Speech Understanding (CSL7770)
Assignment 1 - Question 1
Speech Based Medical Diagnosis

Soham Niraj Deshmukh (B21EE067)
Shubh Goyal (B21CS073)

The code can be found here: GitHub Repository

1 Introduction

Speech understanding and related techniques (speech recognition, identification, synthesis, etc.) is being currently used in a wide range of domains, including speech assistants, automotive industry, education, smart customer service, etc. Other than this, speech understanding can also be used as an important tool in medical diagnosis, which could make medical care easier, efficient, reliable and less invasive. Through the analysis of speech patterns, acoustic features, and linguistic content, it could be possible to identify early markers of various medical conditions, such as neurological disorders like Parkinson's and Alzheimer's disease, respiratory, thoracic, and cardio-vascular conditions. This could facilitate a prompt diagnosis of medical disorders, reducing the need for elaborate medical tests over long periods and also minimizing the risks caused due to late diagnosis. Thus, this could help improve the quality of life for patients through early prevention.

This report explores the application and current developments of speech understanding techniques in medical diagnosis, focusing on machine learning, deep learning, and spectrogram-based analysis methods.

2 Current Works

So far, ample research has been primarily done on four key areas: the detection of Alzheimer's disease, Parkinson's disease; respiratory and thoracic conditions; and cardiovascular conditions.

This section first explores the currently available task related datasets, and then moves towards the advancements in current research in the areas.

2.1 Datasets

The section explores the task-related datasets available and researched upon in past four to five years. This datasets are used to create the foundational models for diagnosing and predicting various medical conditions. The datasets are as follows:

1. ***COUGHVID*** [8] It is a crowdsourcing dataset containing 25,000 cough recordings across a wide range of ages, genders, geographic locations. It is hand-labelled by four experienced physicians.
2. ***ADReSS*** [1] It is benchmark dataset of spontaneous speech for the detection of cognitive impairment and Alzheimer's Dementia (AD). It contains 1955 audios from 78 non-AD subjects and 2122 audios from 78 AD subjects.

3. ***Saarbrücken Voice Database*** [2] This database contains the vocal audio files of normal and diseased patients. The audio files have been collected for two vocal diseases, Laryngozele and Vox senilis, and for normal people.
4. ***Parkinson Detection from Vocal*** [5] This database contains 197 data points with 22 features each, extracted from vocal sounds, and can be used for prediction of Parkinson's disease.
5. ***Pascal Heart Sound Database*** [3] The dataset is separated into two datasets (176 and 656 wav files). Each can be used for heart sound segmentation. One of the subdataset has 3 categories and the other has 4 categories out of which one category is for normal people and other for some early symptoms of cardio vascular diseases.

2.2 Current Research

This section explores the current SOTA models in the four major fields stated earlier. We explore the model architecture and pipeline, it's strengths, it's limitations and the metrics obtained during it's evaluation as stated in the referenced articles and papers corresponding to the SOTA of the field. They are as follows:

- **Cardiovascular Diseases.** [4]
  Method - The pipeline used extracts features through two methods: Mel-Frequency Cepstral Coefficients (MFCCs) for vocal tract characteristics and glottal features using the Quasi-Closed Phase (QCP) method. These features are then made to go through four machine learning models, including SVM, Extra Trees (from tree-based models family), AdaBoost, and Feed Forward Neural Network, which classify speech as either heart failure (HF) or healthy. Feature selection techniques and grid search optimization were applied to improve performance and reduce redundancy.

  Metric - The study tested models under both speaker-dependent and speaker-independent settings, using a leave-five-speakers-out cross-validation strategy. Accuracy was used as the evaluation metric in this case. The Neural Network model achieved 95.020% accuracy in speaker-dependent settings and 81.51% accuracy in speaker-independent setting for heart failure detection.

  Strengths - This is non-invasive, and allows for early heart failure detection. Combining MFCC features with glottal features, helps enhance the classification performance, indicating that these features obtained by both methods reflect complementary aspects of HF-related changes in speech. Feature selection using Gini impurity from Extra Trees classifiers enhances generalization and computation in the model. The model is also resilient to noise, under mild to moderate noise conditions of 15-30 dB SNR.

  Limitations - The dataset used was very small (45 speakers: 25 healthy, 20 HF) with limited generalization to larger populations where demography changes with location. The speaker-independent accuracy drops considerably for the model, reaching even as low as 81.51%, indicating overfitting to specific speaker characteristics. Although the model is noise-robust, performance deteriorates dramatically under low SNR conditions or with non-stationary noise such as traffic noise.

- **Alzeihmer Disease** [6]
  Method - GP-Net, it is a Feature Purification Network with a backbone of Transformer encoder, removing non-discriminative features for better classification of Alzheimer. It

is divided into G-Net, which extracts common features using a Gradient Reversal Layer (GRL), and P-Net, which purifies the remaining features for classification. The model optimizes two loss functions, namely Lossc for common feature extraction and Lossp for purified feature learning, to enhance feature separability. Unlike traditional models, GP-Net proactively removes irrelevant information, therefore claiming to generate comparatively robust feature representation.

Metric - The evaluation metric used was classification accuracy. GP-Net achieved state-of-the-art (SOTA) performance in Alzheimer detection from speech, with an accuracy of 93.5% on the Pitt dataset, 78.6% on ADReSS, and 83.7% on iFLY. It outperformed existing deep learning models of the time.

Strengths - It purifies features, which leads to a more discriminative feature space than traditional Transformer models. It claims to be computationally efficient for deployment in mobile apps or online screening tools, making Alzheimer detection more accessible.

