spectrum, the Fisher ratio (F-ratio) is used and computed between the modulation spectrograms of the two groups. To calculate the F-ratio, two estimates of the variance are made, including the between and within group variances, where the F-ratio is given by :

$$F - ratio = \frac{MS_{between}}{MS_{within}}, \tag{3.4}$$

where $MS_{between}$ represents the between group variance and MS_{within} represents the within group variance for each of the 400 entries of the modulation spectrogram (MS). Higher F-ratio values suggest increased group discrimination.

The top-left plot in Fig. 3.8 shows the F-ratio map indicating which parts of the modulation spectrogram provide the most discriminatory information. F-ratio analysis suggests increased discrimination around $f_m = 8 - 11$ Hz and $f = 0.4 - 1.2$ kHz, as well as between $f_m = 1 - 6$ Hz and $f = 0 - 400$ Hz, and to a lesser extent between $f = 6 - 8$ kHz. Similar F-ratio computations are also made with the eight spectral shape descriptors computed across the 20 conventional frequency bins (top-right plot in Fig. 3.8) and the descriptors computed across the 20 modulation frequencies (bottom-left plot in Fig. 3.8). Lastly, four regions are found that overlap between the three MSF features classes, hence suggesting that these are the modulation frequency ranges providing the most discriminatory potential for COVID-19 detection from speech (bottom-right plot

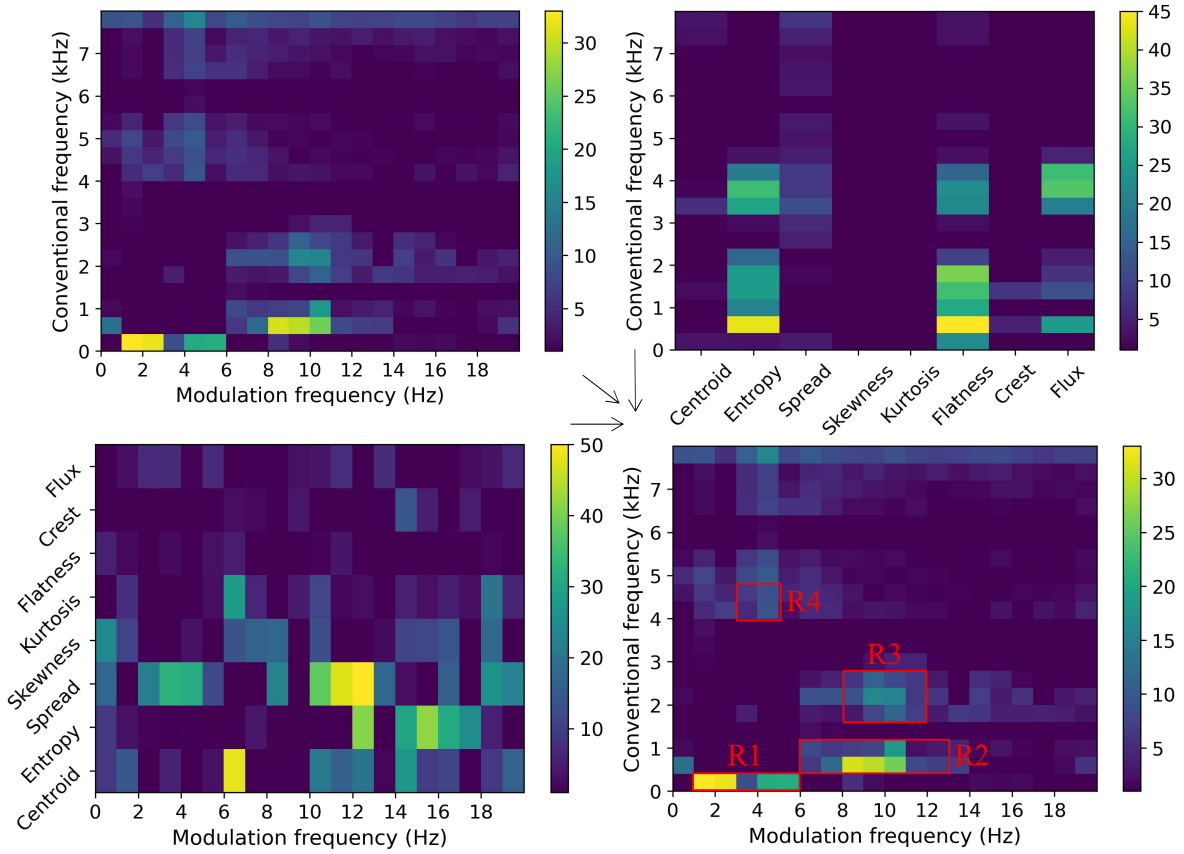


FIGURE 3.8 : F-ratio plots of modulation spectrogram energies (top-left), spectral shape descriptors computed across conventional frequency (top-right) and modulation frequency (bottom-left), and their four overlapping regions (bottom-right).

in Fig. 3.8). These ranges include : (1) $f_m = 1 - 6$ Hz and $f = 0 - 0.4$ kHz, (2) $f_m = 6 - 13$ Hz, $f = 0.4 - 1.2$ kHz, (3) $f_m = 8 - 12$ Hz, $f = 2 - 2.4$ kHz, and (4) $f_m = 3 - 5$ Hz, $f = 4 - 4.8$ kHz.

For the proposed LP features, a Welch t-test is used to test for significance between the two groups. The Welch's t-test is a parametric test that compares the means between two independent groups without assuming equal population variances (Welch, 1947), thus, is favored when the sample sizes of two groups are unequal. The t-statistic and p-value are used to gauge the discriminatory potential of the LP features. Table 3.5 reports the statistical test results for the 20 features obtained from the voiced segments. A significance level of 99% ($p \leq 0.01$) is used and seen in 7 of the 18 LPCs. In particular, it is observed that the 6th, 12th, and 16th LPC of COVID-19 speech are significantly higher than those from non-COVID speech, whereas the 1st, 11th, 13th, and 17th LPC are higher for non-COVID speech. No significant differences are observed for the two pitch descriptors.

Table 3.6 shows the test results for the residual features of the voiced and unvoiced segments. All residual features, except the mean, show significant differences between the two groups. In particular, COVID-19 speech demonstrates significantly higher values for average kurtosis and

TABLEAU 3.5 : Welch's t-test results of LP features for voiced segments. An $*$ indicates features that achieved $p \leq$0.01.

Feature	t	p
LPC-1*	-3.1714	0.0016
LPC-2	1.8291	0.0682
LPC-3	-0.9879	0.3239
LPC-4	1.3276	0.1852
LPC-5	-2.3710	0.0185
LPC-6*	2.8390	0.0050
LPC-7	-2.2956	0.0226
LPC-8	2.1144	0.0356
LPC-9	-2.1762	0.0307
LPC-10	2.4304	0.0160
LPC-11*	-2.9794	0.0032
LPC-12*	3.3049	0.0011
LPC-13*	-2.9700	0.0033
LPC-14	2.0232	0.0441
LPC-15	-1.7128	0.0878
LPC-16*	2.6769	0.0078
LPC-17*	-3.0076	0.0028
LPC-18	1.5792	0.1152
Pitch (Mean)	-1.1187	0.2640
Pitch (Std)	-1.1843	0.2370

skewness, as well as the variation of kurtosis and skewness of voiced residual segments. Meanwhile, the non-COVID group shows higher variations of voiced residual signal values. Similar to the voiced residual features, COVID-19 speech shows significantly higher value of average kurtosis and skewness, while non-COVID speech demonstrates higher cross-frame variations of unvoiced residual signal values.

One common symptom of COVID-19 is hoarseness due to inflammation of the vocal folds. In severe cases, the voice can become almost a whisper. Previous work on the modulation spectral analysis of whispered speech has shown that whispers can be manifested below $f = 1$ kHz and between 3 kHz$\leq f \leq 4.5$ kHz and $f_m \geq 10$ Hz (Falk et al., 2012). These findings suggest that some of the modulation spectral changes seen could be due to increased hoarseness of COVID-19 speech. Moreover, temporary neuromuscular impairments have been reported with COVID-19 (Helms et al., 2020), hence potentially affecting muscular control, and ultimately speech production. The seven LPC coefficients that show significant differences could be linked to such neuromuscular deficiencies. Moreover, as mentioned in (Quatieri et al., 2020), speech production can be modeled as a combination of airflow from the lungs, passing into the larynx, where coordinated coupling with phonation is achieved, followed by vocal tract shaping during articulation. As such, the proposed MSFs can also be seen as a correlate of the breathing amplitude modulations, especially those in the lower f_m range. Plots of COVID-19 modulation spectrograms in Fig. 3.7 demonstrate pronounced effects in this range, potentially due to breathing issues caused by impaired lung function

TABLEAU 3.6 : Welch's t-test results of LP residual features for voiced and unvoiced segments. An $*$ indicates features that achieved $p \leq$0.01, while $**$ indicates $p \leq$0.001.

Segments	Feature	t	p
Voiced	Mean (mean)	-0.1935	0.8466
	Std (mean)*	-2.8676	0.0043
	Kurtosis (mean)**	6.3310	0.0000
	Skewness (mean)**	6.1813	0.0000
	Mean (std)**	-4.2438	0.0000
	Std (std)*	-3.0236	0.0026
	Kurtosis (std)**	5.0586	0.0000
	Skewness (std)**	5.5291	0.0000
Unvoiced	Mean (mean)	-0.1169	0.9070
	Std (mean)	-2.5331	0.0116
	Kurtosis (mean)**	4.3677	0.0000
	Skewness (mean)*	2.6003	0.0097
	Mean (std)**	-4.0154	0.0001
	Std (std)**	-4.0621	0.0001
	Kurtosis (std)	1.6488	0.1000
	Skewness (std)	2.0208	0.0440

(Chowdhury et al., 2020). Lastly, the significant changes seen in the LP residuals for both voiced and unvoiced segments also point towards phonation impairments, potentially also caused by inflammation of the larynx. For example, the increase in the LP residual kurtosis values in COVID-19 speech could be indicative of higher levels of vocal harshness, whereas the lower variability of LP residuals' values could suggest more breathy signals.

Lastly, we examined the complementarity between the baseline openSMILE and our proposed MSF features. The two feature sets were concatenated then MRMR was performed to select the top-10 features. This process had been repeated for CSS, DiCOVA2, and Cambridge set. It is noted that two features consistently appeared in the top-10 list for each dataset, including 'spectral kurtosis across modulation frequencies at $f = 2 - 2.4$kHz' (MSF) and 'pcm_fftmag_spectral entropy' (openSMILE). As previously discussed, the changes to MSFs could be due to increasing vocal hoarseness of COVID patients, possibly caused by inflammation of the respiratory system. Additionally, existing literature has shown that an increase in spectral entropy is seen in muffled speech compared with clean speech Memon (2020), as well as in reduced speech quality due to poor manner of articulation Llanos et al. (2017). As the same pattern is observed here for COVID speech, it could be associated with the difficulty in articulation and oral-facial muscle fatigue in COVID-19 Helms et al. (2020), as well as nasal congestion causing speech to sound e.g., more muffled.

3.6 Phase based cough features

3.6.1 Cough segmentation

Different from existing cough features which follow a speech-like feature extraction pipeline, we propose to segment cough recordings into three different phases (i.e., inhalation, compression, and expulsion) and to extract features from each phase separately to explore their usefulness for COVID-19 detection. However, since the phase information is not labeled by existing datasets, we performed manual segmentation of the cough recordings to obtain precise phase segmentation.

Cough recordings from the ComParE and DiCOVA2 datasets were aggregated and randomly assigned to three annotators without class labels or meta-data information. Two annotators labelled half of the samples while the third annotator labelled all samples. Inconsistent annotations were discussed among annotators at the end for a final decision. Since recordings were collected in uncontrolled environments, cough signals could be mixed with other unwanted artifacts, such as background noise. Meanwhile, it was observed that other articulatory sounds could appear along with coughs, such as the sound of a gag reflex or a throat clearing. Hence, annotators were asked to assign one of the following six labels to each audio event : (1) inhalation, (2) compression, (3) expulsion (cough), (4) noise, (5) silence, and (6) other (which includes all types of articulatory sounds other than cough sounds).

A detailed annotation procedure can be summarized as follows. First, all audio files were imported to the PRAAT software (Boersma et al., 2001) for visualization of the corresponding waveform and spectrogram. Annotators started by identifying the onset of the cough (expulsion) phases since the start of a bursting sound can be more easily located visually. As the majority of the recordings were collected indoors, such explosive sounds could lead to reverberant tails, which made it challenging to accurately find the offset of the cough phases. As such, the offset was determined empirically once the amplitude decreased to 5% of the maximal expulsion phase amplitude and the reverberation tail was labelled as noise.

The inhalation phases, in turn, were more difficult to locate due to their low amplitude, thus could often be fully masked by background noise. Hence, only the audible inhalation phases were labelled. Once cough and inhalation phases were determined, the segment between the end of the inhalation and the start of the cough was labelled as a compression phase. Articulatory sounds were annotated using a similar procedure. Other unwanted segments were labelled as either noise or silence depending on their loudness. An example of a cough recording annotation can be found in Fig.3.9.

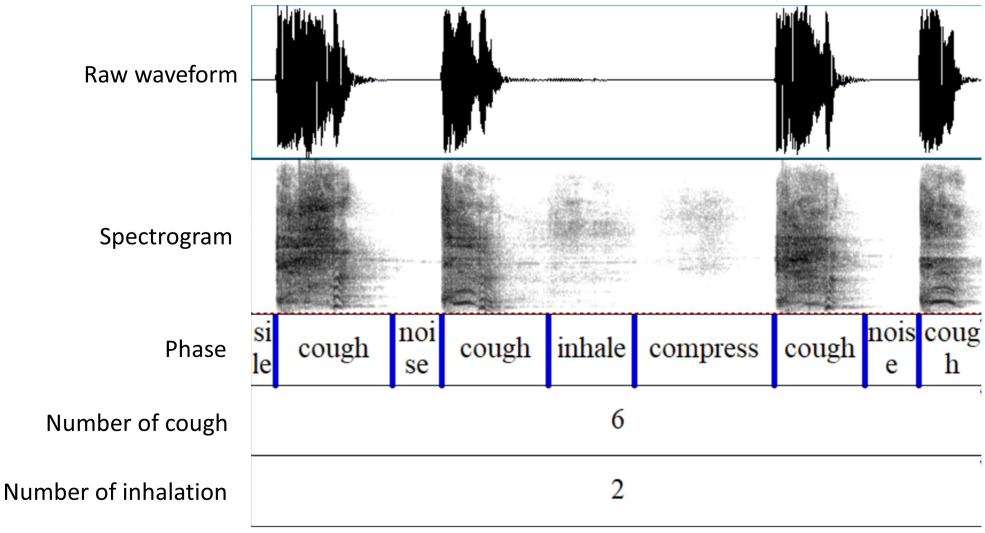


FIGURE 3.9 : Example cough annotation excerpt. Blue vertical lines represent the onset and offset of each phase.

3.6.2 Cough processing pipeline

Extracted features can be divided into two categories : (1) traditional acoustic features (over 6,000 features) computed using the OPENSMILE toolkit (Eyben et al., 2010) and (2) hand-crafted cough temporal features (henceforth referred to as “temporal features”). The acoustic features are computed across three different segments : (1) the entire cough recording, as in the ComParE baseline system (Coppock et al., 2022a); (2) only during cough phases; and (3) only during inhalation phases. The temporal features, in turn, capture cough properties not available within openSMILE. A total of 11 features are computed, including : the frequency of coughs, inhalations, and other articulatory sounds; the average and standard deviation of the duration of cough, inhalation, and compression phases; the ratio of inhalation duration and compression duration to cough duration. A description of these features can be found in Table 3.7.

These temporal features are engineered based on clinical cough measures and pathological insights. For example, the duration of compression phases has been linked to the presence and location of secretion in the airways (Piirila et al., 1995). Inhalation and cough features, in turn, can be indicative of inflammation in different respiratory components (Shannon et al., 2004; Widdicombe, 1954). Lastly, combinations of these feature sets are explored to examine their complementarity; the five different combinations tested are depicted by Fig. 3.10. To test the impact of the different feature sets, and not necessarily of the classifier, each feature set is trained with both the SVM with a linear kernel and the random forest (RF) classifier. The best performance achieved and corresponding classifiers will be reported herein.

TABLEAU 3.7 : Description of cough temporal features.

Abbreviation	Cough temporal features
Num_{cou}	Number of coughs/sec
Num_{inh}	Number of inhalations/sec
Num_{com}	Number of articulatory sound/sec
Dur_{cou_ave},Dur_{cou_std}	Ave and std of cough duration
Dur_{inh_ave},Dur_{inh_std}	Ave and std of inhalation duration
Dur_{com_ave},Dur_{com_std}	Ave and std of compression duration
$Ratio_{inh\&cou}$	Ratio of Dur_{inh_ave} to Dur_{cou_ave}
$Ratio_{com\&cough}$	Ratio of Dur_{com_ave} to Dur_{cou_ave}

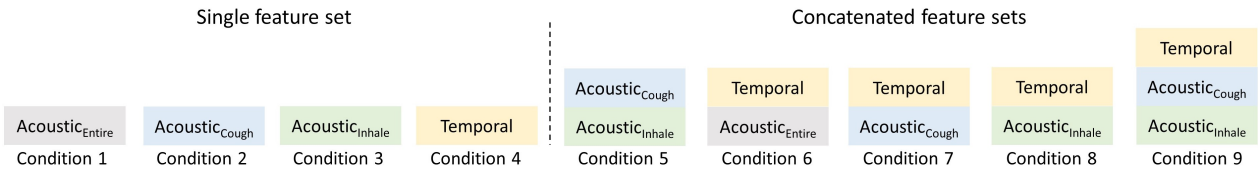


FIGURE 3.10 : Various feature sets and their combinations evaluated in this study (Acoustic : openSMILE, Temporal : proposed).

3.7 Evaluation of cough features

3.7.1 Model performance

Here, we report the performance achieved by different feature set combinations (Table 3.8). As can be seen, for within-dataset tests, the top-performer feature set on ComParE is the fusion of cough and inhalation acoustic features with temporal features (condition 9), achieving an average AUC-ROC of 0.679. For DiCOVA2, in turn, the fusion of baseline acoustic features (computed across the entire recording) and temporal features (condition 5) achieved the highest AUC-ROC of 0.710, with the temporal features only providing a slight increase. The results also show that the importance of cough phases for COVID-19 detection varies across datasets, suggesting poor generalizability. For example, the inhalation features outperformed all other single feature sets on ComParE, while only achieving chance-level performance on DiCOVA2. A possible explanation is that COVID-positive recordings in DiCOVA2 are noisier, thus leading to degraded inhalation phase segments. Moreover, while the proposed temporal features alone do not achieve state-of-the-art performance, their fusion to acoustic ones consistently improved accuracy suggesting their complementarity.

As highlighted in (Zhu et al., ress), existing COVID-19 detection systems based on speech have poor generalizability across datasets. Results for cross-dataset tests suggest this is also the case for coughs where cross-dataset AUC-ROC values were substantially lower than those achieved with within-dataset tests. For sub-condition C-to-D, the fusion of cough acoustic and temporal features (condition 6) showed the highest accuracy with an average AUC-ROC of 0.622. For sub-condition D-to-C, on the other hand, the 11-dimensional proposed temporal feature set

TABLEAU 3.8 : Performance comparison for different feature set combinations. Average AUC-ROC scores are reported with 95% CIs. Bold values indicate best performance for a given condition. C-to-D : train on ComParE and test on DiCOVA2. D-to-C : train on DiCOVA2 and test on ComParE.

Cond	Dim	Within-dataset		Dim	Cross-dataset	
		ComParE	DiCOVA2		CtoD	DtoC
1	6373	.619 (.481-.757)	.704 (.629-.778)	300	.506 (.405-.608)	.547 (.491-.604)
2	6373	.528 (.400-.654)	.647 (.616-.679)	300	.532 (.419-.645)	.523 (.374-.672)
3	6373	.637 (.515-.751)	.515 (.438-.593)	300	.433 (.322-.543)	.501 (.397-.605)
4	11	.532 (.417-.646)	.595 (.491-.700)	11	.548 (.455-.642)	**.576 (.520-.633)**
5	6384	.618 (.479-.756)	**.710 (.640-.779)**	300	.558 (.460-.659)	.549 (.483-.615)
6	6384	.528 (.401-.654)	.655 (.607-.703)	300	**.622 (.559-.686)**	.537 (.474-.599)
7	6384	.644 (.528-.760)	.569(.444-.694)	300	.438 (.324-552)	.518 (.415-.622)
8	12746	.648 (.550-.755)	.648 (.600-.697)	300	.556 (.460-.651)	.565 (.471-.659)
9	12757	**.679 (.558-.781)**	.650 (.603-.698)	300	.576 (.506-.647)	.566 (.480-.651)

alone achieved the best overall accuracy. The feature sets based on inhalation phase features showed the greatest impact dropping to levels below chance. This is likely due to the effect that environmental noise can have on such low-amplitude segments.

3.7.2 Interpretation of cough features

To evaluate the group-level difference in temporal features between COVID-positive and COVID-negative coughs, a Welch's t-test (Delacre et al., 2017) is performed to examine the statistical significance. Among the 11 temporal features, eight were found to be significantly different (*p-value* < 0.05), of which four were with highly significant difference (*p-value* < 0.01). Of these, the frequency of cough/expulsion and inhalation phases were found to be significantly lower for COVID-positive coughs, suggesting that COVID-19 patients made fewer (forced) coughs than healthy individuals during the same time duration. Studies have shown that voluntary coughs rely more on the larynx (Piirila et al., 1995; Korpáš et al., 1996) compared to involuntary coughs, and could be associated with airway clearance ability (Pitts et al., 2008). Considering that subjects were asked to make forced coughs in both datasets, such finding indicates that COVID-19 patients might have decreased control in lower airways, which could be possibly caused by inflammation and decreased neuromotor control of larynx muscles (Naunheim et al., 2020).

In turn, inhalation and compression phases of COVID-positive coughs were shown to be longer and with higher variations. These two phases are at the preparation stage of a cough and have been reported to be longer when secretions are present in the smaller airways (Macklem, 1974; Macklem et al., 1968). Similar patterns have also been found in chronic bronchitis and emphysema (Piirila et al., 1995), of which mucus and shortness of breath are two major symptoms. Interestingly, no significant difference was found in cough duration, which has been shown to increase by 50-100% when inflammation exists in the vocal chord (Korpáš et al., 1996). Notwithstanding,

the obtained findings suggest that the proposed temporal feature set can be a good candidate to capture interpretable and pathological origins of COVID-19 coughs.

3.8 Conclusion

In this chapter, we proposed and evaluated the generalizability and interpretability of two physiology-inspired feature sets for COVID-19 speech and cough, respectively. Regarding speech features, changes in the modulation spectrum, linear prediction coefficients, and linear prediction residuals are investigated and the top-ranked features were found closely related to COVID-19 symptoms. With the obtained insights, a two-stage COVID-19 prediction system is then proposed and tested on three different COVID-19 speech datasets. Experimental results show the proposed two-stage feature fusion system outperforming several benchmarks, especially those based on complex deep neural networks. The two-stage system is tested in a cross-dataset evaluation scheme and shown to achieve greater generalizability across unseen conditions.

For engineering the cough features, we first curated a cough phase segmented dataset based on 1,259 cough recordings. Based on these fine-grained annotations, new cough temporal features are proposed and fused with conventional acoustic features computed separately for different phases. Within- and cross-dataset experiments have shown the importance of the different cough phases for COVID-19 detection, the complementarity of the cough temporal features to acoustic ones, as well as improved generalizability and interpretability.

4 GENERALIZABLE AND INTERPRETABLE DEEP REPRESENTATION LEARNING

4.1 Preamble

This Chapter is compiled from materials extracted from the manuscripts published at Computer Speech & Language (Zhu et al., 2024b), the Advances in Neural Information Processing Systems (NeurIPS) 2024 Zhu et al. (2024d), and the ASVspoof5 Workshop (Zhu et al., 2024c).

4.2 Introduction

In the previous Chapter, we demonstrated that knowledge-based features can benefit decision interpretation, while outperforming some DL models in the low-data regime, such as pathological voice detection. With that being said, the capability of DL models has not been fully explored due to data scarcity. To tackle this issue, a substantial amount of effort has been carried out by researchers around the world that made more pathological speech data available (Xia et al., 2021; Coppock et al., 2022b). Such data expansion motivates the learning of generalizable representations in an end-to-end approach by utilizing larger models. One typical strategy is to use pre-trained large speech models as upstream representation encoders, and append lightweight downstream classification heads for fine-tuning.

While such an approach has shown to outperform knowledge-based ML systems in several tasks (Latif et al., 2020b; Yang et al., 2021a), studies have demonstrated that the improvement is more significant for in-domain tests (i.e., train and test using the same dataset), while generalizing to unseen data remains as a challenging task. Furthermore, since end-to-end models learn representations in a data-driven manner, the output is usually hard to be interpreted by humans. One way to improve the generalizability and interpretability of deep representation learning schemes is to inject domain knowledge into the learning strategy design (Dash et al., 2022). For example, diagnostic models are expected to not place focus on speech content or background noise for prediction. In-the-wild voice data, however, could make it difficult for models to filter these redundant details without guidance. By applying certain constraints during the end-to-end learning process, we can force the models to put less emphasis on such biases and focus on task-related attributes. By doing so, models can still learn representations that carry more information than hand-crafted features, while following task-specific expectations. In this chapter, we introduce two deep representation learning approaches for health diagnostics and deepfake detection, respectively. Since these two tasks differ in nature, they will be described separately in two sections, i.e., Section 4.4 for health diagnostics, and Section 4.5 for deepfake detection.

Overall, the main contributions of this chapter are summarized as follows :

1. We show how domain knowledge can be integrated in the design of different representation learning frameworks for better generalizability and interpretability. We focus on two learning schemes, supervised learning for health diagnostics and self-supervised learning for deepfake detection. In both scenarios, the proposed systems obtain SOTA performance in generalization tests with interpretable results;

2. In Section 4.4, we propose a novel diagnostic system that learns biomarkers from the modulation tensorgram representation of speech, and generates a spectral-temporal saliency map to highlight the discriminative regions for interpretation;

3. In Section 4.5, we introduce a deepfake detection framework that learns the dependency between style and linguistic aspects in real speech via self-supervised pre-training. The dependency embeddings are found crucial for discriminating unseen attacks and interpreting downstream model decisions.

4.3 Related work

4.3.1 Voice health representations

As more pathological speech data become available, researchers have started exploring the use of pre-trained deep models for disease diagnosis. Xia et al., for example, reported improvements in COVID-19 detection performance when using VGGish networks pre-trained on large-scale image datasets as a front-end feature extractor (Xia et al., 2021). Similar improvements were seen in other studies (Coppock et al., 2022b), where transformer-based models pre-trained on non-pathological speech datasets were used. These deep models typically take raw waveforms or the spectrograms as input (Sharma et al., 2022; Akman et al., 2021; Deshpande et al., 2020a). While the waveform and spectrogram representations can provide details about linguistic and paralinguistic content (Benzeghiba et al., 2007), it has been shown they are sensitive to environmental artifacts (e.g., background noise and/or room reverberation) and contain disease-irrelevant information that may bias diagnostic models (e.g., speech content, age, and gender (Zhu et al., 2023a)). Such limitation poses a constraint on model generalizability and interpretability, since it is not clear what information these models are relying on for decision-making.

One alternative representation that has been explored in speech applications is the modulation spectrum representation (MSR) (Kingsbury et al., 1998; Haridas et al., 2018; Wu et al., 2011), which was shown useful for COVID-19 detection in the previous Chapter. However, in our previous work, the temporal dynamics were overlooked, as the MSR entails only the frequency and modulation frequency changes. Furthermore, the spectral information of MSR was described by hand-crafted features which may limit its discriminative power. In the first part of this Chapter, we

propose a model that takes a 3-dimensional version of the MSR as input to automatically learn the biomarkers.

4.3.2 Deepfake speech detection methods

Similar to diagnostics, audio deepfake detection (ADD) models follow a similar trend, where pre-trained large speech models are often used as upstream feature extractors, the extracted representations are fed into different types of downstream classifiers to discriminate deepfakes (Yi et al., 2023). Since synthetic samples can be generated effortlessly with off-the-shelf speech generative models, ADD models are often trained in a supervised manner with a large set of labeled data. However, a major issue faced by existing ADD systems is the severe degradation in performance when tested on unseen data (Müller et al., 2022; Shih et al., 2024), which questions their applicability and trustworthiness for real-world scenarios. To address this issue, multiple works have explored methods to improve model generalizability. With added training cost, improvements have been reported when front-ends are fully fine-tuned alongside the back-end classifiers during downstream training (Tak et al., 2022b; Wang et al., 2021b). Further improvements were achieved with data augmentation, such as RawBoost (Tak et al., 2022a,b) and neural vocoding (Wang et al., 2023c). More recent works also show that distilled student models can generalize better than large teacher models (Lu et al., 2024; Wang et al., 2024). Still, large discrepancies between in-domain and out-of-domain performance are common (Yi et al., 2023; Li et al., 2024).

In addition to generalization, existing ADD models also fall short on interpretability. Several studies have shown that current SOTA models may be focusing on artifacts introduced in the frequency spectrum during voice synthesis and/or the artifacts in non-speech segments (Shih et al., 2024; Müller et al., 2021; Liu et al., 2023; Zhang et al., 2023). While some works proposed to tackle this issue by using interpretable knowledge-based features, such as characterizing breaths (Layton et al., 2024) and vocal tract and articulatory movement details (Blue et al., 2022), the overall detection performance was inferior to deep models.

