THE UNIVERSITY
of ADELAIDE

Biomedical Signal Processing For Bioacoustic Features Extraction and Reproducibility Evaluation

Thesis by

Shaykhah A. Almaghrabi

B.Sc. (Biomedical Engineering),
Imam Abdulrahman Bin Faisal University, Saudi Arabia, 2017

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Philosophy

in the

School of Electrical and Electronic Engineering
Faculty of Engineering, Computer and Mathematical Sciences

March, 2022

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Shaykhah A. Almaghrabi

March, 2022

To my parents and my husband with all my love.

Abstract

Human speech produces acoustic waves that carry information about the speaker's gender, physiological condition, and pathophysiological state. Bio-acoustic properties obtained by speech signal processing show promise for the analysis of psychiatric illnesses. Alterations of acoustic measures associated with major depressive disorder (MDD) could potentially provide objective biomarkers for depression detection. Understanding bio-acoustic features stability is essential to best design sampling and analysis frameworks. Still, the impact of sample duration on bio-acoustic features' reproducibility has not been systematically explored. Classification performance of depressed and non-depressed bio-acoustic features measured at different speech durations remains to be investigated.

This thesis evaluates the reproducibility of bio-acoustic features against changes in speech durations and speech tasks in depressed and non-depressed English speakers. It also investigates the classification potential, in a binary manner, of bio-acoustic features quantified at short speech durations for MDD detection. Thus, source, spectral shape, cepstral, prosodic, and formants features were extracted from speech signals. The intraclass correlation coefficients were calculated to measure feature reproducibility. Support vector machines with radial basis function kernel were employed to evaluate the effect of speech duration on classification performance. Experimental results indicate that the number of reproducible features (out of 125) decreased stepwisely with duration reduction in both depressed and non-depressed speakers. Gender differences had a significant impact on the reproducibility of some features (e.g., pitch). The results also showed a slight improvement in the classification performance (accuracy, weighted F1 score, recall, and precision) when shortening the duration.

In conclusion, bio-acoustic characteristics are less reproducible in shorter speech samples and are affected by gender. Classification metrics are also influenced by speech data duration. Designing speech samples and building classification models to potentially assist medical practitioners in depression diagnosis have to consider the duration effects and gender differences.

Acknowledgements

In the Name of Allah, The Most Gracious, The Most Merciful...
All praises are raised to **Almighty Allah** for His countless blessings and continuous guidance; for giving my soul the strength, courage, and patience needed to complete this journey.

I would like to express my sincere gratitude to Associate Professor Mathias Baumert, my principle supervisor, for his generous guidance, support, and motivation. Thank you for providing me the opportunity to expand my knowledge in the field of "signal processing" and helping me to answer important questions facing psychology–speech communities. Your insightful feedback is what pushed me to sharpen my thinking and bring my work to this level.

My thanks also go to Associate Professor Scott Clark at Adelaide Medical School for his constructive feedback throughout the research stages, as well as the great value he has added to this work. I am also thankful to Associate Professor Dominic Thewlis for providing me with the speech samples used in Chapter 3. My sincere gratitude is also owed to my co-supervisor, Professor Derek Abbott.

I wish to thank my dear aunt, Nora, for encouraging me; my sisters, Shoug and Njoud, for the wonderful time we spent talking together, which has always made me smile; and my brothers, Meshari, Khalifah, and Majed, for their love and trust, and for always asking about me.

A warm thanks go to my husband, Abdulmohsen, for his love, support, understanding, and for believing in me. You were always around whenever I thought it would be impossible to continue; you helped me to keep things in perspective.

Last but not least, I would like to extend my thanks and heartfelt love to my father, Abdulaziz, and my mother, Hila. For their prayers for me, I will never be able to thank them enough; and for their unconditional love, motivation, and eternal support, I am incredibly proud to have them as my parents.

Publication

Journal Article:

S. A. Almaghrabi, D. Thewlis, S. Thwaites, N. C. Rogasch, S. Lau, S. R. Clark, and M. Baumert, “The reproducibility of bio-acoustic features is associated with sample duration, speech task and gender," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 167-175, 2022, doi: 10.1109/TNSRE.2022.3143117.

Contents

List of Figures

List of Tables

List of Abbreviations

MFCC	**M**el-**F**requency **C**epstral **C**oefficients
F0	Fundamental frequency
MDD	**M**ajor **D**epressive **D**isorder
WHO	**W**orld **H**ealth **O**rganization
SVM	**S**upport **V**ector **M**achines
KNN	**K**-**N**earest **N**eighbour
CNN	**C**onvolutional **N**eural **N**etwork
DAIC	**D**istress **A**ssessment **I**nterview **C**orpus
TF	**T**ransfer **F**unction
FT	**F**ourier **T**ransform
LTI	**L**inear **T**ime **I**nvariant
GCI	**G**lottal **C**losure **I**nstant
DYPSA	Dynamic Programming Projected Phase-Slope Algorithm
STFT	**S**hort **T**ime **F**ourier **T**ransform
SC	**S**pectral **C**entroid
SS	**S**pectral **S**kewness
SK	**S**pectral **K**urtosis
SE	**S**pectral **E**ntropy
SF	**S**pectral **F**latness
SR	**S**pectral **R**oll-off
IFT	**I**nverse **F**ourier **T**ransform
DFT	**D**iscrete **F**ourier **T**ransform
ACF	**A**uto **C**orrelation **F**unction
NACF	**N**ormalised **A**uto **C**orrelation **F**unction
SPL	**S**ound **P**ressure **L**evel
LL	**L**oudness **L**evel
ZCR	**Z**ero **C**rossing **R**ate
VP	**V**oicing **P**robability
F1	First Formants
F2	Second Formants
LPC	**L**inear **P**rediction **C**oding
APA	**A**merican **P**sychiatric **A**ssociation
DSM	**D**iagnostic and **S**tatistical **M**anual of Mental Disorders
PHQ	**P**atient **H**ealth **Q**uestionnaire
h	**H**our
s	**S**econd
LLD	**L**ow **L**evel **D**escriptors
SD	**S**tandard **D**eviation

DCT	**D**iscrete **C**osine **T**ransform
ICC	**I**ntraclass **C**orrelation **C**oefficients
ANOVA	**An**alysis **o**f **Va**riance
LOO	**L**eave **O**ne **O**ut
PCA	**P**rincipal **C**omponent **A**nalysis
RBF	**R**adial **B**asis **F**unction

Chapter 1

Introduction

Recent progress in computational speech analysis will potentially enable the application of powerful and effective tools for the analysis of mental disorders in clinical settings. Speech disturbances have been linked to many mental disorders and have been used as an objective biomarker. Previous studies show positive steps toward investigating automatic systems that can diagnose an individual's mental state.

1.1 Speech in Clinical Mental Health Settings

The diagnosis and monitoring of mental health disorders routinely involve self-assessment questionnaires and/or clinicians' opinions [1]. These approaches are prone to a wide range of biases and subjective outcomes [2], [3], leading to inconsistencies in the diagnosis. The potential benefits of implementing a reliable method and an accessible tool for objective assessments have been widely studied. Of these benefits, complementing clinical assessments, enhancing health care quality, and facilitating remote patient monitoring [1]. Artificial intelligence techniques combined with sensor-collected health-related data are used to evaluate mental health conditions automatically [1]. Some of these techniques are based on behavioural descriptors, such as speech.

Human speech carries verbal (linguistic) content such as words, along with non-verbal (paralinguistic) information such as speech tone. It also produces acoustic waves that reflect information about an individual's physiological condition and mental state. The generation of these waves requires enough air pressure to vibrate the vocal folds and produce an acoustic source signal. This signal is then filtered and modulated based on the vocal tract's shape, which is determined by the position of

the speech articulators [4]–[6]. The coordination of speech production is facilitated by several brain areas (e.g., Broca's area, Wernicke's area, and angular gyrus) and the musculoskeletal system [7]. Broca's area is responsible for speech production and articulation, Wernicke's area is responsible for language comprehension, and angular gyrus is associated with auditory and visual information [8].

The neurophysiological changes in the brain associated with mental conditions can potentially disrupt the articulators' coordination and affect the controlling ability of the speech production process. Such changes are encoded into acoustic speech signals, which can be measured through acoustical properties such as source, spectral, prosodic, and formants features [9]–[12]. Disturbances in muscular control of the vocal fold vibration affect source feature values (e.g., jitter and shimmer). The relationship between changes in the vocal tract configuration and speech articulator movements is reflected by spectral characteristics [13]. Spectral features, particularly Mel-frequency cepstral coefficients (MFCCs), are useful in distinguishing mood states [14], [15]. Prosodic measurements, including fundamental frequency (F0) and intensity, are sensitive to changes in an individual's mental conditions. Inappropriate positioning of speech articulators impacts the vocal resonance frequencies (formants location) of the vocal tract. These features can be robustly computed using computerised analysis of the speech waveform [16] and analysed using artificial intelligence techniques [1].

Bio-acoustic properties for applications in the diagnose and monitoring of mental disorders has been showing a growing interest recently. The feasibility and validity of automated assessment of major depressive disorder (MDD) [17]–[19], bipolar disorder [20], schizophrenia [21], and Alzheimer's disease [22] were examined in previous studies. Findings clearly showed the effectiveness of analysing bio-acoustic parameters in diagnosing mental disorders. The assessment methods adopted in these studies were based on machine learning models such as K-nearest neighbour (KNN), support vector machines (SVM), and regression models, as well as on deep learning models such as convolutional neural networks (CNN), and long short-term memory networks.

