

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

UNIVERSITY OF NEW ZEALAND

Speech Processing with Deep Learning for Voice-based Respiratory Diagnosis

A thesis presented in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy
in
Computer Science

at Massey University, Albany,
New Zealand.

Zhizhong Ma

2022

Abstract

Voice-based respiratory diagnosis research aims at automatically screening and diagnosing respiratory-related symptoms (e.g., smoking status, COVID-19 infection) from human-generated sounds (e.g., breath, cough, speech). It has the potential to be used as an objective, simple, reliable, and less time-consuming method than traditional biomedical diagnosis methods. In this thesis, we conduct one comprehensive literature review and propose three novel deep learning methods to enrich voice-based respiratory diagnosis research and improve its performance.

Firstly, we conduct a comprehensive investigation of the effects of voice features on the detection of smoking status. Secondly, we propose a novel method that uses the combination of both high-level and low-level acoustic features along with deep neural networks for smoking status identification. Thirdly, we investigate various feature extraction/representation methods and propose a SincNet-based CNN method for feature representations to further improve the performance of smoking status identification. To the best of our knowledge, this is the first systemic study that applies speech processing with deep learning for voice-based smoking status identification.

Moreover, we propose a novel transfer learning scheme and a task-driven feature representation method for diagnosing respiratory diseases (e.g., COVID-19) from human-generated sounds. We find those transfer learning methods using VGGish, wav2vec 2.0 and PASE+, and our proposed task-driven method Sinc-ResNet have achieved competitive performance compared with other work. The findings of this study provide a new perspective and insights for voice-based respiratory disease diagnosis.

The experimental results demonstrate the effectiveness of our proposed methods and show that they have achieved better performances compared to other existing methods.

Acknowledgements

I wish to take this opportunity to express my sincere appreciation to all the people who have supported me on my PhD journey.

Firstly, I would like to pay my sincere gratitude to my main supervisor, Professor Ruili Wang, my co-supervisors, Dr Feng Hou and Professor Johan Potgieter, for their invaluable academic guidance and support throughout my PhD study. They encouraged me in my academic research and daily life, and their insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would particularly like to thank my main supervisor Professor Wang, who dedicated his time and effort in helping me improve my research ability. Apart from his academic guidance, Professor Wang also taught me many life lessons when I encountered difficulties, and without his help it would have been impossible for me to complete my PhD study.

Secondly, I would also like to thank my friends, lab mates, and colleagues in Professor Wang's research team and in the School of Mathematical and Computational Sciences. They provided valuable advice and support through my doctoral research.

Thirdly, I greatly acknowledge the China Scholarship Council and the Catalyst: Strategic - New Zealand - Singapore Data Science Research Programme, which provided financial support to help me finish my PhD study.

In addition, I would like to express my deepest appreciation to my parents and my family, for their unconditional love and support all through my studies and my life.

Publications

The following papers have been published or submitted to international journals and conferences:

- **Zhizhong Ma**, Christopher Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the Objective Speech Assessment of Smoking Status Based on Voice Features: A Review of the Literature. In the *Journal of Voice*, 2021. <https://doi.org/10.1016/j.jvoice.2020.12.014> (Refer to Chapter 2)
- **Zhizhong Ma**, Satwinder Singh, Yuanhang Qiu, Feng Hou, Ruili Wang, Christopher Bullen and Joanna Ting Wai Chu. Automatic Speech-based Smoking Status Identification. In the *Proceedings of Computing Conference*, 2022. (Accepted) (Refer to Chapter 3)
- **Zhizhong Ma**, Yuanhang Qiu, Feng Hou, Ruili Wang, Joanna Ting Wai Chu and Christopher Bullen. Determining the Best Acoustic Features for Smoking Status Identification. In the *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8177-8181. IEEE, 2022. (Refer to Chapter 4)
- **Zhizhong Ma**, Ruili Wang, Feng Hou, Yuanhang Qiu, Satwinder Singh, Joanna Ting Wai Chu and Christopher Bullen. Transfer Learning and Task-driven Feature Representations for COVID-19 Diagnosis from Respiratory Sound Data. In the *ACM Transactions on Speech and Language Processing (TSLP)*. ACM, 2022. (submitted) (Refer to Chapter 5)
- **Zhizhong Ma**, Junbo Ma, Xijuan Liu and Feng Hou. Large Margin Training for Long Short-Term Memory Neural Networks in Neural Language Modeling. In the *Proceedings of the 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 2022. (Accepted)

-
- Yuanhang Qiu, Ruili Wang, Satwinder Singh, **Zhizhong Ma** and Feng Hou. Self-Supervised Learning Based Phone-fortified Speech Enhancement. In the *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 211-215. IEEE, 2021.
 - Yuanhang Qiu, Ruili Wang, Feng Hou, Satwinder Singh, **Zhizhong Ma** and Xiaoyun Jia. Adversarial Multi-task Learning with Inverse Mapping for Speech Enhancement. In the *Applied Soft Computing*, 120, p.108568. 2022.
 - Zhihan Wang, Feng Hou, Yuanhang Qiu, **Zhizhong Ma**, Satwinder Singh, Ruili Wang. CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition. In the *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH)*. IEEE, 2022. (Accepted)

Contents

Abstract	i
Acknowledgements	ii
Publications	iii
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Overview	1
1.2 Motivations	3
1.3 Contributions	4
1.4 Organization of Thesis	5
2 Literature Review of Smoking Status Identification Based on Voice	
Features	11
2.1 Introduction	12
2.2 Methods	16
2.3 Results	17
2.3.1 Fundamental Frequency	17

2.3.2	Jitter	19
2.3.3	Shimmer	20
2.3.4	Harmonics to Noise Ratio	22
2.3.5	Formant Frequencies	23
2.3.6	Other Features	24
2.3.7	Effects of Smoking Cessation on Voice Features	25
2.4	Discussions	26
2.5	Conclusions	29
3	Automatic Speech-based Smoking Status Identification	39
3.1	Introduction	40
3.2	Acoustic Features for Smoking Status Identification	42
3.2.1	MFCC and Fbank	42
3.2.2	Fundamental Frequency	43
3.2.3	Jitter	44
3.2.4	Shimmer	45
3.3	Methodology	45
3.4	Experiments	47
3.4.1	Datasets	47
3.4.2	Implementation Details	48
3.4.3	Evaluation Metrics	49
3.5	Results and Discussions	50
3.6	Conclusions	52
4	Best Acoustic Features for Smoking Status Identification	59
4.1	Introduction	60
4.2	Related Work	62
4.2.1	extended Geneva Minimalistic Acoustic Parameter Set	62
4.2.2	Computational Paralinguistics Challenge set	63

4.2.3	Bag-of-Audio-Words	63
4.3	Methodology	64
4.4	Experiments	65
4.4.1	Datasets	65
4.4.2	Feature Extraction	67
4.4.3	Classification Setups	68
4.5	Results and Discussions	68
4.6	Conclusions	69
5	Transfer Learning and Task-Driven Feature Representations for COVID-19 Diagnosis	77
5.1	Introduction	78
5.2	Feature Extraction	80
5.2.1	VGGish	80
5.2.2	Wav2vec 2.0	81
5.2.3	Problem-Agnostic Speech Encoder	82
5.2.4	SincNet	82
5.3	Methodology	83
5.3.1	Transfer Learning Scheme	84
5.3.2	Task-Driven Feature Representation Network	85
5.3.3	Datasets	86
5.4	Experimental Results and Discussions	87
5.4.1	Results on the COVID-19 Sounds Dataset	87
5.4.2	Results on the Coswara Dataset	88
5.4.3	Discussions	90
5.5	Conclusions	90
6	Summary	97
6.1	Research Summary	97

6.2 Future Work	99
A Statement of Contribution	103

List of Figures

3.1	The architecture of our proposed automatic smoking status identification method.	46
4.1	The architecture of our proposed method.	65
5.1	The proposed audio-based COVID-19 diagnosis pipeline.	84
5.2	An overview of our Sinc-ResNet architecture.	85

List of Tables

2.1	Comparison of biochemical smoking status validation methods.	13
2.2	Comparison of smokers' vs non-smokers' fundamental frequency.	18
2.3	Comparison of smokers' vs non-smokers' jitter.	20
2.4	Comparison of smokers' vs non-smokers' shimmer.	21
2.5	Comparison of smokers' vs non-smokers' HNR.	23
2.6	Comparison of smokers' vs non-smokers' formant frequencies.	24
2.7	Comparison of smokers' vs non-smokers' other features.	25
2.8	A summary of voice features.	28
3.1	Speech features statistics divided by smoking status and gender.	48
3.2	Smoking status identification experiment results.	51
4.1	Summary of feature sets/representations utilised in this study.	62
4.2	The status of the speaker's age in our dataset.	66
4.3	Experimental results of different acoustic feature sets/representations on the test set.	68
5.1	A comparison of different sound types and methods for task 1 of the COVID-19 Sounds dataset.	88
5.2	A comparison of different sound types and methods for task 2 of the COVID-19 Sounds dataset.	89
5.3	A comparison of different sound types and methods for experiments on the Coswara dataset.	89

Chapter 1

Introduction

This chapter provides an overview of this thesis. The background of this research is introduced briefly in Section 1.1. The motivations are explained in Section 1.2. The contributions of this thesis are summarised in Section 1.3. Finally, the organisation of this thesis is listed in Section 1.4.

1.1 Overview

Human vocal architecture is a complex and unique anatomical structure that enables us to vocalise a wide range of acoustic signals that are coordinated and meaningful [1]. The complexity of speech production makes it a suitable indicator for various health conditions [2]. Hence, speech signals can carry a speaker's basic information, such as age, gender, emotional status, psychological status, intoxication level, and smoking status, which are powerful biomarkers for voice-related health diagnosis [3–5]. There is an active and growing area of deep learning research in this voice-based health diagnosis domain, which focuses on developing paradigms to objectively determine such health status. Specifically, the applications of voice-based health

diagnosis research have been studied include intoxication and fatigue [6]; Alzheimer's disease (AD) [7]; upper respiratory tract infection [8] and depression [9].

In the application to intoxication detection, researchers use acoustic features, prosodic features, speech rate features and glottal pulse features to detect the degree of intoxication from the speech signal by means of statistical classification [10]. It is a widely accepted hypothesis that Alzheimer's disease affects verbal fluency, which is reflected by the patient's slow speech rate, or other speech impairments [11]. The speech-based detection of AD with an attention-based hybrid network demonstrates that voice features have the potential of being a useful tool for early screening of AD [12]. An upper respiratory infection affects the upper part of patients' respiratory system and can make their voices distinctive, creating recognizable voice signatures and enabling the training of deep learning algorithms to grade the severity of the disease. Results show that using vocal biomarkers to aid the diagnosis of upper respiratory infections are promising [8]. Voice symptoms seem more frequent in people with high levels of cortisol [13], which is common in patients with depression; therefore, voice features such as Mel-Frequency Cepstrum Coefficient (MFCC) are used to discriminate symptoms of depression in combination with deep neural networks [9].

Voice analysis using speech processing and deep learning opens new opportunities for healthcare. Voice-based health diagnosis has advantages over traditional biochemical measures for diagnosis, risk prediction, and remote monitoring of various clinical outcomes and symptoms, due to the costs and ease of the sample collection process [14]. Voice-based health diagnosis is especially useful under the current COVID-19 pandemic, where movement restrictions may make other methods more difficult or expensive than usual. Voice-based respiratory diagnosis, one of the voice-based health diagnosis applications, has not been fully studied yet. There is an urgent, unmet need for reliable, intelligent voice-based respiratory diagnosis systems based

on artificial intelligence to support objectively assessment of respiratory diseases. Hence, it motives us to conduct research of speech processing and deep learning methods to fill the gap and improve the performance of the voice-based respiratory diagnosis research.

1.2 Motivations

Smoking status is an important indicator of the human health status, since automatic smoking status identification has a variety of applications, including smoking status validation [15], smoking cessation tracking [16] and speaker profiling [17]. However, the way to use speech processing with deep learning techniques to determine smoking status (i.e., smoker or non-smoker) still has not been fully studied. This motivates us to conduct research on smoking status identification studies, by conducting a comprehensive literature review to find theoretical support for smoking status identification research, to specifying identification methods and determining the best acoustic features using deep learning techniques to increase the performance of smoking status identification.

Concurrently, under the current COVID-19 pandemic, voice-based automatic methods for screening and diagnosing respiratory symptoms (e.g., COVID-19) have recently gained increased attention and became an emerging topic in the voice-based health diagnosis research [4, 18, 19]. Multiple deep learning methods have been developed to identify respiratory diseases (e.g., COVID-19) from human-generated sounds (e.g., breath, cough, speech) [5, 20, 21]. However, deep learning methods are typically data-hungry, and the current amount of available COVID-19 labelled data is normally limited. How to tackle the scarcity of well-labelled data, learning effective speech feature representations and improving diagnosis performance are still long-standing and challenging tasks. This motivates us to propose a transfer

learning scheme to identify the COVID-19 disease by fine-tuning the pre-trained representation models (i.e., VGGish, wav2vec 2.0, problem-agnostic speech encoder (PASE+)) on datasets with COVID-19 labels. We also propose a task-driven feature representation network Sinc-ResNet (SincNet as the frontend, with ResNet as the backend) to learn feature representations effectively.

1.3 Contributions

To address the issues mentioned above, one comprehensive literature review and three novel deep learning methods are proposed in this thesis, and summarised below:

- Chapter 2 refers to our published paper "Towards the objective speech assessment of smoking status based on voice features: a review of the literature" in the *Journal of Voice*, 2021. To find theoretical support for the smoking status identification study, we conducted a comprehensive investigation of the effectiveness of voice features for smoking status identification [22]. To the best of our knowledge, this is the first study that has conducted a comprehensive literature review for smoking status identification based on voice features.
- Chapter 3 refers to our published paper "Automatic speech-based smoking status identification" in the *Computing Conference*, 2022. To develop an automatic voice-based smoking status identification system, we proposed a novel method that uses the combination of both high- and low-level acoustic features along with deep neural networks [23]. We also proposed a dataset that can be used for the smoking status identification experiments. In addition, the data augmentation technique (i.e., SpecAugment) is also applied to further improve smoking status identification accuracy. To the best of our knowledge, this is

the first study that has comprehensively explored both high- and low-level acoustic features for smoking status identification from voice.

- Chapter 4 refers to our published paper "Determining the best acoustic features for smoking status identification" in the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. To determine the best acoustic features for smoking status identification, we compared two feature extraction/learning techniques: (i) hand-crafted feature sets including the extended Geneva Minimalistic Acoustic Parameter Set and the Computational Paralinguistics Challenge Set; (ii) the Bag-of-Audio-Words representations; and proposed a novel neural representation method that extracted features from raw waveform signals by SincNet [24]. To the best of our knowledge, this is the first study that has explored acoustic feature extraction/learning techniques for smoking status identification from speech.
- Chapter 5 refers to our submitted paper "Transfer learning and task-driven feature representations for COVID-19 diagnosis from respiratory sound data" in the *ACM Transactions on Speech and Language Processing (TSLP)*, 2022. To address the scarcity of COVID-19 well-labelled data and to learn feature representations effectively, we proposed a novel transfer learning scheme and a task-driven feature representation network for diagnosing respiratory diseases (e.g., COVID-19) from audio signals (i.e., breath, cough, and speech) [25]. The findings of this study provide new perspective and insights for voice-based respiratory diseases (e.g., COVID-19) diagnosis.

1.4 Organization of Thesis

This is a thesis by publications. It contains four individual studies (Chapter 2 to Chapter 5), either published in a journal/conference or submitted

to a journal/conference. All references related to each chapter are listed at the end of each chapter.

Chapter 2 presents a comprehensive literature review of the effects of voice features for smoking status identification.

Chapter 3 presents the proposed novel method that incorporating low-level and high-level acoustic features with data augmentation, and using deep neural network as a classifier for smoking status identification.

Chapter 4 presents the proposed feature extraction/learning method to determine the best acoustic features for smoking status identification.

Chapter 5 presents the proposed transfer learning scheme and task-driven feature representation network for COVID-19 diagnosis from respiratory sound data.

Chapter 6 summarises this thesis and discusses future work.

References

- [1] W. Tecumseh Fitch. The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, 4(7):258–267, 2000.
- [2] Nicholas Cummins, Alice Baird, and Bjoern W Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.
- [3] P Mayorga, C Druzgalski, RL Morelos, OH Gonzalez, and J Vidales. Acoustics based assessment of respiratory diseases using GMM classification. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 6312–6316. IEEE, 2010.

-
- [4] Maude Desjardins, Lucinda Halstead, Annie Simpson, Patrick Flume, and Heather Shaw Bonilha. Voice and respiratory characteristics of men and women seeking treatment for presbyphonia. *Journal of Voice*, 20(6):731–739, 2020.
- [5] Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8328–8332. IEEE, 2021.
- [6] Björn Schuller, Anton Batliner, Stefan Steidl, Florian Schiel, and Jarek Krajewski. The INTERSPEECH 2011 speaker state challenge. In *the Annual Conference of the International Speech Communication Association (INTER-SPEECH)*, pages 3201–3204. IEEE, 2011.
- [7] Francisco Martínez-Sánchez, Juan José G Meilán, Juan Carro, and Olga Ivanova. A prototype for the voice analysis diagnosis of Alzheimer’s disease. *Journal of Alzheimer’s Disease*, 64(2):473–481, 2018.
- [8] Nicholas Cummins, Maximilian Schmitt, Shahin Amiriparian, Jarek Krajewski, and Björn Schuller. “You sound ill, take the day off”: Automatic recognition of speech affected by upper respiratory tract infection. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3806–3809. IEEE, 2017.
- [9] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.
- [10] Meng Ge, Ruixiong Zhang, Wei Zou, Xiangang Li, Cheng Gong, Longbiao Wang, and Jianwu Dang. Order-aware pairwise intoxication detection. In *2021*

- 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2021.
- [11] María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer’s disease and automatic speech analysis: a review. *Expert Systems with Applications*, 150:113213, 2020.
- [12] Jun Chen, Ji Zhu, and Jieping Ye. An attention-based hybrid network for automatic detection of Alzheimer’s disease from narrative speech. In *the Annual Conference of the International Speech Communication Association (INTER-SPEECH)*, pages 4085–4089. IEEE, 2019.
- [13] Sofia Holmqvist-Jämsén, Ada Johansson, Pekka Santtila, Lars Westberg, Bettina von der Pahlen, and Susanna Simberg. Investigating the role of salivary cortisol on vocal symptoms. *Journal of Speech, Language, and Hearing Research*, 60(10):2781–2791, 2017.
- [14] Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5(1):78–88, 2021.
- [15] Dogan Pinar, Hakan Cincik, Evren Erkul, and Atila Gungor. Investigating the effects of smoking on young adult male voice by using multidimensional methods. *Journal of Voice*, 30(6):721–725, 2016.
- [16] Harveen Kaur Ubhi, Susan Michie, Daniel Kotz, Onno CP van Schayck, Abiram Selladurai, and Robert West. Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features. *Translational Behavioral Medicine*, 6(3):410–417, 2016.
- [17] Amir Hossein Poorjam and Mohamad Hasan Bahari. Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 7–12. IEEE, 2014.