In the second part of this Chapter, we introduce a novel self-supervised pre-training stage tailored for ADD, which learns the dependency between different aspects in real speech (e.g., speech content, prosody, emotion, etc.). Our proposed model stems from an assumption that the style-linguistics dependency is challenging to be modeled perfectly by existing TTS and VC models. For the majority of generative speech models, the style and linguistic subspaces are assumed to be independent of each other (Tan et al., 2021; Kaur et al., 2023; Triantafyllopoulos et al., 2023; Mohammadi et al., 2017). For example, VC systems change the voice of an utterance by replacing the source speaker's embeddings with those of the target speaker (Mohammadi et al., 2017; Triantafyllopoulos et al., 2023), assuming that these embeddings contain no linguistics information. Similarly, modern TTS systems rely on independently learned representations to model different speech aspects (e.g., text, speaker, emotion) to synthesize expressive speech (Baevski

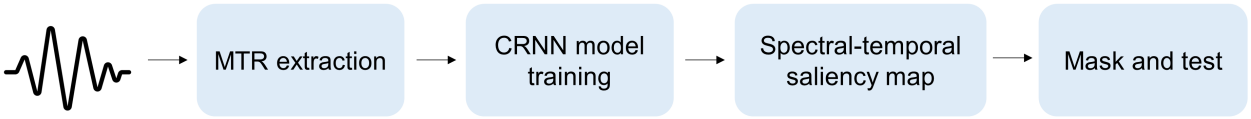


FIGURE 4.1 : Overview of the MTR-CRNN system pipeline.

et al., 2022; Desplanques et al., 2020; Triantafyllopoulos et al., 2024). Because of this disentanglement assumption, a mismatch likely exists between the style and linguistics information in TTS/VC speech that differentiates it from real speech. Since such dependency is difficult to be modelled by existing generative models, the pre-trained representation is by nature a good candidate for discriminating deepfake samples.

4.4 MTR-CRNN : Modulation Tensorgram Representation based Convolutional Recurrent Neural Network

In this Section, we introduce and evaluate a speech-based diagnostic model for COVID-19.

4.4.1 System Overview

The whole pipeline of our proposed system is depicted in Fig. 4.1. We first extract the 3D modulation tensorgram representation (MTR) from the raw waveform, then feed it into the downstream CRNN model. Once the training is completed, a spectral-temporal saliency map is then computed based on the network gradients to capture the important regions in MTR. During the testing stage, the irrelevant regions in MTR are masked out, which forces the model to make decisions using only the salient regions. We explain each step of our pipeline in the following subsections.

4.4.2 Modulation Tensorgram Representation

In the previous Chapter, the steps for obtaining MSR were demonstrated. The 3D MTR is an advanced version of MSR which contains more temporal details. The general processing pipeline for generating MTR is depicted in Fig. 4.2. First, as speech recordings are collected with different devices, the signal amplitude is normalized to remove unwanted amplitude variations caused by different loudness levels. Next, the pre-processed signal $\hat{x}(n)$ is filtered by acoustic filterbanks. For a faster implementation, the signal was first framed with 10ms-windows with the hop length of 2.5ms, leading to a framed signal $\hat{x}(m)$. Regarding the choice of filterbanks, while the gammatone filterbank is commonly used to mimic human perception of sound (Patterson et al., 1987), it is not clear whether such a filterbank remains optimal for processing COVID-19 speech. For example, a recent study showed that a bio-inspired filterbank could outperform the conventional gammatone filterbank when analyzing COVID-19 coughs (Dash et al., 2021). Hence, we experiment with

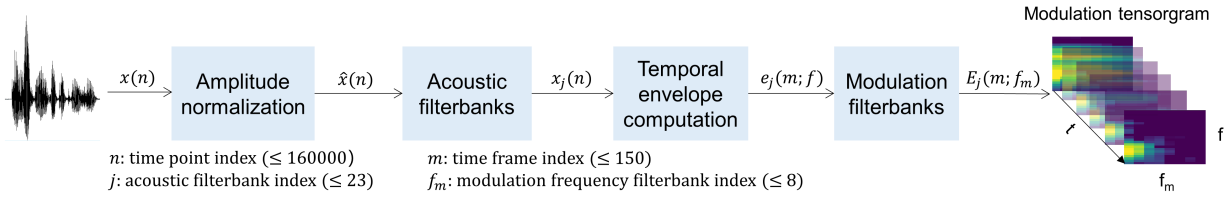


FIGURE 4.2 : Block diagram of the processing steps to compute the 3D modulation tensorgram.

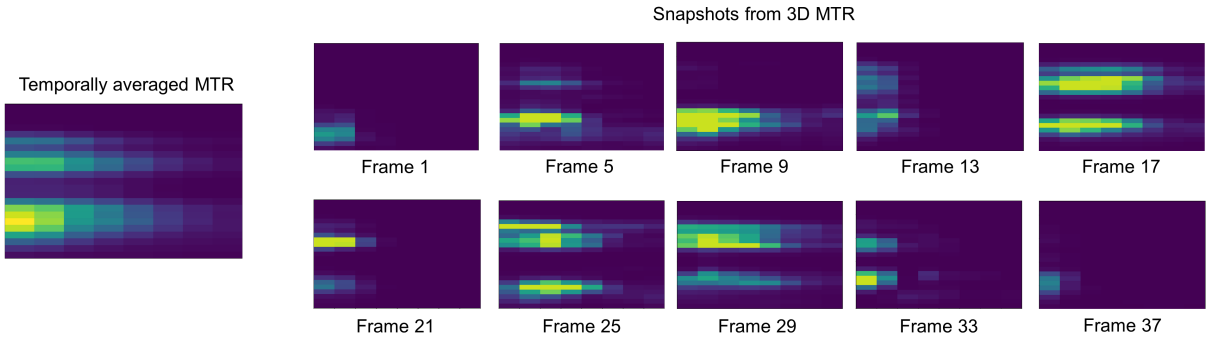


FIGURE 4.3 : Examples of the averaged 2D MTR and MTR snapshots at ten different frames. Examples are generated from the same speech sample.

two different 23-channel filterbanks : linear-scale and gammatone (Slaney et al., 1993). Furthermore, based on insights from (Santos et al., 2013), different lower and upper frequency ranges are explored.

After applying the first filterbank, the temporal envelope $e_j(m)$ is computed from each filtered signal $\hat{x}_j(m)$ via the Hilbert transform :

$$e_j(m) = \sqrt{\hat{x}_j(m)^2 + \mathcal{H}(\hat{x}_j(m))^2}, \tag{4.1}$$

where $\mathcal{H}(\cdot)$ denotes the Hilbert transform and the subscript j denotes the jth filterbank. To obtain temporal dynamics information, each temporal envelope $e_j(m)$ is then windowed with a 256-millisecond (ms) Hamming window and an overlap of 192 ms. Such window length is relatively longer than that used in conventional spectrograms (e.g., 16 ms) and has been shown to provide appropriate resolution in low-frequency modulation frequencies (Falk et al., 2009).

To obtain the modulation spectrum $E_j(m; f_m)$ from each acoustic frequency component, the discrete Fourier transform $\mathcal{DFT}(\cdot)$ is applied to the temporal envelope $e_j(m)$:

$$E_j(m; f_m) = \left|\mathcal{DFT}(e_j(m))\right|, \tag{4.2}$$

where $|\cdot|$ denotes the absolute value operation, m denotes the frame number, and f_m denotes modulation frequency. An 8-channel modulation filterbank is then used to group neighboring modulation frequencies. Similar to the acoustic filterbanks, two different modulation filterbank types are tested, as are different lower/upper modulation frequency values. Table 4.1 summarizes the types of filterbanks tested and upper/lower frequency ranges. The optimal settings found through our ex-

TABLEAU 4.1 : Overview of the Modulation Tensorgram Parameter Search Detailing Types of the Filterbank, Acoustic Frequency (f) Range, and Modulation Frequency (f_m) Range.

Parameter	Range	Step	Optimal
Acoustic filterbank	gammatone, linear	-	gammatone
Modulation filterbank	log, linear	-	log
Lower bound f (kHz)	0-1	.125	.125
Higher bound f (kHz)	6-8	1	8
Lower bound f_m (Hz)	0-3	1	3
Higher bound f_m (Hz)	16-128	8	32

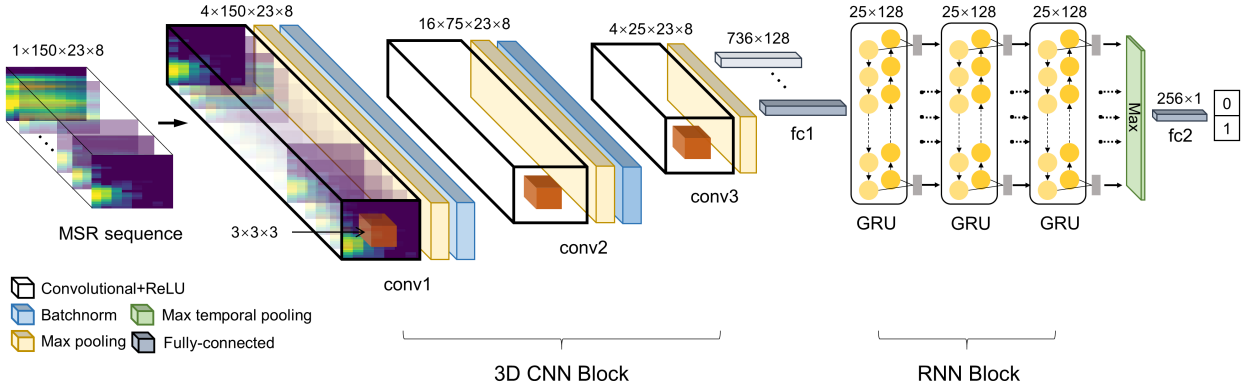


FIGURE 4.4 : Model architecture of the proposed MTR-CRNN system.

perimentation are also reported in the table. Lastly, all MSRs computed per frame are aggregated into a final 3D representation called a "modulation tensorgram" (MTR).

Fig. 4.3 illustrates the importance of using a 3D tensorgram representation, as opposed to an averaged 2D representation used in the previous Chapter. On the far left, the MSR averaged over all frames is shown. On the right, ten different MSR snapshots are shown. As can be seen, the modulation spectral patterns can differ greatly across time frames and such changes may carry important diagnostic cues.

4.4.3 Model Architecture

The model architecture of MTR-CRNN is depicted in Fig. 4.4. The CRNN model is comprised of two blocks, namely the 3D convolutionel neural network (CNN) block and the recurrent neural network (RNN) block. To ensure that the input sequence length is compatible with the 3D maximum pooling layers in the CNN block, speech samples are first unified to 10 s length by right zero-padding shorter recordings and segmenting longer recordings. This leads to a consistent 3D MTR shape $\{1 \times 150 \times 23 \times 8\}$ across all speech samples. Each input 3D MTR is then mean-variance normalized. Each part of the CNN block consists of a convolutional layer, a batch normalization layer, and a max pooling layer. A $\{3 \times 3 \times 3\}$ kernel is used for all three convolutional layers to extract meaningful modulation spectral patterns from neighboring MTR snapshots. The max pooling layer aims to remove the redundant MTR snapshots with relatively low energies, as these

MTR snapshots usually correspond to silent frames. The output of the CNN block is then fed into a fully-connected layer to reduce feature dimensionality, leading to an output sequence of shape $\{25 \times 128\}$.

The subsequent RNN block has three cascaded bi-directional gated recurrent unit (GRU) layers to explore the temporal dependency of neighboring MTR snapshots. The output of the RNN block is then layer-normalized and fed into a pooling layer, which finds the maximal value along the time dimension to generate a sequence-level embedding of shape $\{1 \times 256\}$. Lastly, a fully-connected layer is used to project the 256-dimension embedding to a 1-dimension output, which is then passed through a sigmoid layer to obtain the final COVID-19 probability score. To avoid over-fitting, a dropout factor of 0.7 is applied to the last fully-connected layer.

4.4.4 MTR Saliency Maps

Previous studies have suggested that different regions of the modulation spectrum correspond to different properties of the speech signal (Greenberg et al., 1997; Sarria-Paja et al., 2013; Zhu et al., 2022). A better understanding of the MSR regions being used by the model would allow for better interpretation of the results and could lead to insights about the acoustic properties of COVID-19 speech. To this end, we propose a spectral-temporal saliency map based on the "vanilla gradient" saliency map algorithm originally invented for weakly supervised learning (Simonyan et al., 2013). The vanilla gradient method relies on calculating gradients through backpropagation with respect to each pixel in the input image. The value assigned to each pixel reflects the relevance to the final prediction, where higher values indicate higher relevance (Simonyan et al., 2013). The method has been shown to be more robust than perturbation-based methods, thus is a good candidate for in-the-wild data (Adebayo et al., 2018). However, the original vanilla gradient saliency map is designed for 2D images, where only spatial information is considered. In our case, the inputs are 3D tensorgrams where the 2D spectral patterns change through time. Hence, we designed a spectral-temporal saliency map method which integrates the temporal importance when computing saliency for each pixel in the spectral pattern.

The processing steps used to compute the spectral-temporal saliency maps are depicted in Fig. 4.5. First, the vanilla gradient method is used to compute raw saliency maps from a trained MTR-CRNN. The output map shape remains the same as the input shape, i.e., $\{150 \times 23 \times 8\}$. Although the raw saliency maps suggest attentive regions in each MTR snapshot, the temporal saliency is not well presented. Inspired by an audio-visual fusion saliency model (Koutras et al., 2018), we solve this issue by first transforming the 3D raw saliency maps to a set of 1D temporal saliency coefficients. The transformation procedure is as follows. A 3D filter F is first used to discard

the low-saliency regions in each 2D saliency map $M(t,i,j)$ with a pre-determined threshold :

$$F(t,i,j) = \begin{cases} 1 & \text{for} \quad |M(t,i,j)| \geq 0.2 \max\{M\} \\ 0 & \text{for} \quad |M(t,i,j)| < 0.2 \max\{M\}, \end{cases} \tag{4.3}$$

where M denotes the raw 3D saliency map. Our exploratory analysis showed that higher threshold values (≥ 0.3) led to insignificant difference noticed between COVID-positive and negative samples, while lower values ($\leq$ 0.1) failed to filter out the unwanted regions. Values between 0.1 and 0.3 led to similar outputs. Hence, we set the threshold value to 0.2. Next, the filter is applied to each 2D saliency map and an averaging is used to obtain the temporal saliency coefficient C_t :

$$C_t(t) = \frac{\sum_{i,j}(F(t,i,j) \odot M(t,i,j))M(t,i,j)}{\sum_{i,j} F(t,i,j)M(t,i,j)}. \tag{4.4}$$

Each set of coefficients is of shape $\{150 \times 1\}$, corresponding to the temporal attention scores at each time step. To unify the coefficient range across samples, each set of coefficients is normalized between 0 and 1 and a 1D median filter is applied. Lastly, the temporal saliency coefficients C_t are multiplied with the raw saliency map to obtain the spectral-temporal saliency map M_{ST} :

$$M_{ST} = C_t(t)M(t). \tag{4.5}$$

As the final goal is to localize modulation spectral regions that are most closely related to COVID-19, the 3D spectral-temporal saliency map M_{ST} is then averaged over time, which results in a single 2D saliency map M_{Ave} per sample. To further explore group differences, the Fisher ratio (F-ratio) is computed between two groups (COVID-19 positive and COVID-19 negative) of temporally averaged saliency maps :

$$\textit{F-ratio} = \frac{VAR_b}{VAR_w}, \tag{4.6}$$

where VAR_b represents the between-group variance, and VAR_w represents the within-group variance for each of the 23×8 saliency map values. Fisher ratio scores are then used to highlight the important discriminatory regions in the MTR.

4.4.5 Experiments

Comparison with benchmark models To evaluate the performance of the proposed MTR-CRNN, we compare it to the best performing models on CSS and DiCOVA2 in within-dataset and cross-dataset settings. As can be seen in Table 4.2, the within-dataset results on CSS show the proposed MTR-CRNN system outperforming even the CSS benchmark, resulting in a final average AUC-ROC of 0.770. On the DiCOVA2 dataset, the obtained results were in line with those obtained from the DiCOVA2-optimized benchmark. Overall, systems achieved a somewhat lower accuracy

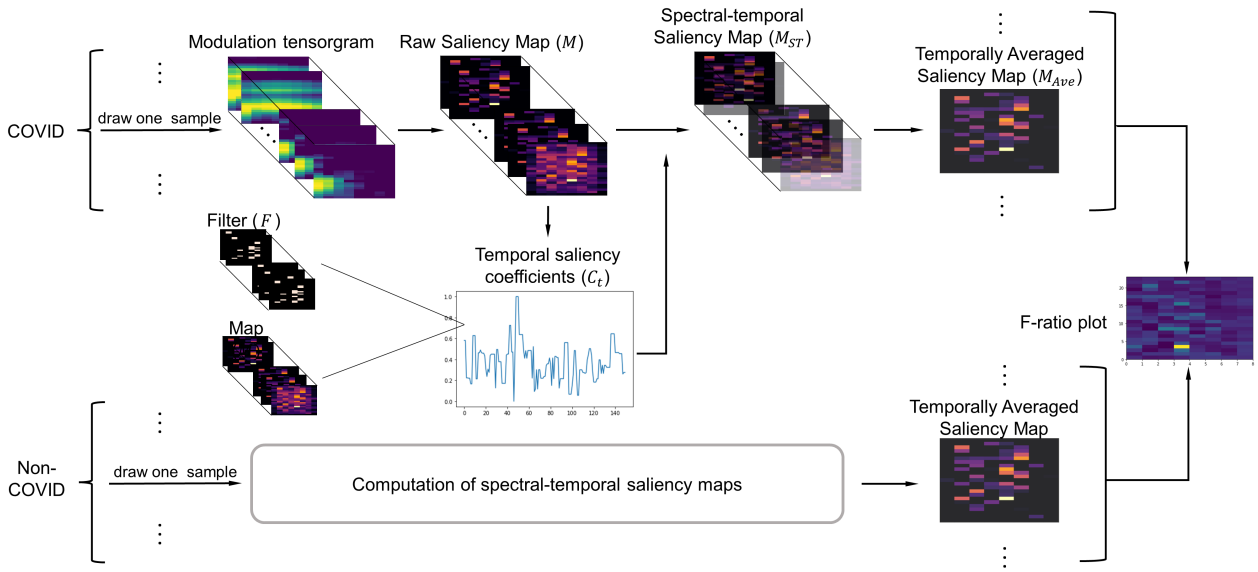


FIGURE 4.5 : Computation of the spectral-temporal saliency maps and the F-ratio plot. Only training data are used.

TABLEAU 4.2 : Task-1 Performance Comparison. Average and Standard Deviation of AUC-ROC Scores Are Calculated From 10 Different Initializations. Bold Values Indicate the Highest AUC-ROC. 'DiC' Corresponds to DiCOVA2 and 'Num_param' to Number of Parameters in the Deep Learning Models.

System	Num_param	Within-dataset		Cross-dataset	
		CSS	DiC	CSS→DiC	DiC→CSS
CRNN	0.9 million	**.770±.019**	.781±.011	**.600±.023**	**.509±.004**
CSS-Benchmark	-	.758±.008	.756±.010	.511±.007	.486±.009
DiC-Benchmark	0.8 million	.714±.015	**.789±.016**	.483±.020	.462±.019

on the CSS dataset, suggesting that COVID-19 detection with CSS could be a more challenging task. This is likely due to the varied language distribution of the dataset, as well as the higher percentage of asymptomatic COVID-19 samples present in CSS. When tested in the cross-database setting, the proposed MTR-CRNN system showed a substantial improvement relative to the benchmarks. As can be seen, the benchmarks dropped to chance levels when tested on unseen sets. As CSS and DiCOVA varied greatly in demographics (e.g., language, gender, age) and speech content, these results suggest that the proposed method achieved greater generalizability and robustness to unseen data, whilst requiring only 0.1 million more parameters than the DiCOVA2 benchmark. Notwithstanding, the drops seen in accuracy suggest that further improvements may be possible.

Interpretation of spectral-temporal saliency maps We have shown that although MTR-CRNN obtains better performance than benchmark models, the generalizability to unseen datasets can be further improved. Furthermore, it is not clear what information the CRNN model relies on for decision-making. These issues are tackled by the spectral-temporal saliency maps obtained from

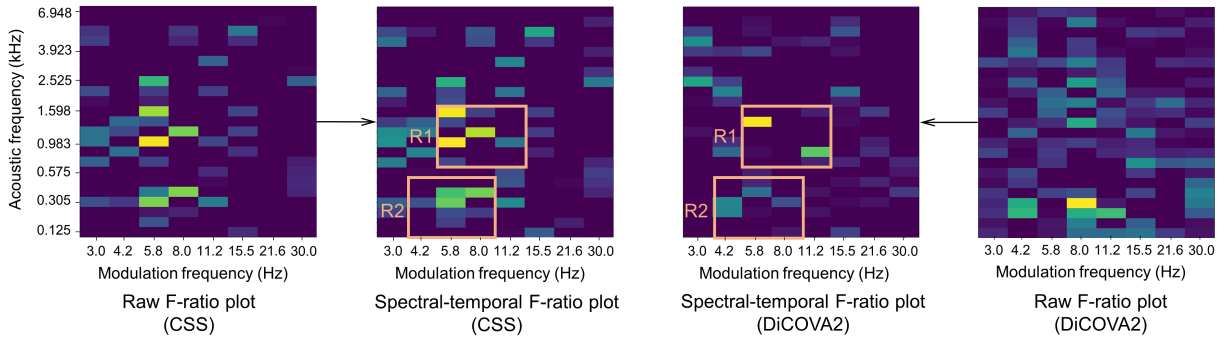


FIGURE 4.6 : Discriminative patches found consistently in the spectral-temporal F-ratio plots. Two middle spectral-temporal F-ratio plots are generated with the spectral-temporal saliency maps, while the raw F-ratio plots are generated directly from the raw saliency maps. Brighter areas represent higher discrimination between positive and negative COVID-19 speech samples.

the pre-trained CRNN models. Specifically, we first generate two saliency maps separately, one from the CRNN trained with CSS and the other from the CRNN trained with DiCOVA2. We then cross-validate the two maps to pinpoint shared salient regions. The saliency maps and regions are depicted in Fig. 4.6.

To avoid measuring energy in individual acoustic-modulation frequency bins, here we propose to group neighboring frequency-frequency bins into patches. We empirically propose patches of shape $\{6 \times 3\}$ where modulation spectral energy values within the patches are summed and min-max normalized. As shown in the figure, two patches, denoted by R1 and R2, are chosen to represent the two most discriminant regions consistently present across the two datasets. R1 corresponds to $f = 650 - 1600$ Hz and $f_m = 5 - 13$ Hz while R2 to $f = 125 - 500$ Hz and $f_m = 3.5 - 10$ Hz. Previous studies have shown that whispered speech is usually manifested at $f < 1$ kHz and $f_m = 5 - 13$ Hz (Sarria-Paja et al., 2013), which partly overlaps with the location of R1 and R2. This finding could be linked to an increased level of vocal hoarseness that has been commonly reported with COVID-19 speech, possibly caused by inflammation of the vocal tract area. In turn, the highest F-ratio values for both datasets were found around $f = 1.6$ kHz and $f_m = 5 - 6$ Hz. This corroborates previous findings where COVID-19 speech showed more centralized spectral energy at around $f = 2$ kHz (Zhu et al., 2022).

Moreover, a direct comparison of the raw and spectral-temporal F-ratio plots shows that both are almost identical for CSS, but a marked difference can be seen for DiCOVA2. In the raw F-ratio plot of DiCOVA2, the highlighted R1 area in the spectral-temporal plot is barely noticeable. Brighter regions appear more often in the higher acoustic and modulation frequency ranges, which have been linked in the past to represent room acoustic effects, such as reverberation (Falk et al., 2009). It is suspected that when the DiCOVA2 data were collected, isolation was still required when testing positive, thus COVID-19 positive speech samples could be affected by e.g., room reverberation, which is more pronounced in enclosed environments. This further shows the importance of aggregating the temporal aspect within the saliency map, thus allowing the model to focus on true

TABLEAU 4.3 : Performance Comparison of Different MTR Masks. The Last Column Reports AUC-ROC Averaged Across All Tasks. Bold Values Indicate the Best System for a Given Task. C-B : Benchmark model on CSS; D-B : Benchmark model on DiCOVA2.

Sys	Patches	Within-dataset		Cross-dataset		Unseen dataset			Ave
		CSS	DiC	C→D	D→C	C→Cam	D→Cam	C+D→Cam	
Ours	R1	.732±.019	.741±.017	**.705±.015**	**.651±.013**	.512±.006	.531±.007	.554±.010	**.632±.091**
	R2	.591±.018	.756±.010	.479±.017	.524±.019	.514±.008	.542±.009	.552±.011	.565±.084
	R1&2	.656±.015	.775±.017	.602±.010	.558±.016	.538±.009	**.556±.010**	**.560±.007**	.606±.078
	Original	**.770±.019**	.781±.011	.600±.023	.509±.004	**.540±.011**	.541±.007	.543±.008	.612±.106
C-B	-	.758±.008	.756±.010	.511±.007	.486±.009	.522±.006	.501±.009	.536±.006	.581±.105
D-B	-	.714±.015	**.789±.016**	.483±.020	.462±.019	.471±.007	.483±.010	.486±.011	.556±.126

COVID-19 discrimination properties and not potential database biases due to, e.g., room properties resulting from quarantine isolation.

Improved generalizability with spectral-temporal masking With these insights and obtained saliency maps, masking is applied to the MTR to extract the two regions only. Table 4.3 (columns 3-6) reports the accuracy achieved when only R1, only R2, and both R1+R2 regions are used in the mask. For comparisons, the original results with the full MTR (as per Table 4.2) are also listed. As can be seen, for the within-dataset test, the original configuration outperformed the masked ones for both datasets. Notwithstanding, the masked version resulted in the highest accuracy in the cross-dataset task, with a substantial margin of improvement relative to the original version.

Interestingly, using only R1 resulted in the highest cross-database accuracy, with no benefits seen by adding information from R2. Columns 7-9 show the accuracy achieved when a model is trained on only the CSS dataset (and tested on the unseen Cambridge set), only DiCOVA2, and on the combined CSS+DiCOVA2 sets, respectively. As can be seen, accuracy drops for all tested algorithms. All proposed solutions outperform the two benchmarks. In this setting, aggregating information from both R1 and R2 regions showed the greatest accuracy across most conditions. Moreover, increasing the training set size by aggregating two datasets showed some improvement in accuracy, but not substantial. For comparisons, within-dataset accuracy on the Cambridge set for the CSS and DiCOVA2 benchmarks of .541 and .543 were achieved. As such, our proposed system with patched input outperforms the two benchmarks even in a more stringent testing condition, which shows the robustness of the patches found in Task-2 and the generalizability across datasets.

4.5 SLIM : Style-Linguistic Mismatch Model for Speech Deepfake Detection

In the previous Section, we focused on incorporating domain knowledge into a supervised training scheme for health diagnostics. In this Section, we show that domain knowledge can also drive a self-supervised learning scheme. More specifically, we model the dependency between style and linguistics aspects of real speech via self-supervised contrastive learning, and introduce a deepfake speech detection model that focuses on style-linguistics mismatch (SLIM).

4.5.1 Quantifying mismatch by CCA analysis

As is discussed in Section 4.3.2, a mismatch likely exists between the style and linguistics information in TTS/VC speech that differentiates it from real speech. To study this hypothesis, we conduct a proof-of-concept experiment on a sample subset of `ASVSpoof2019` (Todisco et al., 2019). Following previous research (Raghu et al., 2017; Kornblith et al., 2019; Pasad et al., 2021), we use canonical correlation analysis (CCA) to derive a subspace where the linear projections of the style and linguistics embeddings are maximally correlated for the real class. We choose the last layer output of the pre-trained `wav2vec2-large-xlsr-53-english` model (Grosman, 2021) for linguistics representation, and the pre-trained `ECAPA-TDNN` embeddings (Desplanques et al., 2020) for style representation.

TABLEAU 4.4 : Mean and standard deviation of Pearson correlation coefficients (*CC*) calculated between style and linguistics embeddings for real and TTS/VC samples across 5 unseen speakers. Significant difference (calculated by Welch's t-test) is seen between real speech and all types of generated speech.

Class	Real	A01 (TTS)	A02 (TTS)	A03 (TTS)	A04 (TTS)	A05 (VC)	A06 (VC)
CC	**.308±.025**	.202±.033	.217±.020	.243±.024	.253±.021	.214±.026	.252±.020

We randomly select 100 real speech samples from the `ASVspoof2019` (Todisco et al., 2019) training set to fit 20-dimensional CCA features for both linguistics and style representations. We then apply the CCA projection to 200 audio files from 5 unseen speakers and 6 TTS/VC systems, and compute the correlation values between these projected style and linguistics vectors to quantify the subspace similarities.

Table 4.4 shows the obtained results. A higher *CC* is seen for the real samples, whereas significantly lower correlations are observed for both TTS and VC generated samples. Moreover, TTS-samples on average show lower *r* (0.228) than VC-samples (0.236), indicating that VC-samples are closer to real speech in terms of style-linguistics dependency. This could explain why VC samples were found to be more challenging to detect than TTS samples in the ASVspoof2019 challenge (Liu et al., 2023). While our findings demonstrate the usefulness of CCA for validating the subspace mismatch, its limitations, such as that it only explores the linear composites of the variables (Weenink, 2003), might make it sub-optimal to be used independently for deepfake detection. We therefore develop a detection framework that *explicitly* studies the style-linguistic mismatch and scales to larger amounts of data.