1.2 Depression

Depression is one of the most common mental disorders [2]. It is characterised by physical, emotional, cognitive, and behavioural symptoms due to the difficulties that patients experience in coping with a stressful life [23], [24]. According to the World Health Organization (WHO), more than 264 million people worldwide are

affected by depression [25]. It has also been predicted that depression will occupy the first place in terms of the global disease burden by 2030 [26]. In Australia, depression is considered the third leading cause of disease burden and the leading cause of non-fatal disability [27], [28]. The diagnosis, evaluation, and treatment of depression in its early stages are essential to ensure effective treatment and improve the quality of human life [19], [29], [30].

Similar to most mental disorders, the diagnosis of depression relies almost exclusively on patients' self-reports and clinicians' opinions. These methods are still the gold standard in clinical assessment for depression, which could lead to a wide range of biases and subjective outcomes. Lack of resources, and trained practitioners are also reported as barriers to effective depression diagnosis [26]. The absence of objective clinical examination in this field has motivated researchers to investigate a more systematic analysis approach based on human speech. Extracting and analysing acoustic speech features potentially will assist in diagnosing depressed patients [1], [2], [17], [23]. Automatic detection systems for MDD have been proposed to enhance the diagnostic efficiency and better characterise this disorder.

1.3 Motivation

Given the growing prevalence of MDD worldwide, the creation of valid, reliable, and objective diagnostic biomarkers has recently been a topic of interest in clinical research. Differences in speech characteristics between depressed and non-depressed individuals have been suggested as a potential biomarker [17], [31]–[33]. Advances in computational speech processing have contributed to the more systematic analysis of speech by means of artificial intelligence [34]. Depression detection studies on the association of speech have formed an active research area for many years. Nevertheless, no standardisation of the acquired speech data in terms of sample length and location is observed. For example, studies have used speech samples that differ regarding speech task type and duration [2], [3], [17], [20], [35], which renders comparability very difficult. Obtaining reproducible and repeatable outcome measurements in acoustic analysis is increasingly important.

Still, no comprehensive analysis has been carried out to study the influence of speech task length on the stability of bio-acoustic qualities obtained from depressed and non-depressed speakers. Thus, the motivation behind this thesis is to investigate the effect of speech sample duration on the reproducibility of

bio-acoustic characteristics in depressed and non-depressed individuals. It also investigates how speech sample length impacts the classification performance of these features. This enables us to understand the stability of bio-acoustic parameters and, therefore, best design speech samples, which potentially improve the reproducibility of speech-based future clinical research.

1.4 Thesis Aim and Objectives

1.4.1 Aim

This thesis aims to evaluate the reproducibility of bio-acoustic features against changes in speech task durations and speech task types in depressed and non-depressed English speakers.

1.4.2 Objectives

The specific objectives of this thesis are to:

- Evaluate the reproducibility of bio-acoustic features in depressed and non-depressed English speakers against changes in speech durations and speech task types.
- Find significant bio-acoustic features that could discriminate between speech with and without depression and which of these are affected by duration reduction.
- Investigate the effect of speech data length on the classification performance of depressed and non-depressed bio-acoustic features measured at different durations.

1.5 Thesis Organisation

This thesis encompasses six chapters, including this introduction. A brief outline of the remaining chapters is as follows:

Chapter 2: This chapter provides the reader with an introductory background on speech as an objective biomarker, starting from the speech production process, moving to the source-filter model, and then providing a deeper review of bio-acoustic features. It also gives a general review of major depressive disorder, its clinical definition, symptoms, diagnostic and assessment tools, and its effect on bio-acoustic features.

Chapter 3: This chapter is dedicated to examining the reproducibility of bio-acoustic features in normal speakers. It highlights the important studies that investigated the stability of bio-acoustic features on the association of the speech task type, speech task duration, and gender. It also describes the dataset used in this evaluation. Besides, the general methods that have been implemented in this thesis to extract features are explained in detail. This chapter concludes with information that help in understanding the reproducibility of features.

Chapter 4: Bio-acoustic features' reproducibility in depressed and non-depressed English speakers is evaluated in this chapter. The main steps used in this examination are also summarised, including dataset, preprocessing, features extraction, and statistical method. This chapter concludes by presenting and discussing reproducibility results of spontaneous speech task.

Chapter 5: In this chapter, the effects of depression on bio-acoustic features are explored. The influence of speech duration on classification metrics for depression detection are also examined. It starts with a summarisation of speech-based studies that investigate depression detection mainly from short durations. Then, it describes the general methods used to build a depression detection system. This chapter concludes by presenting the most significant bio-acoustic features, showing the classification results, and discussing these findings.

Chapter 6: Summary of the main contributions of this thesis is presented in this chapter. Future directions for future researches from this thesis are discussed.

Chapter 2

Literature Review

2.1 The Biological Process of Speech Production

The complexity of neural processing involved in speech production makes speech sensitive to slight changes in the speaker's physiological condition and pathophysiological state [8], [11]. The use of speech-based data as a potential biomarker in MDD diagnostic systems has received considerable attention for many years. Understanding human speech production mechanism is necessary to comprehend speech parameters and then the development of speech-based diagnostic systems.

2.1.1 Speech Production

Human speech production is a complex activity. It involves a three-level process, including cognitive planning, muscular actions, and sound generation (Fig. 2.1). The process starts cognitively by formulating the message and setting up phonetic and prosodic information in the speaker's brain. Phonetic information represents the changes in voice quality. Prosodic information characterises the style and manner of speech. These information, with the help of other brain areas, contributes to establishing the vocalisation plan. This plan is then conveyed to the precentral gyrus in the brain's motor cortex to coordinate muscle activity [7], [36].

Motoric muscular actions require the coordination of around 100 muscles with significant temporal precision. This coordination is uniquely advanced by the control signals transmitted from the motor cortex to motor nuclei (in the brainstem and spinal cord), which moves the articulatory organs in a manner that is consistent with the desired speech sounds. The speech articulators, including the

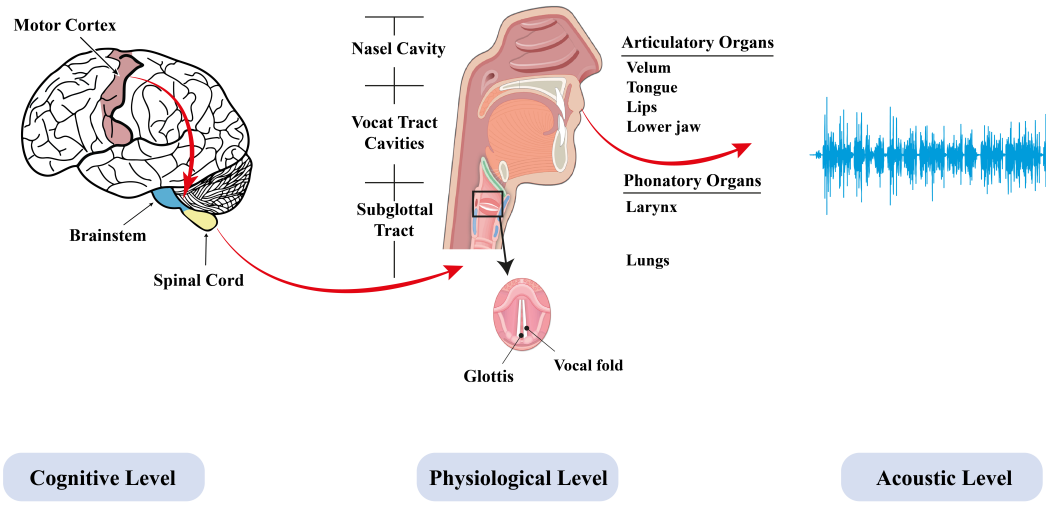

Fig. 2.1 Speech production entails a three-level process: cognitive planning level [37], physiological level (muscular actions) [38], and acoustic level (sound generation). The vocalisation plan is originated in the speaker's brain. Motor nuclei in the brainstem and spinal cord transmit the instructions necessary for muscles coordination activity. The vocal folds vibration generates source sound, which is modulated in the vocal tract area—an area include speech articulators. Finally, speech sounds are radiated from the lips.

lower jaw, tongue, lips, and velum, give resonances to the sound *source* by changing the shape of the vocal tract [7], [36], [39].

The sound *source* is produced as a result of quasi-periodically modulating and balancing the pressurised air across the glottis (in the larynx) by vibrating the vocal fold [36], [39]. This vibration happens when the air expelled from the lungs builds up pressure behind the closed glottis until it crosses a threshold; the glottis is pushed apart, rushing out the air. The released air vibrates the fold in a pulsed manner known as glottal flow pulses, pitch pulses, or vocal excitation pulses. The lungs and larynx are phonatory organs that adjust voice quality and the prosody of speech [36]. Both the phonatory and articulatory organs mutually impact each other during speech production [39]. The resulting speech sound (acoustic wave) is radiated from the oral and nasal cavities after the sound source is shaped in the vocal tract [40]. A schematic drawing of the speech production process is shown in Fig. 2.1.