-
- [18] Neeraj Kumar Sharma, Ananya Muguli, Prashant Krishnan, Rohit Kumar, Srikanth Raj Chetupalli, and Sriram Ganapathy. Towards sound based testing of COVID-19—Summary of the first Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge. *Computer Speech & Language*, 73:101320, 2022.
- [19] Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, and Lukas Stappen. The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. *arXiv preprint arXiv:2102.13468*, 2021.
- [20] Amir Vahedian-Azimi, Abdalsamad Keramatfar, Maral Asiaee, Seyed Shahab Atashi, and Mandana Nourbakhsh. Do you have COVID-19? An artificial intelligence-based screening tool for COVID-19 using acoustic parameters. *The Journal of the Acoustical Society of America*, 150(3):1945–1953, 2021.
- [21] John Harvill, Yash R Wani, Mark Hasegawa-Johnson, Narendra Ahuja, David Beiser, and David Chestek. Classification of COVID-19 from cough using autoregressive predictive coding pretraining and spectral data augmentation. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4261–4265. IEEE, 2021.
- [22] Zhizhong Ma, Christopher Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. *Journal of Voice*, 36(6), 2021.
- [23] Zhizhong Ma, Satwinder Singh, Yuanhang Qiu, Feng Hou, Ruili Wang, Christopher Bullen, and Joanna Ting Wai Chu. Automatic speech-based smoking status identification. In *Computing Conference 2022*. IEEE, 2022.
- [24] Zhizhong Ma, Yuanhang Qiu, Feng Hou, Ruili Wang, Joanna Ting Wai Chu, and Christopher Bullen. Determining the best acoustic features for smoking status identification. In *IEEE International Conference on Acoustics, Speech*

- and Signal Processing (ICASSP)*, pages 8177–8181. IEEE, 2022.
- [25] Zhizhong Ma, Ruili Wang, Feng Hou, Yuanhang Qiu, Satwinder Singh, Joanna Ting Wai Chu, and Christopher Bullen. Transfer learning and task-driven feature representations for covid-19 diagnosis from respiratory sound data. In *ACM Transactions on Speech and Language Processing (TSLP)*. ACM, 2022.

Chapter 2

Literature Review of Smoking Status Identification Based on Voice Features

In smoking cessation clinical research and practice, objective validation of self-reported smoking status is crucial for ensuring the reliability of the primary outcome. Speech signals convey important information about a speaker, such as age, gender, body size, emotional state, and health state. We investigated (1) if smoking could measurably alter voice features, (2) if smoking cessation could lead to changes in voice, and therefore (3) if the voice-based smoking status identification the potential to be used as an objective smoking cessation validation method. We found that fundamental frequency, jitter, shimmer, harmonics to noise ratio, and other voice features are affected by smoking and could be used to assess smoking status. Speech assessment of smoking status based on voice features has potential as a smoking status validation method. Furthermore, this study provides recommendations for future research on the objective speech assessment of smoking status based on voice features.

2.1 Introduction

Smoking remains as one of the leading preventable causes of illness worldwide [1]. Conversely, stopping smoking (smoking cessation) dramatically reduces the risks of future disease and premature death. In smoking cessation research, the proportion of smokers who remain abstinent from smoking for a sustained period of time (typically measured at 6 months after the date of quitting) is the primary outcome measure used to evaluate the effectiveness of new interventions, policies, and practices for reducing smoking [2, 3]. It is regarded as best practice in clinical trials and in the clinical treatment of smoking that self-reported abstinence should be confirmed with biochemical verification to validate smoking status. Several biological markers can be used to validate smoking status using biological samples. Different biological means of validating smoking status will be suitable in different studies, depending on the sample collection methods available, exclusion and/or inclusion of alternative nicotine delivery, the relevant timeframes for assessing smoking status, and availability of technical expertise. A consensus of experts in the field of smoking cessation suggest that carbon monoxide measures in expired breath and cotinine assay measured in urine or saliva are the most valid and feasible methods to validate smoking status [4], due to their relatively low cost and ease of use compared to other validation methods (see Table 2.1).

However, for some studies, the available biological methods for validating smoking status may not be feasible due to costs, remoteness, and sample collection. In such cases, self-report measures are widely used as alternative methods for estimate information concerning the smoking status [5, 6]. However, self-report measures can be subject to bias and misreporting and may compromise the validity of the study findings and affect the trial outcome. Specifically, self-report measures in smoking cessation trials may be subject to social desirability bias, and participants may be biased to report non-smoker status because of a desire to “help” the researcher be

TABLE 2.1: Comparison of biochemical smoking status validation methods.

Validation Methods	Types of Sample	Maximum Detection Window	Advantages	Disadvantages
Nicotine	Blood, saliva, urine	8-12 hours	High specificity, samples can be sent for testing.	Short half-life, expensive, technical difficulty, not suitable for Nicotine Replacement Therapy (NRT) trials.
Carbon Monoxide	Expired air, blood	12-24 hours	Relatively inexpensive, commercially available instruments, simple, portable, immediate feedback.	Marginal utility for cigarette use, sensitive to environmental sources of CO (i.e. pollution, second-hand smoke, marijuana use), short detection window, in person testing, not suitable for non-combustible tobacco products.
Cotinine (Nicotine Metabolite)	Blood, saliva, urine	80-100 hours	High specificity, longer half-life, inexpensive, can be assessed remotely (i.e., saliva test-strips).	Not suitable for NRT trials, biohazard risk constraints for collection and carriage of a specimen, half-life varies across groups
Minor Tobacco Alkaloids (Anabasine, Anatabine)	Urine	50-80 hours	High specificity, differentiates NRT.	Expensive, technical difficulty.

a “good” participant and avoid stigma [7]. Research suggests that there is a high rate of misclassification for self-reported abstinence, which varies across populations [8, 9]. Due to demand characteristics, misclassification may be highest among ethnic minority groups and lower socioeconomic groups, where the burden of harm from tobacco is highest [10, 11].

As such, there is a need to develop new methods of validating smoking status that are simple, noninvasive, low-cost, able to be used across a widely dispersed population and do not require face-to-face contact. The development of such methods would revolutionise the way that smoking cessation studies are evaluated and population smoking trends monitored, thus improving the feasibility of smoking status validation, particularly in large studies.

Speech signals convey a speaker’s important information such as age, gender, body size, emotional state, and health state [12–16]. The physiological effects of cigarette smoking have been well-documented [17–19] and include pharyngeal diseases and disorders resulting from prolonged effects of a variety of harmful chemicals in the cigarette smoke. Exposure to tobacco smoke can affect throat tissues, causing inflammation to vocal-folds and malfunction of the vocal cords [20–22], along with degrading lung function and thereby decreasing the airflow through the vocal cords [21, 23, 24]. The changes in the vocal tract can eventually lead to a dramatic variation in the speaker’s speech signals. This raises the possibility that smoking status validation via speech analysis (as well as smoking behaviour detection) could be used to identify if a person is a smoker from a given speech signal by comparing his and/or her voice features with other smokers’ and/or non-smokers’ voice features. Advances in speech signal processing and machine learning could enable such analyses to be done in real time. There has been a set of research works on machine learning methods applied to the broad voice analysis research area such as pathological voice detection [25, 26], and voice activity detection [27, 28]. A number of studies have

investigated voice analysis based on specific machine learning algorithms such as decision trees [29], support vector machine [30, 31], hidden Markov model [32, 33], Gaussian mixture model [34, 35], and artificial neural networks [36, 37], which have reported high accuracy and performance [38–40].

We hypothesised that when a person changes their smoking behaviour, the change in their voice could be used to validate smoking status. The idea of identifying a smoker from a given speech signal by comparing his and/or her voice features with other smokers' and/or non-smokers' voice features has been explored in previous studies that concluded there is a correlation between a smoker's speech signals and his/her smoking status but were limited by a number of methodological features. Yet the question remains regarding: what kind of voice features will be affected by smoking? Voice features can be divided into two categories: linguistic features and acoustic [41]. While linguistic features involve the analysis of language form, language meaning, and language in context, acoustic features are the acoustic components present in a speech that are capable of being experimentally observed, recorded, and reproduced; they include fundamental frequency (F_0), pitch, jitter, shimmer, and harmonics to noise ratio (HNR). We were interested in the acoustic features that enable the analysis of speech signals because these features contain speakers' discriminative information that can be extracted for further classification.

This chapter is organized as follows. The methods we utilised in this literature review is given in Section 2.2. The results of voice features known to be affected by smoking are introduced in Section 2.3. Sections 2.4 gives our discussions, and the conclusions are shown in Section 2.5.

2.2 Methods

In this research, we aimed to evaluate the effects of smoking and smoking cessation on acoustic voice parameters. We sought to answer the following four questions:

1. Does smoking affect a speakers' voice quality?
2. Which voice features are altered due to smoking and by how much?
3. After cessation of smoking, do these features recover to a normal level, and over what period?
4. Is it possible to detect if a person is an active smoker or an ex-smoker from these voice changes?

The following digital databases below were searched for relevant articles:

- ACM Digital Library [<http://dl.acm.org/>]
- IEEE eXplore [<http://ieeexplore.ieee.org/>]
- ScienceDirect [<https://www.sciencedirect.com/>]
- SpringerLink [<https://link.springer.com/>]
- MDPI (Multidisciplinary Digital Publishing Institute) [<https://www.mdpi.com/>]
- arXiv [<https://arxiv.org/>]
- Taylor & Francis Online [<https://www.tandfonline.com/>]
- PubMed [<https://pubmed.ncbi.nlm.nih.gov/>]
- Google Scholar [<https://scholar.google.com/>]

The following search terms were used and linked: (smoking **OR** cigarette smoking **OR** tobacco smoking) **AND** (voice features **OR** voice parameters **OR** speech signal) **AND/OR** (voice analysis **OR** assessment) **AND/OR** (smoking cessation **OR** quit smoking). The scope of the study was restricted from any period to 2020.

Initially, a total of 17 papers were found. We then widened the search with additional keywords (smoker detection, voice analysis, acoustic analysis, etc.), leading to a total of 34 articles included for review. The following section summarizes the key findings from the review in relation to the research questions.

2.3 Results

In the following sections, we list the voice features of a speech signal known to be affected by smoking.

2.3.1 Fundamental Frequency

Fundamental frequency (F_0) is an important acoustic feature of speech signals. F_0 is the lowest, and usually, the strongest frequency produced by the complex vocal fold vibrations, measured in Hertz (Hz). It is generally considered to be the fundamental tone of sound and represents how high or low the frequency of a person's voice sounds. The F_0 is calculated by using the period T of the speech signal:

$$F_0 = \frac{1}{T}. \quad (2.1)$$

However, for the speech signal, the period T is not constant since the input speech signal contains amplitude and frequency perturbations [42]. Several improved algorithms such as RAPT [43], SWIPE [44], YIN [45], and pYIN [46] have been proposed

TABLE 2.2: Comparison of smokers' vs non-smokers' fundamental frequency.

Authors	Methods	Smokers		Non-smoker	
		Male	Female	Male	Female
Horii and Sorenson [50]	Oral reading				
	Average	105.65	182.7	115.95	186.45
	25-32	114.62	189.93	123.27	199.58
	33-41	106.71	197.67	118.49	178.32
Gonzalez and Carpi [20]	42-49	95.76	159.88	107.42	208.11
	Sustained	Male	Female	Male	Female
	vowels	119.4	192.4	125.4	206.4
Lee et al. [51]	Sustained	Female		Female	
	vowels	229		234	

to estimate the F_0 based on acoustic features. Fundamental frequency values obtained in speech signals are typically less than 300 Hz for children and greater than 100 Hz for adults, 120 Hz for men and 210 Hz for women [47–49].

Studies (Table 2.2) have consistently found lower F_0 in smokers compared to age- and sex-matched non-smokers: in a study of 80 participants aged 25-49 years, half of whom were smokers [50], F_0 was measured for oral reading and spontaneous speech. The mean F_0 values were lower in the smoker group than the non-smoker group for males (105.65 Hz smokers vs 115.95 Hz non-smokers) and females (182.70 Hz smokers vs 186.45 Hz non-smokers). Gonzalez and Carpi evaluated the effect of cigarette smoking on voice features in young adults (n=134) who had smoked less than 10 years [20]. F_0 was lower in smokers than in non-smokers, but the difference was only statistically significant in females (192.4 Hz smokers vs 206.4 Hz non-smokers, $P < 0.01$). There was a dose-response effect with the number of cigarettes smoked. The findings suggest that the effect of smoking on F_0 is more significant for women than for men. Lee et al. compared the voice features of non-smoking women that were exposed to second-hand smoke (i.e., passive smoking) to those that were not exposed [51]. There was no significant difference between passive smokers and non-smokers in F_0 (229 Hz passive smokers vs 234 Hz non-smokers), or for any of the other voice features.

2.3.2 Jitter

Jitter (measured in microseconds or % jitter) is a common perturbation measure of the cycle-to-cycle frequency variation or instability of a speech signal, expressed as:

$$Jitter = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}|. \quad (2.2)$$

where T_i is the extracted period of the i^{th} speech signal segment and N is the number of extracted speech signal segments [52].

Studies have found higher jitter measures in smokers compared to non-smokers (Table 2.3). Gonzalez and Carpi found jitter increased between non-smokers and smokers of less than 10 years, but the difference was only significant in men (47.67 ms non-smoker male vs 62.78 ms smoker male, $P < 0.05$). This suggests that changes in jitter may be related to long-term smoking [20]. In one study [53], authors confirmed that the jitter value was higher in women aged 18-24 years who smoked compared to nonsmoking women, but the difference was not significant. Women in the smoking group had a relatively short history of smoking (3.5 years on average), which may account for the study findings.

Three studies evaluated the voice changes over different smoking frequencies and smoking histories. In one study [21], 32 adults (12 smokers without voice problems, eight smokers with voice problems, and 12 non-smokers) were evaluated by the phonatory tasks. The results of the jitter analyses shown that smokers with voice problems present statistically significant higher jitter values for all speech tasks than non-smokers. In another study [54], jitter values were significantly higher in smoking males who smoked at least five cigarettes a day for five or more years (0.364% smokers vs 0.283% non-smokers) than in non-smoking males. In an evaluation of voice features comparing women who never smoked with women who smoked less

TABLE 2.3: Comparison of smokers' vs non-smokers' jitter.

Authors	Methods	Smokers		Non-smokers	
Gonzalez and Carpi [20]	Sustained vowels	Male 62.78	Female 55.11	Male 47.67	Female 45.6
Awan [53]	Sustained vowels	Female 0.40 ± 0.17		Female 0.37 ± 0.15	
Guimarães and Abberton [21]	% Jitter	Mixed		Mixed	
	/a/	1.1		0.52	
	/i/	0.86		0.47	
	/u/	0.73		0.51	
Chai et al. [54]	Sustained Vowels	Male 0.364		Male 0.283	
Vincent and Gilbert [55]	Sustained Vowels	< 10 years 0.92	\geq 10 years 1.11	Non-smokers 0.69	

than 10 years, and women who smoked 10 or more years [55] the investigators found that jitter value was increased in women who smoked compared to non-smokers, but only the jitter difference between non-smokers and smokers who smoked 10 or more years was significant (1.11% smoker \geq 10 years vs 0.92% smoker < 10 years vs 0.69% non-smoker). However, the authors also noted that the fact that women who had a longer smoking habit also smoked more cigarettes per day and were older than the other groups, could account for the difference in voice perturbation.

2.3.3 Shimmer

Shimmer (measured in decibels [dB] or % shimmer) is another common perturbation measure in the acoustic analysis, which is a measure of amplitude variation of a speech signal, and can be expressed as:

$$Shimmer = \frac{1}{N-1} \sum_{i=1}^{N-1} |20 \log(A_{i+1}/A_i)|. \quad (2.3)$$

where A_i is the extracted peak-to-peak amplitude of the i^{th} cycle of the speech signal and N is the number of extracted cycles of the speech signal [52].

TABLE 2.4: Comparison of smokers' vs non-smokers' shimmer.

Authors	Methods	Smokers		Non-smokers
Chai et al. [54]	Sustained vowels	Male		Male
		4.569		2.497
Vincent and Gilbert [55]	/a/ /i/ /u/	< 10 years	≥ 10 years	Female
		0.31	0.38	0.23
		0.2	0.34	0.2
		0.2	0.36	0.18
Zealouk et al. [56]	/a/ /i/ /u/	Male		Male
		0.648		0.355
		0.51		0.379
		0.551		0.401
Tuhanioglu et al. [57]	Sustained vowels % Shimmer Shimmer dB	cigarettes	e-cigarettes	Male
		3.81±2.71	2.60±0.95	2.67±0.83
		0.34±0.24	0.22±0.08	0.22±0.16

Studies have found higher shimmer values in smokers compared to non-smokers (Table 2.4). In one study [54], the percentage of shimmer was significantly higher in male smokers when compared to male non-smokers (4.57% vs 2.50%). Similarly, in another study [55], the shimmer was significantly higher for female smokers who smoked more than 10 years than for either non-smokers and smokers who smoked less than 10 years (0.37 dB smoker \geq 10 years vs 0.25 dB smoker $<$ 10 years vs 0.21 dB non-smoker). Zealouk et al. examined the voice features of 40 male adults, and 20 smokers with a median duration of 13 years [56]. Both jitter and shimmer values were significantly higher for smokers when compared to non-smokers (jitter: 51.997 ms smokers vs 36.989 ms non-smokers, $P < 0.05$; shimmer: 0.570 dB smokers vs 0.378 dB non-smokers, $P < 0.01$).

In a study [57] with 81 men, 21 of whom were former cigarette smokers that had been using e-cigarettes for one to three years, 30 were users of conventional cigarettes with a smoking history of one to five years, and 30 were non-smokers, the absolute shimmer was significantly different between conventional cigarette smokers and e-cigarette smokers and non-smokers, with an increased shimmer in the conventional

cigarette users (0.34 dB conventional cigarette vs 0.22 dB e-cigarette smokers vs 0.22 dB non-smokers), however, there was no significant difference between groups for F_0 or jitter.

2.3.4 Harmonics to Noise Ratio

Harmonics to noise ratio (HNR), expressed in dB, represents the degree of acoustic periodicity. It is the ratio between a periodic component and a non-periodic component of a speech, which is a measure that quantifies the amount of additive noise in the voice signal. HNR is also used as a measure for the signal-to-noise ratio (SNR) of a periodic signal to determine the voice quality [58]. HNR can be formulated as the following equation according to [59]:

$$HNR = 10 * \log_{10} \frac{AC_v(T)}{AC_v(0) - AC_v(T)}. \quad (2.4)$$

where $AC_v(0)$ is the autocorrelation coefficient at the origin consistent in all energy of the speech signal. The $AC_v(T)$ is the component of the autocorrelation corresponding to the fundamental period.