4.5.2 Model architecture

Our two-stage Style-LInguistics Mismatch (SLIM) learning framework is outlined in Fig. 4.7. The first stage operates on the real class only and employs self-supervised learning to build style and linguistic representations and their dependencies for real speech. In the second stage, a classifier

is fit onto the learned representations via supervised training over deepfake datasets with binary (real/fake) labels.

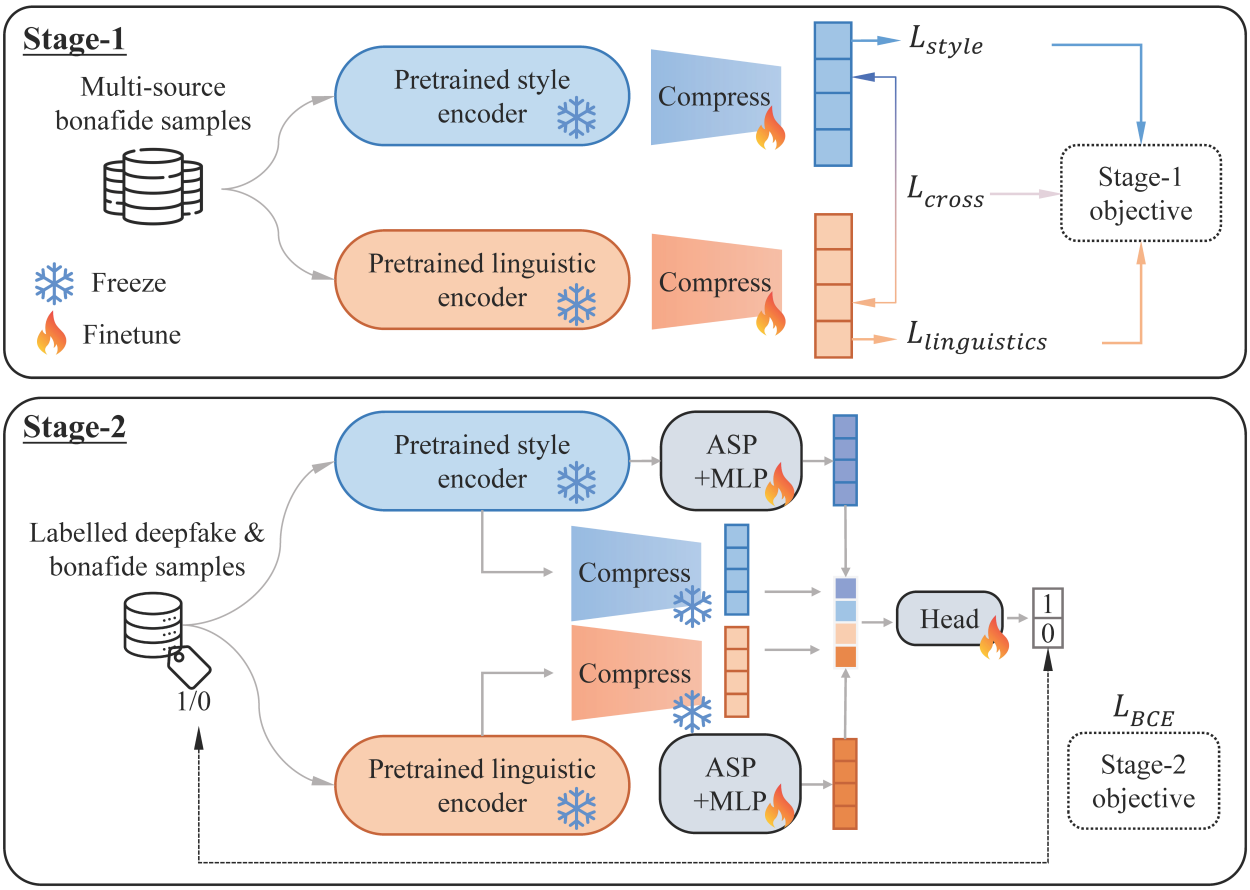


FIGURE 4.7 : SLIM : A two-stage training framework for ADD. Stage 1 extracts style and linguistics representations from frozen SSL encoders, compresses them, and aims to minimize the distance between the compressed representations ($\mathcal{L}_{cross}$), as well as the intra-subspace redundancy ($\mathcal{L}_{style}$ and $\mathcal{L}_{linguistics}$). The Stage 1 features and the original subspace representations (pre-trained SSL embeddings) are combined in Stage 2 to learn a classifier via supervised training.

4.5.2.1 Stage 1 : One-class self-supervised contrastive training

The goal of the first stage is to learn pairs of dependency features from style and linguistics subspaces, which are expected to be highly correlated for real samples and minimally correlated for deepfakes. Since only real samples are needed, we incorporate other open-source speech datasets to diversify the style variations. Given a speech sample, we first extract the style and linguistics representations separately using pre-trained networks. Since recent SSL models achieve superior performance on multiple speech downstream tasks compared to conventional speech representations (e.g., ECAPA-TDNN) (Baevski et al., 2020; Chen et al., 2022; Hsu et al., 2021; Yang et al., 2024a; Pepino et al., 2021) we select a group of SSL models fine-tuned for paralinguistics and linguistics tasks as candidate encoders (Fan et al., 2021; Vaessen et al., 2022; Pepino et al.,

2021; Wang et al., 2021c; Babu et al., 2021). In addition, it has been shown that early to middle model layers carry paralinguistics information, while later layers encode linguistics content (Pasad et al., 2021; Ashihara et al., 2024; Lin et al., 2023; Shah et al., 2021). We, thus, conducted thorough analyses to examine the cross-correlation between pre-trained SSL model layers and chose layer 0-10's output from `Wav2vec-XLSR` fine-tuned for speech emotion recognition to represent style, and layer 14-21's output from `Wav2vec-XLSR` fine-tuned for automatic speech recognition, to represent linguistics information.

Both style ($\mathbf{X}_S$) and linguistics ($\mathbf{X}_L$) embeddings are three-dimensional tensors $\in \mathbb{R}^{K\times F\times T}$, where K denotes the transformer layer index, F denotes the feature size, and T denotes the number of time steps. These subspace embeddings are sent into compression modules $\mathcal{C}(\cdot)$, which average the transformer layer outputs and reduce the feature size from 1024 to 256. We refer to the output from the compression modules as dependency features : $\mathbf{S}_{f,t} = \mathcal{C}(\mathbf{X}_S)$ for style and $\mathbf{L}_{f,t} = \mathcal{C}(\mathbf{X}_L)$ for linguistics, and their temporally averaged versions are denoted $\bar{\mathbf{S}}_f$ and $\bar{\mathbf{L}}_f$. These dependency features are learned by minimizing the self-contrastive loss $\mathcal{L}_{con}$, defined as :

$$\mathcal{L}_{con} = \mathcal{L}_{cross} + \lambda\mathcal{L}_{intra}, \qquad \mathcal{L}_{intra} = \mathcal{L}_{style} + \mathcal{L}_{linguistics} \tag{4.7}$$

$$\mathcal{L}_{cross} = \frac{1}{T}\sum_{t=0}^{T} \|\mathbf{S}_{f,t} - \mathbf{L}_{f,t}\|_{\mathbf{F}}^2, \qquad \mathcal{L}_{intra} = \|\bar{\mathbf{S}}_f\bar{\mathbf{S}}_f^\intercal - \mathbb{I}\|_{\mathbf{F}}^2 + \|\bar{\mathbf{L}}_f\bar{\mathbf{L}}_f^\intercal - \mathbb{I}\|_{\mathbf{F}}^2 \tag{4.8}$$

$\mathcal{L}_{cross}$ denotes the cross-subspace loss; $\mathcal{L}_{intra}$ is the intra-subspace loss, defined in terms of $\mathcal{L}_{style}$ and $\mathcal{L}_{linguistics}$ (Figure 4.7); $\lambda \in [0, 1]$ is a hyperparameter that weighs the two loss terms, T is the number of time steps; and $\|(.)\|_{\mathbf{F}}^2$ is the Frobenius norm. The $\mathcal{L}_{cross}$ term reduces distance between the compressed style and linguistic embeddings, while the $\mathcal{L}_{intra}$ term reduces redundancy within the (temporally averaged) style and linguistic features by pushing off-diagonal elements to zero. The learned dependency features from Stage 1 can be used to quantify whether a mismatch exists between the style and linguistics of an audio. We further demonstrate this in Section 4.5.5.

4.5.2.2 Stage 2 : Supervised training

The second stage of SLIM follows a standard supervised training scheme, where the dependency features and subspace representations are concatenated and fed into a classification head to generate a binary real/fake outcome. As shown in Figure 4.7, the subspace SSL encoders and compression modules are obtained from Stage 1 and are all frozen during Stage 2. Since the dependency features are specifically designed to capture the style-linguistics mismatch alone, we complement them with the original embeddings in order to capture other artifacts that can help separate real samples from the fake class. The original embedding dimensions are reduced from 1024 to 256 through an attentive statistics pooling (ASP) layer and a multi-layer perceptron (MLP)

network. The projected subspace embeddings when concatenated with dependency features result in 1024-dimensional vectors. The classification head consists of two fully-connected layers and a dropout layer. Binary cross-entropy loss is used to jointly train the ASP and MLP modules alongside the classification head.

4.5.3 Training and evaluation details

Stage 1 training Unlike benchmark models which are trained end-to-end in a supervised manner, our model relies on two-stage training where each stage requires different training data to avoid information leakage. Since only real samples are needed in Stage 1, we take advantage of open-source speech datasets by aggregating subsets from the `Common Voice` (Ardila et al., 2020) and `RAVDESS` (Livingstone et al., 2018) as training data and use a small portion of real samples from the `ASVspoof2019 LA train` for validation. Both `Common Voice` and `RAVDESS` cover a variety of speaker traits. The former is a crowdsourced dataset collected online from numerous speakers with uncontrolled acoustic environments, while the latter is an emotional speech corpus with large variations in prosodic patterns. Such data variety enables our model to learn a wider range of style-linguistics combinations.

Stage 2 training and evaluation For a fair comparison with existing works, we adopt the standard train-test partition, where only the `ASVspoof2019` logical access (LA) training and development sets are used for training and validation. For evaluation, we use the test split from `ASVspoof2019 LA` (Todisco et al., 2019) and `ASVspoof2021 DF`(Liu et al., 2023). `ASVspoof2019 LA` and `ASVspoof2021 DF` have been used as standard datasets for evaluating deepfake detection models, where real speech recordings originate from the `VCTK` and `VCC` datasets (Yamagishi et al., 2019; Lorenzo-Trueba et al., 2018; Yi et al., 2020) and the spoofed ones are generated with a variety of TTS and VC systems. In addition, we assess our model's generalizability on out-of-domain data : `In-the-wild` (Müller et al., 2022), and the English subset from `MLAAD v3` (Müller et al., 2024).

4.5.4 Datasets

Table 4.5 describes the details of datasets used for Stage 1 and Stage 2 training and evaluation. Further details about the synthesized datasets can be found in Section 2.4.2.

4.5.5 Experiment results

Detection performance. Table 4.6 summarizes the detection performance of all models and compares the number of trainable parameters. We discuss the models with *frozen front-end* here,

TABLEAU 4.5 : Summary of datasets used for Stage 1 and Stage 2 training and evaluation.

Stage 1 datasets							
Name	Split	#Sample	#Real	#Fake	#Attacks	Speech type	Environment
Common Voice	Train	3k	3k	–	–	Scripted	Crowdsourced
RAVDESS	Train	3k	3k	–	–	Scripted	Studio
19 LA train	Valid	500	500	–	–	Scripted	Studio
Stage 2 datasets							
Name	Split	#Sample	#Real	#Fake	#Attacks	Speech type	Environment
19 LA train	Train	25380	2580	22800	6	Scripted	Studio
19 LA dev	Valid	24884	2548	22336	6	Scripted	Studio
19 LA eval	Test	71237	7355	63882	17	Scripted	Studio
21 DF eval	Test	611829	22617	589212	100+	Scripted	Studio
In-the-wild	Test	31779	11816	19963	N/A	Spontaneous	In-the-wild
MLAAD-EN	Test	37998	18999	18999	25	Scripted	Studio

and compare the models with *fine-tuned front-ends*. ASVspoof2019 eval set contains 19 types of attacks, out of which 6 are seen during training. This makes it the simplest of the four test datasets. We see that a majority of the models achieve near-perfect performance, with several including SLIM reporting EER below 1%. As expected, degradation is seen when models are tested on ASVspoof2021, where the majority of attacks are unseen. Both W2V-LCNN and SLIM are top-performers, with no significant difference between the two.

With the out-of-domain datasets (In-the-wild and MLAAD-EN), more severe degradation is observed, where the majority report EERs over 20%. SLIM, however, outperforms the others with EER of 12.9% and 13.5% on In-the-wild and MLAAD-EN, respectively. It should be noted that although ASVspoof2021 is often used as a standard dataset to evaluate model generalizability to unseen attacks (Liu et al., 2023), part of the real samples in ASVspoof2021 originate from the same dataset (the VCTK corpus (Yamagishi et al., 2019)) as the ASVspoof2019 training data (Todisco et al., 2019; Chintha et al., 2020; Tak et al., 2021; Xie et al., 2023; Wang et al., 2021b). As a result, the real samples from ASVspoof2019 and ASVspoof2021 share a similar distribution, whereas the In-the-wild and MLAAD-EN samples share nearly no overlap with ASVspoof. Generalization to In-the-wild and MLAAD-EN is therefore more challenging than to ASVspoof2021. The large gains reported by SLIM demonstrates how the style-linguistics mismatch helps with generalization to unseen data.

In Table 4.6, we also demonstrate the benefits of introducing Stage 1 by considering features from SLIM variants as inputs to Stage 2 : dependency features, the style and linguistics embeddings (Enc_{sty} and Enc_{ling}), as well as their combination. The architecture of the classification head is kept the same, except for the number of neurons in the input layer. The dependency features outperform the rest on the two out-of-domain datasets, while the subspace embeddings perform better on ASVspoof2021. Simply concatenating the style and linguistics embeddings does not yield signi-

ficant improvements when compared to benchmark models. This suggests that the style-linguistics dependency may not be fully captured by supervised training methods without explicit guidance.

TABLEAU 4.6 : Detection performance on different deepfake datasets. Experiments were repeated three times with different random seeds, and average metric values are reported. #Param refers to the number of trainable parameters (in millions). For SLIM, we sum up parameters trained at both stages. A few models do not make their code open-source, we therefore include the metrics reported in their papers and skip parameter calculation (N/A). Lowest EERs are bolded per category.

Category	Model	ASVspoof19		ASVspoof21		In-the-wild		MLAAD-EN		#Param (million)
		EER↓	F1↑	EER↓	F1↑	EER↓	F1↑	EER↓	F1↑	
Frozen frontend	LCNN (Chintha et al., 2020)	3.7	.834	25.5	.197	65.6	.373	37.2	.654	4
	RawNet2 (Tak et al., 2021)	3.0	.875	22.3	.213	37.8	.602	33.9	.676	4
	PS3DT (Yadav et al., 2023)	4.5	–	–	–	29.7	–	–	–	N/A
	W2V-ASP	3.3	.858	19.6	.233	30.2	.705	29.1	.715	9
	WLM-ASP	**0.3**	.983	9.0	.426	25.4	.751	30.3	.709	9
	HUB-ASP	0.5	.975	15.4	.289	29.9	.718	31.0	.702	9
	W2V-LLGF (Wang et al., 2021b)	2.3	.936	9.4	.402	25.1	.756	27.8	.731	10
	W2V-LCNN (Xie et al., 2023)	0.6	–	**8.1**	–	24.5	–	–	–	N/A
	W2V+WLM	1.8	.916	22.5	.203	30.3	.704	27.0	.739	9
	W2V+HUB	0.9	.956	14.2	.310	27.9	.737	27.6	.732	9
	WLM+HUB	0.8	.963	16.7	.269	29.2	.724	28.5	.720	9
	SSL-Fusion (Yang et al., 2024b)	**0.3**	.981	8.9	.419	24.2	.765	26.5	.739	10
	SLIM variants (ours)									
	Enc_{sty}	6.7	.740	8.6	.438	29.2	.724	25.4	.756	9
	Enc_{ling}	5.9	.764	9.3	.407	30.4	.708	25.0	.760	9
	$\text{Enc}_{style+ling}$	3.5	.834	9.0	.429	25.1	.757	23.9	.772	10
	Dependency	2.8	.897	20.5	.234	25.8	.750	19.8	.811	9
	Full	0.6	.969	**8.3**	.451	**12.9**	.895	**13.5**	.865	11
Finetuned frontend	W2V-ASP (Martín-Doñas et al., 2022)	0.3	.984	4.5	.646	18.6	.836	19.2	.817	317
	W2V-AASIST (Tak et al., 2022b)	**0.2**	.991	**3.6**	.707	17.5	.847	14.5	.856	317
	SLIM (ours)	**0.2**	.989	4.4	.651	**12.5**	.898	**10.7**	.892	253

Style-linguistics mismatch of deepfakes. Figure 4.8 shows the distribution of cosine distances between the style and linguistics dependency features for the real and fake classes; larger distances indicate a higher mismatch. Since the distance values approximately follow a Gaussian distribution with unequal variances, we further conduct a Welch's t-test (Ahad et al., 2014) to examine the statistical significance of the difference between real and fake samples. For all three datasets, the average cosine distance is found to be significantly lower for real speech than for deepfake samples ($p < 1\text{e}^{-5}$). This further corroborates our hypothesis that a higher style-linguistics mismatch exists for fakes. On the other hand, the distance distributions of real and fake samples still share a large overlap, indicating that dependency features alone are not sufficient for perfectly discriminating between the two classes.

Analysis of style-linguistics dependency features. Table 4.6 demonstrates that style-linguistics dependency features can provide better generalizability than the subspace embeddings (Table 4.6

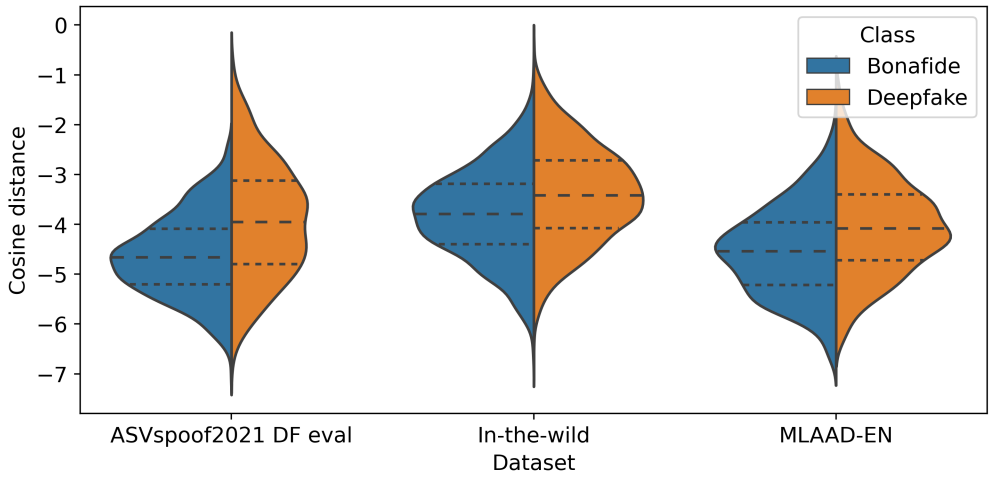


FIGURE 4.8 : Cosine distance (log scale) calculated between the style and linguistics dependency features for `ASVspoof2021 DF eval`, `In-the-wild`, and `MLAAD-EN`. Whiskers from top to bottom represent the 75% quartile, median, and 25% quartile of the distribution.

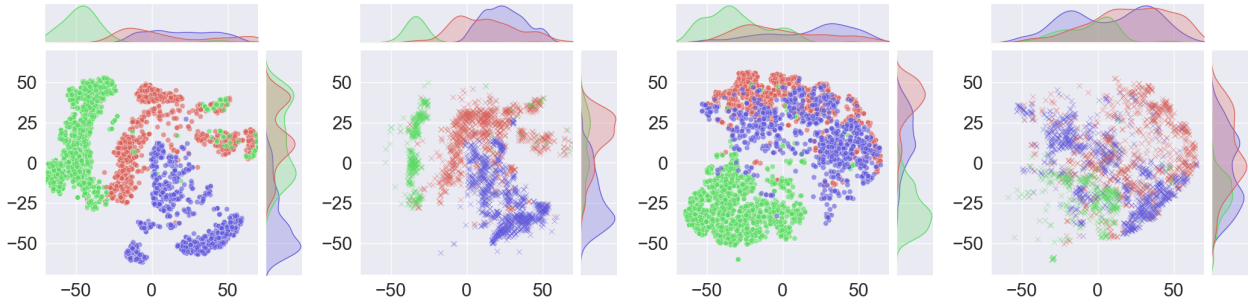


FIGURE 4.9 : Projected embeddings using t-SNE for style-linguistic representations : (a) subspace embeddings - real class, (b) subspace embeddings - fake class, (c) dependency features - real class, (d) dependency features - fake class. Data distributions are visualized on the upper and right side of the embedding plots. Red : `ASVspoof2021`; Green : `In-the-wild`; Blue : `MLAAD-EN`.

SLIM variants, rows 1–4). To examine these results, we first aggregate `ASVspoof2021`, `In-the-wild`, and `MLAAD-EN`, and project the dependency features as well as the concatenated subspace embeddings to a 2-dim space using t-SNE for visualization (Figure 6.3). Since we use frozen front-ends, the embeddings input to Stage 2 training are not affected by backpropagation. Ideal embeddings would exhibit maximal separation between the real and fake classes, while showing minimal shift within each class for different dataset distributions. In Figure 6.3, we see that the dependency features show larger discrimination between real and fakes (4.9c and 4.9d) than the concatenated subspace embeddings (4.9a and 4.9b), and also a smaller shift between datasets : fake and real samples from the same dataset (color) clusters have less overlap in distribution in the plots.

Interpretation of model decisions. Next, we perform a qualitative evaluation of the model decisions. Figure 4.10 shows the mel-spectrograms of four samples selected from `In-the-wild`. These four demonstrate typical acoustic characteristics that represent a larger group of recordings : (1)

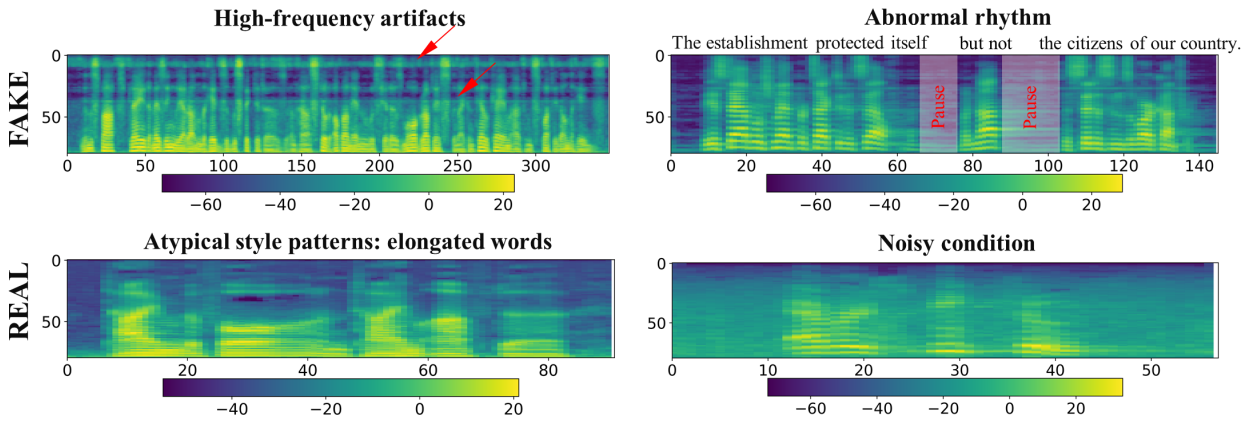


FIGURE 4.10 : Mel-spectrograms of select samples from `In-the-wild`. SLIM classifies all four correctly, and when reporting fakes, provides guidance on abnormalities in style and/or linguistics. Also, the dependency and subspace features in SLIM are complementary to each other. Left : samples missed by dependency features but correctly identified by the style and linguistic features; right : vice versa.

top-left is a *fake* sample with audible artifacts at high-frequency region; (2) top-right is a *fake* sample with unnaturally long pauses heard before and after the phrase “but not”; (3) bottom left is a *real* sample with an atypical speech style where the word pronunciations are elongated; (4) bottom right is a *real* speech recorded in a noisy condition. We find that among the top-performing systems shown in Table 4.6, only SLIM classified all four samples correctly (both frozen and fine-tuned versions; with all features), while others mostly failed on (2) and (4). Findings here suggest that SLIM provides guidance when abnormalities in style and linguistics occur. Such guidance can be complemented via *post-hoc* analysis tools such as human evaluations or saliency maps (Arrieta et al., 2020) for further interpretation.

Additionally, we note that the decisions made by dependency features and the original subspace representations are complementary to each other. Samples in the right column are correctly identified as fake by the dependency features but missed by the original subspace representations, and vice-versa (left column missed by dependency features). These results corroborate the nature of the two feature types. The dependency features are learned by modelling the general style-linguistics relationship seen in real speech, therefore samples with mismatched style-linguistics pattern are likely to be flagged as “unreal.” The original style and linguistics embeddings, on the other hand, are sensitive to signal artifacts, which could be the deepfake imperfections generated during speech synthesis (Shih et al., 2024), or the amount of background noise and device artifacts. By combining the two features, SLIM captures a variety of abnormalities and achieves improved classification.

4.6 Conclusion

In this chapter, we showed two novel deep representation learning strategies with guidance from domain knowledge. For health diagnostics, the disease biomarkers are captured by identi-

fying discriminative regions in the modulation tensorgram representation using the CRNN spectral-temporal saliency map. These regions have been previously shown related to different respiratory and articulatory abnormalities, hence explaining the better generalization performance achieved across different COVID-19 datasets. Regarding deepfake detection, we designed a self-supervised contrastive learning method that captures the style-linguistic dependency using real speech samples only. The proposed model achieves SOTA performance on unseen deepfake datasets without the need of front-end fine-tuning. The cosine distance between the disentangled style and linguistic embeddings is found to be higher for fake samples in three datasets, suggesting a mismatch of style and linguistic information in synthesized speech. Together, we show that domain knowledge guided deep representations can outperform representations learned by pure black-box models in terms of generalization to unseen data, while providing interpretable results.

5 PRIVACY-PRESERVING SPEECH APPLICATIONS VIA VOICE ANONYMIZATION

5.1 Preamble

This Chapter is compiled from materials extracted from the manuscripts published at the IEEE Transactions in Information Forensics and Security (Zhu et al., 2023c) and the 3rd Symposium on Security and Privacy in Speech Communication (Zhu et al., 2023b).

5.2 Introduction

In the previous chapter, we have shown that representations encoded by large models can significantly improve accuracy over knowledge-based features. These large models typically contain hundreds of millions to billions of parameters, which require a significant amount of computation for training, as well as inference. Take remote health diagnostics as an example, usually the model weights are not stored locally on mobile devices (Wang et al., 2018a) and speech data are sent and processed in the cloud. Decisions are then transmitted back to the user device. This transmission of speech data over the cloud, however, could pose serious threats to user privacy, since the user's voice could be linked with sensitive medical information, such as health status(Latif et al., 2020a), disease progression(Harel et al., 2004), or mental state (Low et al., 2020), just to name a few. As such, several approaches have been proposed to address the privacy leakage threat, including federated learning for decentralized model training (Li et al., 2020), model pruning and quantization for efficient local deployment and inference (Zhu et al., 2017; Choudhary et al., 2020), and voice anonymization to remove user identity information (Tomashenko et al., 2022b,a). Different from the first two methods, voice anonymization does not require changes in the model training strategy and architecture. Conventionally, voice anonymization aims at keeping speech content intact while obfuscating speaker identity (Tomashenko et al., 2022b,a). This limits its usage to ASR applications where only linguistic information is needed. For paralinguistic tasks, such as health diagnostics, deepfake detection, or emotion recognition, the feasibility of voice anonymization remains underexplored.

The main contributions of this chapter are summarized as follows :

1. We comprehensively evaluate the impact of different voice anonymization techniques on existing diagnostic models in various settings and demonstrate the limitations of existing anonymization methods;

2. We propose a new anonymization model tailored for diagnostic tasks, which minimizes the loss of health information while effectively obfuscates user identities;

3. We show that existing deepfake detection models, while performing well on healthy voice samples, are observed to be vulnerable to the generated pathological speech samples.

5.3 Related work

Anonymization techniques can be generally divided into two categories : speech transformation and speech conversion. The former refers to modifications directly to the original speech, such as pitch shifting and warping (Stylianou, 2009; Srivastava et al., 2020b), to remove personal identifiable information from the speech signal. The latter, in turn, converts one's voice to sound like that of another without changes in linguistic content (Mohammadi et al., 2017). As voice privacy concerns are on the rise, voice anonymization has gained popularity recently and, in 2020, the Voice Privacy Challenge (VPC) was created (Tomashenko et al., 2022b).

A popular method from the 2020 and 2022 VPCs employs the so-called McAdams coefficients (Tomashenko et al., 2022b,a), where shifts in the pole positions derived from linear predictive coding (LPC) analysis of speech signals (McAdams, 1984) are used to achieve anonymization. Another popular voice transformation method is termed voicemask (Qian et al., 2017), where certain frequency components are compressed (or stretched) to generate a lower-pitched (or higher-pitched) voice signal. Voice conversion systems, on the other hand, have usually relied on modifications to speaker embeddings, such as the x-vector (Snyder et al., 2018) and the ECAPA-TDNN embeddings (Desplanques et al., 2020), which are assumed to only carry nonverbal information that pertains to the speaker identity alone. The modified speaker embeddings are then input with speech content sequence to a speech synthesis module to reconstruct a new speech waveform (Fang et al., 2019). Several innovations have been proposed to the speech synthesis module to make the outcome sound more natural and of greater quality and intelligibility (Kong et al., 2020; Srivastava et al., 2020a; Meyer et al., 2022b,a).