2.1.2 Acoustic Theory of Voice Production

To conduct a detailed analysis of the speech production process and vocal acoustics, a source-filter model was devised as a two-stage process [36]. The first

stage assumes that the excitation signal $e(t)$, a periodic pulse train with pulse spacing τ_p, is a model of the sound *source* originated at the glottis. In the second stage, this signal is amplified and attenuated by a *filter* [23], [36], [40] with a continuous impulse response, peaking at specific resonances. The filter response represents the transfer function (TF) of the vocal tract $v(t)$ resonant properties [36], [40]. The resulting speech signal $s(t)$, which is periodic with period τ_p, is be obtained by convolving $e(t)$ with $v(t)$ in the time-domain as follows:

$$s(t) = e(t) * v(t). \tag{2.1}$$

In the frequency domain, this involves multiplying the Fourier transform (FT) of the excitation signal and the FT of the vocal tract as follows:

$$S(j\omega) = E(j\omega) \cdot V(j\omega). \tag{2.2}$$

The shape of the vocal tract determines its frequency response, which in turn specifies the frequency spacing in the line spectrum and the envelope of the speech signal [40]. Thus, the excitation/source signal (input) is passed through the vocal tract/filter (linear-time invariant (LTI) system). The speech signal is the output of the LTI system [41]. An illustration of the model is shown in Fig. 2.2.

2.1.3 Bio-acoustic Features

Bio-acoustic speech features can be divided into four groups: source, spectral, prosodic, and formants features [1]. These features are described in detail in this section.

2.1.3.1 Source Features

Source features reflect information about the sound *source* at the glottis during natural voice production [23]. There are two basic categories of source features: glottal features, which model the glottal flow; and voice quality features, which measure vocal fold vibration [1]. Jitter and shimmer are well-established voice quality measures for objectively, non-invasively, and quantitatively evaluating different physiological characteristics of the vocal folds [44]. The cycle-to-cycle variations in glottal period duration (jitter) and amplitude (shimmer) measure the micro-perturbations in vocal fold vibration [45], [46] (Fig. 2.3). This vibration is affected by biomechanical factors (asymmetric vocal cords), neurogenic factors (involuntary movement of the larynx muscles), and aerodynamic factors (airflow

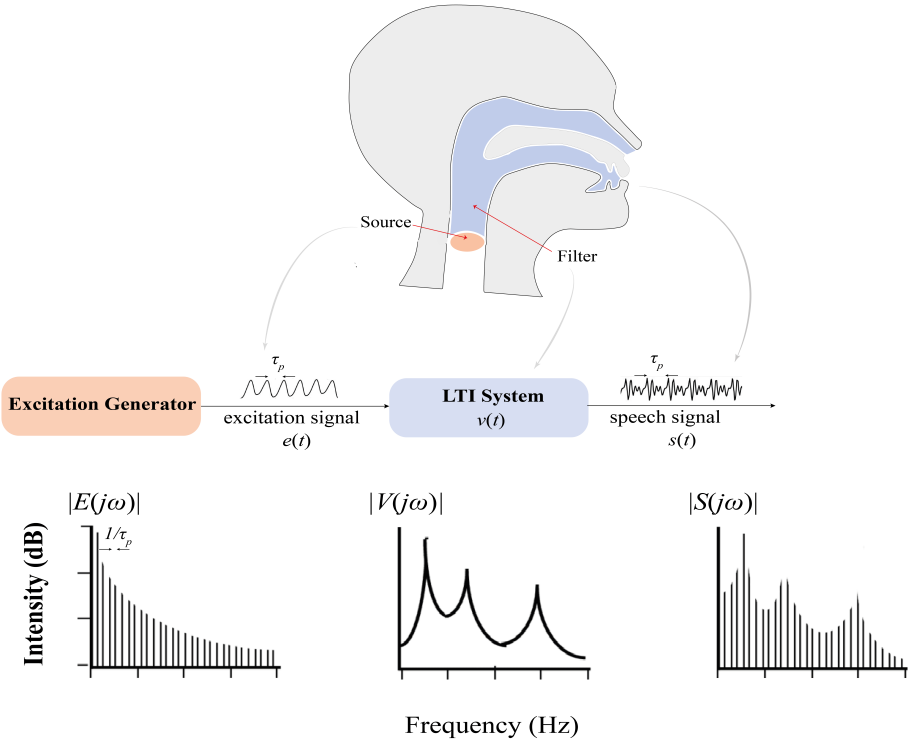

Fig. 2.2 The source-filter model of speech production comprises the glottis as the source of the excitation signal and the vocal tract [42] (i.e., nasal and oral cavities) as the filter. The model also shows the temporal and spectral representations [43] of the source, vocal tract, and resulting speech signal [40].

and sub-glottal pressure fluctuations) [39]. The lack of control over vocal fold vibration influences jitter, while glottal resistance and mass lesions on the vocal folds influence shimmer [47]. A typical jitter during sustained voiced sound in adults ranges between 0.50% and 1.00% [47] and shimmer ranges between 0.05 dB and 0.22 dB [48].

Estimation of voice perturbation is generally based on pitch-mark detection [49]; temporal location of short-time peaks in each glottal cycle (pitch period) [50]. Pitch-marks are positioned pitch-synchronously [50] to define cycle boundaries. Some of the common time-domain pitch marking approaches are *waveform-matching*, which estimates the time of best-matching cycle-to-cycle waveforms, and *peak-picking*, which locates the instantaneous peaks of the waveform [49]. The consensus is in favour of waveform-matching due to its robustness against noise variations compared to peak-picking [51]–[53]. However, the waveform-matching approach implicitly assumes periodicity constraints that are insufficient to characterise the low-periodicity of pathological and breathy voices [54].

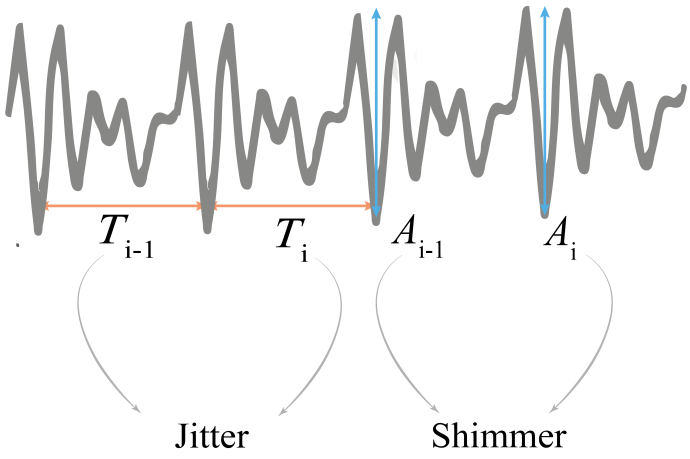

Fig. 2.3 Representation of jitter and shimmer measures; cycle-to-cycle variations in glottal period duration (T_i) and amplitude (A_i), respectively.

The detection of the *Glottal Closure Instant (GCI)* offers opportunities for a reliable approach to localise pitch-marks [36], [49], [55] while maintaining pitch-synchronisation and ignoring periodicity variations of the source signal [54], [56]. Physiologically, GCI is the closing moment of the glottis [36], which marks the glottal closure completion phase [55]. It is also known as the prominent peaks of the time-derivative of the glottal flow signal [57] (Fig. 2.4). The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) is a popular algorithm for automatic GCI estimation, which correctly identified more than 93% of GCIs [58]. This identification is based on the complex dynamics of speech, such as quasi-stationarity [36], [58]. Knowledge of GCIs location, as pitch-marks, is the foundation of jitter and shimmer measurements in this research.

2.1.3.2 Spectral Features

Spectral features capture prosodic, phonetic, and articulatory information associated with speech motor control [23]. They also reflect the relationship between vocal tract configuration and the movement of speech articulators [13], [59]. Changes in these features are correlated with the speaker's mental state [59], relating to the disturbances in muscle tension and psychomotor retardation [60]. Spectral features characterise the *speech spectrum*; the spectral distribution of the speech waveform for a given time [1]. This spectrum is calculated either along the frequency-domain with a linear scale (Hz) or along the Mel-bands with a non-linear scale (Mel) [61]. Spectral features, in this research, are divided into two

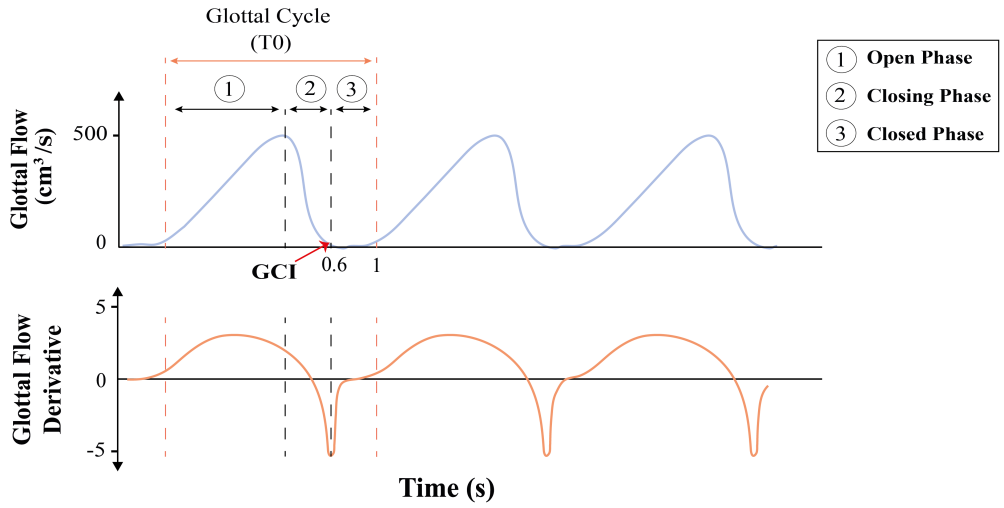

Fig. 2.4 Glottal flow waveform (top) with marking three phases (opening, closing and closed phase) of the glottal cycle, and Glottal Closure Instant (GCI). The derivative of the glottal flow (bottom) is also shown.

main categories: frequency-domain (physical) features and cepstral-domain (perceptual) features.