Although HNR has been labelled as an index of vocal ageing [60], studies have found that HNR was lower among smokers in comparison with non-smokers (Table 2.5). In one study [61], Braun found that the HNR value in the smoker group (9.4 dB) was lower than the non-smokers group (11.4 dB). Díaz et al. found that the amount of noise presented in the smokers' voices was evidently higher than the amount of noise in the voice of the non-smokers [62]. Studies have also found that the HNR value is affected by the duration of smoking. Pinar et al. [24] evaluated 109 young adult men among whom were 58 smokers (52 of these had smoked for less than ten years). The results indicated the smokers' HNR (25.01 dB) was slightly

TABLE 2.5: Comparison of smokers' vs non-smokers' HNR.

Authors	Methods	Smokers	Non-smokers
Braun [61]	Oral reading	Male	Male
	/a/	9.4	11.4
Díaz et al. [62]	Sustained	Mixed	Mixed
	vowels	25.22	36.1
Pinar et al. [24]	Sustained	Male	Male
	vowels	25.01	25.74
		Female	Female
Tafiadis et al. [63]	/a/	24.65	24.94
	/e/	25.3	25.65
Pintoa et al. [64]		Mixed	Mixed
	Sustained vowels	0.051	0.016

lower than non-smokers' (25.74 dB). In a study involving 210 young (average age 22 years) adult females attending smoking status identification [63], with average years of smoking only 2.16 (SD: 1.29) and average number of cigarettes smoked daily 13.19 (SD: 6.65), no statistically significant differences were noted for HNR values for smokers with short smoking years but heavy daily smoking habits compared to non-smokers. Pintoa and Crespob investigated HNR as features to classify smokers voice and non-smokers voice [64]. 40 smokers with an average smoking duration of 30 years and 40 non-smokers were measured, and the HNR value of smokers was different from that of non-smokers (0.051% vs 0.016%).

2.3.5 Formant Frequencies

A formant frequency is a concentration of acoustic energy around a particular frequency in the speech wave, which is a distinctive frequency component of the acoustic signal produced by speech [65]. A formant frequency with the lowest frequency is named F_1 , the second F_2 , the third F_3 , and the fourth F_4 . Each vowel in English

TABLE 2.6: Comparison of smokers' vs non-smokers' formant frequencies.

Authors	Methods	Smokers				Non-smokers			
		F_1	F_2	F_3	F_4	F_1	F_2	F_3	F_4
Zealouk et al. [56]	/a/	850	1600	2600	3500	900	2000	3050	4100
	/i/	400	1900	2700	3900	500	2100	3050	4300
	/u/	450	1500	2500	3700	550	1400	2950	4050

has three formant frequencies and most often, F_1 and F_2 are enough to determine a vowel.

Zealouk et al. [56] reported that smokers' formant frequencies F_1 and F_2 were close to those of non-smokers, and smokers' F_3 and F_4 were lower than that in non-smokers, as shown in Table 2.6. On the other hand, F_1 , F_2 , and F_3 values dramatically decreased with age increasing, and these values for men were lower than those for women.

2.3.6 Other Features

Pitch is the feature to judge sounds as its highness and lowness, which depends on the vibrational frequency produced by the vocal cords during sound production. Pitch can be quantified using fundamental frequency (F_0), as it is correlated with the physical feature of F_0 [66]. Both pitch and F_0 are often used interchangeably in the literature. Nonetheless, a few studies found that smokers had lower pitch values than those of non-smokers (Table 2.7) [21, 56]. In one study [56], Zealouk et al. found that the pitch value for smokers was statistically lower than non-smokers.

Correlation dimension (D_2) is a nonlinear dynamic quantitative measurement that can be applied to voice signals [67]. Chai et al. indicated that D_2 values were sensitive to cigarette smoking and smokers had significantly higher D_2 value than non-smokers [54].

TABLE 2.7: Comparison of smokers' vs non-smokers' other features.

Authors	Features	Smokers	Non-smokers
	Pitch (Hz)	Male	Male
Zealouk et al. [56]	/a/	143	168
	/i/	140	159
	/u/	147	164
		Male	Male
Chai et al. [54]	SNR	18.076	21.863
	ERR	0.324	0.03
	D_2	2.205	1.681

2.3.7 Effects of Smoking Cessation on Voice Features

Findings from a large longitudinal study suggest that changes in the fundamental frequency (F_0) are reversible when individuals quit smoking. Berg et al. evaluated voice frequency in 2274 adults aged 40-79 years classified as non-smokers, former smokers, and current smokers [68]. Regression analysis found significant differences in F_0 , but found only marginal differences between former smokers and non-smokers regardless of gender. F_0 was significantly lower in current smokers compared to non-smokers (103.7 Hz male smokers vs 112.5 Hz male non-smokers, $P < 0.001$; 159.1 Hz female smokers vs 170.7 Hz female non-smokers, $P < 0.001$) and former smokers (103.7 Hz male smokers vs 114.7 Hz male former smokers, $P < 0.001$; 159.1 Hz female smokers vs 166.1 Hz female non-smokers, $P < 0.001$). However, the study did not report on the abstinence duration. Therefore, it is unclear how soon changes in F_0 can be detected after quitting.

Three studies evaluated changes in voice features over short periods of abstinence. In the first of these studies [22], F_0 was measured before, during, and after a 40-hour period of abstinence in two smokers and two non-smokers. There was a small increase in F_0 for smokers after 40 hours abstinence, with no changes in F_0 for the control subjects. In the second study [69], voice features were measured before abstinence, one-week abstinence, and one-month abstinence in 18 smokers (5 female, 13 male).

On average, F_0 was higher during abstinence than before abstinence for both males (103.27 Hz pre-abstinence vs 107.08 Hz one-week vs 109.71 Hz one-month) and females (187.37 Hz pre-abstinence vs 192.91 Hz one-week vs 207.19 Hz one-month) but the difference was not significant. In the third study, Dirk and Braun also found a decrease in jitter across the abstinence period, but the difference was only significant between the pre-abstinence and one-month abstinence time points (0.50% pre-abstinence vs 0.21% one month, $P = 0.00954$). There was also a significant decrease in shimmer across the abstinence period, with the largest difference between the pre-abstinence and one-week abstinence time points (5.98% pre-abstinence vs 4.60% one week, $P = 0.03189$; 5.98% vs 4.64% one month, $P = 0.00988$). In another study of 20 female smokers (duration and quantity of smoking not given) before and after smoking cessation for 6 months, the results were compared with 40 age-matched non-smokers [70]. The results found that an increase in F_0 was present after Reinke's Edema microsurgery and 6-months of smoking cessation, but not fully reversible to normal voice quality (as in non-smokers) due to the vocal alterations caused by smoking.

2.4 Discussions

Fundamental frequency (F_0), jitter, and shimmer are the voice features that have been most used to analyse and discriminate between smokers and non-smokers. Harmonics-to-noise ratio (HNR) acts as a supplement for the smoking voice analysis tasks. A few papers have extracted formant frequency, pitch, and correlation dimension (D_2) as a measurement. Although there are a number of limitations in the literature (such as small samples, gender imbalance, and limited statistical analysis), there is sufficient evidence to support our contention that smoking consistently affects various voice features in specific ways. We also found evidence that the effects

of smoking on voice features such as F_0 , jitter, and shimmer may be reversed after a period of smoking cessation but are not fully reversible.

Overall, it appears that F_0 is affected by smoking in a relatively early stage of smoking history. A period of smoking abstinence would also affect F_0 , especially in women. Findings also suggest that jitter and shimmer may be sensitive to the duration of smoking history. Significant increases in both jitter and shimmer have been found in studies where subjects have a long history of smoking but are mixed in studies where subjects have a shorter history of smoking. Some studies also suggest that perturbation measures may be sensitive to the number of cigarettes smoked per day.

Table 2.8 below provides a summary of the strengths and the weaknesses of these main voice features that have been evaluated in relation to smoking. In addition, there is research to suggest a degree of specificity in detecting active versus passive smoking [51], and traditional cigarettes versus e-cigarettes, non-combustible tobacco, and water pipe smoking [57, 71, 72].

More work is needed to develop automated speech assessment for smoking status based on voice features as an alternative objective smoking status validation method to the point where a computer provided with the voice features extracted from speech recordings of smokers and non-smokers may be able to discriminate whether a given speech sample is from a smoker or non-smoker. Automated speech assessment for smoking status would have many advantages, including quantitative and objective assessment, able to be performed remotely, reducing analysis cost and time, and readily integrated into screening and remote health monitoring applications. In a future project, we will use F_0 , jitter, and shimmer to distinguish smokers from non-smokers by applying combinations of speech signal processing and machine learning techniques.

TABLE 2.8: A summary of voice features.

Features	Advantages	Disadvantages
Fundamental Frequency	Sensitivity to smoking, especially for heavy smokers	Affected by age and other factors
	Specificity for active smoking Reversal with long-term abstinence	Not a valid criterion to distinguish all smokers
Jitter	Sensitivity to smoking cessation	Changes slower over a longer period of time
	Specificity for combustible tobacco	Affected by voice disorders
Shimmer	Significantly affected by smoking	Sensitivity to the length of smoking history Sensitivity to the duration of smoking
	Sensitivity to smoking cessation Significantly affected by smoking	Changes slower over a longer period
Formant Frequencies	Sensitivity to smoking cessation	Changes with ageing
	Sensitivity to gender differences	Affected by age Degrades during signal transmission Affected by human noise

Although this literature review has revealed the effects of smoking on different voice features of speech, the relationship of these voice features with speaker smoking behaviour is complex. Smoking frequency, smoking history (duration and type of product, including the use of newer products such as e-cigarettes and heat-not-burn tobacco devices) need to be considered. Other factors, such as age, gender, presence of chronic respiratory disease, use of inhaled steroids, and alcohol use should also be accounted for. We found limited data on the length of time after quitting smoking that it takes to measure a change, and the level of smoking reduction required to create a change. An important issue is the lack of a reference database to establish the methodologies for smoking status classification from speech signal analysis. We aim to build a long-term smoking cessation voice recording corpus based on this

study and implement our voice-based smoking status validation model in a mobile application to provide an objective self-report measure method in smoking cessation trials and clinical practice.

2.5 Conclusions

In this study, we conducted a comprehensive investigation of the effects of voice features in the detection of smoker/non-smoker speech signals. This paper has presented a comparative review of smoker's voice features affected by smoking. We conclude that acoustic voice parameters appear to be influenced by smoking and smoking cessation: Fundamental frequency (F_0), jitter, shimmer, and harmonics to noise ratio (HNR) are affected by cigarette smoking. Smokers have a lower fundamental frequency than non-smokers in both gender and age groups. Smokers present higher jitter values for all vowels. Smokers' shimmer values are higher than the values of non-smokers. During smoking cessation, HNR value increases dramatically. Moreover, jitter and shimmer decrease significantly. F_0 value rises during smoking abstinence and decreases again after resuming smoking. However, more research with larger samples is needed to refine the sensitivity and specificity of this method to be able to translate it into a real-time tool.

This chapter has been published as follows:

Zhizhong Ma, Christopher Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. In the *Journal of Voice*, 2021. <https://doi.org/10.1016/j.jvoice.2020.12.014>

References

- [1] Mohammad H Forouzanfar, Ashkan Afshin, Lily T Alexander, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1659–1724, 2016.
- [2] Megan E Piper, Christopher Bullen, Suchitra Krishnan-Sarin, Nancy A Rigotti, Marc L Steinberg, Joanna M Streck, and Anne M Joseph. Defining and measuring abstinence in clinical trials of smoking cessation interventions: an updated review. *Nicotine and Tobacco Research*, 22(7):1098–1106, 2020.
- [3] Robert West, Peter Hajek, Lindsay Stead, and John Stapleton. Outcome criteria in smoking cessation trials: proposal for a common standard. *Addiction*, 100(3):299–303, 2005.
- [4] Kei Long Cheung, Dennis de Ruijter, Mickaël Hiligsmann, Iman Elfeddali, Ciska Hoving, Silvia MAA Evers, and Hein de Vries. Exploring consensus on how to measure smoking cessation. A Delphi study. *BMC Public Health*, 17(1):1–10, 2017.
- [5] Bożena Wiskirska-Woźnica and Waldemar Wojnowski. The smokers voice self assessment based on Voice Handicap Index. *Przegląd Lekarski*, 66(10):565–566, 2009.
- [6] Dionysios Tafiadis, Spyridon K Chronopoulos, Evangelia I Kosma, Louiza Voniati, Vasilis Raptis, Vasiliki Siafaka, and Nausica Ziavra. Using receiver operating characteristic curve to define the cutoff points of voice handicap index applied to young adult male smokers. *Journal of Voice*, 32(4):443–448, 2018.
- [7] Howard J Shaffer, Gabriel B Eber, Matthew N Hall, and Joni Vander Bilt. Smoking behavior among casino employees: Self-report validation using plasma cotinine. *Addictive Behaviors*, 25(5):693–704, 2000.

-
- [8] Neal L Benowitz, John T Bernert, Jonathan Foulds, Stephen S Hecht, Peyton Jacob III, Martin J Jarvis, Anne Joseph, Cheryl Oncken, and Megan E Piper. Biochemical verification of tobacco use and abstinence: 2019 update. *Nicotine and Tobacco Research*, 22(7):1086–1097, 2020.
- [9] Taneisha S Scheuermann, Kimber P Richter, Nancy A Rigotti, et al. Accuracy of self-reported smoking abstinence in clinical trials of hospital-initiated smoking interventions. *Addiction*, 112(12):2227–2236, 2017.
- [10] Jessica L Reid, David Hammond, Christian Boudreau, et al. Socioeconomic disparities in quit intentions, quit attempts, and smoking abstinence among smokers in four western countries: findings from the International Tobacco Control Four Country Survey. *Nicotine & Tobacco Research*, 12(1):20–33, 2010.
- [11] Thomas K Houston, Isabel C Scarinci, Sharina D Person, and Paul G Greene. Patient smoking cessation advice by health care providers: the role of ethnicity, socioeconomic status, and health. *American Journal of Public Health*, 95(6):1056–1061, 2005.
- [12] Rita Singh, Joseph Keshet, Deniz Gencaga, and Bhiksha Raj. The relationship of voice onset time and voice offset time to physical age. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5390–5394. IEEE, 2016.
- [13] David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5214–5218. IEEE, 2018.
- [14] Iosif Mporas and Todor Ganchev. Estimation of unknown speaker’s height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.
- [15] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

- [16] Amir Hossein Poorjam, Max A Little, Jesper Rindom Jensen, and Mads Græsbøll Christensen. A parametric approach for classification of distortions in pathological voices. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE, 2018.
- [17] Centers for Disease Control, Prevention, et al. How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general. *Publications and Reports of the Surgeon General*, 1(1):2145–2162, 2010.
- [18] Dario Marcotullio, Giuseppe Magliulo, and Tiziana Pezone. Reinke’s edema and risk factors: clinical and histopathologic aspects. *American Journal of Otolaryngology*, 23(2):81–84, 2002.
- [19] Dilyara G Yanbaeva, Mieke A Dentener, Eva C Creutzberg, Geertjan Wesseling, and Emiel FM Wouters. Systemic effects of smoking. *Chest*, 131(5):1557–1566, 2007.
- [20] Julio Gonzalez and Amparo Carpi. Early effects of smoking on the voice: a multidimensional study. *Medical Science Monitor*, 10(12):656, 2004.
- [21] Isabel Guimarães and Evelyn Abberton. Health and voice quality in smokers: an exploratory investigation. *Logopedics Phoniatrics Vocology*, 30(3-4):185–191, 2005.
- [22] Christopher H Murphy and Philip C Doyle. The effects of cigarette smoking on voice-fundamental frequency. *Otolaryngology—Head and Neck Surgery*, 97(4):376–380, 1987.
- [23] Shaheen N Awan and Danelle L Morrow. Videostroboscopic characteristics of young adult female smokers vs. nonsmokers. *Journal of Voice*, 21(2):211–223, 2007.
- [24] Dogan Pinar, Hakan Cincik, Evren Erkul, and Atila Gungor. Investigating the effects of smoking on young adult male voice by using multidimensional methods. *Journal of Voice*, 30(6):721–725, 2016.