5.4 Privacy-preserving diagnostics systems by voice anonymization

5.4.1 System overview

Figure 5.1 depicts the diagram of an anonymized speech diagnostics (SD) system. Conventionally, the original voice of user X is input to a diagnostic system that will generate a positive or negative output for the tested disease and/or symptom. If an automatic speaker verification (ASV) system was trained with data from user X, the ASV system would be able to detect user X's voice. In practice, SD systems are complex and models are often stored on the cloud, thus requiring the

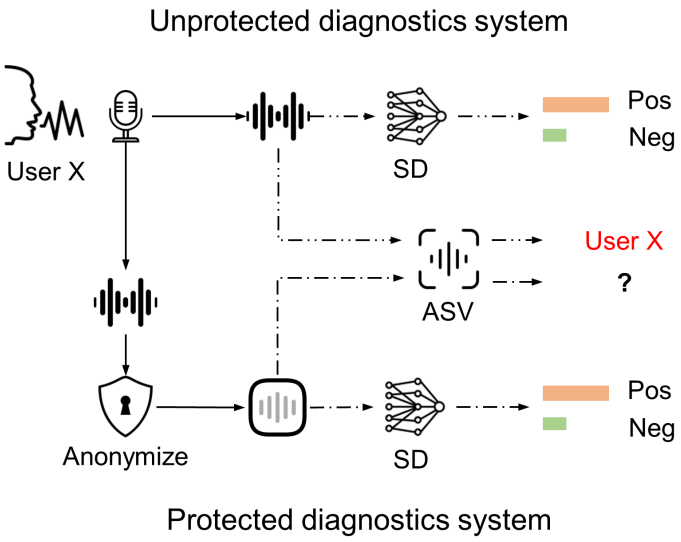


FIGURE 5.1 : Block diagram of a speech-based diagnostics system with (protected) and without (unprotected) anonymization. 'SD' stands for speech-based diagnostic system and 'ASV' for automatic speaker verification.

user's voice (or features) to be uploaded to the cloud. This transmission of data could result in privacy concerns. To overcome this, voice anonymization can be employed locally and anonymized data (or features) are sent to the cloud. In this case, user X would not be identified by the ASV system and speech-based diagnostics could proceed in a more secure and private manner.

5.4.2 Anonymization methods

McAdams coefficient This approach uses a classical signal processing technique and does not require model training. It employs the so-called McAdams coefficient method (McAdams, 1984) to shift the position of formants measured using linear predictive coding (O'Shaughnessy, 1988). For each short-time speech frame, the method first separates the linear prediction residuals and linear prediction coefficients. The LP coefficients are then converted to pole positions in the z-plane by polynomial root-finding, where each pole position represents the position of one formant. The phase of the poles with imaginary parts is then raised to the power of the McAdams coefficient α. The new set of poles is then converted back to LP coefficients. Together with the original residuals, a new speech frame can be synthesized.

Ling-GAN For voice conversion, we implemented two systems based on generative adversarial networks (GAN). The overall architecture of these systems can be found in Figure 5.2. The first system, abbreviated as 'Ling-GAN', was an off-the-shelf anonymizer from (Meyer et al., 2022a), where all modules were already trained and applied to COVID-19 data without any fine-tuning. In general, it preserves the linguistic content (i.e., phoneme sequence) and uses a generator to generate fake, yet realistic speaker embeddings to substitute for the original speaker embeddings. The original speech is first input to an automatic speech recognition (ASR) model to extract the

phone sequence. The ASR model used here is based on the hybrid CTC (Connectionist Temporal Classification)/attention architecture (Watanabe et al., 2017) with a Conformer encoder (Gulati et al., 2020) and a Transformer decoder.

It should be emphasized that the output of the ASR is a phoneme sequence, detailing not only the phonemes uttered but also the pauses. In our exploratory analysis, we found that the removal of these pauses would change the rhythm of the generated speech and lead to degraded diagnostic performance. We hence kept all pauses in the extracted phoneme sequences. The ASR model used here supports English as the default input language, hence may lead to erroneous transcriptions when other languages are used. Although such issue can be potentially tackled by replacing with other multi-language ASR models, their compatibility with the anonymization and synthesizer blocks has not been tested. Hence, we remain using the same architecture as is proposed in (Meyer et al., 2022a), and leave the language compatibility for future investigation.

The anonymization is divided into two stages. During the first stage, the 512-dimensional x-vector (Snyder et al., 2018) and the 192-dimensional ECAPA-TDNN vector (Desplanques et al., 2020) are extracted using the SpeechBrain toolkit (Ravanelli et al., 2021a) and concatenated as the final speaker embeddings. At the second stage, a Wasserstein GAN with Quadratic Transport Cost (WGAN-QC) (Liu et al., 2019) is used to generate a pool of 5,000 'converted' speaker embeddings and saved for later use. When a new recording is input to the system, the model iteratively looks through the pool, and stops when it finds one with a cosine distance above 0.3 with the original speaker embeddings. This set of new embeddings are then used to substitute the original one for synthesis. The 0.3 threshold value of cosine distance was suggested from (Meyer et al., 2022a), which ensures sufficient difference in speaker traits while maintaining the naturalness. Finally, the FastSpeech 2 model (Ren et al., 2020) is used to synthesize the phone sequence into a spectrogram, followed by a HiFiGAN vocoder (Kong et al., 2020) to convert the spectrogram into a final speech waveform. The synthesizer is conditioned on the anonymized speaker embedding, hence keeping the linguistic content while obfuscating the speaker identity.

It is important to emphasize that this off-the-shelf GAN has not seen pathological speech data during its training (Meyer et al., 2022a). As a consequence, the generated speaker embeddings may not encapsulate health-related attributes, thus affecting diagnostic accuracy. The last anonymization system used overcomes this limitation, as detailed next.

Ling-Pros-GAN The second GAN-based system, abbreviated as 'Ling-Pros-GAN', was modified from (Meyer et al., 2023) which can be seen as a more advanced version of the Ling-GAN. While sharing similar architecture, such as the ASR module and the synthesizer, the Ling-Pros-GAN further preserves prosody (i.e., pitch, energy, and duration) during anonymization and uses the style embeddings from (Wang et al., 2018b) to represent speaker attributes. In addition, we fine-tuned the generator and discriminator using the aggregated training set data from all three COVID-

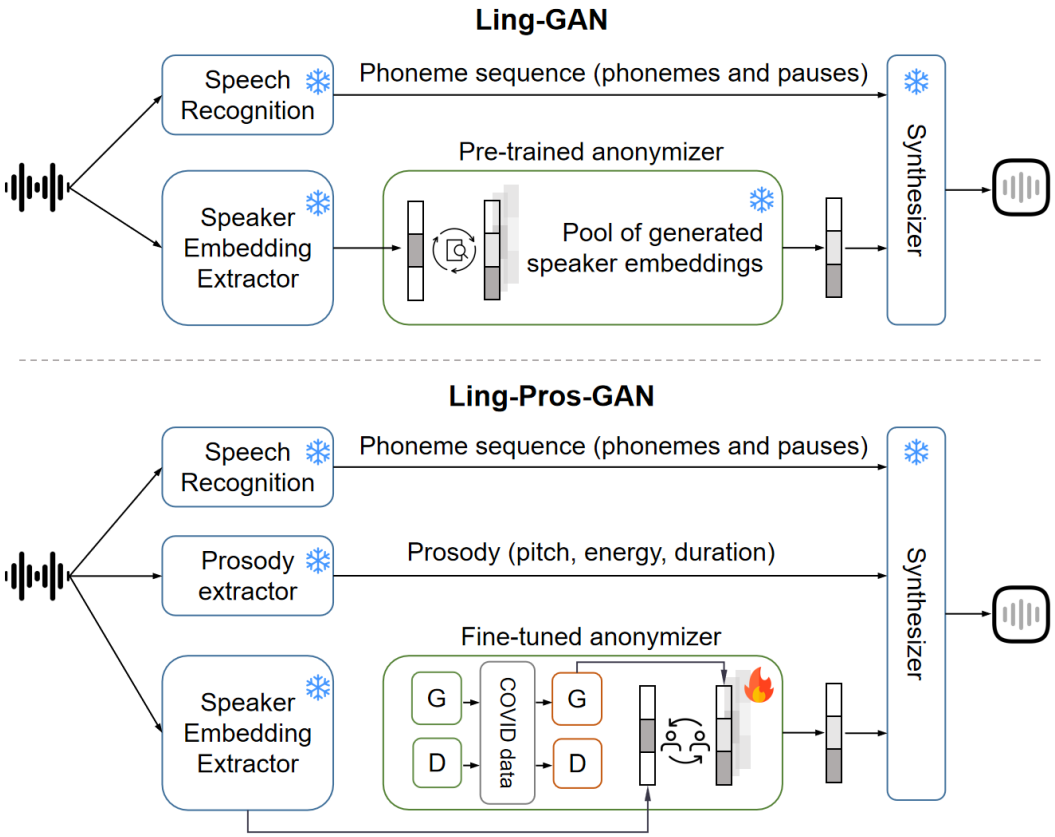


FIGURE 5.2 : Diagram of the two GAN-based anonymizers implemented in this study. Compared to the Ling-GAN, the Ling-Pros-GAN not only preserves the original prosody, but also has the generator and discriminator fine-tuned with COVID-19 speech data, enabling it to generate more COVID-like speaker embeddings.

19 datasets employed in this study, namely CSS, DiCOVA2, and the Cambridge speech set. The goal of fine-tuning was to enable the GAN to generate COVID-aware speaker embeddings.

The generator and discriminator were jointly trained via 2,000 iterations, with the batch size of 128 and learning rate of .00005. Other fine-tuning hyperparameters remained the same as reported in (Meyer et al., 2023), which can also be found in our code repository. Figure 5.3 depicts the t-distributed stochastic neighbor embedding (t-SNE) plots (Van der Maaten et al., 2008) showing a 2-dimensional representation of the speaker embeddings in the COVID-19 datasets (red dots), those produced by the generator without fine-tuning (blue), and after fine-tuning (green). As can be seen, using just the pre-trained generator is not sufficient to model the COVID-19 speaker embedding distribution. With 2,000 iterations of fine-tuning, the generator was able to generate embeddings following a similar distribution of the COVID-19 embeddings.

Different from the original implementation in (Meyer et al., 2023), where a pre-generated pool of speaker embeddings were used, we modified Ling-Pros-GAN in a way that it randomly generates a small set of different speaker embeddings each time it receives a new recording, then chooses which embeddings to swap by iteratively examining the cosine similarity. In other words, Ling-Pros-GAN is guaranteed to generate an unseen version of anonymized speech even with the exact same input recording. In contrast, since Ling-GAN always chooses embeddings from a pre-generated

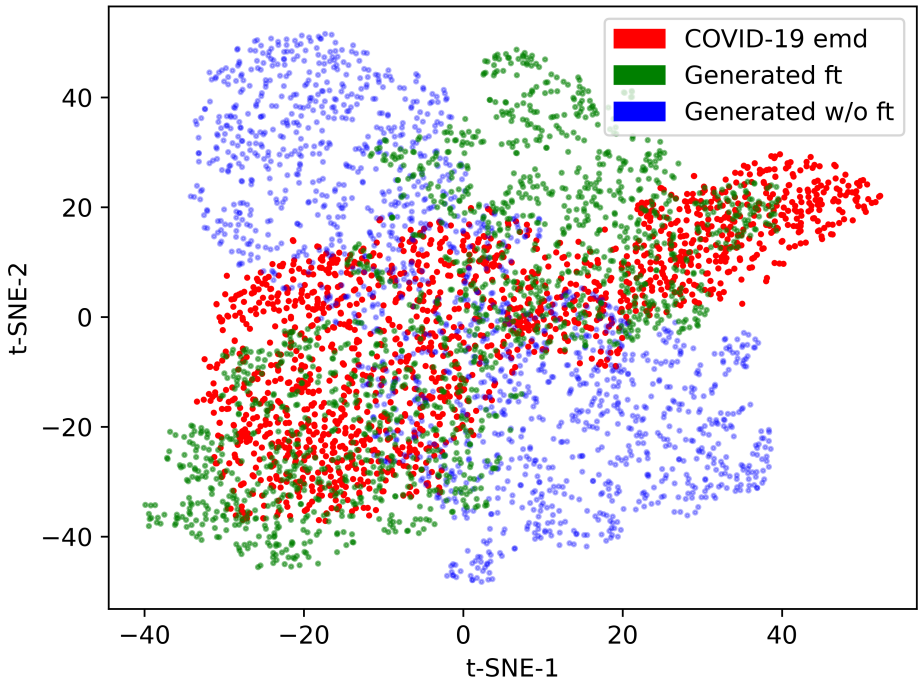


FIGURE 5.3 : Distribution of speaker embeddings ('emd') generated by Ling-Pros-GAN with and without fine-tuning ('ft'). Embeddings are projected to the 2-dimensional space using t-SNE.

pool, there is a slight chance that two recordings may be anonymized with the same generated embeddings. Such possibility becomes higher when the number of speakers increases. While such modification to Ling-Pros-GAN improves privacy, the computing time increases simultaneously due to the online generation process of speaker embeddings.

5.4.3 Diagnostic models

Based on previous experiments on COVID-19 detection (Chapters 3 and 4), the five top-performing diagnostics systems are explored herein :

openSMILE+SVM A total of 6,373 static acoustics features were firstly extracted using the openSMILE toolbox (Eyben et al., 2010), which were then input to a SVM classifier with a linear kernel. This system was used as the benchmark in the 2021 ComParE COVID-19 Speech Sub-challenge (Coppock et al., 2022a).

openSMILE+PCA+SVM The high dimensionality of the openSMILE features can be problematic for smaller datasets. In (Xia et al., 2021), principal component analysis (PCA) (Wold et al., 1987) was used to compress the 6,000+ features into 300 components. Here, the number of principal components was treated as a hyper-parameter and a value of 100 was found to strike a good balance in accuracy and dimensionality.

MSR+SVM The MSR features have been used in Chapter 3 and shown to outperform openSMILE-based systems and to provide improved generalizability across datasets. It is shown to capture the abnormalities in respiration and articulation by focusing on long-term dynamics of speech. Each modulation spectrum comprises 23 frequency bins and 8 modulation frequency bins, which is then flattened into a vector and used as input to a linear SVM classifier.

MSR+PCA+SVM For more direct comparisons with the openSMILE system, here we also explore the compression of the 184-dimensional (23×8) vector via PCA, resulting in a final 100-dimensional vector for classification.

LogmelSpec+BiLSTM The winning system in the DiCOVA2 Challenge was employed (Sharma et al., 2022) as a benchmark. This system adopts the conventional log-mel-spectrogram (logmelspec) with first-and second-order deltas as input, along with a BiLSTM as the classifier. More details about the network architecture can be found in (Sharma et al., 2022).

5.4.4 Evaluation of anonymization efficacy and efficiency

As is shown in Figure 5.4, for each speech recording, the speaker embeddings were extracted separately from the original version, the McAdams-anonymized version, the Ling-GAN anonymized version, and the Ling-Pros-GAN anonymized version. Cosine similarity was then computed between the embeddings of each two signals, where higher cosine similarity values represented higher resemblance between two speech samples. Meanwhile, we employed the pre-trained ECAPA-TDDN speaker verification model from SpeechBrain (Ravanelli et al., 2021a) to detect if two recordings are from the same speaker, then evaluated the misclassification rate, where higher values suggest more successful anonymization. Since multiple evaluation scenarios were considered in this study, where training and test data were processed with different anonymization methods, the cosine similarity and the misclassification rate were computed between not only the clean and anonymized data, but also data processed by different anonymization methods. Additionally, we measured the computation time spent by the three methods per recording, and calculated the average and standard deviation for each dataset. This helps to quantify and compare the time efficiency of the three anonymization methods.

The average cosine similarity scores between the speaker embeddings of speech files anonymized by the different methods together with the misclassification rates are shown in Figure 5.5. As can be seen, near perfect anonymization performance was achieved with both GAN-based methods (misclassification rates), with almost no similarity with either the original speech or the speech anonymized by other methods. On the other hand, nearly half of the McAdams-anonymized samples can be successfully detected, suggesting some speaker-unique information still remained.

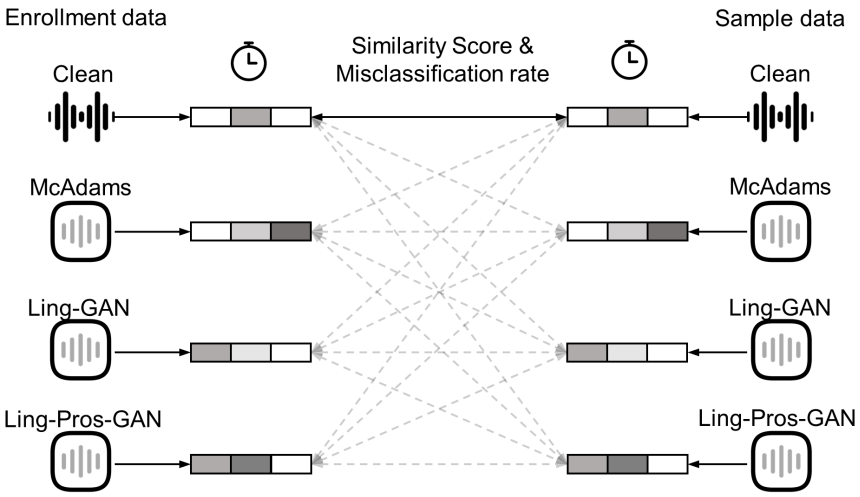


FIGURE 5.4 : Evaluation of the effectiveness of different voice anonymization methods, as well as their computational complexity.

TABLEAU 5.1 : Average computation time per speech file (second) with standard deviations using different anonymization methods for the three datasets.

Method	CSS	DiC	Cam	Ave
McAdams Coef	0.87±0.10	1.15±0.13	0.88±0.92	0.97
Ling-GAN	8.52±2.93	10.22±3.56	9.58±2.70	9.44
Ling-Pros-GAN	26.49±20.53	24.9±11.61	19.47±11.61	23.62

The computational complexity of the three anonymization methods is presented in Table 5.1 for all three datasets. While the GAN-based methods are shown to provide better anonymization effectiveness, it requires computational times approximately 10-20 times longer than using the McAdams coefficient method. The longest time was seen with Ling-Pros-GAN, since it requires extra time to extract prosody, which involves an online training loop, and to generate and find embeddings in real-time. As model loading time was not taken into account, the computation footprint of the GAN-based methods could be larger in real-world settings. Additionally, the GAN-based methods rely on several pre-trained neural networks with millions of parameters (e.g., 22.3 million for ECAPA-TDNN embedding extractor; 10 million for the generator), which could make it challenging to be deployed on mobile devices.

5.4.5 Evaluation of diagnostic performance

Evaluation scenarios : Training and test data could be anonymized using different methods. To mimic a realistic setting, we explore four different scenarios, as detailed below. Table 5.2 summarizes these conditions.

Scenario-A : Unprotected : Here, both training and test data are original, thus anonymization is not performed. This encompasses the traditional diagnostic system evaluation and serves as a baseline of the maximum diagnostics accuracy that can be achieved by each model.

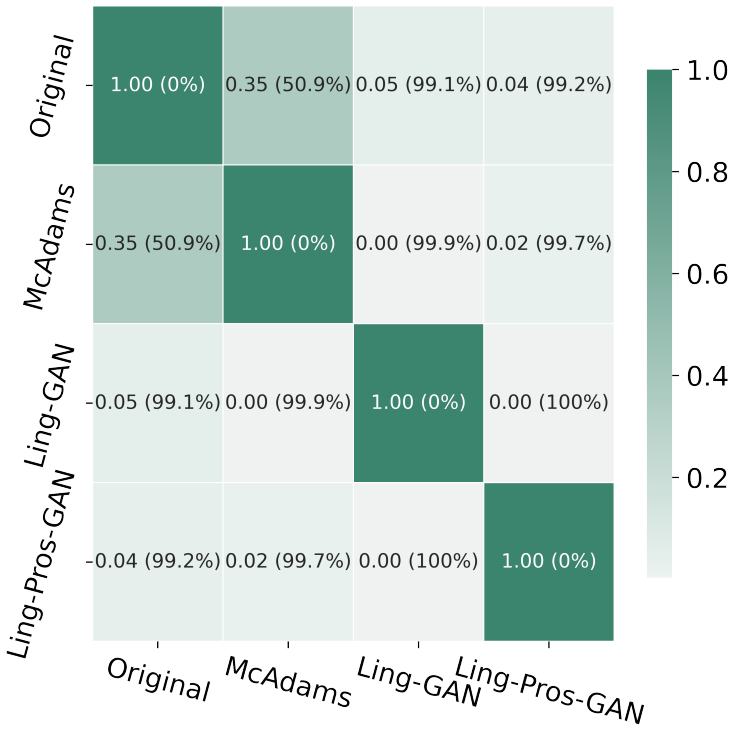


FIGURE 5.5 : Cosine similarity between speech signals under different anonymization conditions averaged across three datasets. Values in the parentheses are the corresponding misclassification rates.

Scenario-B : (Anonymization) Ignorant : In this scenario, the training data are original, and only the test data are anonymized. This scenario can be further separated into three cases : test data are anonymized using the McAdams coefficient (scenario **B1**), the Ling-GAN (scenario **B2**), and the Ling-Pros-GAN (scenario **B3**). This scenario exemplifies the case where new anonymization methods are proposed and tested against legacy original diagnostic systems.

Scenario-C : Semi-informed : In this scenario, anonymized data are seen during training, but from a method different from that used for testing. Six combinations were possible out of the three systems, namely : training set comprised of McAdams coefficient anonymization and test set with Ling-GAN (**C1**), training set with McAdams anonymizer and test set with Ling-Pros-GAN (**C2**), training set with Ling-GAN and test set with McAdams anonymizer (**C3**), training set with Ling-GAN and test with Ling-Pros-GAN (**C4**), training set with Ling-Pros-GAN and test with McAdams anonymizer (**C5**), and training set with Ling-Pros-GAN and test with Ling-GAN (**C6**). This scenario exemplifies the case where new anonymization methods are proposed and tested against legacy or different anonymized systems.

Scenario-D : Fully-informed : In this setting, training and test data are both anonymized using the same method and parameters, with three cases : both are anonymized with the McAdams coefficient (**D1**) method, both with Ling-GAN (**D2**), and both with Ling-Pros-GAN (**D3**).

Within-dataset evaluation The within-dataset performance of the five diagnostics systems under different anonymization scenarios is demonstrated in Figure 5.6. As can be seen from the

TABLEAU 5.2 : Training/test set details for the different conditions and scenarios explored.

Scenarios	Sub-condition	Training anonym.	Test anonym.
Unprotected	A	Clean	Clean
Ignorant	B1	Clean	McAdams Coefs
	B2	Clean	Ling-GAN
	B3	Clean	Ling-Pros-GAN
Semi-informed	C1	McAdams Coef	Ling-GAN
	C2	McAdams Coef	Ling-Pros-GAN
	C3	Ling-GAN	McAdams Coefs
	C4	Ling-GAN	Ling-Pros-GAN
	C5	Ling-Pros-GAN	McAdams Coefs
	C6	Ling-Pros-GAN	Ling-GAN
Fully-informed	D1	McAdams Coefs	McAdams Coefs
	D2	Ling-GAN	Ling-GAN
	D3	Ling-Pros-GAN	Ling-Pros-GAN

average AUC-ROC scores per scenario, the highest performance is achieved under scenario A, i.e., when anonymization is not performed. When the test data are anonymized using the McAdams coefficient (scenario B1), the average AUC-ROC score over all systems dropped by 8.9% (CSS), 5.9% (DiCOVA2), and 6.3% (Cambridge) relative to scenario A. A substantial decrease was observed when using both the Ling-GAN and Ling-Pros-GAN anonymizers (scenario B2 and B3), where an average relative drop 22.5% and 18.1% was achieved respectively. Moreover, nearly all systems degraded to chance levels under scenario C where models were trained with data anonymized by one method and tested with data anonymized with another, suggesting that anonymization may drastically remove COVID-19 speech information. Diagnostic performance in the fully-informed scenarios is shown to be close to scenario A. Among the three anonymizers, McAdams anonymization leads to higher diagnostic performance on average in scenario D. Compared to the Ling-GAN, Ling-Pros-GAN shows higher performance on the English datasets (DiCOVA2 and Cambridge) and lower performance on the multilingual one (CSS).

Next, we evaluate the sensitivity of different diagnostics systems to anonymization and explore the relative drop in accuracy from scenario A to scenario B. Table 5.3 reports the average drops seen per dataset. As can be seen, the two GAN-based methods resulted in a substantially higher degradation relative to the McAdams coefficient method, with the Ling-GAN leading to the most severe decrease. This was expected and corroborates Task-1 results, where speaker embeddings of the GAN-anonymized speech showed practically no similarity to the original speech. Meanwhile, since Ling-Pros-GAN leaves the prosody intact and generates more COVID-like embeddings, it is likely to preserve more COVID-19 attributes than the Ling-GAN, thus rendering higher anonymized diagnostic performance. Previous studies have shown that speaker embeddings (e.g., x-vector) also contain other nonverbal information and can be used for speech para-linguistic tasks (Raj et al., 2019; van Son et al., 2021), such as speech emotion recognition (Pappagari et al., 2020b)

TABLEAU 5.3 : Drop in within-dataset AUC-ROC (%) from scenario A to scenario B for different anonymization methods.

Anonymization method	CSS	DiC	Cam	Ave
McAdams coefficient	8.9	5.9	6.3	7.0
Ling-GAN	27.3	30.5	9.8	22.5
Ling-Pros-GAN	25.2	20.4	8.7	18.1

TABLEAU 5.4 : Drop in within-dataset AUC-ROC (%) from scenario A to the average of all sub-conditions under scenario C for different diagnostics systems.

Diagnostics system	CSS	DiC	Cam	Ave
openSMILE+PCA-SVM	31.7	26.8	6.7	21.7
MSR+PCA-SVM	27.7	32.4	16.7	25.6
LogMelSpec+BiLSTM	18.8	34.0	12.1	21.6

and disease detection (Moro-Velazquez et al., 2020; Pappagari et al., 2020a). While the GAN-based anonymizers substitute the original speaker embedding with a dissimilar speaker embedding, the obtained results suggest that health-related vocal characteristics are likely also discarded, thus resulting in significant drops in diagnostics accuracy.

Lastly, we use scenario A as the baseline and calculate the average drop in accuracy for scenario C, showing the impact that training models completely on anonymized data would have. For both openSMILE and MSR methods, we use the PCA-SVM pipeline to avoid the effects of difference in the number of features. The comparative results are reported in Table 5.4. As can be seen, all three diagnostic systems show degradaded performance, with the logmelspec+BiLSTM system shown to be on average more robust (21.6%) to the semi-informed anonymization scenario. Notwithstanding, it should be highlighted that the logmelspec+BiLSTM system achieved the lowest AUC-ROC in scenario A. Interestingly, with the CSS dataset, the diagnostic system based on a BiLSTM and log-mel spectrogram input resulted in substantially lower degradation percentage compared to the two other systems based on traditional engineered features and classifiers. CSS is a multilingual dataset, thus hand-crafted features (e.g., syllabic rate, speech production features) used in these models may show more sensitivity to language.

Cross-dataset evaluation Figure 5.7 shows the cross-dataset performance under the thirteen different testing scenarios. In line with previous studies (Zhu et al., ress; Coppock et al., 2021), all five diagnostics systems demonstrated significantly lower performance relative to within-dataset results; the logmelspec+BiLSTM achieved the greatest drop in performance. Interestingly, in a few scenarios anonymization helped systems become more generalizable relative to the unprotected setting (e.g., scenarios B2 and C3 for the CSS-DiCOVA2 cross-database experiment). Figure 5.8 depicts the average change in accuracy relative to scenario A for all scenarios and diagnostic systems. While on average a 6.6% drop in accuracy was seen across all five systems, an increase of 2% and 5% was achieved with MSR+SVM and logmelspec+BiLSTM systems for scenarios C4

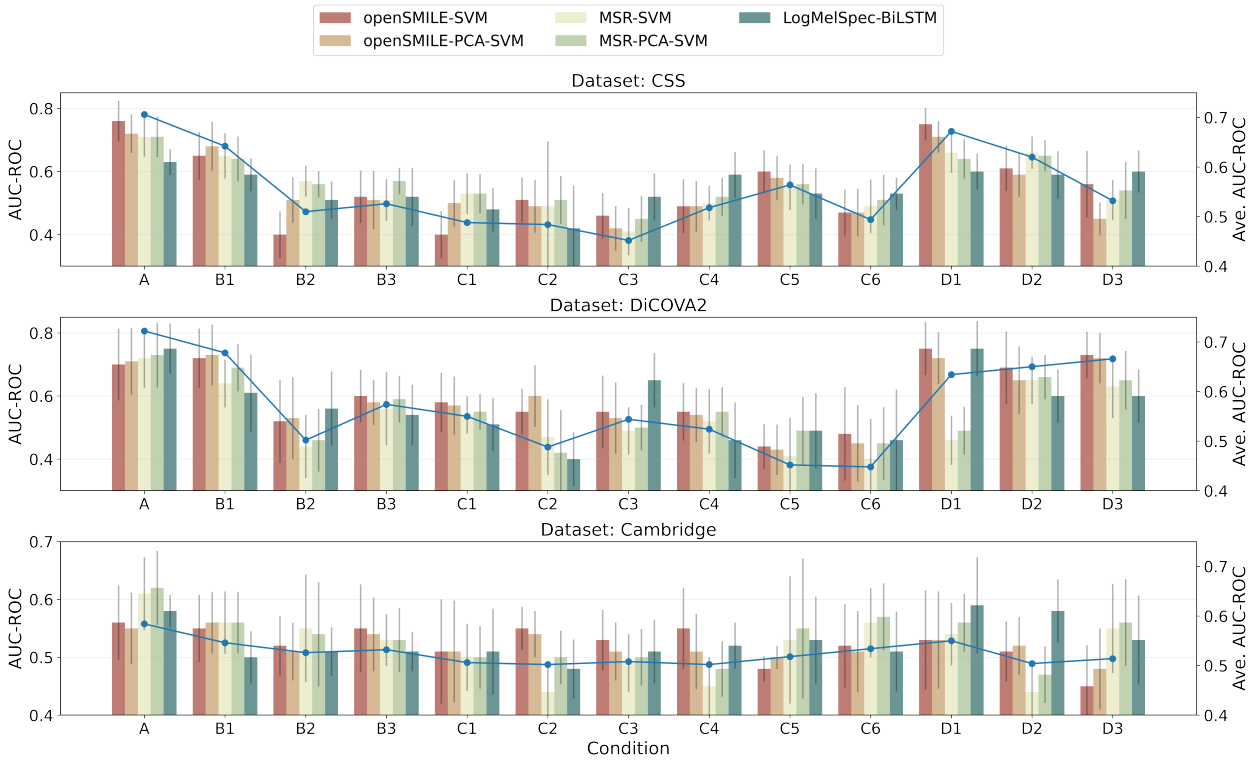


FIGURE 5.6 : Within-dataset performance under different anonymization scenarios. Error bars represent the 95% CIs. The line plot values correspond to the average AUC-ROC scores over the five diagnostic systems calculated per scenario.

and C2, respectively. It is important to note that both scenarios involved GAN-based anonymized test data, thus had typically the lower cross-dataset results to start off with.