Frequency-domain features (e.g., spectral shape descriptors) are computed by converting a time-domain signal into a frequency-domain signal using short-time Fourier transform (STFT). *Spectral shape descriptors*, also called timbrel attributes, are one of the major frequency-domain features [62]. They describe the shape of the magnitude spectrum [63]. Most of these descriptors are related to the timbrel characteristics of the speech signal [61], [64]. Timbre is the perceptual attribute of auditory sensation that helps a listener discriminate between sounds, even if they have a similar pitch and loudness [65]. In psychoacoustics, timbre is often referred to as the colour, or quality of sound [63]. The following spectral shape descriptors are employed in this thesis:

- *Spectral centroid (SC)*: This descriptors is well-correlated with the perceived timbrel *brightness* of a sound signal [62]. Higher centroid values correspond to brighter sounds (or more nasal sounds) with more energy in the high-frequencies [63], [66]. This measure is included in the MPEG-7 standard (i.e., contains a set of low-level audio descriptors useful in describing audio and designing higher-level audio applications) for audio descriptors [67].
- *Spectral skewness (SS)*: This refers to spectral tilt in phonetic terms. SS is used with other spectral moments to distinguish the articulation place. It is a

measure of a distribution's symmetry around its centroid [68]. The SS value drops to zero during pauses and increases for voiced parts [62]. A skewness of zero describes a symmetrical distribution, whereas positive skewness describes a further extension of the distribution's right "tail" than the left tail; negative skewness describes a further extension of the distribution's left "tail" than the right tail [68].

- *Spectral kurtosis (SK)*: This descriptor is used to identify the manners of articulation. SK represents the "peakedness" of a distribution. Positive values indicate a distribution with a relatively acute peak, negative values indicate a relatively flat distribution, and zero indicates a Gaussian distribution [68].
- *Spectral entropy (SE)*: This is a method used to measure the amount of information present in a signal (Shannon's information theory). It is also used to describe the peakiness of the spectral distribution [28], [69]. Speech segments have lower SE values compared to noise or non-speech segments, which is attributable to the non-uniform distribution of energy across the frequency (concentrated in specific frequency bands). Noise or non-speech segments have higher SE values because the segment's energy is flat (i.e., uniformly distributed).
- *Spectral flatness (SF)*: SF refers to the perceptual quality of *tone-likeness*. It is a method for quantifying the noise-like or tone-like aspects of sound. This quantity also reflects the stability of the speech signal [70]. A high SF value (close to one) indicates noise, while a lower value (close to zero) indicates tonality [62]. This measure is part of the MPEG-7 audio descriptors standard [67].
- *Spectral roll-off points (SR)*: This descriptors is the frequency below which a certain amount of energy in a signal is concentrated. In a speech signal, the energy tends to be lower at high frequencies [28]. A low scalar value of SR indicates the presence of a tone, while a high value indicates the presence of noise-filled pauses [63]. Thus, the SR descriptor is useful for discriminating between voiced and unvoiced sounds [66]. Voiced sounds are generated by vibrating the vocal folds during the phoneme pronunciation, while unvoiced sounds the vocal cords do not engage.

Cepstral domain analysis, based on homomorphic analysis, has been largely employed in speech-related applications. It is a method for temporal separation of the source-filter model components, described by a convolution relation (Equ. 2.1). By exploiting properties of the FT, this convolution can be expressed by

multiplication in the frequency-domain (as shown in Equ. 2.2). Taking the magnitude of the speech spectrum $S(j\omega)$ yields to:

$$|S(j\omega)| = |E(j\omega)| \cdot |V(j\omega)|. \tag{2.3}$$

To linearly combine $E(j\omega)$ and $V(j\omega)$ in the frequency-domain, the logarithm is applied to both sides of Equ. 2.3 as follows:

$$\log|S(j\omega)| = \log|E(j\omega)| + \log|V(j\omega)|. \tag{2.4}$$

Hence, the source and filter components can be separated by taking the inverse of Fourier transform (IFT), denoted by $\mathcal{F}^{-1}\{\cdot\}$, of the above equation:

$$\mathcal{F}^{-1}\{\log|S(j\omega)|\} = \mathcal{F}^{-1}\{\log|E(j\omega)|\} + \mathcal{F}^{-1}\{|V(j\omega)|\}. \tag{2.5}$$

Obtaining the IFT of the log spectrum is called "Cepstrum". The new cepstral representation domain is called *quefrency domain* [71]. Cepstrum can be defined as a complex, power, phase, and real cepstrum. Cepstrum for a power spectrum is the most relevant to the speech signal processing [62]. The following relationship can define it:

$$C_p = |\mathcal{F}^{-1}\{\log\Big(|\mathcal{F}\{s(t)\}|^2\Big)\}|^2, \tag{2.6}$$

where C_p is a power cepstrum, $\mathcal{F}\{\cdot\}$ indicates FT or a discrete Fourier transform (DFT), and $s(t)$ is a speech signal in the time-domain.

In the quefrency domain, *liftering* operation is performed on the speech cepstrum to independently extract the vocal tract impulse response, represented by low quefrency components, and the excitation signal, represented by high quefrency components [72]. Fig. 2.5 shows the speech signal components in the quefrency domain. Accordingly, this technique can be used to describe the vocal tract coordination that corresponds to the magnitude of the speech cepstrum [73], [74].

MFCC is the most common short-term cepstral features [1], [76]. It is proposed to create a perceptually relevant representation of the vocal tract spectral shape at any specific time [73], [74]. MFCC is based on the physiological properties of the human auditory system. Human hearing has a non-linear relationship with frequencies higher than 1 kHz [40]. Similarly, MFCC maps a linear frequency scale (Hz) into a non-linear Mel-Scale (Mels) [66]. Measuring MFCC filters out not audible frequencies by removing redundant information in STFT [39], [74]. Sensitivity of MFCC to additive noise is one of the most notable downside [77] because it uses both formant and non-formant regions of the power spectrum [78].

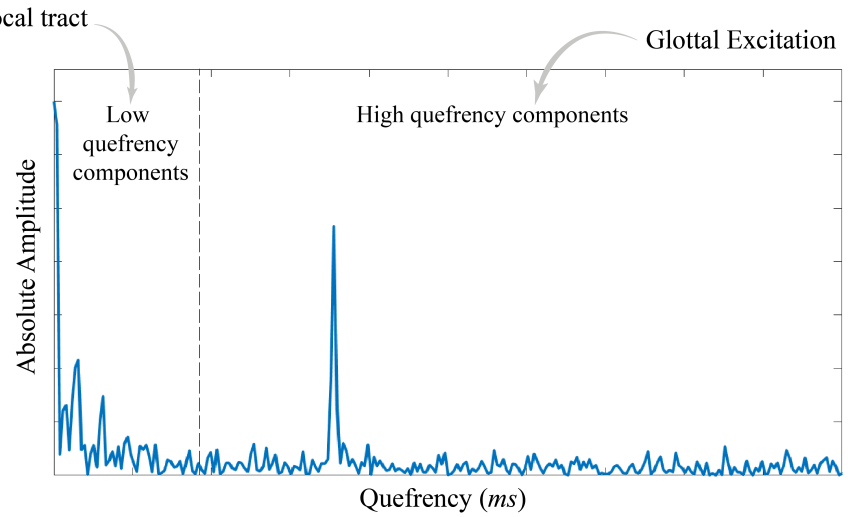

Fig. 2.5 Cepstrum of a speech segment shows the excitation and vocal tract components of speech [75].

The computation steps of MFCCs is described in Section 3.2.2.2.

Since speech is inherently a dynamic signal, regularly changes over time, the time derivative (first- and second-order derivatives) of cepstrum coefficients proposed to capture temporal dynamic information [79]. These derivatives are expressed as MFCC delta and MFCC delta-delta, denoted as ΔMFCC and ΔΔMFCC, respectively. ΔMFCC reflect information about the speech rate, while ΔΔMFCC provides information about speech acceleration.

2.1.3.3 Prosodic Features

Prosody is responsible for controlling variations in pitch, loudness, stress, and rhythmic speech organisation [80]. Prosodic features (or long-term features) describe these variations [81] and reflect the differences in an individual's speaking style [23]. Prosody is adversely influenced by the neurological conditions and psychological states of a speaker [82]. Pitch and loudness are the most commonly used perceptual prosodic descriptors.