- [25] Sarika Hegde, Surendra Shetty, Smitha Rai, and Thejaswi Dodderi. A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6):947–969, 2019.
- [26] Timothy J Wroge, Yasin Özkanca, Cenk Demiroglu, Dong Si, David C Atkins, and Reza Hosseini Ghomi. Parkinson’s disease diagnosis using machine learning and voice. In *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–7. IEEE, 2018.
- [27] Juntae Kim, Jaeseok Kim, Seunghyung Lee, Jinuk Park, and Minsoo Hahn. Vowel based voice activity detection with LSTM recurrent neural network. In *Proceedings of the 8th International Conference on Signal Processing Systems (ICSPS)*, pages 134–137, 2016.
- [28] R Johny Elton, P Vasuki, and J Mohanalin. Voice activity detection using fuzzy entropy and support vector machine. *Entropy*, 18(8):298, 2016.
- [29] Daria Hemmerling, Andrzej Skalski, and Janusz Gajda. Voice data mining for laryngeal pathology assessment. *Computers in Biology and Medicine*, 69(1):270–276, 2016.
- [30] Virgilijus Uloza, Antanas Verikas, Marija Bacauskiene, Adas Gelzinis, Ruta Pribuisiene, Marius Kasetas, and Viktoras Saferis. Categorizing normal and pathological voices: automated and perceptual categorization. *Journal of Voice*, 25(6):700–708, 2011.
- [31] Nafise Erfanian Saeedi, Farshad Almasganj, and Farhad Torabinejad. Support vector wavelet adaptation for pathological voice assessment. *Computers in Biology and Medicine*, 41(9):822–828, 2011.
- [32] Akira Sasou. Voice-pathology analysis based on AR-HMM. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.
- [33] Tan Lee, Yuanyuan Liu, Yu Ting Yeung, Thomas KT Law, and Kathy YS Lee. Predicting Severity of Voice Disorder from DNN-HMM Acoustic Posteriors. In

- the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 97–101. IEEE, 2016.
- [34] Fethi Amara, Mohamed Fezari, and Hocine Bourouba. An improved GMM-SVM system based on distance metric for voice pathology detection. *Applied Mathematics Information Sciences*, 10(3):1061–1070, 2016.
- [35] Ryszard Makowski and Robert Hossa. Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise. *Applied Acoustics*, 166(1):1073–1144, 2020.
- [36] Hui-Ling Chen, Gang Wang, Chao Ma, Zhen-Nao Cai, Wen-Bin Liu, and Su-Jing Wang. An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson’s disease. *Neurocomputing*, 184(1):131–144, 2016.
- [37] Ouhmida Asmae, Raihani Abdelhadi, Cherradi Bouchaib, et al. Parkinson’s disease identification using KNN and ANN algorithms based on voice disorder. In *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–6. IEEE, 2020.
- [38] Manoela Kohler, Marley MBR Vellasco, and Edson Cataldo. Analysis and classification of voice pathologies using glottal signal parameters. *Journal of Voice*, 30(5):549–556, 2016.
- [39] Christina Raichel Francis, Vrinda V Nair, and Salini Radhika. A scale invariant technique for detection of voice disorders using Modified Mellin Transform. In *2016 International Conference on Emerging Technological Trends (ICETT)*, pages 1–6. IEEE, 2016.
- [40] Rimah Amami and Abir Smiti. An incremental method combining density clustering and support vector machines for voice pathology detection. *Computers & Electrical Engineering*, 57(1):257–265, 2017.
- [41] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support

- vector machine-belief network architecture. In *2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 569–577. IEEE, 2004.
- [42] Shyh-Kuang Ueng, Cheng-Ming Luo, Tsung-Yu Tsai, and Hsuan-Chen Yeh. Human voice quality measurement in noisy environments. *Technology and Health Care*, 24(1):313–324, 2016.
- [43] David Talkin and W Bastiaan Kleijn. A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis*, 495(1):518, 1995.
- [44] Arturo Camacho and John G Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652, 2008.
- [45] Alain De Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.
- [46] Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663. IEEE, 2014.
- [47] R Fernández Liesa, D Damborenea Tajada, P Rueda Gormedino, E García y García, J Leache Pueyo, MA Campos del Alamo, E Llorente Arenas, and MJ Naya Gálvez. Acoustic analysis of the normal voice in nonsmoking adults. *Acta Otorrinolaringologica Espanola*, 50(2):134–141, 1999.
- [48] Jiangping Kong. A study on jitter, shimmer and f0 of mandarin infant voice by developing an applied method of voice signal processing. In *2008 Congress on Image and Signal Processing*, pages 314–318. IEEE, 2008.
- [49] Kumar Rakesh, Subhangi Dutta, and Kumara Shama. Gender Recognition using speech processing techniques in LABVIEW. *International Journal of Advances in Engineering & Technology*, 1(2):51, 2011.

- [50] David Sorensen and Yoshiyuki Horii. Cigarette smoking and voice fundamental frequency. *Journal of Communication Disorders*, 15(2):135–144, 1982.
- [51] Linda Lee, Joseph C Stemple, Diane Geiger, and Rebecca Goldwasser. Effects of environmental tobacco smoke on objective measures of voice production. *The Laryngoscope*, 109(9):1531–1534, 1999.
- [52] Mireia Farrús, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. In *8th Annual Conference of the International Speech Communication Association (ISCA)*, pages 778–781, 2007.
- [53] Shaheen N Awan. The effect of smoking on the dysphonia severity index in females. *Folia Phoniatrica et Logopaedica*, 63(2):65–71, 2011.
- [54] Lingying Chai, Alicia J Sprecher, Yi Zhang, Yufang Liang, Huijun Chen, and Jack J Jiang. Perturbation and nonlinear dynamic analysis of adult male smokers. *Journal of Voice*, 25(3):342–347, 2011.
- [55] Irena Vincent and Harvey R Gilbert. The effects of cigarette smoking on the female voice. *Logopedics Phoniatrics Vocology*, 37(1):22–32, 2012.
- [56] Ouissam Zealouk, Hassan Satori, Mohamed Hamidi, Naouar Laaidi, and Khalid Satori. Vocal parameters analysis of smoker using Amazigh language. *International Journal of Speech Technology*, 21(1):85–91, 2018.
- [57] Birgül Tuhanioglu, Sanem Okşan Erkan, Talih Özdaş, Çağrı Derici, Kemal Tüzün, and Özgül Akın Şenkal. The effect of electronic cigarettes on voice quality. *Journal of Voice*, 33(5):811–819, 2019.
- [58] Eiji Yumoto, Wilbur J Gould, and Thomas Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71(6):1544–1550, 1982.
- [59] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, pages 97–110. Citeseer, 1993.

- [60] Carole T Ferrand. Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice*, 16(4):480–487, 2002.
- [61] Angelika Braun. The effect of cigarette smoking on vocal parameters. In *Automatic Speaker Recognition, Identification and Verification*, 1994.
- [62] José Antonio Díaz, A Antonio Arroyo, and Howard B Rothman. Study and proposal of parameters for the objective assessment of voice quality in smokers. *Revista Ingenieria UC*, 21(3):7–16, 2014.
- [63] Dionysios Tafiadis, Eugenia I Toki, Kevin J Miller, and Nausica Ziavra. Effects of early smoking habits on young adult female voices in Greece. *Journal of Voice*, 31(6):728–732, 2017.
- [64] Aline Gomes Lustosa Pinto, Agrício Nubiato Crespo, and Lucia Figueiredo Mourão. Influence of smoking isolated and associated to multifactorial aspects in vocal acoustic parameters. *Brazilian Journal of Otorhinolaryngology*, 80(1):60–67, 2014.
- [65] Ralph O Coleman. Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research*, 14(3):565–577, 1971.
- [66] David Gerhard. *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, SK, Canada, 2003.
- [67] Yu Zhang, Jack J Jiang, Stephanie M Wallace, and Liang Zhou. Comparison of nonlinear dynamic methods and perturbation methods for voice analysis. *The Journal of the Acoustical Society of America*, 118(4):2551–2560, 2005.
- [68] Martin Berg, Michael Fuchs, Kerstin Wirkner, Markus Loeffler, Christoph Engel, and Thomas Berger. The speaking voice in the general population: Normative data and associations to sociodemographic and lifestyle factors. *Journal of Voice*, 31(2):257–263, 2017.
- [69] Louise Dirk and Angelika Braun. Voice parameter changes in smokers during

- abstinence from cigarette smoking. In *the International Congress of Phonetic Sciences (ICPhS)*, pages 588–590, 2011.
- [70] Regina Helena Garcia Martins, Elaine Lara Mendes Tavares, and Adriana Bueno Benito Pessin. Are vocal alterations caused by smoking in Reinke’s edema in women entirely reversible after microsurgery and smoking cessation? *Journal of Voice*, 31(3):380–386, 2017.
- [71] Marie Reine Ayoub, Pauline Larrouy-Maestri, and Dominique Morsomme. The effect of smoking on the fundamental frequency of the speaking voice. *Journal of Voice*, 33(5):802–811, 2019.
- [72] Abdul-latif Hamdan, Abla Sibai, Dima Oubari, Jihad Ashkar, and Nabil Fuleihan. Laryngeal findings and acoustic changes in hubble-bubble smokers. *European Archives of Otorhinolaryngology*, 267(10):1587–1592, 2010.

Chapter 3

Automatic Speech-based Smoking Status Identification

Previous research on smoking status identification mainly focuses on employing the speaker's low-level acoustic features such as fundamental frequency (F_0), jitter, and shimmer. However, the use of high-level acoustic features, such as Mel Frequency Cepstral Coefficients (MFCC) and filter bank (Fbank) for smoking status identification, has rarely been explored. In this study, we utilise both high-level acoustic features (i.e., MFCC, Fbank) and low-level acoustic features (i.e., F_0 , jitter, shimmer) for smoking status identification. Furthermore, we propose a deep neural network method for smoking status identification by employing ResNet along with both high-level and low-level acoustic features. We also apply a data augmentation technique in smoking status identification to further improve the performance. Finally, we present a comparison of identification accuracy results for each feature setting, and obtain the best accuracy of 82.3%, a relative improvement of 29.8% on the rule-based method.

3.1 Introduction

Automatic smoking status identification used is to identify a speaker's smoking status by extracting and analysing the acoustic features that can be affected by cigarette smoking based on the spoken utterances. Speech signals carry a speaker's basic information, such as age, gender, emotional status, psychological status, intoxication level, and smoking status [1]. Compared to traditional biochemical smoking status validation methods (such as biochemical testing of urine or saliva for the nicotine metabolite cotinine, or exhaled breath carbon monoxide) and on-site speech assessments operated by experts, automatic smoking status identification from speech signal is a simple, non-invasive, low-cost method that can be applied across a large population and does not require face-to-face contact.

Automatic smoking status identification has a variety of applications such as smoking status validation, smoking cessation tracking, and speaker profiling. Smoking cessation tracking applications are implicitly or explicitly employing smoking status information to record users' quit smoking timelines. In speaker profiling systems, knowledge of smoking status can be utilised for the normalisation of acoustic features to increase the system's performance. In general, automatic smoking status identification from speech is essential for improving the flexibility of smoking status validation and the performance of speaker profiling systems.

There is a rich literature on the effects of cigarette smoking on a smoker's throat tissues, including their vocal cords [2-5]. Smoking can also degrade lung function by decreasing the airflow through the smoker's vocal cords [6-10]. The signs of laryngeal irritation and disturbed phonatory physiology caused by smoking occur even in young smokers and affects women's voices more than men's voices [4, 8, 11, 12]. Changes in the vocal tract can result in a significant variation in the speaker's speech signals. Previous studies on smoking status identification have concluded

that there is a relationship between a smoker’s speech signals and the corresponding smoking status.

Research shows that the primary acoustic features affected by smoking are fundamental frequency (F_0), jitter, and shimmer [10, 13–15]. The typical method to identifying smoking status has focused on the low-level acoustic features (e.g., F_0 , jitter, and shimmer), such as mean, maximum, minimum, and standard deviation (SD) from the on-site speech assessment including sustained vowels, oral reading, and spontaneous speech tasks. Recent research focused on adopting high-level acoustic features such as Mel-Frequency Cepstral Coefficient (MFCC) as the input in smoking status identification models [16].

Recently, the performance of audio classification tasks such as emotion recognition [17], acoustic event detection [18] and speaker verification [19] has been improved by using a specific type of deep neural networks (DNNs) - Residual Network (ResNet). ResNet [20] was initially designed for image classification and has shown more reliable performance than shallower convolutional neural network (CNN) architectures. Inspired by Google’s recent work on audio classification [18], we adapted ResNet for smoking status identification. To the best of our knowledge, our work is the first use of ResNet in smoking status identification.

Our contributions include (i) the combination of both high- and low-level acoustic features used for automatic speech-based smoking status identification for the first time; (ii) deep learning is used for automatic speech-based smoking status identification for the first time; and (iii) we developed a new smoking status identification dataset based on two existing corpora.

This paper is organised hereon as follows: Section 3.2 introduces the various acoustic features for smoking status identification. Section 3.3 explains our proposed method. Section 3.4 describes the dataset we used and explains our experimental

setup. Section 3.5 presents our experimental results. Finally, the conclusion and future directions are described in Section 3.6.

3.2 Acoustic Features for Smoking Status Identification

The acoustic features are the acoustic components present in a speech that are capable of being experimentally observed, recorded, and reproduced. The following features will be used in our method, which includes both high- and low-level acoustic features.

3.2.1 MFCC and Fbank

Mel-Frequency Cepstral Coefficient (MFCC) and filter bank (Fbank) are two standard high-level acoustic features that are widely utilised in audio classification tasks [21–23] and typically develop from a sub-band spectrum.

MFCC is a method for converting the real cepstral of a windowed short-time speech signal derived from the Fast Fourier Transform (FFT) technique into parameters according to the Mel Scale [24]. It represents short-term spectral features of a speech signal [25].

Filter bank (Fbank) feature is a common alternative to MFCC [26], and has become a trend in acoustic feature learning for very deep neural networks because it contains additional information such as short-range temporal correlations [27].

3.2.2 Fundamental Frequency

The fundamental frequency (F_0) is an important low-level acoustic feature of speech signals. F_0 is the lowest, and typically the strongest frequency produced by the complex vocal fold vibrations measured in Hertz (Hz).

Typical F_0 values captured in the speech signal were 120 Hz for men and 210 Hz for women [3]. Studies have consistently shown that lower F_0 values existed in smokers in comparison to the age and sex-matched non-smokers. In [7], F_0 was assessed through oral reading and spontaneous speech for 80 individuals, half of whom were classified as smokers. The results indicated that the average F_0 values for smokers were lower than for non-smokers. However, the differences between the F_0 values of the female smokers and female non-smokers (182.70 Hz smokers vs 186.45 Hz non-smokers) were not as significant as the male group (105.65 Hz smokers vs 115.95 Hz non-smokers), but the same trends were noted.

Guimarães et al. [5] selected 32 adult subjects (20 smokers and 12 non-smokers) based on their age, gender, and smoking history. The smokers were aged between 27 and 51 years, with a mean age of 37 years. The non-smokers ranged in age from 20 to 51 years, with a mean age of 32 years. The smokers in the study were all regular smokers when they underwent the speech assessment. With the exception of one subject, who had quit smoking ten years earlier, all non-smokers had a non-smoking history. The speech assessment included oral reading tasks, sustained vowels tasks and conversation tasks. The results indicate that a lower mean F_0 value for all speech assessments was found for the smoker group.

3.2.3 Jitter

Jitter (measured in microseconds or % jitter) is a common low-level acoustic feature used in the smoking status validation. It is a measure of the cycle-to-cycle frequency variation or instability of a speech signal, which is mainly affected by the lack of control of vocal fold vibrations.

Many studies have shown higher jitter values in smokers than non-smokers. In [28], male non-smokers had substantially lower jitter values than male smokers who smoked a minimum of five cigarettes per day for five years or longer (0.364% smokers vs 0.283% non-smokers). Gonzalez and Carpi [4] indicated differences in jitter between male non-smokers and male smokers who had been smoking for less than 10 years (47.67 μ s non-smokers vs 62.78 μ s smokers), implying that changes in jitter are also associated with long-term smoking. A more recent study [29] found that smoking women aged 18-24 years had a higher jitter value than non-smoking women. However, the jitter difference was not significant due to the smokers' smoking history being relatively short (3.5 years on average).

In [6], an increasing trend of jitter was found in female smokers compared to female non-smokers, but there were also differences between smokers who had smoked for more than 10 years and those who had smoked for less than 10 years (1.11% smoker ≥ 10 years vs 0.92% smoker < 10 years vs 0.69% non-smoker). However, the authors also observed that the women with a longer smoking habit smoked more cigarettes per day and were older than the other groups, which might explain the difference in voice perturbation.

3.2.4 Shimmer

Another common low-level acoustic feature used in the smoking status analysis is shimmer (measured in decibels [dB] or % shimmer), a measure of amplitude instability of the sound wave. Studies have also found smokers have higher shimmer values than non-smokers [6, 10, 28].

When compared to male non-smokers, male smokers had a considerably higher shimmer (4.57% smokers vs 2.50% non-smokers) [28]. Likewise, the shimmer was substantially higher for female smokers who had smoked for more than 10 years than for either non-smokers and smokers who had smoked for less than 10 years (0.37 dB smokers \geq 10 years vs 0.25 dB smokers $<$ 10 years vs 0.21 dB non-smokers) [6]. Zealouk et al. [10] studied the vocal characteristics of 40 male subjects, 20 of whom were smokers with an average smoking history of 13 years. Smokers had substantially higher shimmer values than non-smokers (0.570 dB smokers vs 0.378 dB non-smokers).

3.3 Methodology

The smoking status identification task is typically considered as an audio classification problem in the speech processing domain. Previous studies utilised either the low-level acoustic features (e.g., fundamental frequency (F_0), jitter, and shimmer) based on the on-site speech assessments [10, 13–15] or the high-level acoustic features such as Mel-Frequency Cepstral Coefficient (MFCC) with an i-vector framework that was designed for speaker recognition tasks [16]. However, no studies have combined both low- and high-level acoustic features for smoking status identification. Furthermore, the deep neural network (DNN) method has not been used to model smoking status information from speech signals as far as we know. Our proposed method

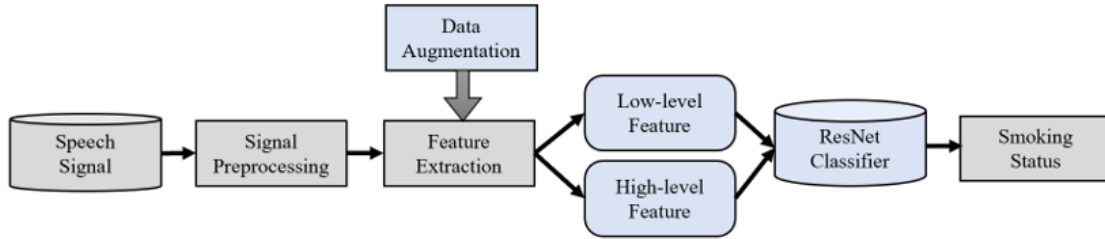


FIGURE 3.1: The architecture of our proposed automatic smoking status identification method.

utilises both low- and high-level acoustic features to distinguish smokers from non-smokers along with deep neural network techniques. Our proposed architecture, as illustrated in Figure 3.1, takes a speech recording as the input, and subsequently passes it to the signal preprocessing module. Speech signal preprocessing is utilised to reduce the influence of acoustic noise and silence on acoustic feature extraction, allowing for more accurate identification of the smoking status output. For further processing, the feature extraction module extracts acoustic features such as MFCC, filter bank (Fbank), F_0 , jitter and shimmer from speech signal inputs. Data augmentation (i.e., SpecAugment [30]) is utilised in the feature extraction process to increase the diversity of the training set and further improve robustness.

In our method, we use ResNet-18 as the deep learning network to be trained with the labeled data and obtain a classifier to determine the smoker/non-smoker label of the unlabeled input test speech data (for more details please refer to below Section 3.4.2). We chose ResNet-18 rather than other variants of ResNet (e.g., ResNet-34, ResNet-50, ResNet-101) because our dataset is a relatively small dataset that contains approximately 45 hours of speech data and ResNet-18 can provide a better trade-off between layers and performance for such dataset.

3.4 Experiments

3.4.1 Datasets

In the absence of large-scale, well-designed datasets specifically for smoking status identification experiments, we collected and created a new smoking status identification dataset based on two corpora, which are available at the Linguistic Data Consortium (LDC): (1) the Mixer 4 and 5 Speech Corpus [31]; and (2) the Mixer 6 Speech Corpus [32]. The speech recordings in the Mixer 4 and 5 Speech Corpus were used in 2008 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE)¹. The speech recordings in the Mixer 6 Speech Corpus were used in 2010 NIST SRE².

Both corpora comprise recordings made via the public telephone network and multiple microphones in office-room settings. The main difference in the setting is that most of the 616 distinct speakers in the Mixer 4 and 5 Speech Corpus have English as their native language, and the 594 distinct speakers in the Mixer 6 Speech Corpus all have English as their native language.