5.4.6 Discussion on the impact of anonymization on diagnostics

Reasons behind the degradation caused by anonymization While our study shows that typical anonymization systems lead to degraded diagnostic performance, it is unclear why different systems caused different levels of degradation and why some diagnostic models could still perform decently after anonymization. To answer these questions, we performed a comprehensive evaluation of the impact of different speech aspects on diagnostic performance, including the linguistic content, speaker representation, and prosody. Similar to the experimental setup of Task-1, we now compare the within-dataset performance obtained by three categories of speech features, namely (1) the phoneme-level features, including the number of mispronunciations (as opposed to the speech script), number of pauses, and number of phonemes uttered per second; (2) the speaker representation extracted by concatenating the pre-trained x-vector and ECAPA-TDNN embeddings (Desplanques et al., 2020); and (3) prosodic features, such as the low-level descriptors of the F0 contour.

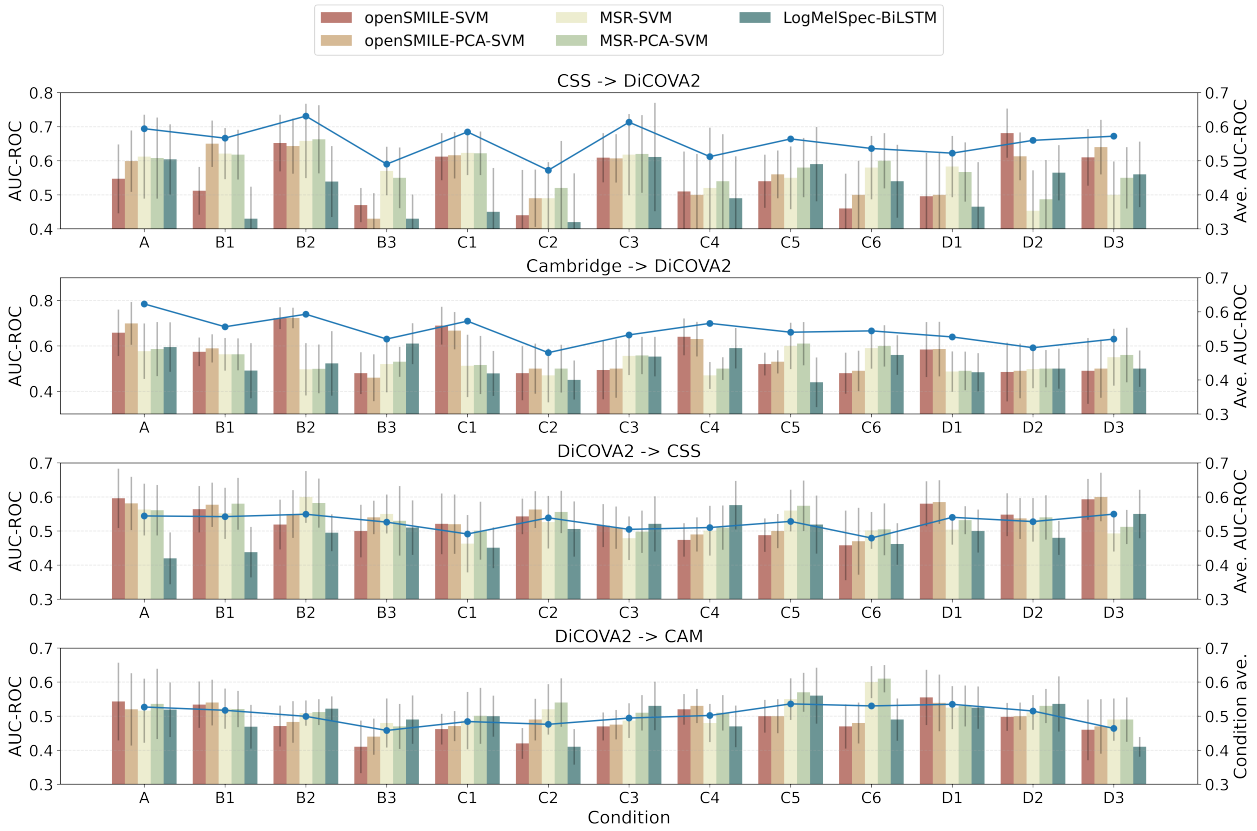


FIGURE 5.7 : Cross-dataset performance under different anonymization scenarios. Error bars represent the 95% CIs. The line plot values correspond to the average AUC-ROC scores over the five diagnostic systems calculated per scenario.

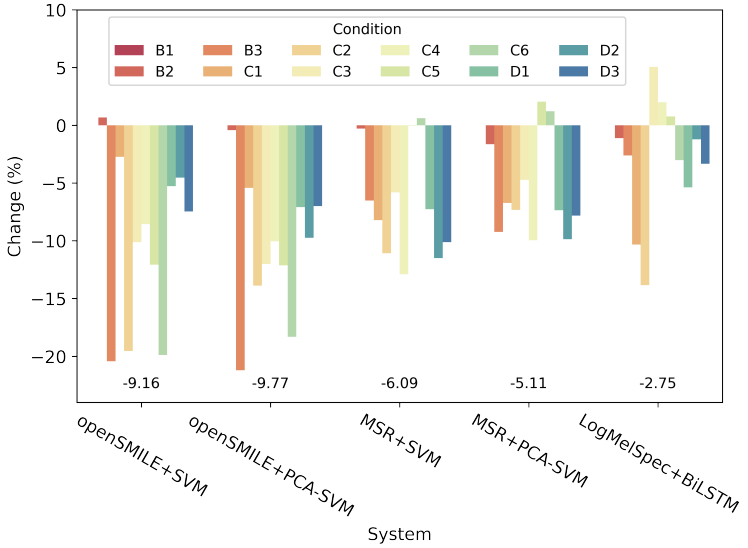


FIGURE 5.8 : Relative changes in the AUC-ROC under different anonymization scenarios for all diagnostics systems in the cross-dataset experiment.

TABLEAU 5.5 : Diagnostic performance achieved by different categories of speech features

Feature	AUC-ROC		
	CSS	DiCOVA2	Cambridge
Linguistic	.561	.632	.555
Speaker	**.739**	**.697**	**.571**
Prosodic	.541	.564	.520

A Linear Discriminant Analysis (LDA) classifier is applied on top of each of the feature sets for classification. The results achieved by these features are reported in Table 5.5. Among the three feature sets, speaker embeddings appear to be the most crucial features for all datasets, corroborating with Task-1 results where the GANs suffered the most severe degradation, where the original speaker embeddings were entirely substituted. Such finding also suggests that speaker-unique attributes and health-related information are highly entangled in the speaker embeddings. Considering that existing anonymization systems rely heavily on these off-the-shelf speaker embeddings, it remains challenging to preserve the health information while altering only the speaker identifier.

While a group of studies reported prosody as a key biomarker to characterize speech disorders, such as dysarthria (Vyas et al., 2016; Kadi et al., 2013; Ramos et al., 2020), our results show that phoneme-level linguistic features outperform prosodic features for COVID-19 detection. Specifically, we found the number of pauses and number of mispronunciations to be the most important phoneme-level features, with COVID-positive samples demonstrating more mispronunciations and fewer pauses. While the correlation between phoneme-level features and COVID-19 status has not been systematically studied, similar features have been examined for other diseases affecting speech production. For example, (Darling-White et al., 2020) shows that individuals with Parkinson's disease produced fewer pauses at syntactic boundaries; the statistics of pauses have been shown crucial for diagnosing neuromuscular disorders, such as dysarthria (Noffs et al., 2018). Since GAN-based systems left linguistic content intact during anonymization, these findings help explain why the diagnostic models could perform above chance-level even when only the phoneme sequences were preserved during anonymization.

Visualizing speech processed by different anonymizers To better understand the impact of different anonymization methods on speech characteristics, we first visualize the waveform of the speech processed by the three anonymizers (see Fig. 5.9) for a direct comparison. As can be seen, those processed by the McAdams anonymizer and Ling-Pros-GAN share higher similarities in the waveform envelope shape with the original signal compared to the one generated by Ling-GAN. The difference seen in the plot is in line with the architecture design of different anonymizers. Among the three, Ling-GAN loses prosody and most of the speaker attributes, hence is expected to cause the highest amount of changes in the anonymized speech. The Ling-Pros-GAN and

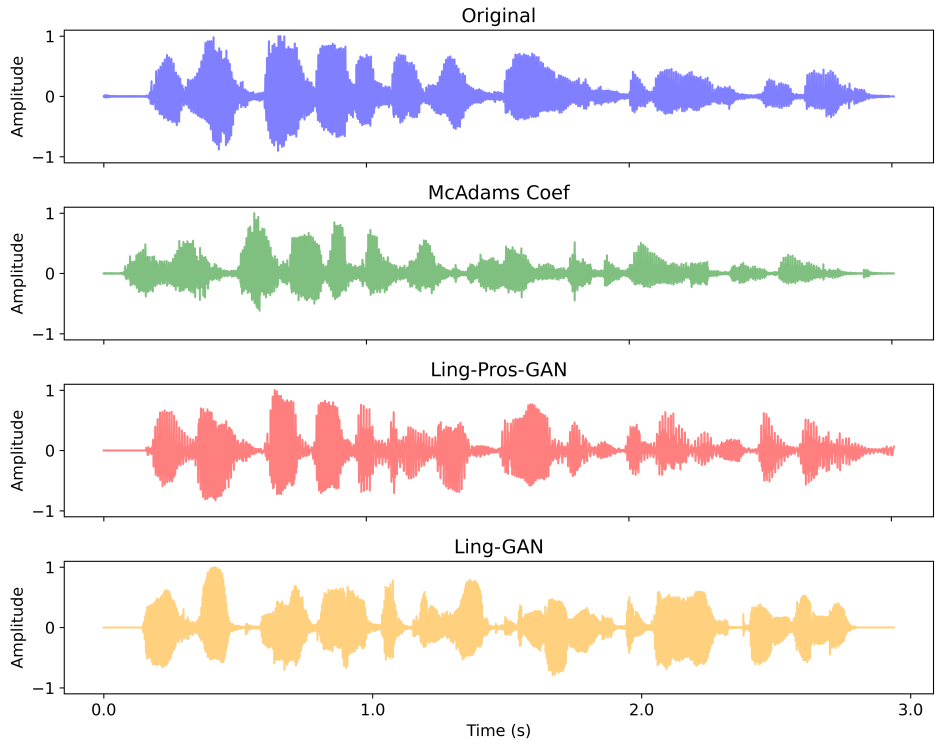


FIGURE 5.9 : A comparison of the waveforms processed by the three anonymizers and the original speech.

McAdams anonymizer, in turn, leave the speech rhythm untouched (i.e., duration and energy of phonemes), hence leading to higher resemblance in the waveform envelope.

Next, t-SNE plots are used to visualize the distribution of the speech features in two dimensions. Figure 5.10 shows the clusters of speech anonymized with different methods (computed from the training and validation data) and for the three features modalities explored herein : openSMILE (subplots a), MSR (b), and logmelspec (c). As can be seen, for all three feature sets, the distribution of clean speech (blue) is closer to that of the McAdams anonymized speech (orange) and Ling-Pros-GAN anonymized speech (red), while the Ling-GAN anonymized speech (green) shows the least similarity with the other two, corroborating findings from Tasks 1 and 2.

Moreover, it can be seen from Figure 5.10a and Figure 5.10b that the clusters computed from openSMILE and MSR features show little overlap, while clusters of the logmelspec features show great overlap (Figure 5.10c). Together with Task-2 results, this shift in the feature space is likely the main cause of the higher decrease observed in the openSMILE and MSR systems under different anonymization settings. Meanwhile, since all anonymization methods keep the speech content intact and change only the nonverbal attributes, a greater shift in feature space may indicate a stronger correlation with the para-linguistic aspect and less with the linguistic aspect. This echoes with previous studies which showed that openSMILE and MSR features are preferred over logmelspec features in characterizing emotional and unnatural speech (Falk et al., 2012; Eyben, 2015).

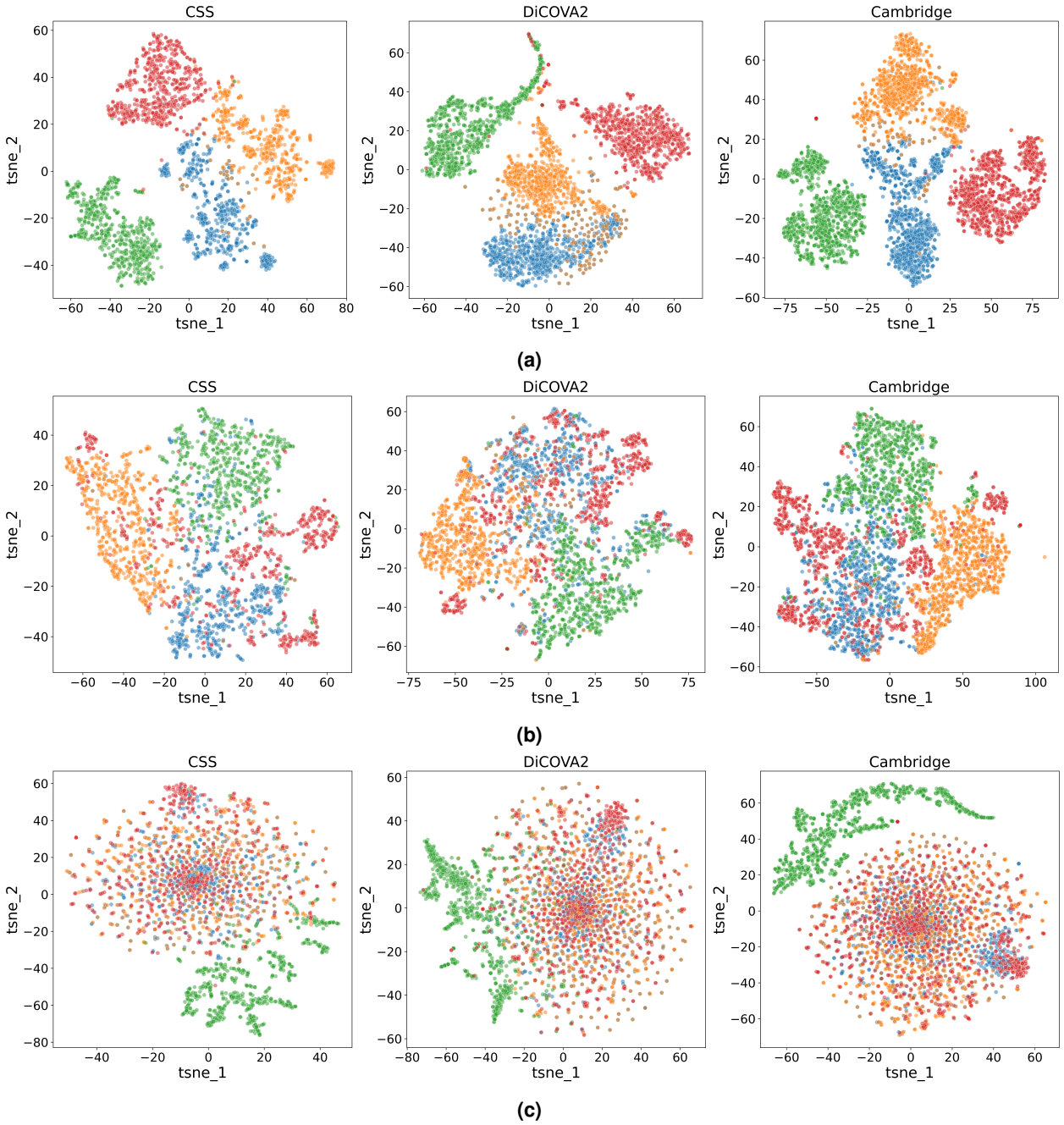


FIGURE 5.10 : t-SNE clusters of anonymized speech features for different feature sets, namely : (a) openSMILE, (b) MSR, and (c) logmelspec. Blue dots corresponds to original speech; orange to McAdams coefficient anonymized speech; red to Ling-Pros-GAN anonymized speech; and green to Ling-GAN anonymized speech.

Exacerbated biases in anonymized diagnostic models Beside a general degradation seen in overall performance, we further noticed that existing biases in diagnostic models can be exacerbated when anonymization is applied. For example, while openSMILE+SVM achieved similar performance with the original test set in the unanonymized and McAdams-anonymized conditions, the difference between unweighted average recall obtained with high-SNR and low-SNR sets exceeds 10% in the latter. Similar pattern is also seen with MSR+SVM and logmelspec+BiLSTM

when comparing the unanonymized and GAN-anonymized conditions. These findings suggest that although overall diagnostics performance can be maintained after anonymization (or even be improved in a few cases), the discrepancy between different subgroups may increase, thus further accentuating the impact of biases.

Previous studies on automatic speaker verification demonstrated that due to underrepresented subgroups of data (e.g., female) in the speaker datasets, some subgroups are not as well represented by the extracted speaker embeddings as others (Hutiri et al., 2022). Since our implemented GAN anonymizer integrated a pre-trained ASV model as the speaker embedding extractor, a bias from the upstream speaker dataset may likely have been inherited and transferred to the anonymized diagnostics system. This may explain the increased gap between subgroups after anonymization. In fact, such accumulation of bias from pre-trained models is also shown in other domains. For example, bias against individuals with disabilities has been previously reported with pre-trained large language models (Venkit et al., 2022). Such issue should be handled more carefully when building voice anonymization algorithms for healthcare applications, as it may lead to misdiagnosis.

5.5 Detection of synthesized pathological voice

5.5.1 Introduction

In the previous section, we have demonstrated the use of two TTS models for voice anonymization. Since the goal was to obfuscate the user identity, original speaker embeddings were swapped randomly with candidate embeddings from a pre-generated pool under a constraint on cosine similarity (as depicted by Fig.5.2). After fine-tuning the GAN on pathological speech data, we observed that the speaker embeddings also entail health information, which benefits the downstream diagnostics system. From a generative perspective, the anonymized speech can be regarded as a special type of 'deepfake' where health information is preserved while speaker identity is modified. When the target user's speaker embeddings are obtained, this method can be used to convert target user's voice to a different health state (e.g., healthy to pathological). While deepfake detection can be a potential countermeasure, existing models are trained with healthy speech, while the performance with pathological speech and deepfakes remains unknown. In this section, we explore whether existing deepfake detection models can accurately detect these 'pathological deepfakes' that are generated using the anonymizers proposed in the previous section.

5.5.2 Method

The same anonymizers employed in the previous section are used as deepfake generators, namely the McAdams method, Ling-GAN, and Ling-Pros-GAN. This resulted in three types of deep-

TABLEAU 5.6 : Detection performance obtained with pathological speech datasets. 1 : Original+McAdams; 2 : Original+Ling-GAN; 3 : Original+Ling-Pros-GAN.

Model	CSS-1	CSS-2	CSS-3	DiCO-1	DiCO-2	DiCO-3	Cam-1	Cam-2	Cam-3
Baseline detector	20.7	2.9	15.4	23.2	0.1	24.4	22.3	2.8	13.3
SLIM	25.5	2.0	12.1	30.1	0.2	23.2	24.2	3.0	18.4

fakes, where the speaker identity is changed while in most cases the health information remains intact. We then fed these generated (anonymized) speech samples into two detection models : one is the SLIM model introduced in Chapter 4 and the other is a WavLM-based detector which was benchmarked against SLIM. Both detection models were trained with ASVspoof5 data and had obtained decent test performance on the ASVspoof5 evaluation set ($\leq$ 10% EER). We performed a zero-shot evaluation of these two detectors on three COVID-19 datasets (i.e., CSS, DiCOVA2, and Cambridge), where both models were not fine-tuned on the pathological datasets. Similar to the evaluation metrics used in Chapter 4, we used EER to gauge model robustness, where higher EER reflects higher vulnerability. Original speech samples from the three pathological speech datasets are labeled as real speech, and the anonymized ones are labeled as fake.

5.5.3 Results

The detection performance is summarized in Table 5.6. The McAdams-anonymized speech samples are shown to be the most challenging ones, with the highest EER obtained for all three datasets. This is likely because the conversion pipeline of the McAdams method is drastically different from modern TTS and VC systems, where no deep models were needed. As a result, the anonymized speech would not entail the artifacts seen from vocoders and other neural network modules. Although the McAdams-anonymized samples can fool detection models, it is also difficult to utilize this method to generate an attack target at specific users, as the conversion process relies mainly on formant shifting while the other vocal attributes are not well modeled.

On the other side, while the GAN-anonymized samples were expected to be accurately detected, as they are based on a widely-used TTS model, a large gap is seen between the Ling-GAN and the Ling-Pros-GAN, where the former can be easily detected while the latter shows to be more challenging. It should be emphasized that the Ling-Pros-GAN was fine-tuned on COVID-19 data and was shown to preserve more health information compared to the Ling-GAN. Therefore, the gap observed here likely indicates that the preservation of health and prosody attributes reduced the discrimination between original and anonymized speech. While the anonymized samples are not crafted to mimic specific speakers, our findings suggest that health-preserving anonymizer-generated speech can lead to a significant drop in detection performance.

5.6 Conclusion

In this chapter, we comprehensively evaluated the impact of three voice anonymization methods on the accuracy of five leading COVID-19 detection systems as well as the anonymization efficacy. All anonymization methods showed to degrade diagnostics accuracy, where the most severe degradation was seen with the systems that directly altered speaker embeddings. Our findings suggest that existing methods lack the capability of effectively preserving diagnostic information while obfuscating speaker identifiers. Additionally, we explored the use of anonymized external data as a data augmentation tool and promising results were obtained. Finally, we showed that some biases can be exacerbated after anonymization and lead to a biased diagnostics model. When these anonymizers are used as pathological deepfake generators, we observed a significant drop in detection performance, especially when the health and prosody information were preserved. This suggests that existing deepfake detection models are not yet robust to this unique type of deepfake. While health is overlooked in existing generative models, future research should take the pathological deepfakes into consideration to further improve model generalizability.

6 A TASK-AGNOSTIC, EXPLAINABLE, AND PRIVACY-PRESERVING MODEL

6.1 Preamble

This Chapter is compiled from material extracted from the manuscripts published at the IEEE Journal in Biomedical and Health Informatics (Zhu et al., 2024a) and the International Conference on Acoustics, Speech and Signal Processing Workshop 2023 (Zhu et al., 2023d).

6.2 Introduction

In previous chapters, we showcased a variety of methods to improve the generalizability, interpretability, and privacy-preserving ability of speech applications. While demonstrating promising results, each of these techniques tackles only one aspect of the problem. For example, models proposed in Chapter 4 aim at improving generalization and do not take user privacy into consideration, therefore voice anonymization methods introduced in Chapter 5 are needed to bridge the gap. However, this would complicate the overall system design and require careful tuning of each component to avoid performance degradation. Similar issues exist in many real-world speech applications, where aggregation of multiple task-specific models is needed to satisfy different needs.

One way to tackle this issue is to investigate models that can be task-agnostic. As such, speech universal representations, such as Wav2vec2 (Baevski et al., 2020), WavLM (Chen et al., 2022), and HuBERT (Hsu et al., 2021), have been proposed and gained attention due to their generalizability to different downstream tasks (Yang et al., 2021a). However, they have been mostly evaluated on conventional speech tasks, such as speech recognition, speaker identification, and emotion recognition. With the emerging interests in health diagnostics and deepfake detection, it is not clear whether the universal representations are optimal for these tasks. Furthermore, given that these representations are built to entail both linguistic and paralinguistic information, it is difficult to pinpoint the acoustic attributes used for each specific task. Finally, since universal representations are known to carry user identity information, their usage in privacy-sensitive applications can be limited. In this section, we explored the possibility of having a model architecture that can be applied to a variety of speech tasks (i.e., task-agnostic), and can learn an interpretable and speaker-disentangled representation.

The main contributions of this chapter are summarized as follows :

1. We propose a novel speech classification model architecture, termed WavRx, that achieves SOTA performance on a variety of disease diagnostic tasks and synthesized speech detection tasks;

2. The representation learned by WavRx contains significantly less speaker identity information compared to other widely used representations (e.g., Wav2vec2, WavLM, HuBERT), which makes it suitable for privacy-preserving applications;

3. The representation learned by WavRx is interpretable. We showcase how it can be used to pinpoint biomarkers in pathological speech, as well as vocoder-related patterns in synthesized speech.

6.3 Motivation

Speech is generated by airflow from the lungs, which results in voiced and unvoiced speech given the different states of the vocal folds. The vibration is then transmitted through the vocal tract and modulated by the articulatory movement and respiration, generating the speech hearable by humans (Casserly et al., 2010; Ohala, 1990). Typical speech analysis has been relying on short time frames. For example, the window size for the short-time Fourier transform (STFT) is usually 8 to $32\,\mathrm{ms}$ (Rabiner et al., 2007). Speech modulation, in turn, changes at a much lower rate due to the limit of human physiology. For articulatory movement, for instance, between 2 and 10 syllables are being uttered per second for most of the recorded languages (Hilton et al., 2011; Pellegrino et al., 2011; Hermansky, 1998). However, such underlying modulation is not well isolated in the spectrogram (Hermansky, 1998), and descriptors such as delta and double-delta cepstral parameters have been used for decades as measures of velocity and acceleration of changes in the cepstral parameters over somewhat larger window durations.

To address this issue, several researchers have relied on the modulation spectrum (e.g., (Greenberg et al., 1997; Falk et al., 2010a)), which applies a second STFT to each frequency component obtained from the spectrogram. This extends the conventional spectrogram to a 3-dimensional space with an added modulation frequency axis. With a window size of over $128\,\mathrm{ms}$, the modulation spectrum analyzes the long-term dynamics of human speech. While most of the linguistic content is lost in the modulation frequency domain, other vocal characteristics such as speaking rate (Hermansky, 1998), vocal hoarseness (Markaki et al., 2011), and whispering (Sarria-Paja et al., 2013) may be better manifested. Features derived from such representation have been previously applied in the detection of dysarthric speech (Falk et al., 2012), whispered speech (Sarria-Paja et al., 2013), voice pathologies (Markaki et al., 2011), COVID-19 (Zhu et al., 2023f), and emotional speech (Wu et al., 2011), to name a few. Motivated by the idea of the modulation spectrum, we here applied the modulation transformation to the universal representations to better capture utterance-level patterns.

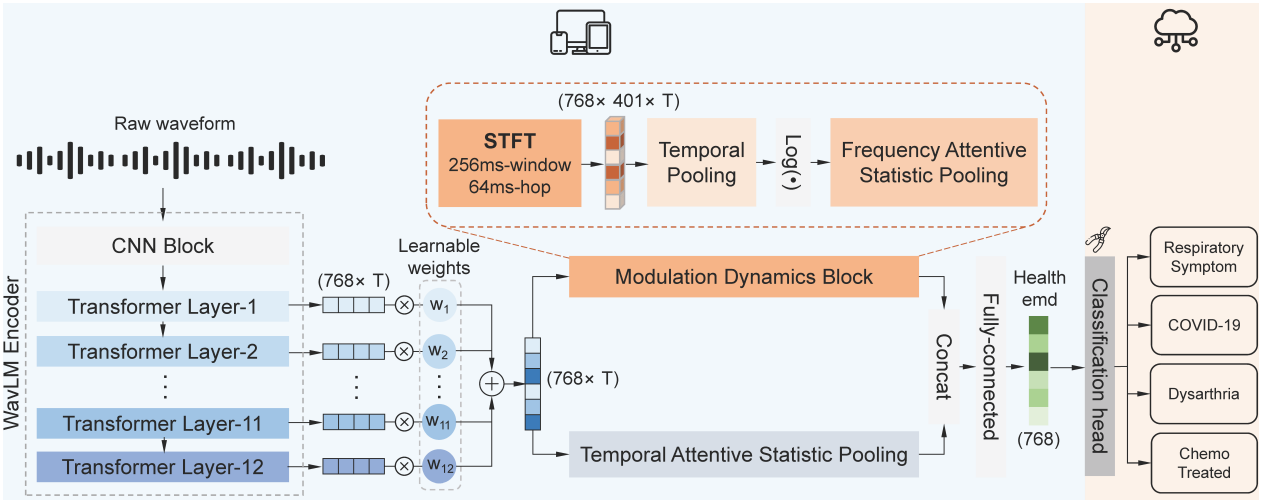


FIGURE 6.1 : Architecture of the proposed WavRx model. The raw waveform firstly passes through a pretrained WavLM encoder that generates a *temporal* representation. The *temporal* representation is then fed into the modulation dynamics block to extract the long-term *dynamics* caused by respiration and articulation, which are then fused with the *temporal* information to obtain the health embeddings. The embedding extraction is performed locally to avoid identity leakage. Only downstream classifier parameters are updated in the cloud.