Pitch is a subjective psycho-acoustical attribute of sound. It characterises the glottal excitation rate [14]. Physically, it is known as F0 represented by the number of vocal fold vibration cycles per second [80]. Paralinguistically, it is responsible for the expressiveness of speech and known as major carrier of prosodic information [83]. In adult speakers, the measure is usually higher in women (200–220 Hz), who typically have short and thin vocal folds, compared to men (100–120 Hz), who have long and thick vocal folds [84]. Differences in F0 values mainly depend on vocal fold anatomy, larynx size, and aerodynamic adjustment factors [39]. Several

algorithms in the time-domain, frequency-domain, and time-frequency domain, have been proposed to estimate the F0 value and detect the periodicity of a given speech frame. Auto-correlation function (ACF) is a well-known time-domain method for estimating F0. It measures the signal's self-similarity at given discrete-time lags or delays (τ) [41]. The ACF for an infinite signal length (N) is given by:

$$\mathrm{ACF}(\tau) = \sum_{n=0}^{N-1} x(n+\tau)x(n). \tag{2.7}$$

The robustness of this method against a white noisy environment has been reported in previous research [85]. However, ACF may lead to pitch halving or pitch doubling errors when the first formant masking effects impact the identification of F0 [86]. The error rate of pitch estimation using this method is also significantly influenced by the characteristics of the vocal tract [85]. To overcome these limitations, the normalised correlation function (NACF) for pitch estimation with higher accuracy was proposed, where peaks are more prominent and less impacted by rapid changes in speech signal amplitude [86]. Other popular methods for pitch estimation include, but are not limited to, short time cepstrum analysis, peak picking, and sub-harmonic summation. In this work, NACF method, a good candidate for estimating F0, is employed.

Loudness represents the perceived intensity of a speech signal—loudness is a perceptual attribute, while intensity is a physical attribute. It is governed by physiological characteristics (e.g., glottal excitation strength, sub-glottal pressure, and the vocal tract's resonant properties) and the speaker's behavioural characteristics (e.g., mood state) [36], [87]. Human perception of loudness is non-linear with respect to changes in frequency and intensity [63], as given in Fig. 2.6. Each curve, known as equal-loudness contour, represents the sound pressure level (SPL) in dB and intensity in $\mathrm{W/m^2}$ at which sounds of various frequencies are equally loud. The curves are labelled with the loudness level (LL) in phons, which is numerically equal to SPL at 1 kHz [40], [88]. The perceived loudness of sounds (in sones) is not directly equal to LL. It is defined as having a value of 1 sones at a LL of 40 phons. Loudness, as a function of LL, can be approximated above 40 phons by the following [40]:

$$\text{Loudness (sones)} = 2^{(\mathrm{LL}-40)/10}. \tag{2.8}$$

An earlier set of equal-loudness contours for the auditory sensation was published in 1933 by Fletcher and Munson [89]. Later, these curves have been standardised under an international standard (ISO 226) and called "isophonic curves".

Examination of these curves in Fig. 2.6 indicates that the ear seems to be more sensitive within the frequency range of 3 kHz and 4 kHz [88], [90]. Generally, the sound level of a conversation in normal speech is about 60 dB [40].

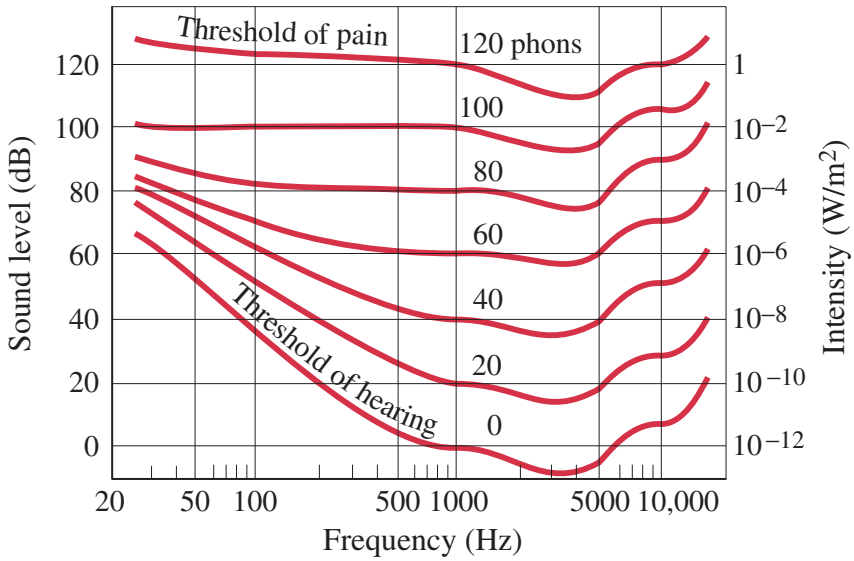

Fig. 2.6 Sensitivity of the human ear as a function of frequency. Each curve represents the sound level (dB) and intensity (W/m^2) as a function of frequency for a fixed loudness level (phons) for pure tones [88].

Several loudness measures have been proposed based on physiological and acoustic characteristics. Of these measures, SPL and sub-glottal pressure levels have been shown to be strongly correlated with perceived loudness. Other measures such as maximum flow declination rate, which is derived primarily from the acoustic signal, utilise the characteristics of vocal fold vibrations for loudness estimation [87]. Loudness estimation, in this thesis, is based on SPL measurement.

Another time-domain prosodic feature is the zero-crossing rate (ZCR) [80]. ZCR identifies the presence of human speech in a speech sample, where unvoiced portions have a high ZCR and voiced segments usually show a low ZCR [62]. This feature can be used to estimate the frequency at which the speech energy is concentrated in the spectrum. It is also considered a good indicator of short and loud sounds [91]. Voicing probability (VP) determines the speech-silence pattern in the participants' speech by estimating the percentage of voiced and unvoiced energy for each harmonic.

2.1.3.4 Formant Features

Vocal tract resonances, which spectrally shape the glottal excitation signal in the speech production process, vary over several pitch periods. The dynamics of vocal resonance are governed by speech articulators and reflected in formant frequency locations in the speech spectrum [14]. Tracking resonant frequencies mainly captures information about the coordination of speech articulators [73].

Formant (or filter) features are distinctive frequency components (peaks) at which the acoustic energy of a speech signal is concentrated [28], [92]. Theoretically, an acoustic signal contains an infinite number of formants, but only three to four formants are within the range of human hearing [93]. The first- and second-formant frequencies, or simply F1 and F2 play a key role in determining the vowel quality [39], while the upper three (F3, F4, and F5) determine the colour of an individual voice [94]. Formant locations do not disrupt with additive white-noise, leading to a higher signal-to-noise ratio in formant regions than in non-formant regions [78]. Fig. 2.7, as an example, shows six distinct formant peaks of a short-time speech. Formant frequencies provide information pertaining to the vocal tract's shape, while formant amplitudes reflect vocal intensity levels. Most often, differences in vocal anatomy between adult men and women affect the position of formant frequencies. Accordingly, the formant feature is a well-known gender-dependent acoustic measure [84].

Typical approaches to track and estimate formant frequencies involve peak-picking of speech spectral representations, usually from STFT, cepstrum, and linear prediction coding (LPC) analysis. LPC (or all-pole filter) is capable of providing an accurate estimation of the speech spectral envelope using a linear prediction model, which is not the case in the STFT and cepstrum methods. It predicts the next values by linearly combining the previously known coefficients. However, increasing the amount of noise in the speech signal significantly influences LPC [62], [95]. The most recent method using deep learning algorithms to obtain formants is proposed. In this method, the speech signal is represented by either spectrogram or cepstral coefficients derived from LPC, as well as quasi-pitch-synchronous [96] (detailed in Section 3.2.2.2).

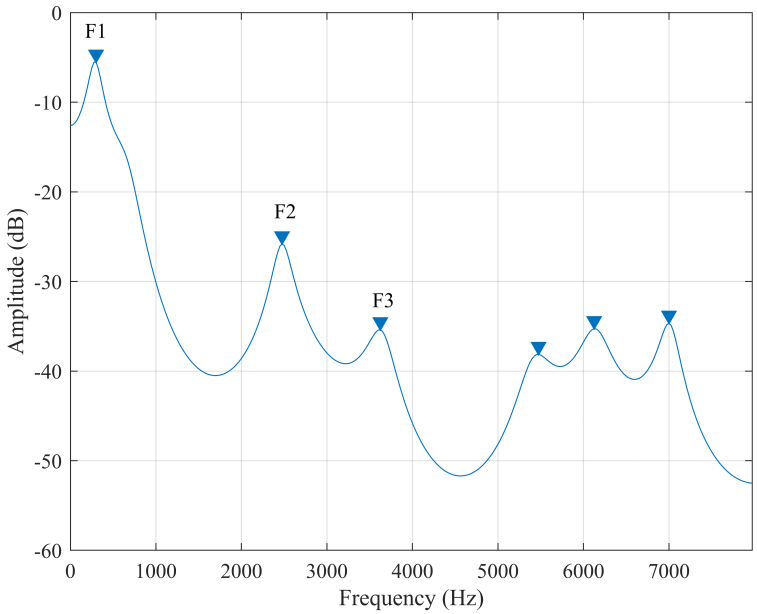

Fig. 2.7 Spectrum shows the distinct formant peaks of a speech frame with 20 ms duration of woman voice ('a' vowel) and computed using LPC method. It also shows that spectral energy falls between two adjacent formants.