In Mixer 4 and 5 Speech Corpus, only 89 of 616 speakers have valid smoking status labels. There are 40 female smokers, 8 female non-smokers, 37 male smokers, and 4 male non-smokers. In Mixer 6 Speech, 589 of 594 speakers have valid smoking status labels. There are 48 female smokers, 252 female non-smokers, 70 male smokers, and 219 male non-smokers. For balanced data training purposes, 200 speakers (50 female smokers, 50 female non-smokers, 50 male smokers, and 50 male non-smokers) from both corpora jointly are selected for experiments. Most of the speakers have two to three 12 mins transcripts reading audio segments; a few of them only have one 12 mins transcript reading audio segment. We split the training set, validation set and

¹<https://catalog.ldc.upenn.edu/LDC2020S03>

²<https://catalog.ldc.upenn.edu/LDC2013S03>

TABLE 3.1: Speech features statistics divided by smoking status and gender.

		F_0 (Hz)	Jitter (μ s)	Shimmer (%)
Male Smokers	Min	92.528	23.422	5.642
	Max	220.618	59.754	13.624
	Mean	108.287	35.199	8.979
	SD	25.749	1.337	4.325
Male Non-smokers	Min	97.183	20.278	4.971
	Max	249.035	47.925	12.48
	Mean	116.592	24.734	6.491
	SD	25.153	0.942	2.903
Female Smokers	Min	124.078	29.351	8.472
	Max	277.399	53.201	13.172
	Mean	181.021	33.635	11.716
	SD	22.572	1.473	2.673
Female Non-smokers	Min	126.334	20.653	5.416
	Max	297.481	39.714	11.669
	Mean	210.359	23.786	7.695
	SD	21.437	1.274	2.351

test set following the 8:1:1 ratio. We chose 5 female smokers, 5 female non-smokers, 5 male smokers, 5 male non-smokers as the test set and the rest of the speakers as the training set. The fundamental frequency (F_0), jitter, and shimmer statistics for smokers and non-smokers in the training set are shown in Table 3.1.

3.4.2 Implementation Details

The input features are either 40-dimensional MFCC features or 40-dimensional log Mel-filterbank features with a frame-length of 40 ms with 50% overlap. We extracted the fundamental frequency (F_0) of each speech in the dataset using Praat [33], an open-source toolbox. Jitter and shimmer were calculated upon frames of 40 ms with a time-shift of 20 ms by using the DisVoice toolkit [34].

Before being fed into the ResNet, the input features are mean-normalised along the time-axis, and nonspeech frames (silent speech frame) are removed using an energy-based voice activity detection (VAD) method. We chose SpecAugment in

this study because it is a novel data augmentation method that is applied directly to the feature inputs of a neural network (i.e., MFCC, Fbank). Other traditional data augmentation methods that deformed the raw waveform by speeding it up or slowing it down are not suitable for smoking status identification. In training, we use ResNet-18 with 16-32-64-128 channels for each residual block. The model with the best validation loss was selected for testing. In testing, the entire speech is evaluated at once.

The models are implemented using PyTorch [35] and optimised by a stochastic gradient descent (SGD) optimiser [36] with a momentum of 0.9. The mini-batch size is 64, and the weight decay parameter is 0.0001. We set the initial learning rate to 0.1 and decay it by a factor of 10 until convergence. All the models were trained for 100 epochs.

3.4.3 Evaluation Metrics

In order to validate our proposed method, we evaluate our proposed method using two metrics, which include accuracy and F_1 -score.

The confusion matrix for smoking identification uses a 2×2 matrix where one axis of the matrix is the predicted class (smoker, non-smoker) and the actual class. Each box in the matrix shows the number of True Smokers (TS), True Non-smokers (TN), False Smokers (FS), and False Non-smokers (FN). Accuracy (ACC) is given by the following equation (3.1):

$$\text{ACC} = \frac{TS + TN}{TS + TN + FS + FN}. \quad (3.1)$$

Accuracy can be further analysed as precision (3.2) and recall (3.3). High precision indicates a low degree of false positives, while high recall indicates a high degree of

class recognition. The following equations are examples of precision and recall for the validation of smokers.

$$\mathbf{Precision} = \frac{TS}{TS + FS}. \quad (3.2)$$

$$\mathbf{Recall} = \frac{TS}{TS + FN}. \quad (3.3)$$

The F_1 -score can be used to evaluate the accuracy of each given class label using the following equation:

$$\mathbf{F}_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (3.4)$$

3.5 Results and Discussions

According to the statistics of the acoustic features in Table 3.1 above, there is a considerable variation within fundamental frequency (F_0), jitter, and shimmer between smokers and non-smokers for both genders. The mean F_0 , jitter, and shimmer values show that the most significant difference is between smokers and non-smokers.

We developed the following rule to act as our baseline, based on the difference within the mean F_0 , jitter, and shimmer between smokers and non-smokers: if the mean F_0 of the speaker is closer to the average F_0 of male smokers than to the mean F_0 of female smokers, and if both means of jitter and shimmer of the speaker are closer to the mean jitter and shimmer of smokers than to the non-smokers, we identify the speaker as a smoker. Otherwise, we identify the speaker as a non-smoker. We employed this simple classification rule to classify our test dataset, and it achieved an accuracy of 63.4%. Although this rule requires knowledge of the ground truth

TABLE 3.2: Smoking status identification experiment results.

Features	Accuracy	F ₁ -score
Rule-based	0.634	0.617
MFCC		
w/o SpecAugment	0.714	0.714
with SpecAugment	0.734	0.745
Fbank		
w/o SpecAugment	0.73	0.724
with SpecAugment	0.77	0.766
MFCC + F_0 + jitter + shimmer		
w/o SpecAugment	0.754	0.754
with SpecAugment	0.769	0.765
Fbank + F_0 + jitter + shimmer		
w/o SpecAugment	0.787	0.795
with SpecAugment	0.823	0.823

mean F_0 for smokers and non-smokers, it indicates that a simple rule may identify smoking status from speech signals.

In the rest of Table 3.2, the models are trained on the ResNet-18 described in Section 3.4.2. The experimental results are presented with and without (w/o) the SpecAugment in two types of feature settings (high-level acoustic features only and a combination of both high- and low-level acoustic features) as inputs.

We obtained better smoking status identification accuracy and F_1 -score results by using the data augmentation technique for different acoustic feature settings. MFCC with SpecAugment achieved a relative improvement of 2.8% than without SpecAugment. Fbank with SpecAugment achieved a relative improvement of 5.5% than without SpecAugment.

On the other hand, we can see that Fbank always yielded better performance than MFCC with or without other acoustic features. Without SpecAugment settings, Fbank outperformed MFCC either by using itself or jointly with acoustic features. A combination of Fbank with SpecAugment and low-level acoustic features (i.e., F_0 ,

jitter, and shimmer) provides the best accuracy of 82.3%, which is a relative improvement of 12.7% and 29.8% on the initial Fbank-only method (without SpecAugment and low-level acoustic features) and rule-based method, respectively.

3.6 Conclusions

Based on our experimental result, it is indicated that Fbank outperforms MFCC if we only utilise high-level acoustic features. We have demonstrated for the first time that the combination of both high- and low-level acoustic features along with the deep neural network technique can achieve high performance in smoking status identification. The data augmentation technique (i.e., SpecAugment) can further improve the smoking status identification accuracy. The proposed automatic smoking status identification model could be an alternative solution to obtain an accurate and objective smoking status when the biological verification methods are not feasible.

Our proposed method has outperformed the rule-based method and obtained the best accuracy of 82.3%, which is a relative improvement of 12.7% and 29.8% on the initial high-level acoustic features only method (without data augmentation and low-level acoustic features) and rule-based method, respectively.

In future, we will build a long-term smoking status related speech recording corpus. Additional features such as age, gender, smoking history and smoking frequency will also be considered in the data collection and smoking status identification process. We will also explore the smoking status identification deep neural network model to further improve performance.

This chapter has been published as follows:

Zhizhong Ma, Satwinder Singh, Yuanhang Qiu, Feng Hou, Ruili Wang, Christopher Bullen and Joanna Ting Wai Chu. Automatic speech-based smoking status identification,. In the Computing Conference, 2022. (Accepted)

References

- [1] Amir Hossein Poorjam and Mohamad Hasan Bahari. Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 7–12. IEEE, 2014.
- [2] Christopher H Murphy and Philip C Doyle. The effects of cigarette smoking on voice-fundamental frequency. *Otolaryngology—Head and Neck Surgery*, 97(4):376–380, 1987.
- [3] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. *Stockholms Universitet*, 11, 1995.
- [4] Julio Gonzalez and Amparo Carpi. Early effects of smoking on the voice: a multidimensional study. *Medical Science Monitor*, 10(12):656, 2004.
- [5] Isabel Guimarães and Evelyn Abberton. Health and voice quality in smokers: an exploratory investigation. *Logopedics Phoniatrics Vocology*, 30(3):185–191, 2005.
- [6] Irena Vincent and Harvey R Gilbert. The effects of cigarette smoking on the female voice. *Logopedics Phoniatrics Vocology*, 37(1):22–32, 2012.
- [7] Horii and Sorenson. Cigarette smoking and voice fundamental frequency. *Journal of Communication Disorders*, 15(2):135–144, 1982.
- [8] Shaheen N Awan and Danelle L Morrow. Videostroboscopic characteristics of young adult female smokers vs. nonsmokers. *Journal of Voice*, 21(2):211–223, 2007.

- [9] Louise Dirk and Angelika Braun. Voice parameter changes in smokers during abstinence from cigarette smoking. In *the International Congress of Phonetic Sciences (ICPhS)*, pages 588–590, 2011.
- [10] Ouissam Zealouk, Hassan Satori, Mohamed Hamidi, Naouar Laaidi, and Khalid Satori. Vocal parameters analysis of smoker using Amazigh language. *International Journal of Speech Technology*, 21(1):85–91, 2018.
- [11] Dogan Pinar, Hakan Cincik, Evren Erkul, and Atila Gungor. Investigating the effects of smoking on young adult male voice by using multidimensional methods. *Journal of Voice*, 30(6):721–725, 2016.
- [12] Susanna Simberg, Hanna Udd, and Pekka Santtila. Gender differences in the prevalence of vocal symptoms in smokers. *Journal of Voice*, 29(5):588–591, 2015.
- [13] Linda Lee, Joseph C Stemple, Diane Geiger, and Rebecca Goldwasser. Effects of environmental tobacco smoke on objective measures of voice production. *The Laryngoscope*, 109(9):1531–1534, 1999.
- [14] Angelika Braun. The effect of cigarette smoking on vocal parameters. In *Automatic Speaker Recognition, Identification and Verification*, 1994.
- [15] Zhizhong Ma, Christopher Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. *Journal of Voice*, 36(6), 2021.
- [16] Amir Hossein Poorjam, Soheila Hesaraki, Saeid Safavi, Hugo van Hamme, and Mohamad Hasan Bahari. Automatic smoker detection from telephone speech signals. In *International Conference on Speech and Computer*, pages 200–210. Springer, 2017.
- [17] Siqu Han, Feng Leng, and Zitong Jin. Speech emotion recognition with a ResNet-CNN-Transformer parallel neural network. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*,

- pages 803–807. IEEE, 2021.
- [18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, Malcolm Slaney, Ron J Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.
- [19] Ying Liu, Yan Song, Ian McLoughlin, Lin Liu, and Li-rong Dai. An effective deep embedding learning method based on dense-residual networks for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6683–6687. IEEE, 2021.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [21] Jie Pu, Yannis Panagakis, and Maja Pantic. Learning separable time-frequency Filterbanks for audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3000–3004. IEEE, 2021.
- [22] Takuya Fujioka, Takeshi Homma, and Kenji Nagamatsu. Meta-learning for speech emotion recognition considering ambiguity of emotion labels. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2332–2336, 2020.
- [23] Raphael Tang and Jimmy Lin. Deep residual learning for small-footprint keyword spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5484–5488. IEEE, 2018.
- [24] Namrata Dave. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology*, 1(6):1–4, 2013.
- [25] Vinay Kumar Mittal and B Yegnanarayana. Production features for detection of

- shouted speech. In *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, pages 106–111. IEEE, 2013.
- [26] Geoffrey Hinton, Li Deng, Dong Yu, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- [27] Takuya Yoshioka, Anton Ragni, and Mark JF Gales. Investigation of unsupervised adaptation of DNN acoustic models with filter bank input. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE, 2014.
- [28] Lingying Chai, Alicia J Sprecher, Yi Zhang, Yufang Liang, Huijun Chen, and Jack J Jiang. Perturbation and nonlinear dynamic analysis of adult male smokers. *Journal of Voice*, 25(3):342–347, 2011.
- [29] Shaheen N Awan. The effect of smoking on the dysphonia severity index in females. *Folia Phoniatrica et Logopaedica*, 63(2):65–71, 2011.
- [30] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.
- [31] Linda Brandschain, Christopher Cieri, David Graff, Abby Neely, and Kevin Walker. Speaker recognition: Building the Mixer 4 and 5 corpora. In *the International Conference on Language Resources and Evaluation*. Citeseer, 2008.
- [32] Linda Brandschain, D Graff, C Cieri, K Walker, C Caruso, and A Neely. The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition. In *the International Conference on Language Resources and Evaluation*, 2010.
- [33] Paul Boersma. Praat, a system for doing phonetics by computer. *Glott International*, 5(9):341–345, 2001.
- [34] Juan Camilo Vásquez-Correa, JR Orozco-Arroyave, T Bocklet, and E Nöth.

- Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of Communication Disorders*, 76(1):21–36, 2018.
- [35] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- [36] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Chapter 4

Best Acoustic Features for Smoking Status Identification

Speech-based automatic smoking status identification (also known as smoker/non-smoker classification) aims to identify speakers' smoking status from their speech. This study focuses on determining the best acoustic features for smoking status identification. In this paper, we investigate the performance of four acoustic feature sets/representations that were extracted using three feature extraction/learning techniques: (i) hand-crafted feature sets including the extended Geneva Minimalistic Acoustic Parameter Set and the Computational Paralinguistics Challenge Set; (ii) the Bag-of-Audio-Words representations; and (iii) the neural representations extracted from raw waveform signals by SincNet. Experimental results show that: (i) SincNet feature representations are the most effective for smoking status identification and outperform the MFCC baseline features by 16% in absolute accuracy; (ii) the performance of hand-crafted feature sets and the Bag-of-Audio-Words representations rely on the scale of the dimensions of feature vectors.

4.1 Introduction

Speech-based automatic smoking status identification (also known as smoker/non-smoker classification) aims to identify a speaker’s smoking status from his or her speech data. Automatic smoking status identification has a variety of applications including smoking status validation [1], smoking cessation tracking [2] and speaker profiling [3]. Speech-based smoking status identification has advantages over traditional biochemical measures for determining if an individual has successfully stopped smoking, because of the costs and the ease of the sample collection process. Speech-based automatic smoking status identification is especially useful for smoking cessation research under the current COVID-19 pandemic where movement restrictions may make other methods more difficult or expensive than usual. Many studies have shown that cigarette smoking negatively affects smokers’ vocal tissues and permanently alters the acoustic properties of smokers’ speech compared with non-smokers [4–7]. Such alterations are confirmed by assessing acoustic features like fundamental frequency (F_0), jitter and shimmer [8–11]. There has been some previous work in the speech-based automatic smoking status identification field [3, 12]. In these two papers, the authors utilised Mel Frequency Cepstral Coefficients (MFCC) to identify smokers from their spontaneous speech. Recently, in the speech-health analysis related field, hand-crafted feature sets and learned neural representations, which were not considered for smoking status identification, have proven to be more effective acoustic features than MFCC [13–16].

The hand-crafted feature sets (e.g., eGeMAPS [17] and ComParE [18]) and the Bag-of-Audio-Words (BoAW) [19] representations have been used successfully for speech-health analysis related tasks [13, 14, 16, 20]. Furthermore, learning task-driven features directly from the raw waveform by deep neural networks (DNNs) has proven to be an effective feature extractor [21] for a variety of applications, such as speech recognition [22], speaker recognition [23] and emotion recognition [24].

For example, SincNet [23] is a Convolutional Neural Network (CNN) for learning feature representations from raw waveforms. Compared with hand-crafted feature sets, SincNet is more effective in learning the most suitable feature representations for the given tasks [15, 23].

We hypothesise that the quality of the acoustic features is crucial for the performance of speech-based smoking status identification systems. In this study, we aim to identify speakers' smoking status by using more advanced feature extraction/learning techniques. We compare the four acoustic feature sets/representations extracted/learned by using three feature extraction/learning techniques: (i) hand-crafted feature sets, i.e., eGeMAPS and ComParE; (ii) the BoAW representations quantising acoustic low-level descriptors (LLDs); and (iii) the neural representations extracted from raw waveform signals by SincNet.

However, there are just a few publicly available datasets for smoking status identification tasks. The dataset we utilise is derived from the two corpora (i.e., the Mixer 4 and 5 Speech Corpus [25], and the Mixer 6 Speech Corpus [26]) which include rich metadata regarding speakers' smoking status, age, height, weight, etc., making them applicable for smoking status identification experiments.

The main contributions of our paper are as follows:

- (i) We identify that the most effective acoustic features are the feature representations learned by using deep neural networks.
- (ii) We compare the effectiveness and generalisability of acoustic features extracted by using three different feature extraction/learning techniques for smoking status identification.
- (iii) We propose a new dataset for smoking status identification experiments based on two existing corpora.

TABLE 4.1: Summary of feature sets/representations utilised in this study.

Name	Type	No. of Features
eGeMAPS	Hand-crafted	88
ComParE	Hand-crafted	6373
BoAW	Hand-crafted+BoAW	1000
SincNet	Raw Waveform	2048

The rest of this paper is structured as follows. Section 4.2 presents the related work of the acoustic feature sets/representations we utilise in this paper. Section 4.3 presents our dataset. The design and methods are provided in Section 4.4. Section 4.5 describes the experimental results, and the conclusions and proposals for future work are discussed in Section 4.6.

4.2 Related Work

4.2.1 extended Geneva Minimalistic Acoustic Parameter Set

extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) is a low-dimensional, frame-level, knowledge-inspired acoustic feature set containing a wide range of standardised relevant acoustic features [17]. eGeMAPS is extracted on two levels: (i) low-level descriptors (LLDs); and (ii) statistical functionals.

eGeMAPS includes 88 acoustic features derived from 23 LLDs that cover spectral, cepstral, prosodic and voice quality information of the speech, as shown in Table 4.1. The efficiency of eGeMAPS has been proven successful in various areas of clinical and paralinguistic speech analysis, including Alzheimer’s Dementia detection [13], speech intelligibility assessment [14] and speech emotion recognition [27].