6.4 WavRx

WavRx comprises three main components : (1) a pre-trained encoder to extract temporal representations from the raw waveform; (2) the modulation dynamics block to capture long-term dynamics of the encoded temporal representations; and (3) attentive statistic pooling and output layers to fuse representations from the previous two blocks and generate a final decision. Details about each component are described in the following subsections. It should be noted that while the general model architecture remains the same for diagnostics and deepfake detection tasks, a slightly different classification backend is used for the latter. Details can be found in the following descriptions. For simplicity, we used the same name WavRx for both tasks in the following text.

The model architecture is depicted in Figure 6.1. Considering the privacy requirement in real-world applications, WavRx encoder can be deployed locally to extract intermediate embeddings, which are then uploaded to a central cloud server for decision-making. In later sections, we show that the WavRx embeddings entail minimal speaker identity information, hence preventing the leakage of user identity.

6.4.1 Temporal representation encoder

The proposed model builds on top of the pre-trained SSL mdoel as the temporal representation encoder. For both tasks, we found WavLM (Chen et al., 2022) to be a better base encoder than the others (i.e., Wav2vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021)). WavLM takes a raw speech waveform as input and firstly feeds it into a CNN block comprised of 7 temporal CNN

layers with 512 channels, cascaded by layer normalization and a GELU activation layer. Each time step in the output from the CNN block represents $25\,\mathrm{ms}$ of audio with $20\,\mathrm{ms}$ hop length. The CNN output is then sent into a transformer backbone, which comprises 13 layers with 768-dimensional hidden states. We employed the WavLM Base+ version which was pre-trained on 60k hours of Libri-light (Kahn et al., 2020), 10k hours of Gigaspeech (Chen et al., 2021a), and 24k hours of VoxPopuli (Wang et al., 2021a).

Previous studies have shown that later transformer layers in WavLM carry more linguistic content, while early layers are likely to encode paralinguistic information (Pasad et al., 2023). For diagnostics and deepfake detection, it remains unclear which layers are more crucial. Hence, we aggregated outputs from all 12 layers (with the first input layer excluded) by assigning weights to each of them. These weights were learned through supervised training on downstream tasks. The layer-weighted output from the WavLM encoder is a time by feature representation $\{\mathbf{T} \times \mathbf{F}\}$, which can be seen as a temporal representation showing how each feature changes over time. Given the temporal pooling configurations of the CNN layers, the resultant temporal representation has a temporal resolution of $20\,\mathrm{ms}$. However, speech production is modulated at a lower rate and the temporal representation may carry redundant linguistic information that is less essential for disease diagnosis. Thus, we proposed the modulation dynamics block to provide complementary information that is missing from the temporal representation.

6.4.2 Modulation dynamics block

A visual demonstration of the modulation dynamics block is provided in Fig. 6.2. Given an output $\mathbf{T}(\mathbf{m}, \mathbf{n})$ from WavLM (i.e., the weight sum of twelve transformer layer outputs), where $\mathbf{m}$ represents the number of time windows and $\mathbf{n}$ represents the number of features, we applied a short-time Fourier transform (STFT) to each feature channel, leading to a 3-dimensional modulation dynamics representation $\mathbf{D_n}(\mathbf{j}, \mathbf{f_j})$:

$$\mathbf{D_n}(\mathbf{j}, \mathbf{f_j}) = |\mathcal{STFT}(\mathbf{T}(\mathbf{m}, \mathbf{n}))|^2 , \tag{6.1}$$

where $\mathbf{j}$ refers to the number of time frames used for STFT and $\mathbf{f_j}$ to the number of modulation frequency channels. The results of the STFT include both real and imaginary parts; here, we take the absolute value operation (denoted by $|\cdot|$) and calculate the power.

For phoneme-level speech applications (e.g., speech recognition), the STFT usually relies on short time windows (e.g.,16-$32\,\mathrm{ms}$) (Rabiner et al., 2007), enabling the temporal resolution high enough to discriminate transitory events. The articulatory movement and respiration, on the other hand, are relatively steady and change at a much lower rate than the vibration of vocal cords. Therefore, we extended the window length to $\geq 128\,\mathrm{ms}$ with a hop length $\geq 32\,\mathrm{ms}$ to capture the dynamics at a wider range. To search for the optimal performance, we experimented with window

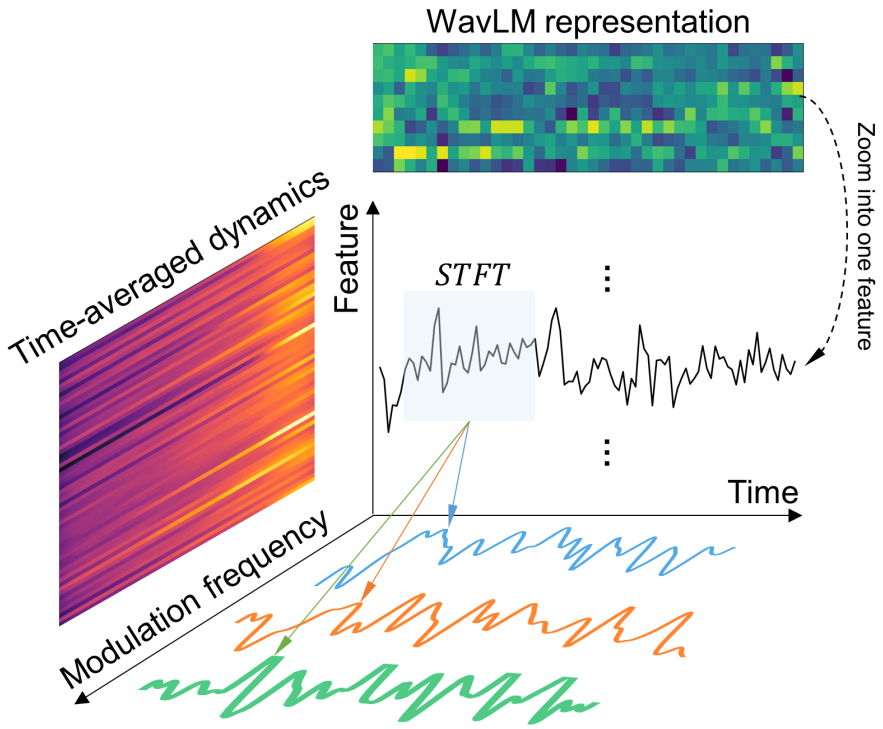


FIGURE 6.2 : The modulation dynamics block takes the weighted sum of hidden states from the WavLM transformer backbone and applies STFT to each feature channel. This operation decomposes the $\{\mathbf{Time} \times \mathbf{Feature}\}$ representation into $\{\mathbf{Time} \times \mathbf{Frequency} \times \mathbf{Feature}\}$, which models the long-term temporal modulation of each feature sequence.

length values from $128\,\mathrm{ms}$ to $1\,\mathrm{s}$ with 25% hop length, and found the best to be around $256\,\mathrm{ms}$. The resultant dynamics has three different axes, namely feature, time, and modulation frequency, where each slice along the time axis carries the decomposed modulation frequency values for all features.

6.4.3 Downstream components

Similar to speaker embeddings, we assume that the WavRx embeddings correspond to utterance-level characteristics. Thus, both temporal and dynamics representations require a temporal pooling operation to obtain the time-invariant embeddings. We compared average pooling and attentive statistic pooling (ASP) and found the former to be better suited for deepfake detection while the latter is better for diagnostics. The average pooling can be simply seen as an average of features over the time axis. The original ASP aims to integrate the frame-level attention when calculating mean and standard deviation as follows :

$$\mu = \sum_{t}^{T} \alpha_t h_t, \tag{6.2}$$

TABLEAU 6.1 : Employed pathological speech datasets.

Pathology	Dataset	Lang	#hours	#spk	#utt	ave_dur (s)	Data split			
							Pos/Neg‡	Train	Valid	Test
Resp symptom	CS-Res	EN	31.3	6,623	9,456	11.93±4.66	1.05	6,648	1,914	894
	CS-Res-L	EN	123.1	24,134	37,140	11.94±4.97	0.78	22,308	7,969	3,863
COVID-19	DiCOVA2	EN	3.93	975	975	14.33±4.15	0.20	617	154	193
Dysarthria	TORGO	EN	8.1	15	9,417	3.09±2.13	0.51	4,564	1,753	3,100
	Nemours	EN	1.5	12	1,628	3.35± 2.79	0.38	1,184	148	296
Chemo treated	NCSC	NL	1.4	55	1647	3.13± 1.74	1.27	701	200	746

‡ Ratio of number of positive samples to number of negative samples.

$$\sigma = \sqrt{\sum_{t}^{T} \alpha_t h_t \odot h_t - \mu \odot \mu}, \tag{6.3}$$

where α_t represents the weight assigned to the tth time frame.

The ASP can be used directly on temporal features to flatten them into a 1-dimensional vector. With modulation dynamics, we first computed the average along the time axis, which leads to the shape $\{\mathbf{Freq} \times \mathbf{F}\}$, where $\mathbf{Freq}$ stands for frequency and $\mathbf{F}$ for features. We then applied attention to different frequency channels, and calculated the attentive mean and standard deviation.

The temporal and dynamics vectors were firstly concatenated then fed into a fully-connected (FC) layer to map into a 768-dimension vector, which was used as the WavRx embeddings. A dropout layer and a LeakyReLU with the negative slope of 0.1 were appended after. The second FC layer maps the WavRx embeddings to a single value as the final decision. For diagnostics tasks, we further applied pruning on top of the last FC layer, where the percentage of neurons to be pruned was set as a hyperparameter.

6.5 Experimental setup

6.5.1 Diagnostics

Datasets To diversify the types of speech pathologies to be tested, we used six publicly available datasets covering four different speech-related abnormalities; a summary of these six datasets can be found in Table 6.1. The data collection protocol of these datasets can be found in Chapter 2.4.1.

Training and evaluation details For training efficiency, we limited all input recordings to be within $10\,\mathrm{s}$ by cutting off the over-length part. For those with left and right channels, we took the average to obtain a single-channel audio. All recordings were re-sampled to $16\,\mathrm{kHz}$ and the amplitude was normalized between -1 and 1. Since the STFT operation in the modulation block requires

a minimum of $1\,\mathrm{s}$-signal, short audios were zero-padded to $1\,\mathrm{s}$. The aforementioned pre-processing was achieved using the Torchaudio library (Yang et al., 2021b).

Regarding data augmentation, we injected two types of environmental corruptions in each training batch, namely noise (SNR between 0 to 15 dB) and reverberation (reverberation scale factor between 1 and 1.5), and concatenated the augmented samples with the original samples. Furthermore, we added speed perturbations by slightly speeding up (105%) and down (95%) the signal. These approaches were implemented via the SpeechBrain toolkit (Ravanelli et al., 2021b).

We used the same hyperparameters for training WavRx on all six datasets, changing only the data augmentation and pruning parameters. Data augmentation was only used when trained on DiCOVA2 and TORGO; the optimal pruning percentage was set to 90% for DiCOVA2 and NCSC, and 0% for the others. With the baseline models, we employed the same data augmentation methods used to train WavRx, and tuned the hyperparameters separately for each one of them.

We used AUC-ROC and F1 score as the evaluation metrics. With both metrics, we calculated for each class and took the unweighted mean (i.e., *macro*). Furthermore, we found that a model could perform decently on the test set but poorly on the validation set (or vice versa). As such, we report F1 scores achieved with both test and validation sets, where the difference between these two can indicate the model robustness.

Experiments were conducted on the Compute Canada platform (Baldwin, 2012) with four NVIDIA V100-SXM2 ($32\,\mathrm{GB}$ RAM per GPU). The training time with WavRx was approximately 3-4 hours for CS-Res and CS-Res-L, and less than 2 hours for the other datasets.

6.5.2 Deepfake detection

Datasets We use ASVspoof2019 and 2021 for evaluating model performance, and use WaveFake as an out-of-domain dataset to interpret the modulation dynamics representation. Details of these datasets can be found in Chapter 2.4.2.

Training and evaluation details With the baseline systems, we followed the same pipelines as described in the ASVspoof challenge code repositories with the model hyperparameters unchanged. Interested readers are encouraged to refer to (Todisco et al., 2019) and (Yamagishi et al., 2021) for more details. For a fair comparison between different variants of the universal representation based systems, we adopted the exact same training strategies as well as model hyperparameters.

Similar to the pre-processing steps performed for diagnostics, we limited all speech to be within $8\,\mathrm{s}$ for computation efficiency, and normalized the amplitude between -1 and 1. Since samples in each batch need to be padded to the same length, we firstly rearranged the whole dataset by duration, so that samples in each batch would share similar lengths. We then performed zero-

padding at a batch-level to align with the maximal duration within each batch. By doing so, we could minimize the length of zero-padding while maintaining high training efficiency.

The universal representation encoders were frozen during training, only the learnable weights assigned to the encoder transformer layers and downstream classification layers were updated. This resulted in a total of fewer than $5\,\mathrm{million}$ parameters for all models. The binary cross entropy (BCE) loss was used, where the weights assigned to genuine samples were set to 10 to tackle the class imbalance. For computational efficiency, the batch size was set to 1 and the maximal training epochs were set to 20. The learning rate was decreased linearly from .0001 to .00005. To avoid over-fitting, the training process was stopped when the EER values did not decrease for three consecutive epochs. The FC layer dropout value was set to 0.25. Model training and inference were conducted on the Compute Canada cluster (Baldwin, 2012) with four V100l GPUs. Average computing time was around $8\,\mathrm{h}$ per model.

Similar to the two ASVspoof challenges (Todisco et al., 2019; Yamagishi et al., 2021), the EER was used as the main metric for model evaluation, which is the rate where the false acceptance rate is equal to the false rejection rate. The lower the EER value is, the more accurate the detection system is. Meanwhile, we also report the F1 score, which is a commonly used metric for unbalanced binary classification tasks. To measure if two compared models performed significantly differently in EERs, we employed the method from (Bengio et al., 2004), which conducts a pair-wise statistical analysis.

6.6 Experiment results

6.6.1 Disease diagnostics

In-domain diagnostic This task aims to compare the proposed WavRx to the other baseline models in an in-domain setting. Models were trained and evaluated within each of the six datasets (i.e., the same disease was seen during training and testing). An ablation study is also conducted to demonstrate the effects of different model components of WavRx. In-domain diagnostics usually indicates the highest performance that can be achieved by each model in an ideal setting, where training and evaluation data share the same distribution. As shown in Table 6.2, the proposed WavRx obtains the highest test F1 scores in 4 out of 6 datasets, along with the highest average F1 score of 0.744 (combining test and validation) among all models. With the three datasets that were released with official baseline systems (i.e., CS-Res, DiCOVA2, and NCSC), WavRx markedly outperforms the baselines. When using only the modulation dynamics branch for detection, while the overall performance is not competitive as other benchmarks, it is shown to be the top-performer in the Nemours dysarthria detection task. Together, these results suggest that the dynamics of universal representations is crucial for disease detection.

TABLEAU 6.2 : Comparison of model performance on six speech diagnostics datasets. Note that only CS-Res, DiCOVA2, and NCSC had official baselines. For all three metrics, higher values suggest better performance. Highlighted values represent the best performing model (s) for the metric.

Model	Respiratory abnormality						COVID-19			Dysarthria						Cancer			Ave[4]
	CS-Res			CS-Res-L			DiCOVA2			TORGO			Nemours			NCSC			
	ROC[1]	$F1_t$[2]	$F1_v$[3]	ROC	$F1_t$	$F1_v$	ROC	$F1_t$	$F1_v$	ROC	$F1_t$	$F1_v$	ROC	$F1_t$	$F1_v$	ROC	$F1_t$	$F1_v$	
WavRx	**.815**	**.730**	**.725**	**.694**	**.655**	**.624**	**.878**	**.600**	.524	.918	.767	**.756**	.939	.959	.946	**.774**	**.737**	.910	**.744**
$WavRx_{mod}$	.740	.645	.650	.620	.571	.504	.801	.550	.466	.810	.659	.636	**.961**	**.980**	.932	.753	.716	.804	.676
$WavRx_{tem}$	.807	.721	.720	.691	.649	.594	.861	.589	.478	.918	.768	**.756**	.855	.872	.916	.735	.695	.910	.722
Wav2vec	.798	.712	.707	.682	.640	.591	.841	.576	.480	.827	.677	.624	.945	.966	**.959**	.759	.721	.905	.713
Hubert	.796	.711	.707	.689	.645	.592	.829	.568	.479	**.931**	**.782**	.687	.843	.858	.973	.705	.666	**.928**	.717
AST_{speech}	.683	.582	.595	.610	.554	.507	.738	.510	**.571**	.762	.611	.612	.727	.736	.676	.636	.598	.804	.613
AST_{audio}	.722	.625	.661	.613	.548	.584	.539	.383	.462	.770	.620	.603	.902	.919	.703	.639	.610	.822	.628
ECAPA-TDNN	.687	.571	.638	.582	.523	.555	.755	.498	.543	.636	.487	.499	.640	.636	.637	.703	.663	.712	.580
Baselines[5]	.695	.594	.620	–	–	–	.817	.561	.544	–	–	–	–	–	–	.699	.658	.710	–

[1]AUC-ROC score. [2]F1 score obtained with the test set. [3]F1 score obtained with the validation set. [4]The average of validation and test F1 scores across all datasets. [5]CS-Res baseline : MelSpec+vGGish networks. DiCOVA2 baseline : Melspec&deltas+BiLSTM. NCSC baseline : openSMILE+random forest.

When comparing different model categories, SSL models (i.e., WavRx, Wav2vec, Hubert) in general outperform those pre-trained in a supervised manner (i.e., AST, ECAPA-TDNN), though both did not include pathological speech during the pre-training. This again demonstrates the benefits of SSL pre-training when evaluated on a variety of downstream tasks. Interestingly, as the only backbone that was pre-trained not on speech data, the AST_{audio} outperforms its speech version. Since existing speech foundation models are usually trained with only speech data, potential improvement might be achieved when adding audio data to the pre-training stage, such as music and other non-speech acoustic events.

Zero-shot diagnostic When applied in real-world settings, the amount of data collected from one disease is usually quite limited, as can be seen from the size of the existing pathological speech datasets (Xia et al., 2021; Sharma et al., 2022; Menendez-Pidal et al., 1996; Schuller et al., 2017, 2012, 2021). This task investigates the model generalizability in a stringent setting, where models were trained on one dataset and made predictions on unseen datasets (i.e., test disease was not seen during training). During inference, both the health embedding encoder and classification head were fixed. This task emulates a scenario where no training data is available from the target domain (e.g., an unseen disease). Hence, it can be beneficial when a diagnostic model can generalize to unseen diseases with similar symptoms or pathological origins. As the top-performers in Task 1, we systematically tested WavRx as well as its two individual branches in a cross-dataset setting. Table 6.3 reports the AUC-ROC scores achieved for the model trained on one disease and tested across unseen diseases, as well as the average over all unseen diseases.

As can be seen, when comparing different test diseases, respiratory abnormality is shown as the pathology that is distinct from the others, which can be seen from the lowest AUC-ROC score (the bottom row in each sub-table). The two dysarthric speech datasets, on the other hand, can lead to good generalization to each other (drops in AUC-ROC of 0.1555 and 0.005 respectively with TORGO and Nemours obtained by WavRx), although the speech content and data collection protocols differ. Models trained with dysarthric speech can also benefit the detection of COVID-19,

TABLEAU 6.3 : Cross-disease zero-shot prediction performance using different representations. Values reported are AUC-ROC scores. There are a total of 25 train-test combinations (5 train sets $\times$ 5 test sets); for each train-test disease combination, the most generalizable representation is color-shaded. Scores without significant difference between the three or below chance-level are ignored. The diagonal values represent the in-domain diagnostic performance (i.e., same datasets used for training and testing).

		Test set					Ave
		Resp	COVID	Dys-1	Dys-2	Cancer	
Dyn+Tem	Resp	.815	.369	.489	.836	.554	.613
	COVID	.493	.878	.567	.693	.454	.617
	Dys-1	.504	.684	.918	.984	.690	.756
	Dys-2	.542	.659	.763	.989	.614	.713
	Cancer	.478	.504	.638	.878	.774	.654
	Ave	.566	.619	.675	.876	.617	
Dynamics	Resp	.700	.447	.652	.708	.631	.628
	COVID	.498	.827	.635	.934	.346	.648
	Dys-1	.510	.759	.821	.978	.631	.740
	Dys-2	.533	.798	.750	.998	.495	.715
	Cancer	.490	.337	.419	.391	.753	.647
	Ave	.546	.634	.655	.802	.571	
Temporal	Resp	.721	.598	.568	.756	.552	.639
	COVID	.492	.861	.580	.783	.437	.631
	Dys-1	.522	.600	.916	.993	.679	.742
	Dys-2	.550	.406	.682	.968	.563	.634
	Cancer	.495	.378	.598	.760	.746	.595
	Ave	.556	.569	.669	.852	.595	

as well as chemo-treated speech, which indicates that neuromuscular deficiency can be a shared characteristic among these three pathologies. When comparing the test performance between the three WavRx variants ('Ave' row at the bottom of the three sub-tables), better performance is achieved with four out of the five test sets when fusing temporal and dynamics representations (namely Resp, Dys-1, Dys-2, and Cancer). Together with Task 1 results, findings here suggest that integrating modulation dynamics of universal representations can help capture the disease-related biomarkers and improve the model generalizability to diseases sharing similar pathological origins.

Privacy of health embeddings This task examines if the speaker identity is concealed in the WavRx health embeddings by running an automatic speaker verification (ASV) task on top. Since ASV requires multiple recordings from each single individual, TORGO (15 speakers) and Nemours (10 speakers) were selected for this task. With each individual, 10% of the speech samples were used for training and the remaining 90% were used for testing. We first extracted the health embeddings using the pre-trained WavRx from Task 1, then applied LDA as the speaker classifier. The WavLM model fine-tuned on Voxceleb 1&2 (Nagrani et al., 2017) was used as the baseline speaker embedding encoder for comparison purposes. Given the system shown in Fig. 6.1, the

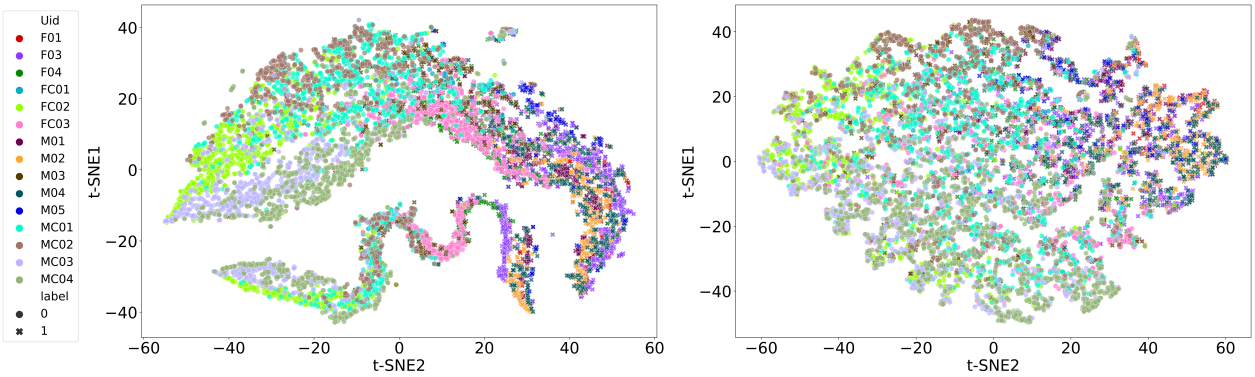


FIGURE 6.3 : Projected health embeddings learned from temporal representations (left) and dynamic representations (right). While health and pathological samples are well separated in both plots, speakers are better separated in the left plot, suggesting that speaker information is entangled with health attributes for temporal representations.

health embeddings encoded by the local model are expected to carry minimal speaker identity attributes while maximally representing the health information.

In this task, we investigate if the health embeddings encoded by the temporal representation alone carry speaker identities, and if the modulation dynamics block can help tackle this issue. The speaker verification accuracies and diagnostic AUC-ROC scores are shown side-by-side in Table 6.4. Ideally, privacy-preserving health embeddings should have a low ASV accuracy and a high diagnostic AUC-ROC score. As can be seen from rows 4 and 7, when relying on only the temporal representation, the learned health embeddings carry a higher amount of speaker identity information than the baseline speaker embeddings. This is likely because pathological speech follows a different feature distribution than the healthy speech in Voxceleb, hence leading to suboptimal performance of the pre-trained speaker embeddings. The health embeddings obtained from the temporal representation may encode both speaker identity and health attributes, therefore resulting in high ASV accuracies. The modulation dynamics representation, on the other hand, decreases the ASV accuracies by an average rate of 31.9% and 13.5% for TORGO and Nemours respectively (rows 3 and 6). When fusing the two branches together, the resultant health embeddings lead to the best diagnostic performance, while maintaining the leakage of speaker identity at a lower level than the baseline speaker embeddings (rows 2 and 5).

We further visualize the health embeddings learned from temporal and dynamics representations, which are shown, respectively, in the left and right plots in Fig. 6.3. Colors represent different speakers and marker types represent disease states. While in both plots, the positive and negative classes can be well separated, a more clear distinction between speaker clusters can be seen in the left plot (temporal) than the right one (modulation dynamics). This observation is in line with the lower speaker recognition accuracies achieved by WavRx, which provides supporting evidence of the disentanglement between speaker identity and health attributes in the WavRx dynamics representation.

TABLEAU 6.4 : Speaker verification accuracy and diagnostic AUC-ROC scores obtained by different representations. For ideal health embeddings, we expect lower speaker accuracy and higher diagnostic score.

Representation ($\text{Model}_{\text{finetuned dataset}}$)	TORGO		Nemours	
	$\text{ACC}_{\text{spk}}\downarrow$	$\text{AUC}_{\text{Diag}}\uparrow$	$\text{ACC}_{\text{spk}}\downarrow$	$\text{AUC}_{\text{Diag}}\uparrow$
$\text{WavLM}_{\text{Voxceleb}}$	.715	-	.951	-
$\text{WavRx}_{\text{TORGO}}$	.711	.918	.898	.984
$\text{WavRx-dynamics}_{\text{TORGO}}$	.602	.821	.831	.978
$\text{WavRx-temporal}_{\text{TORGO}}$	.902	.916	.990	.991
$\text{WavRx}_{\text{Nemours}}$	.609	.763	.873	.989
$\text{WavRx-dynamics}_{\text{Nemours}}$	.594	.750	.847	.998
$\text{WavRx-temporal}_{\text{Nemours}}$	.857	.682	.955	.967

Analysis/interpretability of the modulation dynamics block While the modulation dynamics branch is shown to improve the diagnostic performance and generalizability, it is crucial to investigate the characteristics of such representation to understand the reasons behind the improvements. To this end, we start by extracting the modulation dynamic representations from both positive and negative classes, then compute the Fisher's F-ratio (Fisher, 1970) between the two groups. F-ratio is a statistic measure that examines whether a difference exists between two groups, where an F-ratio value larger than 1 suggests a significant difference. Since the representation is 2-dimensional (feature by modulation frequency), the F-ratio is calculated per pixel, where the higher value suggests more discrimination between two classes. We further filtered out F-ratio values below 1 since those regions were statistically insignificant. This process was repeated for all six datasets. The F-ratio plots for all tasks can be seen in Fig. 6.4.

With the given hop length of the STFT ($64\,\text{ms}$), the maximal modulation frequency is $8.3\,\text{Hz}$ with the resolution of $0.125\,\text{Hz}$. For all six datasets, the majority of the difference is observed below $2\,\text{Hz}$, with peaks seen between 0.1 - $0.5\,\text{Hz}$, corresponding to a $2\,\text{t}$o $5\,\text{s}$-period modulation. Such slow rate of modulation aligns with our initial hypothesis that long-term dynamics of universal representations are crucial for disease detection. For example, the automatic contraction of respiratory muscles has been shown to take place once every five seconds during dialogues (Rochet-Capellan et al., 2014; Winkworth et al., 1994); an average of 15-25 breathing cycles per minute (equivalent to 2.4 to $4\,\text{s}$ per cycle) has been reported for adults and the elderly (Barrett, 2019).