2.2 Major Depressive Disorder

2.2.1 Clinical Definition and Symptoms

Clinical depression, also known as MDD, is a common and serious mental disorder. The American Psychiatric Association (APA) defines clinical depression as a medical condition that lasts for at least two consecutive weeks consisting of a persistently depressed mood, negativity, feelings of sadness, and/or loss of interest in activities, which causes significant impairment in coping with daily life. Environmental, psychological, biochemical, and genetic factors may all be involved in the pathophysiology of depression [97].

The APA has outlined diagnostic criteria for mental disorders, including depression, to provide a common ground and improve the classification of these disorders. These criteria are defined in the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* [98]. The DSM criteria encompass a wide range of recognisable symptoms of MDD. For a positive diagnosis of depression, five (or more) of the defined symptoms, in combination with either a depressed mood or markedly diminished interest, must be present nearly every day over a two-week period. Some of these symptoms include [98]:

- Psychomotor agitation or retardation,
- Fatigue or loss of energy,
- Feelings of worthlessness or inappropriate guilt,
- Recurrent thoughts of death and suicide,
- Changes in appetite, including significant weight loss or gain unrelated to dieting.

Although the DSM criteria have standardised the diagnosis of psychiatric disorders based on empirical evidence, criticisms have also highlighted the heterogeneity of its definition of depressive syndrome [99], the possibility of cultural bias [100], and the reliability and validity of DSM-based depression diagnoses [101]. Based on these criteria, Østergaard *et al.* calculated at least $1,497$ unique profiles of depression [99]. Freedman *et al.* found fair reliability of MDD diagnoses, with a kappa statistic of 0.28; this indicates a high level of disagreement on MDD diagnoses between clinicians [101]. Singer reported different manifestations of depressive disorder in Western cultures (cognitive symptoms) compared to non-Western cultures (somatic symptoms) [100]. Hence, the complexity level of fitting the clinical profile of a depressed individual into an objective level is considerable [23], which suggests the usefulness of narrowly redefining clinical depressive syndrome [99]. However, these criteria still constitute the most widely used and standard definition available, where a broad range of the existing assessment tools are scored based on depression symptoms listed in the DSM.

2.2.2 Diagnostic and Assessment Tools for Depression

Depression diagnosis in primary care settings primarily relies on patient responses, which are typically impacted the professionals' evaluations or self-assessment questionnaires. This approach is often subject to a significant variation depending on each clinician's expertise and the diagnostic test used [1], [19]. Although these methods have the potential for a wide range of biases and subjective outcomes, they are still the gold standard of clinical assessments of depression [1].

Common assessment tools used for depression diagnosis are clinical interviews, rating scales, and self-assessment reports. The eight-item Patient Health Questionnaire (PHQ-8), a self-reported questionnaire, is established as a valid screening measure of MDD severity [102]. It comprises eight criteria for depression assessments derived from the DSM. Each criterion is scored on a scale from 0 to 3, providing a 0 to 24 depression severity score, where PHQ-8 score ≥ 10 represents a

significant depression level (see Appendix A for details). This test is used in this thesis as a ground truth score of the depressed speech database.

Other diagnostic tools include the Hamilton Rating Scale for Depression (a clinician-rated questionnaire) [103], the Quick Inventory of Depressive Symptoms-Self Report (self-reported questionnaire) [104], and the Beck Depression Inventory (self-reported questionnaire) [105]. These tests rate the severity of depression symptoms by scoring a patient's depression level. Each scale uses a different number of items, a different set of symptoms, and a different weighting scheme, which leads to inconsistencies in depression diagnoses. Accordingly, an objective marker (i.e., speech) is needed for depression diagnose to enhance the available tools and support clinical practice [23]. Several studies investigated biological (e.g., blood and saliva) [106], physiological (e.g., biosignals) [107] and behavioural (e.g., facial expressions) [33] measures to detect depression. Results reported the feasibility of using these biomarkers in diagnosing depression as a complement to the clinical assessments. This thesis is focused on the extraction and analysing of behavioural measures, particularly bio-acoustic features.

2.2.3 Depression and Bio-acoustic Features

Qualitative changes of speech with depression have been reported decades ago [108]. In the clinic, speech behaviours and communicative defects associated with depression are well-documented. Kraepelin, in 1921, described speech characteristics in depressed patients as follows: "patients speak in a low voice, slowly, hesitatingly, monotonously, sometimes stuttering, whispering, try several times before they bring out a word, become mute in the middle of a sentence" [109]. In fact, neural changes in a depressed patient's brain manifest behaviourally during the speech production process, which changes the acoustic quality of the produced speech [31], [73]. The bio-acoustic characteristics of human speech have previously been identified as a possible marker to objectively discriminate between depressed and non-depressed speech [17], [31]–[33].

Studies on vocal-source biomarkers have reported a noticeable acoustic abnormalities in both jitter and shimmer in patients with depression [1], [31], [110]. This abnormality is due to neuro-physiological changes in muscle tension, laryngeal control, and the movement of the vocal folds [31]. Quatieri and Malyska, in 2012, reported a significant increase in jitter and shimmer values with increasing depression severity and psychomotor retardation [31]. Similarly, in 2017, Kiss and Vicsi observed higher jitter and shimmer values in depressed individuals compared to healthy controls [111]. A study in 2019 found an increase in shimmer

values of men and women and an increase in jitter rate values in men only, when depressed to non-depressed individuals were compared [112]. More recently, in 2021, Silva *et al.* showed higher mean values of both jitter and shimmer in patients diagnosed with depression than those without depression [113], which is generally agreed with previous studies [31], [111], [112]. However, increased shimmer and jitter values have not been consistently observed in depressed speech studies. For example, Hönig *et al.* found a strong negative correlation between shimmer and depression levels [110]. Discrepancies across these studies might be due to the differences in signal processing methods for source feature extraction and differences in the utilised speech segments—jitter and shimmer measures are more accurate during steady voices and more periodic signals (e.g., vowel). These signals are produced stable F0 and loudness to maintain a stable articulatory condition for voice quality assessment, avoiding confounds results from interactions between the larynx and vocal tract [114].

The relative shift in energy is one of the main spectral effects investigated in depressed speech studies. This shifting results from increasing muscle tension, limiting articulatory movements, which affects the physiological coordination of the vocal tract and, hence, its resonance properties. France *et al.* reported an energy shift from lower to higher frequency bands, leading to a significant increase in energy in the higher band for a speech with depression. Kiss and Vicsi found that in depressed speech, the mean energy value of the low-frequency region (65–400 Hz) is relatively shifted to a higher value, while that of the high-frequency region (1,330–5,735 Hz) is shifted to a lower value [111]. A similar shifting trend was observed by Yingthawornsuk *et al.*, where the lower band was defined as 0–500 Hz and the higher band was defined as 500–1,000 Hz [115]. Typically, in an adult's voice, the majority of the energy is contained in the frequency band between 0 Hz and 2000 Hz [116]. In addition, a flattening pattern in the speech spectrum was observed in patients experiencing MDD [60]. Spectral centroid and spectral entropy were found to be higher in non-depressed speech, with higher timbre brightness and richer spectral information [69].

MFCCs are significantly impacted by speech content and have been effectively used in speech content characterisation and automatic speech recognition [117], [118]. These coeffecients represent perception-based sound where the same words are not strongly influnced by changes that occur while being voiced [119]. Measuring MFCC parameters, as distinguishing features, was carried out to maximise the depressed and control classification performance with an accuracy of 80% [15]. Taguchi *et al.* reported that while depression is associated with a

significant increase in MFCC2, other MFCC features remained stable [120]. Wang et al., in contrast, found that MFCC5 and MFCC7 can distinguish between people with and without depression; higher mean values are found in healthy individuals [121]. We speculate that the differences between the two studies may be relevant to the language differences and speech task type. Additionally, differences in MFCC3 between depressed and non-depressed speakers depends on speech scenarios (e.g., speech task type and voice expressions) [121].

Changes in prosodic features in relation to depression have been widely studied. An increase in vocal tract tension during depression tightens the vocal folds [83] and alteration in salivation and mucus secretion affects vocal tract and articulatory movements [122], which results in monotone and less variable speech [23]. In 1980, Hollien characterised depressed patient's speech by a reduction in both pitch and speaking intensity [123]. Later, numerous studies found a negative correlation of these descriptors with depression [1], [111], [120], [121], supporting Hollien's findings. However, in 2012, Quatieri and Malyska highlighted a reduction in pitch (variance and average) with decreasing depression severity [31]. In 2021, a higher standard deviation F0 parameter was also found in depressed than in non-depressed individuals [113]. Similarly, Ellgring and Scherer found that minimum F0 decreased in women voices in a recovered state [83], meaning that they had a higher and less variable pitch in depression. In 2015, Hussenbocus and Allen reported a decreased in F0 values for men in depression and an increased on those of depressed women [69]. A potential reason for this difference is the heterogeneity of depression symptoms [23]. Furthermore, Wang *et al.* found a reduction in VP measure. However, this reduction was not significant in most speech scenarios. A similar ZCR was reported when voices of non-depressed and depressed individuals were compared [120], [121].