4.2.2 Computational Paralinguistics Challenge set

Computational Paralinguistics Challenge set (ComParE) is a well-evolved, high-dimensional brute-forced acoustic feature set that is extracted on three levels: (i) low-level descriptors (LLDs); (ii) statistical functionals; and (iii) LLDs deltas [18]. It contains 6373 static features resulting from the computation of various functionals over 65 LLDs. ComParE consists of fundamental frequency (F_0), energy, spectral, cepstral coefficients (MFCCs) and voicing related frame-level features. It also includes zero-crossing rate, jitter, shimmer, harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. The statistical functionals applied to the LLDs include the mean, standard deviation, percentiles and quartiles, linear regression functionals, and local minima/maxima related functionals.

The ComParE feature set has demonstrated its ability and robustness for capturing acoustic information in many speech-health analysis related tasks, including COVID-19 diagnosis [16] and upper respiratory tract infections (URTI) classification [28].

4.2.3 Bag-of-Audio-Words

Bag-of-Audio-Words (BoAW) is extended from the concept of Bag-of-Words [29], a common representation of information in the Natural Language Processing (NLP) field. BoAW is a sparse audio representation that first clusters the input frame-level feature vectors (e.g., MFCC, eGeMAPS, ComParE), replaces each frame-level feature vector by its cluster, and then uses a rich dictionary (i.e., codebook) of these clusters to represent an utterance-level feature vector [19]. The main advantage of BoAW is its capacity of summarising the meaningful information of a variable-length input audio using a fixed-length vector (i.e., the histogram). The histogram represents the distribution of quantised feature vectors from a given audio instance [30].

Recently, the BoAW representation method has become very popular and has demonstrated its suitability in various speech-related fields [20, 30, 31].

4.3 Methodology

In this chapter, we propose a novel SincNet based CNN method for feature representations. SincNet is a novel CNN-based architecture, originally proposed for speaker recognition [23]. Our method adopts a SincNet to only learn low and high cutoff frequencies from raw waveform, instead of learning all elements from each filter in the traditional CNN architecture. SincNet has embedded bandpass filters for extracting features from the raw waveform, and makes it more interpretable and faster to converge.

SincNet has shown improved performance for research in different areas of speech-related tasks, including neurodegenerative related disorder classification [15], speech-based age and cognitive decline estimation [32] and speech emotion recognition [33].

As illustrated in Figure 4.1, all neural representations extracted by SincNet are fed into a CNN classifier. The CNN classifier we implemented was proposed by Ravanelli and Benjio [23], which has two standard convolutional layers, each with 60 filters of length 5 to evaluate the neural representations. For both the input samples and all convolutional layers (including the SincNet input layer), layer normalisation [34] is employed. Following that, three fully-connected layers with a total of 2048 neurons are applied and normalised with batch normalisation [35]. Leaky-ReLU [36] (with variable nonlinearity) have been used in all hidden layers. The neural networks are implemented with PyTorch¹.

¹<https://pytorch.org/>

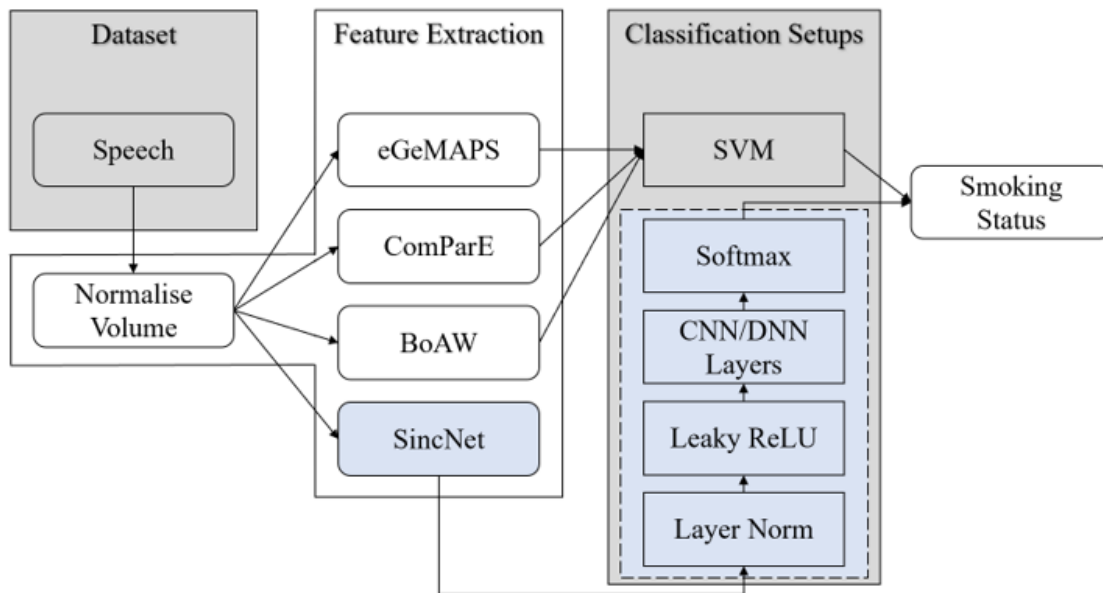


FIGURE 4.1: The architecture of our proposed method.

4.4 Experiments

4.4.1 Datasets

In the absence of large-scale, well-designed datasets expressly for smoking status identification experiments, we collect and create our datasets by extracting from two corpora released through the Linguistic Data Consortium (LDC): (i) the Mixer 4 and 5 Speech Corpus; and (ii) the Mixer 6 Speech Corpus. The speech recordings in the Mixer 4 and 5 Speech Corpus were used in 2008 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE)². The speech recordings in the Mixer 6 Speech Corpus were used in 2010 NIST SRE³. Both corpora comprise conversation recordings made via the public telephone network and multiple microphones in office-room settings. The main difference in the setting is that few of the 616 distinct speakers in the Mixer 4 and 5 Speech Corpus are bilingual

²<https://catalog.ldc.upenn.edu/LDC2020S03>

³<https://catalog.ldc.upenn.edu/LDC2013S03>

TABLE 4.2: The status of the speaker’s age in our dataset.

	Avg	Min	Max
Female Smokers	31.76	18	63
Female Non-Smokers	31.36	17	68
Male Smokers	30.38	19	60
Male Non-Smokers	28.10	19	60

English speakers, while the rest of the speakers of Mixer 4 and 5 Speech Corpus and all 594 distinct speakers in the Mixer 6 Speech Corpus are native English speakers. There is no overlap between the two corpora.

However, not all speakers in these two corpora have a valid smoking status label. In Mixer 4 and 5 Speech Corpus, only 89 of 616 speakers have smoking status labels. There are 40 female smokers, 8 female non-smokers, 37 male smokers, and 4 male non-smokers. In Mixer 6 Speech Corpus, 589 of 594 speakers have smoking status labels. There are 48 female smokers, 252 female non-smokers, 70 male smokers, and 219 male non-smokers. For valid smoking status identification purposes and balancing speakers’ gender and smoking status distribution, only those speakers with valid smoking status labels are considered in our experiments. In the end, 200 speakers (50 female smokers, 50 female non-smokers, 50 male smokers, and 50 male non-smokers) are selected for experiments. The details of the speaker’s status are shown in Table 4.2. Most of the speakers have two or more 12 mins to 15 mins transcript reading audio segments; a few only have one 12 mins transcript reading audio segment. We split the training set, development set and test set following the 8:1:1 ratio. We chose 5 female smokers, 5 female non-smokers, 5 male smokers, 5 male non-smokers as the test set and the rest of the speakers for the training set and the development set. To ensure our smoking status identification experiments are speaker-independent, recordings from speakers who contributed more than one recording are retained in the same division.

4.4.2 Feature Extraction

Before extracting any acoustic features, we normalise the volume of all voice utterances into the range $[-1: +1]$ dBFS. The goal is to improve the smoking status identification’s robustness against diverse recording conditions, such as microphone distance from the subject’s mouth.

We use the openSMILE toolkit [18] and standard configuration files to extract features for eGeMAPS and ComParE standard feature sets, respectively. We also use an MFCC feature set (MFCC12_0_D_A.conf) as our baseline feature set. The openXBOW toolkit [19] is used to generate BoAW representations from the 23 LLDs of eGeMAPS and the 65 LLDs of ComParE with the corresponding deltas, respectively. For each of the LLDs and their deltas, a separate codebook is learnt using random sampling of the LLDs from the training data. We test codebook sizes of $N = 500, 1000$ and 5000 . In order to get rid of the variation of scales between LLDs, which have an influence on the quantisation step, LLDs are normalised to zero mean and unit variance. The parameters mean and standard deviation have been estimated from the training. We observe that the ComParE-BoAW feature representations with a codebook size of 1000 achieved the best performance, hence it is reported in the rest of our experiments.

The SincNet layer is applied to the raw waveform and acts as a feature extractor to generate feature vectors. The raw waveform of each speech recording is chunked using a frame size of 200 ms and fed into the SincNet architecture described in Section 4.4.2.

TABLE 4.3: Experimental results of different acoustic feature sets/representations on the test set.

Features	Accuracy	F1-score
MFCC	0.71	0.70
eGeMAPS	0.78	0.77
ComParE	0.83	0.83
BoAW	0.81	0.80
SincNet	0.87	0.87

4.4.3 Classification Setups

For evaluating the extracted hand-crafted feature sets, a Support Vector Machine (SVM) is utilised because of its high effectiveness in the acoustic-based speech classification fields [13, 28, 37]. For SVM, we set the cost parameter C as 0.01 and use Radial Basis Function (RBF) kernels. The SVM classifier is trained by using the hand-crafted feature sets extracted from the training and development sets. The test set is used for evaluating the performance of the SVM classifier. SVM is implemented with scikit-learn⁴.

4.5 Results and Discussions

A summary of experimental results for smoking identification is provided in Table 4.3. Our results show that all four proposed acoustic feature sets/representations achieve better performances on the dataset than the MFCC baseline features.

For the hand-crafted feature sets, the ComParE feature set (6373 features) achieves the higher classification accuracy with 83%, which is significantly better than the MFCC baseline of 71%. The eGeMAPS feature set (88 features) achieves a classification accuracy of 78%. The performance is higher than the MFCC baseline feature but is slightly lower than the one achieved by ComParE. This indicates that

⁴<https://scikit-learn.org/>

hand-crafted feature sets including fundamental frequency (F_0), jitter, shimmer etc., provide better performance than traditional conventional acoustic features such as MFCC in this task-driven speech classification experiment. It also suggests that the more features in the hand-crafted feature sets are used, the better the classification performance will be.

The BoAW representation method (i.e., ComParE-BoAW) achieves a slightly lower performance with an accuracy of 81% compared with using the ComParE feature set directly. We also test the performance of BoAW built from the eGeMAPS feature set, but the results are consistently lower than using the eGeMAPS feature set directly and are not included in Table 4.3. A key direction for future research is determining the most useful frame-level features for a BoAW model.

Compared with hand-crafted features, the neural representations learned from raw waveform include more information for generating task-driven acoustic features. The best experimental result based on SincNet achieves an accuracy of 87%. This suggests that learning neural representations from raw waveform is capable of providing better performance than most models using domain-knowledge based acoustic feature sets/representations such as eGeMAPS, ComParE and BoAW representations in smoking status identification tasks.

4.6 Conclusions

In this paper, we propose a dataset that can be used for the smoking status identification study, and we investigate the efficiency of different acoustic features extracted/learned using three extraction/learning techniques for smoking status identification. We find that all proposed acoustic features perform better than traditional conventional acoustic features (i.e., MFCC). To the best of our knowledge, this is

the first study that comprehensively explores acoustic features for smoking status identification from speech.

In the future, we will explore the effect of combining different acoustic feature sets/representations and also investigate the performance of using different deep neural networks as the classifiers. We will extend this study to learn how the acoustic properties of smokers' speech alter during the smoking cessation process (e.g., before they have fully stopped smoking, when they have quit smoking for one week, quit smoking for one month, etc.).

This chapter has been published as follows:

Zhizhong Ma, Yuanhang Qiu, Feng Hou, Ruili Wang, Joanna Ting Wai Chu and Christopher Bullen. Determining the best acoustic features for smoking status identification,. In the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8177-8181. IEEE, 2022.

References

- [1] Dogan Pinar, Hakan Cincik, Evren Erkul, and Atila Gungor. Investigating the effects of smoking on young adult male voice by using multidimensional methods. *Journal of Voice*, 30(6):721–725, 2016.
- [2] Harveen Kaur Ubhi, Susan Michie, Daniel Kotz, Onno CP van Schayck, Abiram Selladurai, and Robert West. Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features. *Translational Behavioral Medicine*, 6(3):410–417, 2016.
- [3] Amir Hossein Poorjam and Mohamad Hasan Bahari. Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 7–12. IEEE, 2014.

-
- [4] Christopher H Murphy and Philip C Doyle. The effects of cigarette smoking on voice-fundamental frequency. *Otolaryngology—Head and Neck Surgery*, 97(4):376–380, 1987.
- [5] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. *Stockholms Universitet*, 11, 1995.
- [6] Julio Gonzalez and Amparo Carpi. Early effects of smoking on the voice: a multidimensional study. *Medical Science Monitor*, 10(12):656, 2004.
- [7] Isabel Guimarães and Evelyn Abberton. Health and voice quality in smokers: an exploratory investigation. *Logopedics Phoniatrics Vocology*, 30(3):185–191, 2005.
- [8] Linda Lee, Joseph C Stemple, Diane Geiger, and Rebecca Goldwasser. Effects of environmental tobacco smoke on objective measures of voice production. *The Laryngoscope*, 109(9):1531–1534, 1999.
- [9] Ouissam Zealouk, Hassan Satori, Mohamed Hamidi, Naouar Laaidi, and Khalid Satori. Vocal parameters analysis of smoker using Amazigh language. *International Journal of Speech Technology*, 21(1):85–91, 2018.
- [10] Angelika Braun. The effect of cigarette smoking on vocal parameters. In *Automatic Speaker Recognition, Identification and Verification*, 1994.
- [11] Zhizhong Ma, Christopher Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. *Journal of Voice*, 36(6), 2021.
- [12] Amir Hossein Poorjam, Soheila Hesaraki, Saeid Safavi, Hugo van Hamme, and Mohamad Hasan Bahari. Automatic smoker detection from telephone speech signals. In *International Conference on Speech and Computer*, pages 200–210. Springer, 2017.

-
- [13] Fasih Haider, Sofia De La Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of Alzheimer’s dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2019.
- [14] Wei Xue, Catia Cucchiaroni, RWNM van Hout, and Helmer Strik. Acoustic correlates of speech intelligibility. The usability of the eGeMAPS feature set for atypical speech. *In the 8th Workshop on Speech and Language Technology in Education*, pages 17–25, 2019.
- [15] Yilin Pan, Bahman Mirheidari, Zehai Tu, et al. Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification. *In the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4806–4810. IEEE, 2020.
- [16] Jing Han, Kun Qian, Meishu Song, et al. An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety. *arXiv preprint arXiv:2005.00096*, 2020.
- [17] Florian Eyben, Klaus R Scherer, Björn W Schuller, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2015.
- [18] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in OpenSmile, the Munich open-source multimedia feature extractor. *In the 21st ACM International Conference on Multimedia*, pages 835–838, 2013.
- [19] Maximilian Schmitt and Björn Schuller. Openxbow: introducing the passau open-source crossmodal Bag-of-Words toolkit. *Journal of Machine Learning Research*, 29(1):237–248, 2017.
- [20] Gábor Gosztolya and Róbert Busa-Fekete. Ensemble Bag-of-Audio-Words representation improves paralinguistic classification accuracy. *IEEE Transactions on Audio, Speech, and Language Processing*, 29(1):477–488, 2020.
- [21] Tara Sainath, Ron J Weiss, Kevin Wilson, Andrew W Senior, and Oriol Vinyals.

- Learning the speech front-end with raw waveform CLDNNs. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 121–128. IEEE, 2015.
- [22] Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux. End-to-end speech recognition from the raw waveform. *arXiv preprint arXiv:1806.07098*, 2018.
- [23] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018.
- [24] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak. Emotion identification from raw speech signals using DNNs. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3097–3101. IEEE, 2018.
- [25] Linda Brandschain, Christopher Cieri, David Graff, Abby Neely, and Kevin Walker. Speaker recognition: Building the Mixer 4 and 5 corpora. In *the International Conference on Language Resources and Evaluation*. Citeseer, 2008.
- [26] Linda Brandschain, D Graff, C Cieri, K Walker, C Caruso, and A Neely. The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition. In *the International Conference on Language Resources and Evaluation*, 2010.
- [27] Filip Povolny, Pavel Matejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. Multimodal emotion recognition for AVEC 2016 challenge. In *the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 75–82, 2016.
- [28] Nicholas Cummins, Maximilian Schmitt, Shahin Amiriparian, Jarek Krajewski, and Björn Schuller. “You sound ill, take the day off”: Automatic recognition of speech affected by upper respiratory tract infection. In *2017 39th Annual*

- International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3806–3809. IEEE, 2017.
- [29] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding Bag-of-Words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52, 2010.
- [30] Liwen Zhang, Jiqing Han, and Shiwen Deng. Unsupervised temporal feature learning based on sparse coding embedded BoAW for acoustic event recognition. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3284–3288. IEEE, 2018.
- [31] Aaron Keesing, Yun Sing Koh, and Michael Witbrock. Acoustic features and neural representations for categorical emotion recognition from speech. In *the 22nd Annual Conference of the International Speech Communication Association*, pages 3415–3419, 2021.
- [32] Yilin Pan, Venkata Srikanth Nallanthighal, Daniel Blackburn, Heidi Christensen, and Aki Härmä. Multi-task estimation of age and cognitive decline from speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7258–7262. IEEE, 2021.
- [33] Hong Zeng, Zhenhua Wu, Jiaming Zhang, Chen Yang, Hua Zhang, Guojun Dai, and Wanzeng Kong. EEG emotion classification using an improved SincNet-based deep learning model. *Brain Sciences*, 9(11):326, 2019.
- [34] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [35] César Laurent, Gabriel Pereyra, Philémon Brakel, Ying Zhang, and Yoshua Bengio. Batch normalized recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2657–2661. IEEE, 2016.
- [36] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on*

Machine Learning, volume 30, page 3. Citeseer, 2013.

- [37] Nicholas Cummins, Yilin Pan, and Zhao and others Ren. A comparison of acoustic and linguistics methodologies for Alzheimer’s dementia recognition. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2182–2186. IEEE, 2020.