Another important phenomenon noticed is the sparsity of the F-ratio plots, where only very few features among a total of 768 are shown with statistical significance. Based on this observation, we further calculated the sparsity of the 768-dimensional health embeddings learned from dynamic representations and compared with those learned from temporal representations. The sparsity values below 1% of the per-sample-maximum were thresholded to zeros. The final results are reported in Table 6.5. As can be seen, it is found that the health embeddings learned from temporal representations have an average sparsity of 35.8% across six datasets with a standard deviation of 9.1% across samples, while the average sparsity doubles to 76.7% with only 0.8% standard

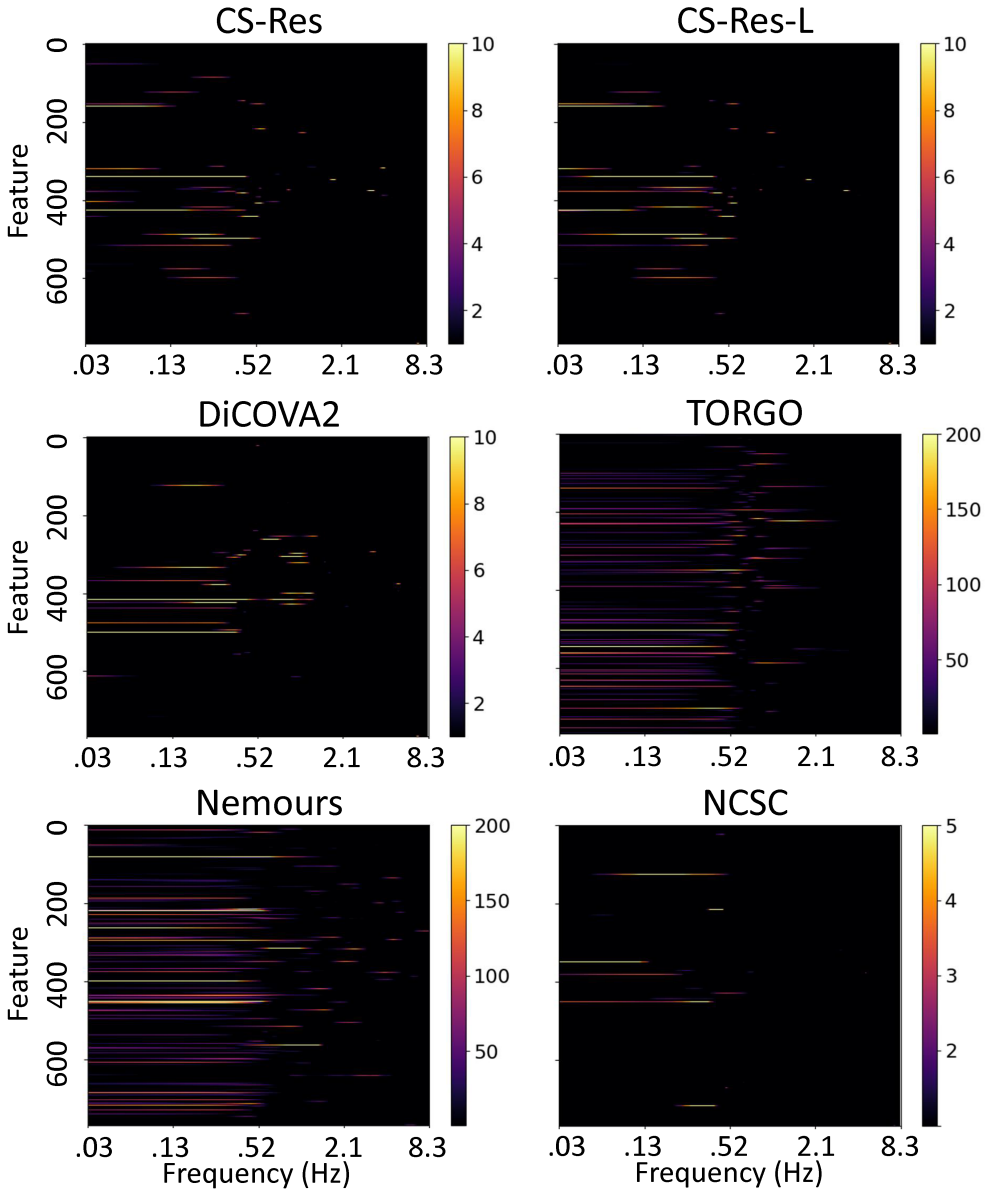


FIGURE 6.4 : F-ratio plots computed between the modulation dynamics of positive and negative samples obtained for each of the six datasets. X-axis shows the modulation frequency (in Hz) and Y-axis represents the feature dimension, which contains 768 features in total.

deviation for those learned from dynamic representations. Fusing the two together leads to an average sparsity of 64.1%.

These findings suggest that disease-related information can be encoded more efficiently by the modulation dynamics, where roughly only half of the features are required for accurately detecting a disease. This not only provides insights into the reasons behind the improved generalizability across diseases, but also helps explain the improved privacy-preserving property of the proposed WavRx model. When learning the health embeddings from the fused representations, health-irrelevant information was likely discarded, which may include speaker attributes, such as gender and age.

TABLEAU 6.5 : Sparsity of health embeddings learned for each dataset. Sparsity is calculated after thresholding the embedding values.

Embedding	Sparsity						Average
	Cam-Res	Cam-Res-L	DiCOVA2	TORGO	Nemours	NCSC	
Temporal	33.6±5.6	48.0±5.0	45.7±10.8	28.3±8.9	38.8±16.3	20.6±8.2	35.8±9.1
Dynamics	**88.5±0.7**	**94.2±0.1**	**86.6±0.9**	**65.9±1.3**	49.4±1.2	**75.4±0.8**	**76.7±0.8**
Combined	72.8±1.7	76.6±1.9	64.2±2.4	58.0±1.4	**61.2±1.9**	52.0±1.6	64.1±1.8

6.6.2 Deepfake detection

Model Performance We first compare the model performance obtained using the ASVspoof 2021 evaluation set (Table 6.6), where the majority of the test samples were generated by unseen algorithms. Among all tested systems, WavRx outperformed all systems with an EER of 9.87% and F1 of 0.403, surpassing the ASVspoof 2021 DF track top-1 result. The Wav2vec2 based systems, on the other hand, achieved only baseline-level performance. When comparing WavRx versus the baseline SSL models, a consistent statistically significant improvement is seen with both Wav2vec2 and WavLM after applying the proposed modulation transformation. Interestingly, while the proposed representation performs better with the evaluation data, the raw universal representations achieve lower EER scores on the validation set, indicating potential model over-fitting to seen attacks. Such a finding corroborates with our initial hypothesis that although universal representations encompass rich temporal details, which can be important for detecting DF speech, part of the information learned is unique to each generative algorithm and may not generalize well to unseen attacks.

We further report the performance achieved with the ASVspoof 2019 evaluation set, where 6 out of the 17 algorithms were seen during training (Table 6.7). In accordance with the ASVspoof 2021 results, WavLM is shown to be a stronger encoder than Wav2vec2. Meanwhile, it can be noticed that both encoders perform markedly better on the ASVspoof 2019 data, suggesting that detecting seen attacks is a much simpler task than detecting unseen attacks. Similar to the ASVspoof 2021 validation results, the top-performer here is shown to be the raw WavLM representation based system. To further quantify the representation generalizability, we calculated the gap between the EERs achieved with ASVspoof 2019 and ASVspoof 2021 data. A significantly lower gap is seen with WavRx compared to the raw one. Together, these findings demonstrate that while universal representations can be directly used to differentiate between genuine and DF speech, generalizability can be further improved by characterizing long-term temporal dynamics, thus leading to more trustworthy speech applications.

Visualizing Dynamic Representations Next, we visualize the proposed representation extracted from different TTS systems. Since no text ground-truth nor speaker info was provided with the utterances from ASVspoof (Yamagishi et al., 2021), we selected one exemplary utterance from

TABLEAU 6.6 : Performance achieved using ASVspoof 2021 DF track data. Statistical significance was measured between EERs of each pair of universal representations (raw and proposed) with the significantly better score highlighted in bold.

Model	Valid		Eval	
	EER (%)	F1	EER (%)	F1
LFCC+GMM	3.76	.834	25.56	.197
RawNet2	3.07	.858	22.38	.213
Top-1 result (Yamagishi et al., 2021)	-	-	15.64	-
Wav2vec2-raw	**4.35**	.819	28.00	.160
WavRx (W2V as base encoder)	9.47	.662	**27.10**	.166
WavLM-raw	.08	.996	10.89	.380
WavRx (WLM as base encoder)	.78	.963	**9.87**	.403

TABLEAU 6.7 : Performance achieved with ASVspoof 2019 LA track data. EER gap suggests the difference in EERs obtained with ASVspoof 2021 and 2019 evaluation data. Statistical significance was measured between EERs and the EER gap of each pair of universal representations.

Model	ASVspoof 2019 Eval		EER gap
	EER (%)	F1	
LFCC+GMM	25.30	.387	.26
RawNet2	7.46	.745	14.72
Wav2vec2-raw	**13.20**	.576	12.44
Wav2vec2-proposed	16.40	.513	11.60
WavLM-raw	**.72**	.966	10.22
WavLM-proposed	2.47	.891	**7.40**

the LJspeech corpus (Ito et al., 2017) and seven different deepfake versions from the WaveFake corpus. All eight utterances share the exact same speech content and correspond to the same speaker. We computed the proposed 2D temporal dynamic representations from all utterances and averaged over the feature axis to highlight the modulation spectral patterns. For better visualization, the spectral pattern of the genuine one is subtracted from all eight utterances; resultant plots are shown in Fig. 6.5.

Regions with higher spectral energies are indicated by warmer colors, suggesting more discrimination compared to the genuine utterance. Since all other vocal attributes are controlled, the difference seen here is likely caused by the generation process. A spectral peak between 7 to $12\,\mathrm{Hz}$ can be observed across all seven algorithms, while some also demonstrate peaks in the higher modulation frequency range (e.g., higher energies above $12\,\mathrm{Hz}$ for MelGAN and HiFiGAN). Such consistency in the modulation spectral pattern corroborates with the improved generalizability obtained by applying the modulation transformation. A possible explanation is that though unseen attacks are generated by models different from those used in the training set, the abnormalities in long-term temporal dynamics can be similar, leading to a consistent modulation spectral pattern that can be used for speech deepfake detection for unseen attacks.

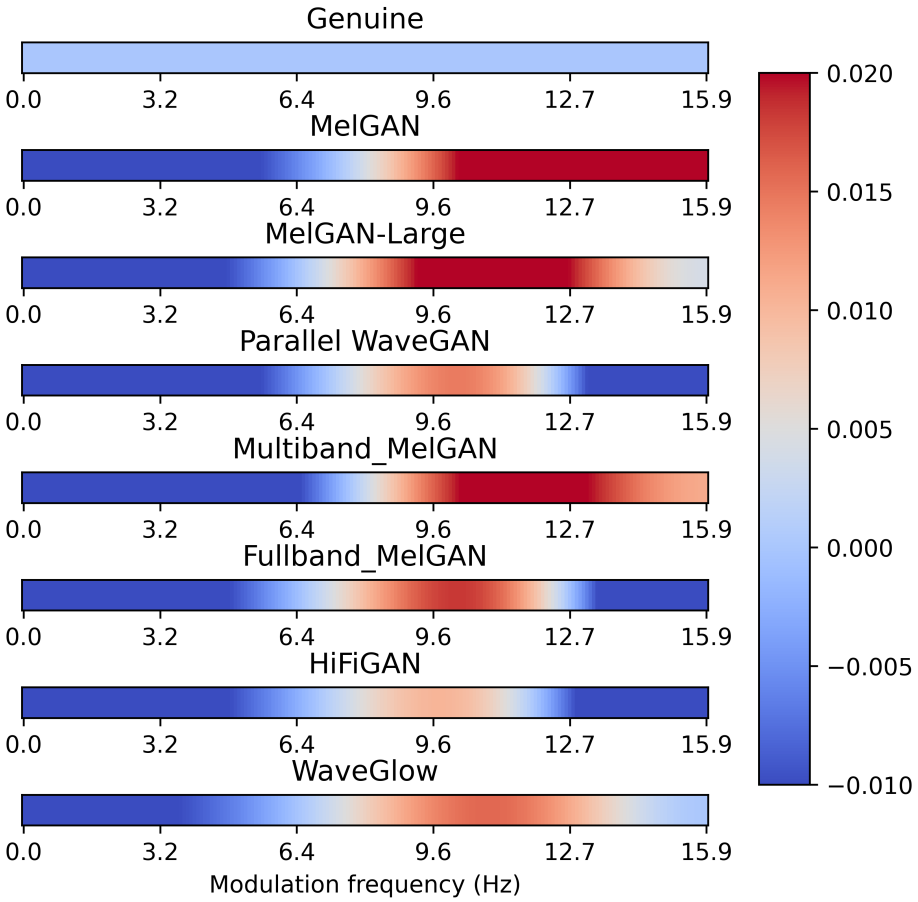


FIGURE 6.5 : Visualization of 1-dimensional compressed version of the proposed representation for a genuine utterance and seven different deepfake versions. The genuine pattern is subtracted from all for better comparisons.

6.7 Conclusion

In this chapter, we presented a task-agnostic model termed WavRx. Experiments on a wide variety of pathological speech datasets as well as synthesized speech datasets have demonstrated that the representation encoded by WavRx outperforms conventional universal representations in terms of generalization to unseen data. Furthermore, we show that by incorporating the frequency dynamics, WavRx disentangles speaker identity without explicit guidance. This makes the model a good candidate for privacy-sensitive applications. Finally, since the WavRx representation is sparse and time-invariant, differences between healthy and pathological speech, as well as genuine and synthesized speech, can be clearly visualized and compared side-by-side. The improvement in interpretability allows a deeper understanding of what information the model relies on for decision-making.

7 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

7.1 Conclusions

In this dissertation, we focused on developing generalizable, interpretable, and privacy-preserving acoustic applications, such as health diagnostics and deepfake detection. The main contributions of this dissertation can be summarized into the following aspects :

Dataset. Existing cough datasets lack cough-phase annotations, which have been used frequently in clinical assessment to help with respiratory disease diagnosis. To bridge this gap, we open-sourced an annotated cough dataset comprising 1000+ manually labeled COVID-19 cough recordings. This is by far the largest cough dataset with fine-grained phase annotations. The release of this dataset will motivate the development of novel cough processing models and enable deeper understanding of cough physiology.

Feature sets. We proposed two knowledge-based feature sets for health diagnostics tasks. The first one combines features extracted from modulation spectrum and linear prediction features. These features aim at capturing articulatory and respiratory abnormalities in pathological speech. We showed their efficacy in COVID-19 detection tasks, outperforming the widely-used openSMILE features in terms of generalization to unseen data. The second feature set is based on our open-sourced annotated cough dataset, which consists of phase-related descriptors that identify abnormalities in lungs, trachea, and vocal folds. We showed that phasic features significantly outperform acoustic features extracted using off-the-shelf speech toolkit.

Model architectures and learning strategies. We proposed several new model architectures and learning strategies with better generalizability and interpretability. These include : (1) MTR-CRNN that learns to capture respiratory-related biomarkers by focusing on specific regions in the modulation tensorgram representation, which are learned from the spectral-temporal saliency map. The cross-dataset generalizability in COVID-19 detection is markedly higher than benchmark VGGish networks, as well as other Challenge-winning knowledge-based systems. (2) SLIM is a novel learning framework that captures the style-linguistic dependency in real speech and uses it for deepfake detection. Different from existing models that rely solely on supervised training, SLIM entails a self-supervised contrastive learning stage to learn real speech patterns that are missing in existing SSL representations. Without full finetuning, we showed that SLIM can significantly outperform other fine-tuned large speech models, especially in unseen attack detection. (3) To protect user privacy for speech diagnostic systems, we evaluated existing voice anonymization methods and proposed an improved method that helps to alleviate the degradation seen in diagnostic ac-

curacies. And finally, (4) we introduced WavRx, a task-agnostic model with interpretability and that preserves user privacy. WavRx not only achieves SOTA on various pathological and synthesized speech datasets, but also generates representations that can be more easily explained and are speaker-independent. These characteristics demonstrate its potential to be deployed in real-world applications.

7.2 Future research directions

In the following, we discuss potential future research directions that can build upon the works described herein.

Multimodal fusion In Chapter 3, we presented two novel feature sets for speech and cough separately. Studies have shown that fusion of the two modalities can lead to better diagnostics accuracy (Deshpande et al., 2020b). However, this requires multiple modalities collected from the same individual, which unfortunately is not the case for most of the datasets available (Chaudhari et al., 2020; Coppock et al., 2023; Deshpande et al., 2020b). In some cases, only a subset of multiple modalities are made available due to privacy concerns (e.g., (Budd et al., 2024)), thus limiting what the data can be used for. Given these limitations on data availability, it remains challenging to evaluate a fusion-based system and its advantages over a uni-modal system. We hope that more high-quality open-source multimodal datasets will become accessible in the future, which will benefit the development of a multimodal diagnostics system.

Multilingual speech models For the majority of the speech applications discussed in this dissertation, we focused on English data due to the scarcity of multilingual datasets at the time of writing. For example, voice anonymization models entail an ASR module to transcribe the speech content. As such, an unknown language would therefore lead to an erroneous text transcription, which could degrade the quality of the anonymized speech. Similar issues exist for diagnostic and deepfake detection models, where models may bias towards English data and underperform on other languages. This has been seen with our cross-dataset evaluation of COVID-19 detection models (Chapter 3). Recently, some attempts have been made to incorporate more languages, as well as accents, to expand language diversity (Müller et al., 2024; Budd et al., 2024). With these data in hand, future research can investigate and improve the robustness of speech models to different languages.

Health foundation model and diagnostic benchmark. In Chapter 6, we showed that WavRx demonstrates decent zero-shot generalizability when predicting unseen diseases. However, since our focus was to explore a novel model architecture, we did not train WavRx with the aggregation of all six datasets, hence it remains unclear whether it would be possible to have a single model

to accurately predict different diseases (i.e., health foundation model). One prerequisite to train a foundation model is to prepare sufficient training data that cover different types of pathologies, such as respiratory diseases, neuromuscular and neuromotor disorders, mental disorders, etc. Fortunately, with larger speech pathological datasets being released to the public recently (Budd et al., 2024), this can unlock the possibility of large-scale pre-training on pathological voice data. Meanwhile, since different pathological speech datasets differ in sample size, data collection protocols, and usually have unbalanced label distribution, it is crucial to design a benchmark to conduct a systematic and fair evaluation of existing diagnostic models. The evaluation procedure conducted for WavRx can be used as a template for developing such benchmark, where models are evaluated in-domain, cross-domain, and in privacy-sensitive conditions with several different metrics to compare their performance.

Pathological voice generation. In Chapter 5, we have demonstrated that when used for pathological voice anonymization, existing voice generative models remove health attributes together with the speaker identity. This is due to the lack of disentanglement between health and other identity-related attributes. In Chapter 6, however, we showed that it is possible to obtain such embeddings that carry health information while containing minimal amount of identity attributes. As such, one possible future research direction can be to explore novel generative model techniques that are built to disentangle speech content, speaker identity, and health information. Successful development of such a model could also benefit health diagnostic model training. Current diagnostic model training relies heavily on high-quality labeled data, which are time-consuming to collect. The pathological voice generative pipeline, on the other hand, can be used to convert healthy speech to different types of pathological speech, which would significantly increase the amount of training data for diagnostic models without having to rely solely on labeled data.

Generalization of deepfake detectors. One challenge faced by speech deepfake detection is the rapidly-evolving field of voice generative modelling, with higher quality samples being generated continuously. As an outcome, deepfake detectors need to be updated regularly to recognize these new types of attack. This usually means retraining models on massive datasets that include the latest type of attacks. Such strategy can be very expensive and time-consuming, while it is extremely difficult to include all different types of attacks. In Chapter 4, we showed a new training framework that can leverage only real speech samples to improve generalization to unseen attacks, which alleviates the need of curating massive-labeled training sets. However, degradation is still seen when tested on out-of-domain data, while the root causes of false positive and negative predictions remain unclear. As such, future research should first focus on investigating the reasons behind false predictions, which would help identify what is missing in existing detectors. Second, it should explore new training strategies that do not solely rely on labeled fake samples.

BIBLIOGRAPHIE

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Adila D, Kang D (2022) Understanding out-of-distribution : A perspective of data dynamics. *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*, PMLR, pages 1–8.

Ahad NA, Yahaya SSS (2014) Sensitivity analysis of welch'st-test. *AIP Conference proceedings*, American Institute of Physics, volume 1605, pages 888–893.

Akman A, Coppock H, Gaskell A, Tzirakis P, Jones L, Schuller BW (2021) Evaluating the COVID-19 identification resnet (cider) on the interspeech COVID-19 from audio challenges. *arXiv preprint arXiv :2107.14549*.

Alqudaihi KS, Aslam N, Khan IU, Almuhaideb AM, Alsunaidi SJ, Ibrahim NMAR, Alhaidari FA, Shaikh FS, Alsenbel YM, Alalharith DM et al. (2021) Cough sound detection and diagnosis using artificial intelligence techniques : challenges and opportunities. *Ieee Access*, 9:102327–102344.

Ardila R, Branson M, Davis K, Kohler M, Meyer J, Henretty M, Morais R, Saunders L, Tyers F, Weber G (2020) Common voice : A massively-multilingual speech corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR (2017) Parkinson's disease and aging : analysis of their effect in phonation and articulation of speech. *Cognitive Computation*, 9:731–748.

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

Ashihara T, Delcroix M, Moriya T, Matsuura K, Asami T, Ijima Y (2024) What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. *arXiv preprint arXiv :2401.17632*.

Avila AR, Akhtar Z, Santos JF, O'Shaughnessy D, Falk TH (2018) Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Transactions on Affective Computing*, 12(1):177–188.

Babu A, Wang C, Tjandra A, Lakhotia K, Xu Q, Goyal N, Singh K, von Platen P, Saraf Y, Pino J et al. (2021) Xls-r : Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv :2111.09296*.

Baevski A, Hsu WN, Xu Q, Babu A, Gu J, Auli M (2022) Data2vec : A general framework for self-supervised learning in speech, vision and language. *International Conference on Machine Learning*, PMLR, pages 1298–1312.

Baevski A, Zhou Y, Mohamed A, Auli M (2020) wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Baldwin S (2012) Compute canada : advancing computational research. *Journal of Physics : Conference Series*, IOP Publishing, volume 341, 012001 pages.

Barrett KE (2019) *Ganong's review of medical physiology*. Numéro 1. : McGraw Hill Education.

Baur S, Nabulsi Z, Weng WH, Garrison J, Blankemeier L, Fishman S, Chen C, Kakarmath S, Maimbolwa M, Sanjase N et al. (2024) HeAR–Health Acoustic Representations. *arXiv preprint arXiv :2403.02522*.

Bedoya S, Katz N, Brian J, O'Shaughnessy D, Falk T (2020) Acoustic and prosodic analysis of vocalizations of 18-month-old toddlers with autism spectrum disorder. *Acoustic Analysis of Pathologies*, De Gruyter, pages 93–126.

Bengio S, Mariéthoz J (2004) A statistical significance test for person authentication. *Proceedings of Odyssey 2004 : The Speaker and Language Recognition Workshop*, numéro CONF.

Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C et al. (2007) Automatic speech recognition and speech variability : A review. *Speech communication*, 49(10-11):763–786.

Blue L, Warren K, Abdullah H, Gibson C, Vargas L, O'Dell J, Butler K, Traynor P (2022) Who are you (I really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. *31st USENIX Security Symposium (USENIX Security 22)*, pages 2691–2708.

Boersma P, Van Heuven V (2001) Speak and unspeak with PRAAT. *Glot International*, 5(9/10): 341–347.

Brabenec L, Mekyska J, Galaz Z, Rektorova I (2017) Speech disorders in parkinson's disease : early diagnostics and effects of medication and brain stimulation. *Journal of neural transmission*, 124:303–334.

Budd J, Baker K, Karoune E, Coppock H, Patel S, Payne R, Tendero Cañadas A, Titcomb A, Hurley D, Egglestone S et al. (2024) A large-scale and PCR-referenced vocal audio dataset for COVID-19. *Scientific Data*, 11(1):700.

Calzada I (2022) Citizens' data privacy in China : The state of the art of the personal information protection law (pipl). *Smart Cities*, 5(3):1129–1150.

Casserly ED, Pisoni DB (2010) Speech perception and production. *Wiley Interdisciplinary Reviews : Cognitive Science*, 1(5):629–647.

Chaudhari G, Jiang X, Fakhry A, Han A, Xiao J, Shen S, Khanzada A (2020) Virufy : Global applicability of crowdsourced and clinical datasets for ai detection of COVID-19 from cough. *arXiv preprint arXiv :2011.13320*.

Chen G, Chai S, Wang G, Du J, Zhang WQ, Weng C, Su D, Povey D, Trmal J, Zhang J et al. (2021a) Gigaspeech : An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv :2106.06909*.

Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F (2021b) Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, Li J, Kanda N, Yoshioka T, Xiao X et al. (2022) Wavlm : Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Chintha A, Thai B, Sohrawardi SJ, Bhatt K, Hickerson A, Wright M, Ptucha R (2020) Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1024–1037.

Choudhary T, Mishra V, Goswami A, Sarangapani J (2020) A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53:5113–5155.

Chowdhury M, Hossain N, Kashem M, Shahid A, Alam A (2020) Immune response in COVID-19 : A review. *Journal of Infection and Public Health*.

Chung KF, Pavord ID (2008) Prevalence, pathogenesis, and causes of chronic cough. *The Lancet*, 371(9621):1364–1374.

Ciotti M, Ciccozzi M, Terrinoni A, Jiang WC, Wang CB, Bernardini S (2020) The COVID-19 pandemic. *Critical reviews in clinical laboratory sciences*, 57(6):365–388.

Clapham RP, van der Molen L, van Son R, van den Brekel MW, Hilgers FJ et al. (2012) Nkiccrt corpus-speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. *LREC*, Citeseer, volume 4, pages 3350–3355.

Coppock H, Akman A, Bergler C, Gerczuk M, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, Xia T et al. (2022a) A summary of the compare COVID-19 challenges. *arXiv preprint arXiv :2202.08981*.

Coppock H, Akman A, Bergler C, Gerczuk M, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, Xia T et al. (2023) A summary of the compare COVID-19 challenges. *Frontiers in Digital Health*, 5:1058163.

Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller B (2021) End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio : a pilot study. *BMJ innovations*, 7(2).

Coppock H, Nicholson G, Kiskin I, Koutra V, Baker K, Budd J, Payne R, Karoune E, Hurley D, Titcomb A et al. (2022b) Audio-based ai classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers. *arXiv preprint arXiv :2212.08570*.

Cox J (2023) *How I Broke Into a Bank Account With an AI-Generated Voice*. `https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice`. Accessed : 2024-04-30.

Cui C, Ma Y, Cao X, Ye W, Zhou Y, Liang K, Chen J, Lu J, Yang Z, Liao KD et al. (2024) A survey on multimodal large language models for autonomous driving. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Darling-White M, Huber JE (2020) The impact of parkinson's disease on breath pauses and their relationship to speech impairment : A longitudinal study. *American Journal of Speech-Language Pathology*, 29(4):1910–1922.

Dash T, Chitlangia S, Ahuja A, Srinivasan A (2022) A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040.

Dash TK, Mishra S, Panda G, Satapathy SC (2021) Detection of COVID-19 from speech signal using bio-inspired based cepstral features. *Pattern Recognition*, 117:107999.

Delacre M, Lakens D, Leys C (2017) Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, 30(1).

Deshpande G, Schuller B (2020a) An overview on audio, signal, speech, & language processing for COVID-19. *arXiv preprint arXiv :2005.08579*.

Deshpande G, Schuller BW (2020b) Audio, speech, language, & signal processing for COVID-19 : A comprehensive overview. *arXiv preprint arXiv :2011.14445*.

Desplanques B, Thienpondt J, Demuynck K (2020) Ecapa-tdnn : Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv :2005.07143*.

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv :1702.08608*.

Drugman T, Bozkurt B, Dutoit T (2012) A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26(1):20–34.

Eyben F (2015) *Real-time speech and music classification by large audio feature space extraction*. Springer.

Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS et al. (2015) The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Eyben F, Wöllmer M, Schuller B (2010) Opensmile : the munich versatile and fast open-source audio feature extractor. *Proc. ACM international conference on Multimedia*, pages 1459–1462.

Fagherazzi G, Fischer A, Ismael M, Despotovic V (2021) Voice for health : the use of vocal biomarkers from research to clinical practice. *Digital biomarkers*, 5(1):78–88.

Falk T, Zheng C, Chan WY (2010a) A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. Audio, Speech, and Language Processing*, 18(7): 1766–1774.

Falk TH, Chan WY (2009) Modulation spectral features for robust far-field speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):90–100.

Falk TH, Chan WY, Sejdic E, Chau T (2010b) Spectro-temporal analysis of auscultatory sounds. *New Developments in Biomedical Engineering*, pages 93–104.

Falk TH, Chan WY, Shein F (2012) Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, 54(5):622–631.

Fan Z, Li M, Zhou S, Xu B (2021) Exploring wav2vec 2.0 on speaker verification and language identification. *Interspeech 2021*.

Fang F, Wang X, Yamagishi J, Echizen I, Todisco M, Evans N, Bonastre JF (2019) Speaker anonymization using x-vector and neural waveform models. *arXiv preprint arXiv :1905.13561*.

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) *Learning from imbalanced data sets*. volume 10. Springer.

Fisher RA (1970) Statistical methods for research workers. *Breakthroughs in statistics : Methodology and distribution*, Springer, pages 66–70.

Fontana GA, Lavorini F (2006) Cough motor mechanisms. *Respiratory physiology & neurobiology*, 152(3):266–281.

Frank J, Schönherr L (2021) Wavefake : A data set to facilitate audio deepfake detection. *arXiv preprint arXiv :2111.02813*.

Greenberg S, Kingsbury BE (1997) The modulation spectrogram : In pursuit of an invariant representation of speech. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, volume 3, pages 1647–1650.

Grosman J (2021) *Fine-tuned XLSR-53 large model for speech recognition in English*. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`.

Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y et al. (2020) Conformer : Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv :2005.08100*.

Han J, Qian K, Song M et al. (2020) An early study on intelligent analysis of speech under COVID-19 : Severity, sleep quality, fatigue, and anxiety. *arXiv :2005.00096*.