Formant's behaviour has been shown to be sensitive to depression, where it reveals the articulatory effort reduction and psychomotor disturbances [124]. Thus, it is considered a significantly distinguishable feature for depression classification [60]. Generally, Kiss and Vicsi observed a reduction trend associated with depression in the mean values of formant frequencies (F1, F2) [111]. Flint's result also indicated a decrease in F2 location during periods of depression [124]. Similarly, Vicsi *et al.* reported a decrease in formants locations of depressed individuals in comparison with healthy controls [125]. The identified decreasing trend of formant frequencies is inconsistent with the results presented by France; his result showed an increase in F1 and F2 locations in the case of depressed participants [60]. The potential cause of this discrepancy might be the complex relationship between source and

Table 2.1 DIRECTION EFFECT OF BIO-ACOUSTIC FEATURES' VALUES WITH DEPRESSION.

Feature [a]	Direction changed			
	Increased	References	Decreased	References
Jitter	✓	[31], [111]–[113]		
Shimmer	✓	[31], [111]–[113]	✓	[110]
Energy shift	✓	[60], [111], [115]	✓	[31]
SC	✓	[69]		
SE	✓	[69]		
MFCC2	✓	[120]	✓	[121]
MFCC5			✓	[121]
MFCC7			✓	[121]
Pitch	✓	[31], [83], [113]	✓	[111], [120], [121], [123]
Intensity			✓	[111], [123]
Loudness			✓	[121]
VP			✓	[121]
F1	✓	[60]	✓	[111], [125]
F1	✓	[60]	✓	[111], [124], [125]

[a] SC: Spectral centroid; SE: spectral entropy; MFCC: Mel-frequency cepstral coefficients; VP: Voicing probability; F1: First formant; F2: Second formant.

filter dynamics [23].

2.2.4 Automated Depression Detection Systems

The application of computational acoustical analysis to mental disorder assessment has gained increasing interest recently [126]. Studies on depression have been performed to identify depressed speech using artificial intelligence techniques [2], [3], [17]–[19]. The discrepancies across these studies that have been observed are attributable to differences in speech samples, the set of bio-acoustic features, and machine learning algorithms, rendering direct comparability between them impossible. In this section, a summary is given of some of the investigations that have been undertaken into the automatic analysis of acoustic characteristics as a predictor of depression.

To identify depression severity, Cohn *et al.* analysed vocal prosodic expression, particularly pitch, and speaker switch duration in a sample of 28 participants [33]. The authors used a logistic regression classifier that achieved an accuracy of 79%. Prosodic qualities were measured with the help of publicly available computer software called Praat [92]. Another study, performed in 2011, investigated speech with depression on a 47-speaker sample (23 depressed patients and 24 healthy controls), where each participant was asked to read a two set of sentences. Some

prosodic features and detailed spectral information were measured using VoiceBox [127]. Using Gaussian mixture models (GMM), classification accuracy approached 80% when MFCC and formant features were analysed [15].

Helfer *et al.*, in 2013, used audio data obtained from 35 subjects (using the *James Mundt 35-speaker Database*) to explore depression classification of GMM and SVM classifiers by analysing the first three formant frequencies and their dynamics. They reported an optimal classification performance with sensitivity/specificity of 0.86/0.64 and 0.77/0.77 for GMMs and SVMs, respectively [10]. An investigation of voice quality features for depression detection was conducted by Scherer *et al.* [128]. Features were extracted from 36 participants, representing a subset of the *Distress analysis interview corpus* (DAIC) [129]. An accuracy of 75% was achieved using the SVM classifier to identify depression. Alghowinem *et al.* used a balanced dataset of 60 subjects (30 depressed and 30 non-depressed) to extract voice parameters using the openSMILE computer software [130]. Their results showed a remarkable performance with an average recall of 81.61% in depression detection when a hybrid classifier consisting of GMM and SVM was used, and the classification performance of four classifiers was compared [131].

In 2016, Valstar *et al.* used clinical interview samples from the DAIC [129] to identify non-verbal indicators of depression [132]. They extracted prosodic, voice quality, and spectral features using the *Cooperative Voice Analysis Repository for Speech Technologies* [133], which is open source and freely available. A linear SVM with stochastic gradient descent was fit on the training set and validated on the development set. They reported the following baseline results for depression classification on a test set of depressed and non-depressed classes, respectively: F1-scores were 0.410 and 0.582; precision values were 0.267 and 0.941; and recall values were 0.889 and 0.421.

Jiang *et al.* investigated the discriminative power of different classifiers (KNN, GMM, and SVM) for depression identification using balanced speech samples of 170 speakers, modelling males and females separately. The authors used openSMILE to quantify several acoustic features [130]. Their results showed that SVM achieved the best classification performance, resulting in accuracies of 65.68% and 65.78% for females and males, respectively [2]. Another study employed a parallel SVM algorithm to examine the classification accuracy of a set of bio-acoustic features obtained from 74 speakers (37 depressed and 37 non-depressed) [134]. Measured bio-acoustic characteristics mainly include source, prosodic, spectral, and formants features. Results showed a high classification performance with an accuracy of 78.02%.

In 2018, Stolar *et al.* investigated adolescent depression identification in a clinical speech data sample consisting of 63 speakers (29 depressed and 34 non-depressed). A set of the extracted acoustic parameters was fed into a SVM classifier to discriminate between depressed and non-depressed speech characteristics. On average, classification accuracies of 82.2% and 70.5% were achieved for males and females, receptively [28]. In 2019, McGinnis *et al.* used speech data of 71 children that was recorded during a speaking task of three minute duration. They implement a different machine learning models (logistic regression, SVM, and random forest) to detect depression, resulting in a good identification accuracy (around 80%) [135].

A study analysed the vocal acoustic of 33 individuals (22 diagnosed with depression and 11 healthy controls) to detect MDD [136]. Voice data were submitted to GNU Octave^{TM}, an open-source computer software for the extraction of vocal features. Mean accuracy of 87.55% was reported using random tree models with 100 trees. Additionally, Saidi *et al.* employed a hybrid model combing CNN and SVM to detect depression [137]. The authors evaluated the model on speech samples from the DAIC and reported an accuracy of 68%, which outperformed the CNN model's accuracy of 58.57%. Aharonson *et al.* implemented two machine learning architectures, trained on acoustic features extracted from DAIC-WOZ speech samples, for depression analysis [138]. First they used a binary classifier followed by a regression model, then compared a five-class classifier followed by a regression model. They reached 78.84% and 82.22% of classification accuracy, respectively.

In 2021, Lee *et al.* developed a voice-based automated diagnostic system for depression screening. Voice samples were derived from reading pre-defined sentences by elderly people (61 diagnosed with depression and 143 without depression) [139]. Acoustic measures were analysed using OpenSMILE computer software. Significant discriminatory performances were found with 86% and 77% accuracy for males and females, respectively [139]. Patil and Wadhai extracted bio-acoustic features using Praat computer software [92], to study the performance of different classifiers for depression detection [140]. They utilised spontaneous speech samples (collected during an interview) of 54 depressed patients and 75 healthy controls. Their results showed that a hybrid classifier (GMM and SVM) achieved the best overall classification accuracy with a mean of around 83% [140].

Recently, in 2022, Rejaibi *et al.* adopted a recurrent neural network structure to identify depression and predict its severity level using acoustic speech

characteristics (e.g., MFCCs), reporting an overall accuracy of 86% [141]. Speech samples was taken from the DAIC-WOZ database.

Table 2.2 A SUMMARY OF AUTOMATED DEPRESSION DETECTION SYSTEMS STUDIES.

Year	Reference	n [a]	Features	Classifier [b]	Metrics[c]	Values	Confounders
2009	Cohn *et al.* [33]	11 ND 17 D	F0 variability Latency	LR	Accuracy	79%	Gender: w, m (ND=2, D=2) Language: 19% non-Caucasian Task: Clinical interview Duration: Avg. 10 min Medication: Anti-depressant or interpersonal psychotherapy Treated over 7 weeks
2011	Cummins *et al.* [15]	24 ND 23 D	Prosodic Spectral	GMM	Accuracy	80%	Gender: 50% m, 50% w Task: Read 20 sentence Duration: 40–60s
2013	Helfer *et al.* [10]	35 D	Formants	GMM SVM	Sen./Spec.	0.86/0.64 0.77/0.77	Gender: 20 w, 15 m Age: mean of 41.8 years Language: English Task: Conversation and sustained vowels Duration: 3–6 min per session Medication: Pharmacotherapy and/or psychotherapy Treated over 6 weeks
	Scherer *et al.* [128]	18 ND 18 D	Voice quality	SVM	Accuracy	75%	Gender: Men and Women Age: $<$ 18 years Language: English Task: Interviews Duration: 5–15min