Chapter 5

Transfer Learning and Task-Driven Feature Representations for COVID-19 Diagnosis

Multiple deep learning methods have been developed to identify respiratory diseases (e.g., COVID-19) from human-generated sounds (e.g., breath, cough, speech). Currently, the amount of available COVID-19 labelled data is normally limited. To address the scarcity of well-labelled data, we propose a transfer learning scheme to identify the COVID-19 disease by fine-tuning the pre-trained representation models (i.e., VGGish, wav2vec 2.0, PASE+) on datasets with COVID-19 labels. We also propose a task-driven feature representation network Sinc-ResNet (SincNet as the frontend, with ResNet as the backend) to learn feature representations effectively. With a ROC-AUC of over 0.8, both proposed methods significantly outperform traditional hand-crafted feature methods (e.g., OpenSMILE+SVM) and provide competitive results as compared with other deep learning methods.

5.1 Introduction

The human voice carries the speaker's information including age, gender, emotional status, psychological status, and health status, which is a powerful indicator for respiratory symptom prediction [1–3]. There has been increasing interest in developing a reliable, accessible, and contactless method for preliminary diagnosis of respiratory diseases including COVID-19. In the current COVID-19 pandemic, a contactless method is more desirable than ever.

Recently, several respiratory sound datasets for COVID-19 research (e.g., Virufy [4], Coswara [5], COUGHVID [6], COVID-19 Sounds [7]) have been developed. With these datasets, several deep learning-based methods have been developed to facilitate automatic audio-based COVID-19 diagnosis research [8–12].

In principle, the diagnosis of the COVID-19 disease is a binary classification task (i.e., either positive or negative). Most previous work is based on acoustic feature extraction in different settings. Differences between individuals who tested COVID-19 positive and negative in various acoustic parameters were found [8, 9]. Amir et al. [8] extracted 25 acoustic features (e.g., fundamental frequency and its perturbation, harmonicity, vocal tract function, airflow sufficiency, and periodicity) from the vowel sustained vowel (i.e., /a/) and used a deep multi-layer feedforward neural network for screening COVID-19 patients, which achieved an accuracy of 89.71%. In addition, Maral et al. [9] demonstrated significant differences between COVID-19 patients and healthy participants in voice quality-related acoustic features (e.g., cepstral peak prominence, maximum phonation time, harmonic-to-noise). In [10], the ComParE 2016 feature set was employed for the diagnosis of COVID-19 with machine learning models such as Random Forest and Support Vector Machines (SVM), which achieved an ROC-AUC score of 0.85 on the first Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge [13]. Besides these, other extracted feature

settings including Mel-frequency cepstral coefficients (MFCCs) and Mel log spectrograms in combination with Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) have also been explored [11, 12].

Previous research has primarily focused on fully-supervised feature extraction algorithms that trained and evaluated on the same dataset. However, the fully-supervised setting limits the selected dataset’s applicability and effectiveness. As a result, the additional datasets cannot be used to improve the classification performance. To address these limitations, we propose to leverage the knowledge learned from pre-training models for audio classification tasks on other large-scale corpora, and then fine-tune the pre-trained models for COVID-19 diagnosis. Such pre-training and fine-tuning paradigms have shown to be a promising technique in such health-related audio-based classification tasks (e.g., Parkinson’s disease identification [14], Alzheimer’s disease detection [15], COVID-19 diagnosis [16]) where the training dataset is limited.

We also propose a task-driven feature encoder based on SincNet [17], to extract more efficient and meaningful features directly from raw input waveform. In particular, for acoustic feature extraction, task-driven feature extraction has outperformed other traditional feature extraction methods under the same limited dataset scenario [18, 19].

In summary, our main contributions are as follows:

- (i) We propose a transfer learning scheme using audio representations extracted from the pre-trained deep neural models (i.e., VGGish [20], wav2vec 2.0 [21], PASE+ [22]) for the task of COVID-19 diagnosis from respiratory sounds.
- (ii) We propose and implement a task-driven feature representation method SincResNet to diagnose the COVID-19 disease. Applying SincNet [17] in the feature extraction phase provides efficient and meaningful feature representations, while

ResNet [23] enhances the performance in the feature encoder phase. To the best of our knowledge, this is the first paper utilising task-driven representation learning for the task of COVID-19 diagnosis from respiratory sounds.

(iii) Through extensive experiments on both the COVID-19 Sounds dataset [7] and the Coswara dataset [5], we carefully evaluate our proposed method, which allows us to draw more general conclusions on the performance and generalisability of our proposed method.

The rest of this paper is structured as follows. Section 5.2 presents the related work of the acoustic feature extraction methods we utilised in this paper. The proposed methods and datasets are provided in Section 5.3. Section 5.4 describes the experimental results and discussions. The conclusions and proposals for future work are discussed in Section 5.5.

5.2 Feature Extraction

In this section, we introduce three pre-trained audio networks investigated in our paper, namely, VGGish [20], wav2vec 2.0 [21] and PASE+ [22]. In addition to the pre-trained audio networks, we also investigate SincNet [17] as a task-driven feature representation method to extract efficient and meaningful features directly from the raw input waveform.

5.2.1 VGGish

VGGish [20] is a modification of the VGG network [24] which is created by training audio embeddings with the AudioSet dataset [25], a dataset of over 2 million human-labelled 10-second YouTube video soundtracks with labels derived from an ontology

of over 600 audio event types for audio classification tasks. VGGish extracts audio input features into a high-level 128-dimensional embedding that can be fed into a downstream audio classification model. All audio input is resampled to a mono channel with a frequency of 16 kHz. The features are extracted from non-overlapping audio patches that last 0.96 seconds and cover 64 mel bands and 96 frames of 10 ms each. VGGish comprises four sets of convolutional and pooling layers. The output of the final pooling layer is flattened, and a fully connected layer that acts as a compact embedding layer is applied next. The dimensionality was reduced using principal component analysis (PCA) [26].

5.2.2 Wav2vec 2.0

Wav2vec 2.0 [21] is a deep neural network representation of audio through self-supervised learning to replace conventional feature extraction methods such as MFCC. It allows us to build more robust audio classification systems with limited training data. In the wav2vec 2.0 model, the raw input waveform is encoded via a multi-layer convolutional neural network to obtain each 25 ms latent audio representation. These representation vectors are fed into the quantiser and transformer [27]. The quantiser selects a phonetic unit from the inventory of learned units as the latent audio representation vector. Subsequently, spans of the resulting latent audio representations are masked before being fed into the transformer. The transformer then models the contextualised representation in around 25 ms and extracts a high-level feature from the input. The transformer adds information from the entire audio sequence, and the output is used to compute the loss function.

5.2.3 Problem-Agnostic Speech Encoder

The problem-agnostic speech encoder (PASE+) model encodes raw input waveform to capture relevant speech information and transmits it to an ensemble of small neural networks (i.e., workers) [22], and is an improved version of the original PASE model [28]. The first layer of the PASE+ model is SincNet, followed by a stack of seven convolutional blocks that include a one-dimensional convolutional layer, batch normalisation, and PReLU activation [29]. PASE+ employs a Quasi-RNN layer (QRNN) [30] to learn long-term dependencies. Each worker contributes extra prior knowledge to the encoder by giving a different view of the input signal. After joint training of the encoder and the workers, PASE+ features are extracted. The final encoder representation is the sum of the linearly projected intermediate features computed by the seven convolutional blocks employing skipped connections and the QRNN output. This enables information transmission between the different levels of abstraction and the enhancement of gradient flows. The encoder's output is fed into twelve workers. Six regression workers are trained to estimate common speech features including the speech waveform itself, log power spectrum (LPC), MFCCs, prosody, FBANKS, and Gammatone. For LPS, MFCC, FBANKS, and Gammatone features, PASE+ further added four workers that estimate features on longer analysis windows (200 ms rather than 25 ms). Lastly, two workers are adopted for binary tasks, namely Local Info Max (LIM) and Global Info Max (GIM).

5.2.4 SincNet

SincNet is a novel CNN-based architecture originally proposed for speaker recognition by Ravanelli and Benjio [17]. SincNet is based on parameterised sinc function bandpass filters for extracting features. Instead of learning all elements from each

filter in the traditional CNN architecture, SincNet only learns those low and high cut-off frequencies from raw input waveform, making it more interpretable and faster to converge. SincNet has shown improved performance for research in different areas of speech-related tasks, including neurodegenerative related disorder classification [31], speech-based age and cognitive decline estimation [32], and smoking status identification [33]. Compared with our previous work [33], we have introduced the ResNet architecture as the backend for the proposed Sinc-ResNet model in this paper (see Section 5.3.2).

5.3 Methodology

The proposed audio-based COVID-19 diagnosis pipeline is illustrated in Figure 5.1. The pipeline uses a transfer learning scheme (i.e., using the pre-trained audio neural network models: VGGish, wav2vec 2.0, PASE+) or task-driven feature extractor (i.e., SincNet) to better extract feature representations. The resulting feature representations are used to train a feature encoder so that it can learn discriminative representations and feed into the binary classifier. The output of this pipeline is the predicted COVID-19 result (i.e., positive, negative). Detailed implementations of our transfer learning scheme and proposed Sinc-ResNet are given in Section 5.3.1 and 5.3.2 respectively.

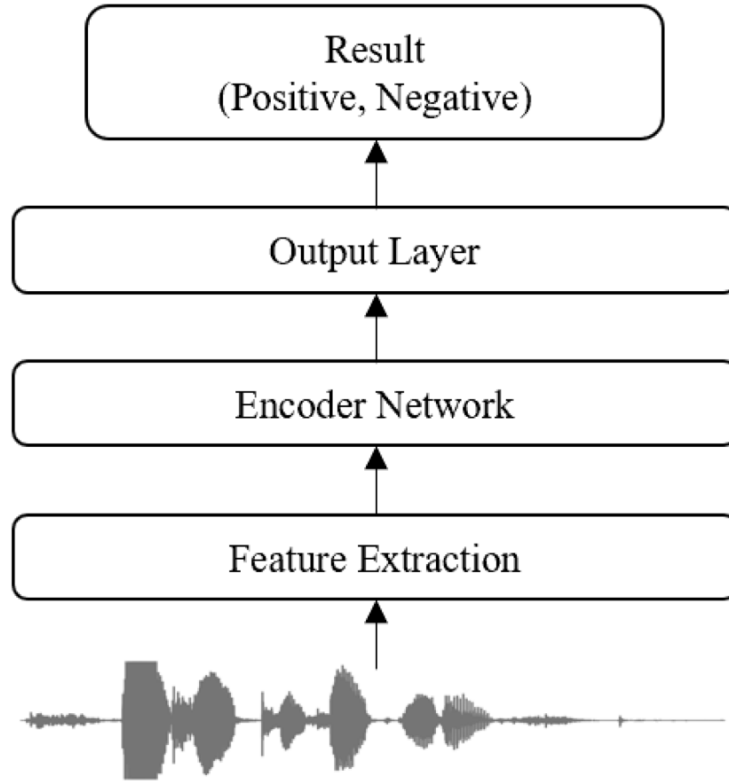


FIGURE 5.1: The proposed audio-based COVID-19 diagnosis pipeline.

5.3.1 Transfer Learning Scheme

The embedding features of raw input waveform are calculated by using three pre-trained models (i.e., VGGish¹, wav2vec 2.0², PASE+³), respectively. An adaptive average pooling layer handles the variable audio duration, resulting in a 128-dimensional feature vector for each sound sample. The pre-trained model and the succeeding fully connected layer are fine-tuned. Finally, the features are utilised for training a linear classifier, followed by a ReLU activation, a batch normalisation layer and a dropout layer.

¹<https://github.com/harritaylor/torchvggish/>

²<https://github.com/huggingface/transformers/>

³<https://github.com/santi-pdp/PASE+/>

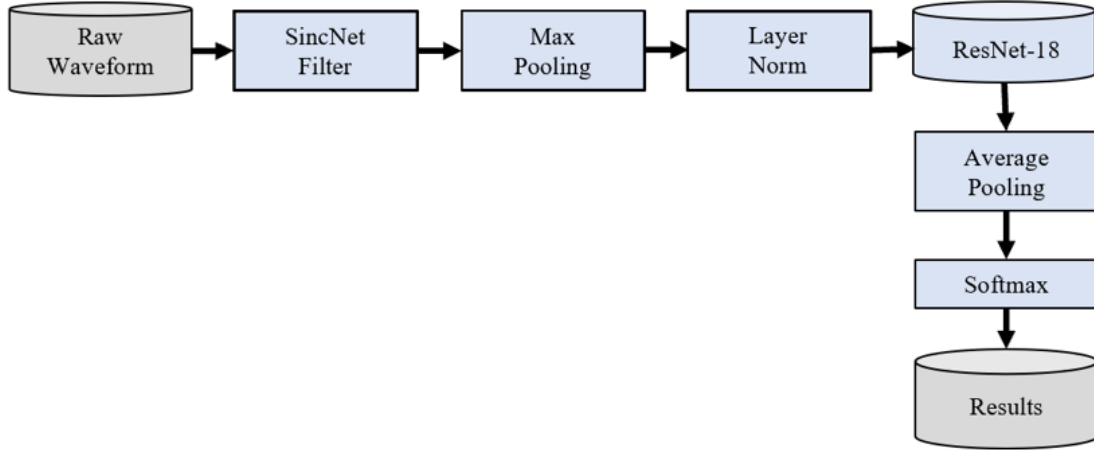


FIGURE 5.2: An overview of our Sinc-ResNet architecture.

5.3.2 Task-Driven Feature Representation Network

We propose a task-driven feature representation network named Sinc-ResNet. An overview of our proposed Sinc-ResNet architecture is illustrated in Figure 5.2. To extract efficient and meaningful task-driven features from the raw input waveform, a SincNet followed by Max pooling layers and layer normalisation is applied as the first functional layer. The SincNet layer is composed of $N=80$ filters of length $L=125$ samples. The output of the SincNet layer is then fed into the ResNet-18 model, a type of ResNet with 18 layers [23], followed by adaptive pooling layers in both the time and frequency dimensions, as the backbone for our Sinc-ResNet model. Finally, the output is passed through two linear layers, followed by a final predictive layer with two neurones and a softmax activation function that predicts if the input audio sample is COVID-19 positive or negative. For testing, the model with the best validation loss was chosen.

5.3.3 Datasets

We evaluate the performance of our proposed methods on two independent datasets: the COVID-19 Sounds dataset [7], and the Coswara dataset [5]. To the best of our knowledge, the COVID-19 Sounds dataset is the largest dataset of COVID-19 respiratory sounds (i.e., more than 550 hours duration sound samples), while the Coswara dataset contains more than 50 hours sound samples which is the second largest dataset.

The COVID-19 Sounds dataset has been developed by Cambridge University. Breath, cough, and speech samples are collected from different countries. The data comes in 2 to 30-second WAV files with up to 48kHz sampling rate. In our experiments, we select the same two curated subsets of the COVID-19 Sounds dataset that are described and evaluated in [7]. The two curated subsets refer to the following two classification tasks: (i) Task 1 is designed to distinguish respiratory abnormalities by detecting the participants' various voice types. Note that Task 1 is a respiratory symptom prediction task, not a COVID-19 diagnosis task, but it could be considered to evaluate the performance and robustness of the proposed COVID-19 diagnosis methods [7]; (ii) Task 2 is a COVID-19 diagnosis task that aims to distinguish COVID-19 status among participants by examining their various voice types. For these two tasks, 6,623 participants with 9,456 samples and 1,000 participants with 1,486 samples are utilised, respectively.

The Coswara dataset has been developed by the Indian Institute of Science, which aims to develop a COVID-19 diagnostic tool based on respiratory, cough, and speech sounds. The audio files are categorised into four sound groups (breath, cough, counting, and sustained phonation of vowel sounds). Since the available categories and recording types of the Coswara dataset are different from the COVID-19 Sounds dataset, in order to remain consistent, we filter the data for COVID-19 positive and

negative participants that have contributed both breath and cough recordings. The filtered samples are split into speaker-independent sets for training, validation, and testing with a ratio of 8:1:1. Overall, 2,508 participants with 7,524 samples (i.e., one breath sample, one shallow cough sample, and one deep cough sample for each participant) are utilised in this task.

5.4 Experimental Results and Discussions

The findings of our experiments to identify the COVID-19 disease from human-generated sounds (i.e., breath, cough, speech) are presented in this section. The area under the receiver operating characteristic curve (ROC-AUC) with 95% confidence intervals (CIs) is reported as the evaluation metric.

5.4.1 Results on the COVID-19 Sounds Dataset

The experimental results on the two tasks of the COVID-19 Sounds dataset are presented in Tables 1 and 2, respectively. For Task 1 respiratory symptom prediction, a ROC-AUC up to 0.81 (0.76-0.85) is achieved. For Task 2 COVID-19 diagnosis, a slightly lower performance of ROC-AUC up to 0.67 (0.60-0.75) is obtained. We considered the Cambridge team’s work in [7] as the baseline. Both our proposed transfer learning methods and task-driven feature extraction network Sinc-ResNet outperform all three baseline methods proposed in [7] for both tasks.

As the results are shown in Table 5.1 and 5.2, the PASE+ model achieves the best scores for breath sounds, while our Sinc-ResNet model achieves the best scores for cough sounds and speech sounds. Meanwhile, both VGGish model and wav2vec 2.0 model show competitive results.

TABLE 5.1: A comparison of different sound types and methods for task 1 of the COVID-19 Sounds dataset.

Types	Methods	ROC-AUC
Breath	OpenSMILE+SVM [7]	0.60 (0.58-0.63)
	Pre-trained VGGish [7]	0.52 (0.50-0.56)
	Fine-tuned VGGish [7]	0.65 (0.63-0.67)
	VGGish	0.66 (0.62-0.69)
	wav2vec 2.0	0.62 (0.60-0.65)
	PASE+	0.69 (0.66-0.71)
	Sinc-ResNet	0.67 (0.62-0.72)
Cough	OpenSMILE+SVM [7]	0.70 (0.67-0.72)
	Pre-trained VGGish [7]	0.66 (0.63-0.68)
	Fine-tuned VGGish [7]	0.74 (0.72-0.76)
	VGGish	0.75 (0.72-0.78)
	wav2vec 2.0	0.74 (0.72-0.76)
	PASE+	0.76 (0.74-0.78)
	Sinc-ResNet	0.81 (0.76-0.85)
Speech	OpenSMILE+SVM [7]	0.63 (0.66-0.71)
	Pre-trained VGGish [7]	0.59 (0.57-0.62)
	Fine-tuned VGGish [7]	0.69 (0.66-0.71)
	VGGish	0.70 (0.69-0.72)
	wav2vec 2.0	0.72 (0.70-0.74)
	PASE+	0.73 (0.70-0.75)
	Sinc-ResNet	0.77 (0.75-0.79)

5.4.2 Results on the Coswara Dataset

The experimental results on the Coswara database are presented in Table 5.3. The PASE+ model outperforms the other two pre-trained models (i.e., VGGish, wav2vec 2.0), reporting a ROC-AUC score of 0.78 (0.72-0.85), and 0.81 (0.75-0.88) for breath sounds and cough sounds, respectively. The Sinc-ResNet model achieves the best results in terms of ROC-AUC score in breath sounds (0.80) and cough sounds (0.83).