Hara S, Hayashi K (2018) Making tree ensembles interpretable : A Bayesian model selection approach. *Intl. conf. artificial intelligence and statistics*, PMLR, pages 77–85.

Harel BT, Cannizzaro MS, Cohen H, Reilly N, Snyder PJ (2004) Acoustic characteristics of parkinsonian speech : a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, 17(6):439–453.

Haridas A, Marimuthu R, Chakraborty B (2018) A novel approach to improve the speech intelligibility using fractional delta-amplitude modulation spectrogram. *Cybern. Systems*, 49(7-8):421–451.

Heckman G, Saari M, McArthur C, Wellens N, Hirdes J (2020) COVID-19 outbreak measures may indirectly lead to greater burden on hospitals. *CMAJ*, 192(14):E384–E384.

Helms J, Kremer S, Merdji H et al. (2020) Neurologic features in severe SARS-CoV-2 infection. *New England Journal of Medicine*, 382(23):2268–2270.

Hermansky H (1998) Modulation spectrum in speech processing. *Signal Analysis and Prediction*, Springer, pages 395–406.

Hilton NH, Schüppert A, Gooskens C (2011) Syllable reduction and articulation rates in danish, norwegian and swedish. *Nordic Journal of Linguistics*, 34(2):215–237.

Honda K (2008) Physiological processes of speech production. *Springer handbook of speech processing*, pages 7–26.

Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A (2021) HuBERT : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hutiri WT, Ding AY (2022) Bias in automated speaker recognition. *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 230–247.

Infante C, Chamberlain D, Fletcher R, Thorat Y, Kodgule R (2017) Use of cough sounds for diagnosis and screening of pulmonary disease. *2017 IEEE global humanitarian technology conference (GHTC)*, IEEE, pages 1–10.

Iqbal U, Bahrami PN, Trimananda R, Cui H, Gamero-Garrido A, Dubois D, Choffnes D, Markopoulou A, Roesner F, Shafiq Z (2022) Your echos are heard : Tracking, profiling, and ad targeting in the amazon smart speaker ecosystem. *arXiv preprint arXiv :2204.10920*.

Ito K, Johnson L (2017) *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/`.

Jin H, Wang S (2018) *Voice-based determination of physical and emotional characteristics of users*. US Patent 10,096,319.

Kadi K, Selouani S, Boudraa B, Boudraa M (2013) Discriminative prosodic features to assess the dysarthria severity levels. *Proceedings of the World Congress on Engineering*, volume 3.

Kahn J, Rivière M, Zheng W, Kharitonov E, Xu Q, Mazaré PE, Karadayi J, Liptchinsky V, Collobert R, Fuegen C et al. (2020) Libri-light : A benchmark for asr with limited or no supervision. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 7669–7673.

Kaloudi N, Li J (2020) The ai-based cyber threat landscape : A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–34.

Kaur N, Singh P (2023) Conventional and contemporary approaches used in text to speech synthesis : A review. *Artificial Intelligence Review*, 56(7):5837–5880.

Khanjani Z, Watson G, Janeja VP (2023) Audio deepfakes : A survey. *Frontiers in Big Data*, 5:1001063.

Kim J, Kong J, Son J (2021) Vits : Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proc. ICML*, pages 5530–5540.

Kingsbury B, Morgan N, Greenberg S (1998) Robust speech recognition using the modulation spectrogram. *Speech commun.*, 25(1-3):117–132.

Knibbs K (2024) *Researchers Say the Deepfake Biden Robocall Was Likely Made With Tools From AI Startup ElevenLabs*. `https://www.wired.com/story/biden-robocall-deepfake-elevenlabs/`. Accessed : 2024-04-30.

Kong J, Kim J, Bae J (2020) Hifi-gan : Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Kornblith S, Norouzi M, Lee H, Hinton G (2019) Similarity of neural network representations revisited. *International conference on machine learning*, PMLR, pages 3519–3529.

Korpas J, Sadlonova J, Salat D, Masarova E (1987) The origin of cough sounds. *Bulletin europeen de physiopathologie respiratoire*, 23:47s–50s.

Korpáš J, Sadloňová J, Vrabec M (1996) Analysis of the cough sound : an overview. *Pulmonary pharmacology*, 9(5-6):261–268.

Koutras P, Panagiotaropoulou G, Tsiami A, Maragos P (2018) Audio-visual temporal saliency modeling validated by fmri data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2000–2010.

Kshirsagar SR, Falk TH (2022) Quality-aware bag of modulation spectrum features for robust speech emotion recognition. *IEEE Transactions on Affective Computing*.

Laguarta J, Hueto F, Subirana B (2020) COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J. Engineering in Medicine and Biology*, 1:275–281.

Latif S, Qadir J, Qayyum A, Usama M, Younis S (2020a) Speech technology for healthcare : Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14: 342–356.

Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller BW (2020b) Deep representation learning in speech processing : Challenges, recent advances, and future trends. *arXiv preprint arXiv :2001.00378*.

Lauriola I, Aiolli F (2020) Mklpy : a python-based framework for multiple kernel learning. *arXiv :2007.09982*.

Layton S, De Andrade T, Olszewski D, Warren K, Gates C, Butler K, Traynor P (2024) Every breath you don't take : Deepfake speech detection using breath. *arXiv preprint arXiv :2404.15143*.

Lee P, Cotterill-Jones C, Eccles R (2002) Voluntary control of cough. *Pulmonary pharmacology & therapeutics*, 15(3):317–320.

Li L, Fan Y, Tse M, Lin KY (2020) A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854.

Li M, Ahmadiadli Y, Zhang XP (2024) Audio anti-spoofing detection : A survey. *arXiv preprint arXiv :2404.13914*.

Lin GT, Feng CL, Huang WP, Tseng Y, Lin TH, Li CA, Lee Hy, Ward NG (2023) On the utility of self-supervised models for prosody-related tasks. *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pages 1104–1111.

Little M, McSharry P, Hunter E, Spielman J, Ramig L (2008) Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *Nature Precedings*, pages 1–1.

Liu H, Gu X, Samaras D (2019) Wasserstein gan with quadratic transport cost. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4832–4841.

Liu X, Wang X, Sahidullah M, Patino J, Delgado H, Kinnunen T, Todisco M, Yamagishi J, Evans N, Nautsch A et al. (2023) Asvspoof 2021 : Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Livingstone SR, Russo FA (2018) The ryerson audio-visual database of emotional speech and song (ravdess) : A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Llanos F, Alexander JM, Stilp CE, Kluender KR (2017) Power spectral entropy as an information-theoretic correlate of manner of articulation in american english. *The Journal of the Acoustical Society of America*, 141(2):EL127–EL133.

Lorenzo-Trueba J, Yamagishi J, Toda T, Saito D, Villavicencio F, Kinnunen T, Ling Z (2018) The voice conversion challenge 2018 : Promoting development of parallel and nonparallel methods. *The Speaker and Language Recognition Workshop*, ISCA, pages 195–202.

Low DM, Bentley KH, Ghosh SS (2020) Automated assessment of psychiatric disorders using speech : A systematic review. *Laryngoscope investigative otolaryngology*, 5(1):96–116.

Lu J, Zhang Y, Wang W, Shang Z, Zhang P (2024) One-class knowledge distillation for spoofing speech detection. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 11251–11255.

Lugović S, Dunđer I, Horvat M (2016) Techniques and applications of emotion recognition in speech. *2016 39th international convention on information and communication technology, electronics and microelectronics (mipro)*, IEEE, pages 1278–1283.

Macklem P (1974) Physiology of cough. *Annals of Otology, Rhinology & Laryngology*, 83(6):761–768.

Macklem PT, Mead J (1968) Factors determining maximum expiratory flow in dogs. *Journal of Applied Physiology*, 25(2):159–169.

Maharana K, Mondal S, Nemade B (2022) A review : Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99.

Makhoul J (1975) Linear prediction : A tutorial review. *Proceedings IEEE*, 63(4):561–580.

Makhoul J (1977) Stable and efficient lattice methods for linear prediction. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 25(5):423–428.

Markaki M, Stylianou Y (2011) Voice pathology detection and discrimination based on modulation spectral features. *IEEE Transactions on audio, speech, and language processing*, 19(7):1938–1948.

Markel J, Gray A (2013) *Linear prediction of speech*. volume 12. Springer Science & Business Media.

Markowitz JA (2000) Voice biometrics. *Communications of the ACM*, 43(9):66–73.

Martín-Doñas JM, Álvarez A (2022) The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9241–9245.

McAdams SE (1984) *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university.

Memon SA (2020) Acoustic Correlates of the Voice Qualifiers : A survey. *arXiv :2010.15869*.

Menendez-Pidal X, Polikoff JB, Peters SM, Leonzio JE, Bunnell HT (1996) The nemours database of dysarthric speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, IEEE, volume 3, pages 1962–1965.

Meyer S, Lux F, Denisov P, Koch J, Tilli P, Vu NT (2022a) Speaker anonymization with phonetic intermediate representations. *arXiv preprint arXiv :2207.04834*.

Meyer S, Lux F, Denisov P, Koch J, Tilli P, Vu NT (2022b) Speaker Anonymization with Phonetic Intermediate Representations. *Proc. Interspeech 2022*, pages 4925–4929.

Meyer S, Lux F, Koch J, Denisov P, Tilli P, Vu NT (2023) Prosody is not identity : A speaker anonymization approach using prosody cloning. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 1–5.

Mohammadi SH, Kain A (2017) An overview of voice conversion systems. *Speech Communication*, 88:65–82.

Morais E, Hoory R, Zhu W, Gat I, Damasceno M, Aronowitz H (2022) Speech emotion recognition using self-supervised features. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 6922–6926.

Moro-Velazquez L, Villalba J, Dehak N (2020) Using x-vectors to automatically detect parkinson's disease from speech. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 1155–1159.

Müller N, Czempin P, Diekmann F, Froghyar A, Böttinger K (2022) Does audio deepfake detection generalize? *Interspeech 2022*.

Müller N, Dieckmann F, Czempin P, Canals R, Böttinger K, Williams J (2021) Speech is silver, silence is golden : What do asvspoof-trained models really learn? *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*.

Müller NM, Kawa P, Choong WH, Casanova E, Gölge E, Müller T, Syga P, Sperl P, Böttinger K (2024) Mlaad : The multi-language audio anti-spoofing dataset. *arXiv preprint arXiv :2401.09512*.

Nagrani A, Chung JS, Zisserman A (2017) Voxceleb : a large-scale speaker identification dataset. *arXiv preprint arXiv :1706.08612*.

Naudé W (2020) Artificial intelligence vs covid-19 : limitations, constraints and pitfalls. *AI & society*, 35:761–765.

Naunheim MR, Zhou AS, Puka E, Franco Jr RA, Carroll TL, Teng SE, Mallur PS, Song PC (2020) Laryngeal complications of COVID-19. *Laryngoscope investigative otolaryngology*, 5(6):1117–1124.

Nessiem MA, Mohamed MM, Coppock H, Gaskell A, Schuller B (2021) Detecting COVID-19 from breathing and coughing sounds using deep neural networks. *IEEE Intl Symposium on Computer-Based Medical Systems*, IEEE, pages 183–188.

Noffs G, Perera T, Kolbe SC, Shanahan CJ, Boonstra FM, Evans A, Butzkueven H, van der Walt A, Vogel AP (2018) What speech can tell us : A systematic review of dysarthria characteristics in multiple sclerosis. *Autoimmunity reviews*, 17(12):1202–1209.

Ohala JJ (1990) Respiratory activity in speech. *Speech production and speech modelling*, Springer, pages 23–53.

O'Shaughnessy D (1988) Linear predictive coding. *IEEE potentials*, 7(1):29–32.

Pahar M, Klopper M, Warren R, Niesler T (2021) COVID-19 cough classification using machine learning and global smartphone recordings. *Computers in Biology and Medicine*, 104572 pages.

Pappagari R, Cho J, Moro-Velazquez L, Dehak N (2020a) Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity. *Interspeech*, pages 2177–2181.

Pappagari R, Wang T, Villalba J, Chen N, Dehak N (2020b) x-vectors meet emotions : A study on dependencies between emotion and speaker recognition. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 7169–7173.

Pasad A, Chou JC, Livescu K (2021) Layer-wise analysis of a self-supervised speech representation model. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, pages 914–921.

Pasad A, Shi B, Livescu K (2023) Comparative layer-wise analysis of self-supervised speech models. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 1–5.

Patterson RD, Nimmo-Smith I, Holdsworth J, Rice P (1987) An efficient auditory filterbank based on the gammatone function. *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, volume 2.

Pedregosa F, Varoquaux G, Gramfort A et al. (2011) Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peeters G (2004) A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Ist Project Report*, 54(0):1–25.

Pellegrino F, Coupé C, Marsico E (2011) A cross-language perspective on speech information rate. *Language*, pages 539–558.

Pender T (2023) *AI Threatens Courts with Fake Evidence, UW Prof Says*. `https://www.jdsupra.com/legalnews/ai-threatens-courts-with-fake-evidence-7371356/`. Accessed : 2024-05-05.

Pepino L, Riera P, Ferrer L (2021) Emotion recognition from speech using wav2vec 2.0 embeddings. *Interspeech 2021*.

Piirila P, Sovijarvi A (1995) Objective assessment of cough. *European Respiratory Journal*, 8(11): 1949–1956.

Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V (2020) SARS-CoV-2 detection from voice. *IEEE Open J. Engineering in Medicine and Biology*, 1:268–274.

Pitts T, Bolser D, Rosenbek J, Troche M, Sapienza C (2008) Voluntary cough production and swallow dysfunction in parkinson's disease. *Dysphagia*, 23(3):297–301.

Qian J, Du H, Hou J, Chen L, Jung T, Li XY, Wang Y, Deng Y (2017) Voicemask : Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv :1711.11460*.

Quatieri T, Talkar T, Palmer J (2020) A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open J. Engineering in Medicine and Biology*, 1:203–206.

Rabiner LR, Schafer RW et al. (2007) Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, 1(1–2):1–194.

Radovic M, Ghalwash M, Filipovic N, Obradovic Z (2017) Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC bioinformatics*, 18(1):1–14.

Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J (2017) Svcca : Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Raj D, Snyder D, Povey D, Khudanpur S (2019) Probing the information encoded in x-vectors. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, pages 726–733.

Ramachandram D, Taylor G (2017) Deep multimodal learning : A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

Ramanarayanan V, Lammert AC, Rowe HP, Quatieri TF, Green JR (2022) Speech as a biomarker : opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups*, 7(1):276–283.

Ramos VM, Hernandez-Diaz HAK, Huici MEHD, Martens H, Van Nuffelen G, De Bodt M (2020) Acoustic features to characterize sentence accent production in dysarthric speech. *Biomedical Signal Processing and Control*, 57:101750.

Ravanelli M, Parcollet T, Plantinga P, Rouhe A, Cornell S, Lugosch L, Subakan C, Dawalatabad N, Heba A, Zhong J et al. (2021a) Speechbrain : A general-purpose speech toolkit. *arXiv preprint arXiv :2106.04624*.

Ravanelli M, Parcollet T, Plantinga P, Rouhe A, Cornell S, Lugosch L, Subakan C, Dawalatabad N, Heba A, Zhong J, Chou JC, Yeh SL, Fu SW, Liao CF, Rastorgueva E, Grondin F, Aris W, Na H, Gao Y, Mori RD, Bengio Y (2021b) *SpeechBrain : A General-Purpose Speech Toolkit*. arXiv :2106.04624.

Ren Y, Hu C, Tan X, Qin T, Zhao S, Zhao Z, Liu TY (2020) Fastspeech 2 : Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv :2006.04558*.

Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L et al. (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217.

Rochet-Capellan A, Fuchs S (2014) Take a breath and take the turn : how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 369(1658):20130399.

Rudzicz F, Namasivayam AK, Wolff T (2012) The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541.

Santos JF, Cosentino S, Hazrati O, Loizou PC, Falk TH (2013) Objective speech intelligibility measurement for cochlear implant users in complex listening environments. *Speech communication*, 55(7-8):815–824.

Sarria-Paja M, Falk TH (2013) Whispered speech detection in noise using auditory-inspired modulation spectrum features. *IEEE Signal Processing Letters*, 20(8):783–786.

Scherer KR (2003) Vocal communication of emotion : A review of research paradigms. *Speech communication*, 40(1-2):227–256.

Schuller B, Steidl S, Batliner A, Bergelson E, Krajewski J, Janott C, Amatuni A, Casillas M, Seidl A, Soderstrom M et al. (2017) The interspeech 2017 computational paralinguistics challenge : Addressee, cold & snoring. *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, pages 3442–3446.

Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, MüLler C, Narayanan S (2013a) Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Schuller B, Steidl S, Batliner A, Nöth E, Vinciarelli A, Burkhardt F, Van Son R, Weninger F, Eyben F, Bocklet T et al. (2012) The interspeech 2012 speaker trait challenge. *INTERSPEECH 2012, Portland, OR, USA*.

Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E et al. (2013b) The interspeech 2013 computational paralinguistics challenge : Social signals, conflict, emotion, autism. *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Schuller BW, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, Kaya H, Amiriparian S, Baird A, Stappen L et al. (2021) The interspeech 2021 computational paralinguistics challenge : COVID-19 cough, COVID-19 speech, escalation & primates. *arXiv preprint arXiv :2102.13468*.

Shah J, Singla YK, Chen C, Shah RR (2021) What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *arXiv preprint arXiv :2101.00387*.

Shannon R, Baekey D, Morris K, Nuding S, Segers L, Lindsey B (2004) Production of reflex cough by brainstem respiratory networks. *Pulmonary pharmacology & therapeutics*, 17(6):369–376.

Sharma NK, Chetupalli SR, Bhattacharya D, Dutta D, Mote P, Ganapathy S (2022) The second dicova challenge : Dataset and performance analysis for diagnosis of COVID-19 using acoustics. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 556–560.

Shih TH, Yeh CY, Chen MS (2024) Does audio deepfake detection rely on artifacts? *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 12446–12450.

Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks : Visualising image classification models and saliency maps. *arXiv preprint arXiv :1312.6034*.

Slack D, Friedler S, Scheidegger C, Roy C (2019) Assessing the local interpretability of machine learning models. *arXiv :1902.03501*.

Slaney M et al. (1993) An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 35(8).

Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors : Robust dnn embeddings for speaker recognition. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pages 5329–5333.

Solomon NP (2006) What is orofacial fatigue and how does it affect function for swallowing and speech? *Seminars in speech and language*, volume 27, pages 268–282.

Srivastava BML, Tomashenko N, Wang X, Vincent E, Yamagishi J, Maouche M, Bellet A, Tommasi M (2020a) Design choices for x-vector based speaker anonymization. *arXiv preprint arXiv :2005.08601*.

Srivastava BML, Vauquier N, Sahidullah M, Bellet A, Tommasi M, Vincent E (2020b) Evaluating voice conversion-based privacy protection against informed attackers. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 2802–2806.

Stevens KN (2000) *Acoustic phonetics*. volume 30. MIT press.

Stupp C (2019) Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. *The Wall Street Journal*, 30(08).

Stylianou Y (2009) Voice transformation : a survey. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pages 3585–3588.

Tak H, Jung JW, Patino J, Kamble M, Todisco M, Evans N (2021) End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *ASVSPOOF 2021, Automatic Speaker Verification and Spoofing Countermeasures Challenge*, ISCA, pages 1–8.

Tak H, Kamble M, Patino J, Todisco M, Evans N (2022a) Rawboost : A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 6382–6386.

Tak H, Todisco M, Wang X, Jung Jw, Yamagishi J, Evans N (2022b) Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *The Speaker and Language Recognition Workshop (Odyssey 2022)*, ISCA.

Tan X (2023) *Neural text-to-speech synthesis*. Springer Nature.

Tan X, Qin T, Soong F, Liu TY (2021) A survey on neural speech synthesis. *arXiv preprint arXiv :2106.15561*.

Tena A, Claria F, Solsona F (2022) Automated detection of COVID-19 cough. *Biomedical Signal Processing and Control*, 71:103175.

Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A, Yamagishi J, Evans N, Kinnunen T, Lee KA (2019) Asvspoof 2019 : Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv :1904.05441*.

Tomashenko N, Wang X, Miao X, Nourtel H, Champion P, Todisco M, Vincent E, Evans N, Yamagishi J, Bonastre JF (2022a) The voiceprivacy 2022 challenge evaluation plan. *arXiv preprint arXiv :2203.12468*.

Tomashenko N, Wang X, Vincent E, Patino J, Srivastava BML, Noé PG, Nautsch A, Evans N, Yamagishi J, O'Brien B et al. (2022b) The voiceprivacy 2020 challenge : Results and findings. *Computer Speech & Language*, 74:101362.

Triantafyllopoulos A, Schuller BW (2024) Expressivity and speech synthesis. *arXiv preprint arXiv :2404.19363*.

Triantafyllopoulos A, Schuller BW, İymen G, Sezgin M, He X, Yang Z, Tzirakis P, Liu S, Mertes S, André E et al. (2023) An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*.

Tu J (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231.

Vaessen N, Van Leeuwen DA (2022) Fine-tuning wav2vec2 for speaker recognition. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 7967–7971.

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

van Son RJ et al. (2021) Measuring voice quality parameters after speaker pseudonymization. *Interspeech*, pages 1019–1023.

Vásquez-Correa JC, Orozco-Arroyave J, Bocklet T, Nöth E (2018) Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of communication disorders*, 76:21–36.

Venkit PN, Srinath M, Wilson S (2022) A study of implicit bias in pretrained language models against people with disabilities. *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332.

Vetter P, Vu D, L'Huillier A, Schibler M, Kaiser L, Jacquerioz F (2020) *Clinical features of COVID-19*.

Vyas G, Dutta MK, Prinosil J, Harár P (2016) An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features. *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, pages 515–518.

Wang C, Chen S, Wu Y, Zhang Z, Zhou L, Liu S, Chen Z, Liu Y, Wang H, Li J et al. (2023a) Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv :2301.02111*.

Wang C, Riviere M, Lee A, Wu A, Talnikar C, Haziza D, Williamson M, Pino J, Dupoux E (2021a) Voxpopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv :2101.00390*.

Wang J, Cao B, Yu P, Sun L, Bao W, Zhu X (2018a) Deep learning towards mobile applications. *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, pages 1385–1393.

Wang J, Lan C, Liu C, Ouyang Y, Qin T, Lu W, Chen Y, Zeng W, Philip SY (2022) Generalizing to unseen domains : A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072.

Wang L, Huang N, Hong Y, Liu L, Guo X, Chen G (2023b) Voice-based ai in call center customer service : A natural field experiment. *Production and Operations Management*, 32(4):1002–1018.

Wang X, Yamagishi J (2021b) Investigating self-supervised front ends for speech spoofing countermeasures. *arXiv preprint arXiv :2111.07725*.

Wang X, Yamagishi J (2023c) Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 1–5.

Wang X, Yamagishi J (2024) Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 10311–10315.

Wang Y, Boumadane A, Heba A (2021c) A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv :2111.02735*.

Wang Y, Stanton D, Zhang Y, Ryan RS, Battenberg E, Shor J, Xiao Y, Jia Y, Ren F, Saurous RA (2018b) Style tokens : Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *International conference on machine learning*, PMLR, pages 5180–5189.

Watanabe S, Hori T, Kim S, Hershey JR, Hayashi T (2017) Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Weenink D (2003) Canonical correlation analysis. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, University of Amsterdam Amsterdam, volume 25, pages 81–99.

Welch BL (1947) The generalization of 'STUDENT'S' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

Weninger F, Eyben F, Schuller B, Mortillaro M, Scherer K (2013) On the acoustics of emotion in audio : what speech, music, and sound have in common. *Frontiers in Psychology*, 4:292.

Widdicombe J (1954) Receptors in the trachea and bronchi of the cat. *The Journal of physiology*, 123(1):71.

Winkworth AL, Davis PJ, Ellis E, Adams RD (1994) Variability and consistency in speech breathing during reading : Lung volumes, speech intensity, and linguistic factors. *Journal of Speech, Language, and Hearing Research*, 37(3):535–556.

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Wu S, Falk T, Chan WY (2011) Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768–785.

Xia T, Spathis D, Ch J, Grammenos A, Han J, Hasthanasombat A, Bondareva E, Dang T, Floto A, Cicuta P et al. (2021) COVID-19 sounds : a large-scale audio dataset for digital respiratory screening. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xie Y, Cheng H, Wang Y, Ye L (2023) Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection. *Proc. INTERSPEECH*, volume 2023, pages 2808–2812.

Yadav AKS, Xiang Z, Bhagtani K, Bestagini P, Tubaro S, Delp EJ (2023) Ps3dt : Synthetic speech detection using patched spectrogram transformer. *2023 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pages 496–503.

Yamagishi J, Veaux C, MacDonald K (2019) *CSTR VCTK Corpus : English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*.

Yamagishi J, Wang X, Todisco M, Sahidullah M, Patino J, Nautsch A, Liu X, Lee KA, Kinnunen T, Evans N et al. (2021) ASVspoof 2021 : accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv :2109.00537*.

Yamin MM, Ullah M, Ullah H, Katt B (2021) Weaponized ai for cyber attacks. *Journal of Information Security and Applications*, 57:102722.

Yang Sw, Chang HJ, Huang Z, Liu AT, Lai CI, Wu H, Shi J, Chang X, Tsai HS, Huang WC et al. (2024a) A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yang Sw, Chi PH, Chuang YS, Lai CIJ, Lakhotia K, Lin YY, Liu AT, Shi J, Chang X, Lin GT et al. (2021a) SUPERB : Speech processing universal performance benchmark. *arXiv preprint arXiv :2105.01051*.

Yang Y, Qin H, Zhou H, Wang C, Guo T, Han K, Wang Y (2024b) A robust audio deepfake detection system via multi-view feature. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 13131–13135.

Yang YY, Hira M, Ni Z, Chourdia A, Astafurov A, Chen C, Yeh CF, Puhrsch C, Pollack D, Genzel D, Greenberg D, Yang EZ, Lian J, Mahadeokar J, Hwang J, Chen J, Goldsborough P, Roy P, Narenthiran S, Watanabe S, Chintala S, Quenneville-Bélair V, Shi Y (2021b) TorchAudio : Building Blocks for Audio and Speech Processing. *arXiv preprint arXiv :2110.15018*.

Yi J, Wang C, Tao J, Zhang X, Zhang CY, Zhao Y (2023) Audio deepfake detection : A survey. *arXiv preprint arXiv :2308.14970*.

Yi Z, Huang WC, Tian X, Yamagishi J, Das RK, Kinnunen T, Ling ZH, Toda T (2020) Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion–. *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, ISCA.

Zahorian S, Hu H (2008) A spectral/temporal method for robust fundamental frequency tracking. *Journal of the Acoustical Society of America*, 123(6):4559–4571.

Zarsky TZ (2016) Incompatible : The gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995.

Zhang Y, Li Z, Lu J, Hua H, Wang W, Zhang P (2023) The impact of silence on speech anti-spoofing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhang Y, Xia T, Han J, Wu Y, Rizos G, Liu Y, Mosuily M, Chauhan J, Mascolo C (2024a) Towards open respiratory acoustic foundation models : Pretraining and benchmarking. *arXiv preprint arXiv :2406.16148*.

Zhang Y, Xia T, Saeed A, Mascolo C (2024b) Respllm : Unifying audio and text with multimodal llms for generalized respiratory health prediction. *arXiv preprint arXiv :2410.05361*.

Zhang Z (2016) Mechanics of human voice production and control. *Journal of the Acoustical Society of America*, 140(4):2614–2635.

Zhu M, Gupta S (2017) To prune, or not to prune : exploring the efficacy of pruning for model compression. *arXiv preprint arXiv :1710.01878*.

Zhu Y, Falk T (2024a) Wavrx : a disease-agnostic, generalizable, and privacy-preserving speech health diagnostic model. *IEEE Journal of Biomedical and Health Informatics*.

Zhu Y, Falk T (in press) How generalizable and interpretable are speech-based COVID-19 detection systems? : A comparative analysis and new system proposal. *2022 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE.

Zhu Y, Falk TH (2022) Fusion of modulation spectral and spectral features with symptom metadata for improved speech-based COVID-19 detection. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 8997–9001.

Zhu Y, Falk TH (2024b) Spectral–temporal saliency masks and modulation tensorgrams for generalizable covid-19 detection. *Computer Speech & Language*, 86:101620.

Zhu Y, Goel C, Koppisetti S, Tran T, Kumar A, Bharaj G (2024c) Learn from real : Reality defender's submission to asvspoof5 challenge. *arXiv preprint arXiv :2410.07379*.

Zhu Y, Imoussaine M, Côté-Lussier C, Falk T (2023a) Investigating biases in COVID-19 diagnostic systems processed with automated speech anonymization algorithms. *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, pages 46–54.

Zhu Y, Imoussaıne-Aıkous M, Côté-Lussier C, Falk TH (2023b) Investigating biases in covid-19 diagnostic systems processed with automated speech anonymization algorithms. *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, pages 46–54.

Zhu Y, Imoussaïne-Aïkous M, Côté-Lussier C, Falk TH (2023c) On the impact of voice anonymization on speech-based COVID-19 detection. *arXiv preprint arXiv :2304.02181*.

Zhu Y, Koppisetti S, Tran T, Bharaj G (2024d) Slim : Style-linguistics mismatch model for generalized audio deepfake detection. *arXiv preprint arXiv :2407.18517*.

Zhu Y, Powar S, Falk TH (2023d) Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection. *arXiv preprint arXiv :2309.08099*.

Zhu Y, Shaik MH, Falk TH (2023e) On the importance of different cough phases for covid-19 detection. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 1–5.

Zhu Y, Tiwari A, Monteiro J, Kshirsagar S, Falk TH (2023f) COVID-19 detection via fusion of modulation spectrum and linear prediction speech features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.