Limitations - Despite improvements, it still inherits the computational complexity of Transformer models ($O(n^2d)$), which makes it costly for long speech transcripts. The model's effectiveness depends on the parameter , used for feature purification, thus it requires careful hyperparameter tuning to avoid over-removing useful features. Additionally, while the purification process reduces noise, it risks losing subtle linguistic information. Did not achieve SOTA results on ADReSS and iFLY.

– **Parkinson Disease** [7]
Method - ISNDAM was proposed an enhancement of the Smallest Normalized Difference Associative Memory (SNDAM). It incorporated a feature selection phase to reduce noise and improve accuracy. The model proposed has three phases: Training (constructs learning matrices and computes similarities), Relevance Identification (removes irrelevant features using wrapper-based selection), and Testing (classifies new instances using the smallest normalized difference metric). ISNDAM model manages input-output relationships via memory matrices, using R and R operators to measure similarity.

Metric - The proposed model achieved a classification accuracy of 99.48% on Dataset 1 and 99.66% on Dataset 2 for Parkinson's disease detection using voice recordings. As stated in the paper, the improvements were statistically significant (p ¡ 0.05), which demonstrated ISNDAM's superiority in classification performance.

Strengths - ISNDAM achieves state-of-the-art accuracy and robustness to noise, making it highly reliable for real-world medical applications. Its feature selection process enhances performance by eliminating irrelevant data, thereby improving classification accuracy. The memory-based approach significantly reduces computational complexity compared to deep learning models, enabling faster training and classification.

Limitations - Despite its high accuracy, ISNDAM suffers from limited generalization, as it was validated only on two controlled datasets. Its wrapper-based feature selection increases computational effort due to testing multiple feature combinations. The model also is a black-box, as associative memories do not intuitively reveal feature importance.

– **Respiratory Diseases** [9]
Method - The proposed RBF-Net here has three components: a CNN-LSTM feature encoder to extract spatial and temporal features, a COVID-19 classifier for disease detection, and a bias predictor (adversarial component) to remove demographic confounders. The model uses a conditional Generative Adversarial Network (c-GAN) approach, training with binary cross-entropy loss (L_C) for classification and adversarial loss (L_B) for bias elimination.

Metric - RBF-Net demonstrates superior bias mitigation in COVID-19 cough classification, achieving a +5.5% improvement in gender-biased settings, +7.7% in age-biased settings, and +8.2% in smoking-biased settings over standard CNN-LSTM models.

Strengths - RBF-Net is actively removing age, gender, and smoking status related bias, thus, preventing misleading associations and improving real-world applicability. The model generalizes well across different demographic subgroups, outperforming traditional CNN-LSTM models under biased training conditions. It was trained on clinically verified cough data.

Limitations - The model has been limited to COVID-19 classification, and still requires validation on other respiratory diseases like COPD and asthma for broader applicability. Adversarial training increases computational complexity, requiring careful hyperparameter tuning to maintain stability. Despite bias correction, imbalanced demographic representation in the dataset (e.g., more male and non-smoker samples) may still introduce subtle biases.

3 Open Problems

Major advancements have been made in speech-based healthcare diagnostics in the recent year, but there still remains several open problems that needs to be addressed in future research which are as follows:

– The first major challenge is availability of clinically verified high quality diverse data. Training models on small, imbalanced datasets that lack sufficient demographic representation, often lead to biased predictions or model availability in only one demography.
– Generalization across diverse populations is another important issue, as models often overfit to specific speaker characteristics, and loose effectiveness in real-world applications.
– Another concern is model's robustness to noise and environmental variations, as speech data collected in controlled settings may not align to noisy, real-world environments such as hospitals or homes. This distribution shift between training and testing data can lead to wrong inferences.
– The computational complexity of certain models', specifically transformer based models like GP-Net makes it difficult to deplot them in resource-constrained environment.
– Another major issue is with the black-box nature of deep learning models. The non-interpretability of the decision made and the lack of reasoning makes it difficult for doctors to trust the models.
– Standardization of evaluation metrics and datasets must also be done, as different studies use varying datasets and methodologies, making it difficult to compare models and assess true performance improvements.

References

1. Alzheimer's dementia recognition through spontaneous speech. https://luzs.gitlab.io/adress/.
2. Saarbrücken voice database. http://stimmdb.coli.uni-saarland.de/). https://www.kaggle.com/datasets/subhajournal/patient-health-detection-using-vocal-audio.
3. P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. http://www.peterjbentley.com/heartchallenge/index.html.
4. M. Kiran Reddy, Pyry Helkkula, Y. Madhu Keerthana, Kasimir Kaitue, Mikko Minkkinen, Heli Tolppanen, Tuomo Nieminen, and Paavo Alku. The automatic detection of heart failure using speech signals. *Computer Speech Language*, 69:101205, 2021.
5. Max Little. Parkinsons. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C59C74.
6. Yuan Z.- Tang Q. Liu, N. Improving alzheimer's disease detection for speech based on feature purification network. *Frontiers in Public Health*, 2022.
7. M.; Uriarte-Arcia A.V.; Rodríguez-Molina A.; Alarcón-Paredes A.; Ventura-Molina E. Luna-Ortiz, I.; Aldape-Pérez. Parkinson's disease detection from voice recordings using associative memories. *healthcare*, 2023.
8. Teijeiro T. Atienza D. Orlandic, L. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. sci data 8, 156 (2021). https://www.nature.com/articles/s41597-021-00937-4, 2021. https://www.kaggle.com/datasets/nasrulhakim86/coughvid-wav.
9. Tabish Saeed, Aneeqa Ijaz, Ismail Sadiq, Haneya N. Qureshi, Ali Rizwan, and Ali Imran. An ai-enabled bias-free respiratory disease diagnosis model using cough audio: A case study for covid-19, 2024.