TABLE 5.2: A comparison of different sound types and methods for task 2 of the COVID-19 Sounds dataset.

Types	Methods	ROC-AUC
Breath	OpenSMILE+SVM [7]	0.56 (0.50-0.61)
	Pre-trained VGGish [7]	0.59 (0.52-0.65)
	Fine-tuned VGGish [7]	0.62 (0.56-0.69)
	VGGish	0.61 (0.53-0.68)
	wav2vec 2.0	0.60 (0.56-0.64)
	PASE+	0.64 (0.58-0.70)
	Sinc-ResNet	0.63 (0.56-0.67)
Cough	OpenSMILE+SVM [7]	0.62 (0.56-0.68)
	Pre-trained VGGish [7]	0.62 (0.56-0.68)
	Fine-tuned VGGish [7]	0.66 (0.59-0.71)
	VGGish	0.63 (0.56-0.70)
	wav2vec 2.0	0.64 (0.58-0.71)
	PASE+	0.67 (0.59-0.75)
	Sinc-ResNet	0.67 (0.60-0.75)
Speech	OpenSMILE+SVM [7]	0.52 (0.45-0.58)
	Pre-trained VGGish [7]	0.61 (0.54-0.67)
	Fine-tuned VGGish [7]	0.61 (0.55-0.67)
	VGGish	0.61 (0.56-0.67)
	wav2vec 2.0	0.58 (0.52-0.64)
	PASE+	0.63 (0.57-0.69)
	Sinc-ResNet	0.64 (0.56-0.72)

TABLE 5.3: A comparison of different sound types and methods for experiments on the Coswara dataset.

Types	Methods	ROC-AUC
Breath	VGGish	0.75 (0.70-0.81)
	wav2vec 2.0	0.76 (0.69-0.84)
	PASE+	0.78 (0.72-0.85)
	Sinc-ResNet	0.80 (0.73-0.87)
Cough	VGGish	0.77 (0.72-0.82)
	wav2vec 2.0	0.78 (0.71-0.85)
	PASE+	0.81 (0.75-0.88)
	Sinc-ResNet	0.83 (0.76-0.89)

5.4.3 Discussions

For transfer learning methods, the PASE+ model achieves the best performance compared with the VGGish model and the wav2vec 2.0 model, since the PASE+ model is the only pre-trained model that adopts SincNet as the first layer of the encoder to extract the raw input waveform. On the other hand, our proposed task-driven feature extraction network Sinc-ResNet provides competitive results as the PASE+ model, and achieved the best performance in most sound types and tasks.

Based on our results, cough is the best sound type to identify the COVID-19 disease. This finding is consistent on both datasets and suggests that cough is an informative indicator in the diagnosis of the COVID-19 disease. In addition, breath and speech can be considered as supplementary resources to build the multi-modality audio-based COVID-19 diagnosis system.

5.5 Conclusions

In this paper, we propose and implement a classification pipeline for diagnosing respiratory diseases (e.g., COVID-19) from audio signals (i.e., breath, cough, and speech). The proposed pipeline is evaluated on the COVID-19 Sounds dataset and the Coswara dataset. We find that transfer learning methods using VGGish, wav2vec 2.0 and PASE+, and our proposed task-driven method Sinc-ResNet have significantly improved the performance. Experimental results show that both transfer learning and task-driven methods achieve competitive performance. For transfer learning methods, the PASE+ model achieves the best performance among all three pre-trained models. We also demonstrate that our proposed task-driven representation network using SincNet as the frontend, with ResNet as the backend achieves

the best performance in most sound types and tasks compared with transfer learning methods. The findings of this study provide a new perspective and insights for audio-based COVID-19 diagnosis. In the future, we will investigate multi-modality feature representation-based methods and more deep neural network architectures for performance improvements.

This chapter has been submitted as follows:

Zhizhong Ma, Ruili Wang, Feng Hou, Yuanhang Qiu, Satwinder Singh, Joanna Ting Wai Chu and Christopher Bullen. Transfer learning and task-driven feature representations for COVID-19 diagnosis from respiratory sound data,. In the ACM Transactions on Speech and Language Processing (TSLP). ACM, 2022. (Submitted)

References

- [1] P Mayorga, C Druzgalski, RL Morelos, OH Gonzalez, and J Vidales. Acoustics based assessment of respiratory diseases using GMM classification. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 6312–6316. IEEE, 2010.
- [2] Maude Desjardins, Lucinda Halstead, Annie Simpson, Patrick Flume, and Heather Shaw Bonilha. Voice and respiratory characteristics of men and women seeking treatment for presbyphonia. *Journal of Voice*, 2020.
- [3] Jing Han, Chloë Brown, Jagmohan Chauhan, et al. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8328–8332. IEEE, 2021.
- [4] Gunvant Chaudhari, Xinyi Jiang, Ahmed Fakhry, Asriel Han, Jaelyn Xiao, Sabrina Shen, and Amil Khanzada. Virufy: Global applicability of crowdsourced

- and clinical datasets for AI detection of COVID-19 from cough. *arXiv preprint arXiv:2011.13320*, 2020.
- [5] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Prasanta Kumar Ghosh, and Sriram Ganapathy. Coswara—a database of breathing, cough, and voice sounds for COVID-19 diagnosis. *arXiv preprint arXiv:2005.10548*, 2020.
- [6] Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):1–10, 2021.
- [7] Tong Xia, Dimitris Spathis, J Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, and Cecilia Mascolo. COVID-19 Sounds: A large-scale audio dataset for digital respiratory screening. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [8] Amir Vahedian-Azimi, Abdalsamad Keramatfar, Maral Asiaee, Seyed Shahab Atashi, and Mandana Nourbakhsh. Do you have COVID-19? An artificial intelligence-based screening tool for COVID-19 using acoustic parameters. *The Journal of the Acoustical Society of America*, 150(3):1945–1953, 2021.
- [9] Maral Asiaee, Amir Vahedian-Azimi, Seyed Shahab Atashi, Abdalsamad Keramatfar, and Mandana Nourbakhsh. Voice quality evaluation in patients with COVID-19: An acoustic analysis. *Journal of Voice*, 2020.
- [10] Isabella Södergren, Maryam Pahlavan Nodeh, Prakash Chandra Chhipa, Konstantina Nikolaidou, and György Kovács. Detecting COVID-19 from audio recording of coughs using Random Forests and Support Vector Machines. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 916–920. IEEE, 2021.
- [11] John Harvill, Yash R Wani, Mark Hasegawa-Johnson, et al. Classification of COVID-19 from cough using autoregressive predictive coding pretraining

- and spectral data augmentation. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4261–4265. IEEE, 2021.
- [12] Swapnil Bhosale, Upasana Tiwari, Rupayan Chakraborty, and Sunil Kumar Kopparapu. Contrastive learning of cough descriptors for automatic COVID-19 preliminary diagnosis. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 946–50. IEEE, 2021.
- [13] Neeraj Kumar Sharma, Ananya Muguli, Prashant Krishnan, et al. Towards sound based testing of COVID-19—Summary of the first Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge. *Computer Speech & Language*, 73(1):1013–1020, 2022.
- [14] Yermiyahu Hauptman, Ruth Aloni-Lavi, Itshak Lapidot, et al. Identifying distinctive acoustic and spectral features in Parkinson’s disease. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2498–2502. IEEE, 2019.
- [15] Anna Pompili, Thomas Rolland, and Alberto Abad. The INESC-ID multi-modal system for the ADReSS 2020 challenge. *arXiv preprint arXiv:2005.14646*, 2020.
- [16] Rubén Solera-Ureña, Catarina Botelho, Francisco Teixeira, et al. Transfer learning-based cough representations for automatic detection of COVID-19. In *the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4336–4340. IEEE, 2021.
- [17] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018.
- [18] Hong Zeng, Zhenhua Wu, Jiaming Zhang, Chen Yang, Hua Zhang, Guojun Dai, and Wanzeng Kong. EEG emotion classification using an improved SincNet-based deep learning model. *Brain Sciences*, 9(11):326, 2019.

- [19] Xuan Shi, Erica Cooper, and Junichi Yamagishi. Use of speaker recognition approaches for learning and evaluating embedding representations of musical instrument sounds. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.
- [20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, et al. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.
- [21] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.
- [22] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE, 2020.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [25] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, et al. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.
- [26] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need.

- Advances in Neural Information Processing Systems*, 30, 2017.
- [28] Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*, 2019.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [30] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. Quasi-recurrent neural networks. *arXiv preprint arXiv:1611.01576*, 2016.
- [31] Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification.
- [32] Yilin Pan, Venkata Srikanth Nallanthighal, Daniel Blackburn, Heidi Christensen, and Aki Härmä. Multi-task estimation of age and cognitive decline from speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7258–7262. IEEE, 2021.
- [33] Zhizhong Ma, Yuanhang Qiu, Feng Hou, Ruili Wang, Joanna Ting Wai Chu, and Christopher Bullen. Determining the best acoustic features for smoking status identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8177–8181. IEEE, 2022.

Chapter 6

Summary

This chapter provides a summary of this thesis. Firstly, we present a summary of the contributions in Section 6.1, which includes a comprehensive literature review of speech assessment of the smoking status based on voice features (Chapter 2); automatic speech-based smoking status identification method (Chapter 3); determining the best acoustic features for smoking status identification (Chapter 4); and transfer learning and task-driven feature representations for COVID-19 diagnosis from respiratory sound data (Chapter 5). Furthermore, we also discuss future work of the voice-based respiratory diagnosis research in Section 6.2.

6.1 Research Summary

In this thesis, we propose one literature review and two novel methods for smoking status identification and aim to fill the gap of the voice-based respiratory diagnosis research, two novel methods for COVID-19 diagnosis from respiratory sound data

to address the scarcity of well-labelled data so as to learn effective speech feature representations. A recap of our methods and contributions is listed as follows:

- Chapter 2 presents a comprehensive investigation of the effects of voice features in the detection of smoker/non-smoker speech signals [1]. We conclude that acoustic voice parameters appear to be influenced by smoking and smoking cessation: smoking permanently alters the acoustic parameters of smokers' speech compared with non-smokers, while smoking cessation will partly undo the permanent effect of smoking on various voice features. Overall, it appears that smokers have a lower fundamental frequency than non-smokers in both gender and age groups. Smokers present higher jitter values for all vowels. Smokers' shimmer values are higher than the values of non-smokers. During smoking cessation, HNR value increases dramatically. Moreover, jitter and shimmer values decrease significantly. F_0 value rises during smoking abstinence and decreases again after resuming smoking.
- Chapter 3 presents a novel method that uses the combination of both high- and low-level acoustic features along with deep neural networks for smoking status identification [2]. We propose a dataset that can be used for smoking status identification study, and the data augmentation technique (i.e., SpecAugment) is implemented to further improve the smoking status identification accuracy. Based on our experimental result, it indicates that Fbank outperforms MFCC if we only utilise high-level acoustic features. Our proposed method has outperformed the rule-based method and obtained the best accuracy of 82.3%, which is a relative improvement of 12.7% and 29.8% on the initial high-level acoustic features only method and rule-based method, respectively. The proposed automatic smoking status identification model could be an alternative solution to obtain an accurate and objective smoking status when biological verification methods are not feasible.

- Chapter 4 presents a novel SincNet based CNN method for feature representations and investigates the performance of three different acoustic feature sets: (i) the extended Geneva Minimalistic Acoustic Parameter Set; (ii) the Computational Paralinguistics Challenge Set; and (iii) the Bag-of-Audio-Words representations. We investigate the efficiency of different acoustic features extracted/learned by using three extraction/learning techniques and find that all proposed acoustic features perform better than traditional conventional acoustic features (i.e., MFCC).
- Chapter 5 presents a classification pipeline and two novel methods for diagnosing respiratory diseases (e.g., COVID-19) from audio signals (i.e., breath, cough, and speech) [3]. The proposed pipeline is evaluated on the COVID-19 Sounds dataset and the Coswara dataset. We found that transfer learning methods using VGGish, wav2vec 2.0 and PASE+, and our proposed task-driven method Sinc-ResNet have significantly improved the performance. Experimental results show that both transfer learning and task-driven method achieve competitive performance. For transfer learning method, the PASE+ model achieves the best performance among all three pre-trained models. We also demonstrate that our proposed task-driven representation network using SincNet as the frontend, with ResNet as the backend achieves the best performance in most sound types and tasks compared with transfer learning methods.

6.2 Future Work

In this section, we propose some future work for the voice-based respiratory diagnosis research.

- **Enlarge experiment scale.** We will further conduct experiments with sufficient speech data by including more realistic scenarios to evaluate the effectiveness and robustness of our methods. The training data should consider as many scenarios as possible to reflect the realistic environments and improve the adaptability of the proposed model.
- **Multiple features fusion for feature representation learning.** Multiple features fusion can provide multiple hierarchies of data representation for model training and mapping learning. In many research areas, feature fusion methods are used to achieve a more robust and effective model [4–7]. Thus, further exploration about multiple features fusion in voice-related respiratory diagnosis will be one of our future projects.
- **Novel neural networks for voice-related health research.** Recently, several novel architectures were proposed which made a breakthrough in many research areas such as attention based transformer architecture [8] and its variants [9–11]. Those models, adopting a revolutionary concept by eliminating recurrent or convolutional portions to improve information learning and result inference, will be applied to the voice-based respiratory diagnosis research in our future work.
- **Applications of voice-based respiratory diagnosis.** Voice-based respiratory diagnosis research has numerous potential applications (e.g., smoking status validation [12], smoking cessation tracking [13] and speaker profiling [14]). We will apply our proposed methods to these applications in future work.

References

- [1] Zhizhong Ma, Christopher Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. *Journal of Voice*, 36(6), 2021.
- [2] Zhizhong Ma, Satwinder Singh, Yuanhang Qiu, Feng Hou, Ruili Wang, Christopher Bullen, and Joanna Ting Wai Chu. Automatic speech-based smoking status identification. In *Computing Conference 2022*. IEEE, 2022.
- [3] Zhizhong Ma, Ruili Wang, Feng Hou, Yuanhang Qiu, Satwinder Singh, Joanna Ting Wai Chu, and Christopher Bullen. Transfer learning and task-driven feature representations for COVID-19 diagnosis from respiratory sound data. In *ACM Transactions on Speech and Language Processing (TSLP)*. ACM, 2022.
- [4] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *the European Conference on Computer Vision*, pages 269–284, 2018.
- [5] Khursheed Aurangzeb, Irfan Haider, Muhammad Attique Khan, et al. Human behavior analysis based on multi-types features fusion and Von Nauman entropy based features reduction. *Journal of Medical Imaging and Health Informatics*, 9(4):662–669, 2019.
- [6] Linhui Sun, Jia Chen, Keli Xie, and Ting Gu. Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition. *International Journal of Speech Technology*, 21(4):931–940, 2018.
- [7] Yongming Huang, Kexin Tian, Ao Wu, and Guobao Zhang. Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *Journal of Ambient Intelligence and Humanized Computing*, 10(5):1787–1798, 2019.
- [8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

- Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.
- [10] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.
- [11] David So, Quoc Le, and Chen Liang. The evolved transformer. In *the International Conference on Machine Learning (ICML)*, pages 5877–5886. PMLR, 2019.
- [12] Dogan Pinar, Hakan Cincik, Evren Erkul, and Atila Gungor. Investigating the effects of smoking on young adult male voice by using multidimensional methods. *Journal of Voice*, 30(6):721–725, 2016.
- [13] Harveen Kaur Ubhi, Susan Michie, Daniel Kotz, Onno CP van Schayck, Abiram Selladurai, and Robert West. Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features. *Translational Behavioral Medicine*, 6(3):410–417, 2016.
- [14] Amir Hossein Poorjam and Mohamad Hasan Bahari. Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 7–12. IEEE, 2014.


Appendix A

Statement of Contribution

I confirm that the “Statement of Contribution to Doctoral Thesis Containing Publications (DRC16)”, have been completed for each published paper within the thesis, and are bound into the thesis and included in the electronic copy.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS


We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhizhong Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work: Chapter 2	
<p>Please select one of the following three options:</p> <p><input checked="" type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Zhizhong Ma, Chris Bullen*, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and Satwinder Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. <i>Journal of Voice</i>, 2021. <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	<i>Zhizhong Ma</i>
Date:	07-Apr-2022
Primary Supervisor's Signature:	Prof Ruili Wang  <small>Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, o=NZ, ou=Massey University, ou=School of Natural and Computational Sciences, email=ruili.wang@massey.ac.nz Date: 2022.04.14 16:31:48 +12'00'</small>
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhizhong Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work: Chapter 3	
<p>Please select one of the following three options:</p> <p><input checked="" type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Zhizhong Ma, Satwinder Singh, Yuanhang Qiu, Feng Hou*, Ruili Wang, Christopher Bullen and Joanna Ting Wai Chu. Automatic speech-based smoking status identification,. In Computing Conference 2022. <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	<i>Zhizhong Ma</i>
Date:	07-Apr-2022
Primary Supervisor's Signature:	Prof Ruili Wang  <small>Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, o=NZ, ou=Massey University, ou=School of Natural and Computational Sciences, email=ruili.wang@massey.ac.nz Date: 2022.04.14 16:33:27 +12'00'</small>
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS


We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhizhong Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 4
<p>Please select one of the following three options:</p> <p><input checked="" type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Zhizhong Ma, Yuanhang Qiu, Feng Hou*, Ruili Wang, Joanna Ting Wai Chu and Christopher Bullen. Determining the best acoustic features for smoker identification,. In 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2022. <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	<i>Zhizhong Ma</i>
Date:	07-Apr-2022
Primary Supervisor's Signature:	Prof Ruili Wang <small>Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, o=NZ, ou=Massey University, ou=School of Natural and Computational Sciences, email=ruili.wang@massey.ac.nz Date: 2022.04.14 16:32:22 +12'00'</small>
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhizhong Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 5
Please select one of the following three options:	
<input type="radio"/> The manuscript/published work is published or in press <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: 	
<input checked="" type="radio"/> The manuscript is currently under review for publication – please indicate: <ul style="list-style-type: none"> • The name of the journal: ACM Transactions on Speech and Language Processing (TSLP) • The percentage of the manuscript/published work that was contributed by the candidate: 75.00 • Describe the contribution that the candidate has made to the manuscript/published work: <ul style="list-style-type: none"> - proposed the idea of transfer learning and task-driven feature representations based COVID-19 diagnosis methods. - implemented the experiments of proposed methods on respiratory sound data 	
<input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal	
Candidate's Signature:	<i>Zhizhong Ma</i>
Date:	06-Jul-2022
Primary Supervisor's Signature:	Prof Ruili Wang  <small>Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University, ou=School of Natural and Computational Sciences, email=ruili.wang@massey.ac.nz Date: 2022.07.09 20:01:41 +1200</small>
Date:	9-Jul-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