Continued on next page

Table 2.2 – continued from previous page

Year	Reference	n [a]	Features	Classifier [b]	Metrics[c]	Values [d]	Confounders
	Alghowinem *et al.* [131]	30 ND 30 D	Voice quality Prosodic Spectral Formants	Hybrid GMM+SVM	Accuracy	75%	Gender: Matched subset Age: Adults Language: English Task: Interview Duration: 92 s
2016	Valstar *et al.* [132]	133 ND 56 D	Voice quality Prosodic Spectral	SVM	F1-scores Precision Recall	0.41 D, 0.58 ND 0.26 D, 0.94 ND 0.88 D, 0.42 ND	Gender: ND (77 m, 56 w), D (25 m, 31 w) Language: English Task: Clinical interviews Duration: Avg. 16 min
2017	Jiang *et al.* [2]	85 ND 85 D	Source Prosodic Spectral	Best was SVM	Accuracy Sen./Spec.	Around 65% Around 61%/70%	Gender: ND (34 m, 51 w), D (32 m, 53w) Age: 18–55 years Language: Chinese Task: interview, picture description, and reading
	Long *et al.* [134]	37 ND 37 D	Source Prosodic Spectral Formants	Parallel SVM	Accuracy	Around 78.02%	Gender: ND (19 m 18 w), D (19 m, 18 w) Age: 18–55 years Language: Chinese Task: Interview, picture description, and reading
2018	Stolar *et al.* [28]	34 ND 29 D	Spectral Formants	SVM	Accuracies	82.2% m 70.5% w	Gender: D (24 w, 5 m), ND (24 w, 10 m) Age: 14–18 years Task: Conversation Duration: Around 1 hour

Continued on next page

Table 2.2 – continued from previous page

Year	Reference	n [a]	Features	Classifier [b]	Metrics[c]	Values [d]	Confounders
2019	McGinnis *et al.* [135]	n=71	Prosodic Spectral Formants	LR SVM	Accuracy Sen./Spec. Accuracy Sen./Spec.	80% 54%/93% 80% 62%/89%	Gender: 63% w Age: children; 3–7 years Language: English Task: Diagnostic interviews Duration: 3 min
2020	Espinola *et al.* [136]	11 ND 22 D	Broad set of vocal features	Best was RF	Accuracy Sen./Spec.	87.55% 0.9149/ 0.8354	Gender: 11 m, 22 w (5 ND, 17 D) Age: 30.1(±12.6) ND, 42.9(±13.0) D Task: Interview Duration: No duration limit
	Saidi *et al.* [137]	133 ND 56 D	Extracted using CNN	hybrid CNN+SVM	Accuracy Precision Recall	68% 0.67 0.71	Gender: ND (77 m, 56 w), D (25 m, 31 w) Language: English Task: Clinical interviews Duration: Avg. 16 min
	Aharonson *et al.* [138]	133 ND 56 D	Prosodic	NN	Accuracy	78.84%– 82.22%	Gender: ND (77 m, 56 w), D (25 m, 31 w). Language: English Task: Clinical interviews Duration: Avg. 16 min
2021	Lee *et al.* [139]	143 ND 61 D	Spectral Energy Prosodic	AdaBoost	Accuracy Sen./Spec. Accuracy Sen./Spec.	86%m 0.95/0.88 m 77% w 0.73/0.86 w	Gender: 70% w Age: 72(±6) years Language: Elderly Koreans Task: Read sentences
	Patil and Wadhai [140]	75 ND 54 D	Source Prosodic Spectral	Hybrid GMM+SVM	Accuracy	Around 83%	Age: Adolescents Task: Spontaneous speech Duration: 11min

Continued on next page

Table 2.2 – continued from previous page

Year	Reference	n [a]	Features	Classifier [b]	Metrics[c]	Values [d]	Confounders
2022	Rejaibi *et al.* [141]	133 ND 56 D	MFCCs Formants F0 Voice quality	RNN	Accuracy	86%	Gender: ND (77 m, 56 w), D (25 m, 31 w) Language: English Task: Clinical interviews Duration: Avg. 16 min

[a] ND: Non-depressed; D: Depressed.
[b] LR: Logistic Regression; GMM: Gaussian Mixture Models; SVM: Support Vector Machine; RF: Random Forest; CNN: Convolutional Neural Network; NN: neural network; RNN: Recurrent Neural Network.
[c] Sen.: Sensitivity; Spec.: Specificity.
[d] m: Men; w: Women.

Chapter 3

Bio-acoustic Features Reproducibility

3.1 Introduction

Experimental protocols and methodologies across studies on the association of the speech with clinical outcomes vary significantly [17], [20], [35], limiting the comparability of results. Studies on speech signal processing use speech samples that differ in speech task type and duration. Some of the studies on depressed individuals, for example, were conducted on three speaking tasks, including an interview, reading-a-story, and picture description, with the overall recording lengths differing 14.5 h, 5.9 h, and 4.5 h, respectively, and average duration of speech recording was 18.3 s [2], [3]. Other researchers used only interview samples with a duration range between 7 to 30 minutes [128], [142]. Therefore, it is critical to determine whether differences in speaking tasks and task duration impact the stability of bio-acoustic feature measurements.

Kiss and Vicsi reported that the measurement of speech features, mainly those calculated over sustained vowels or voiced parts of reading-a-story, is affected by the type of speech task [111]. It was also found that quantifying spectral and cepstral acoustic features, whether from vowel or continuous speech, is dependent on speech content [143]. A study of healthy speakers revealed that different speech types, such as counting, reading passages, and spontaneous speech, impacted the vibration frequency of the vocal folds in connected speech (speaking fundamental frequency) [144].

Vogel and Morgan documented that the length of obtained speech data impacted the measurement accuracy of bio-acoustic features [145]. Although several efforts have been made to explore the accuracy of short-duration speech samples for

detecting a disease or estimating a physical parameter [146]–[149], only a few studies have explored the impact of voice sample length on speech characteristics [150]–[152]. Scherer *et al.* have shown that, in sustained vowel tasks, the stability of perturbation measurements, jitter and shimmer, is affected by the task duration. At least 3 s of speech is required to provide accurate measurement [150]. Another study also found that reducing the speech duration from 60 s to 30 s affects the pitch measurements [153]. Additionally, there is high variability in optimal sample duration across a type of predictive task, reflecting the complexity of the outcome measure. For example, complex neurological phenotypes, such as dementia, may take up to 12 minutes of interview speech [147], and one-minute picture descriptions to distinguish individuals with dementia from healthy controls using only acoustic features [154].

Variation in sampling accuracy may also be influenced by gender. Several differences between men's and women's speech have been found related to the vocal folds' mass and vocal tract length, leading to significant differences in phonetics and the quality of voice [84], [155]. Simpson reported that both vocal fold vibration rate and the formants frequencies are higher in women than in men [84].

The first aim of this chapter is to examine the effect of speech duration on the reproducibility of women and men adults' bio-acoustic features by determining whether there is a difference between the features extracted from a full-duration task and those measured over shorter durations of the same task. The second aim is to investigate the difference in these parameters between different speech tasks, reading a predefined story versus counting.

3.2 Methods

3.2.1 Dataset

The database contained 796 audio recordings of 199 English speakers aged between 18 and 45 years that were collected at the University of Adelaide as part of a larger study. From every participant, four voice recordings were collected over two separate assessment sessions, with sessions spaced at least three days and at most two weeks apart, at a sampling rate of 44.1 kHz and 16-bit sampling depth in uncompressed WAV format. By using a headset type microphone, the distance between the speaker's mouth and the microphone was kept constant. Each of the four recordings contained a different speaking task: reading a pre-selected story, re-telling the story in the participants' own words, counting from 1 to 20, and

Table 3.1 CHARACTERISTICS OF THE PARTICIPANTS ENROLLED IN THE STUDY

Gender	Men ($n = 87$)	Women ($n = 98$)	p-value [a]
Age (years)	26.16 ($\pm$6.66)	27.77 ($\pm$7.11)	0.1192
Mood score (session 1)	2.32 ($\pm$2.04)	2.76 ($\pm$2.28)	0.2006
Mood score (session 2)	2.30 ($\pm$2.02)	2.69 ($\pm$2.25)	0.2470

[a] p-values refer to the sex differences for each variable and measured using Wilcoxon rank sum.

telling the capital cities of Australia loudly. Speech was recorded in a 10×14 m isolated room within a research facility. Only the investigators and participants were present and the door remained closed at all times. Participants also completed the Mood and Feeling Questionnaire (containing 13 items) before speech recording.

Fourteen subjects were removed to match mood score and age between men and women; nine of them had a mood score suggesting depression (>10 points), while others were relatively older (= 45 years). Two types of structured, controlled speech tasks were analysed: reading-a-story, with a mean duration of about 124.4 s (standard deviation = 25.0 s), and counting, with a mean duration of about 26.0 s (standard deviation = 7.7 s). Section 3.2.1 presents the basic characteristics of the participants.

All procedures were approved by the University of Adelaide's Human Research Ethics Committee. All participants provided written informed consent in compliance with the Declaration of the University of Helsinki.

3.2.2 Biomedical Speech Signal Processing

Speech analysis included preprocessing, bio-acoustic feature extraction and statistical analysis of the extracted features to examine reproducibility (Fig. 3.1).

3.2.2.1 Preprocessing Steps

Several preprocessing steps were applied to the speech signals to improve the performance of the feature extraction algorithm [93]. Linear down-mixing was used to convert each recording from two channels (stereo) into a single channel (mono). Silent pauses were eliminated from the input signal to avoid extracting acoustic features from the background acoustical noise [156], [157], by detecting the

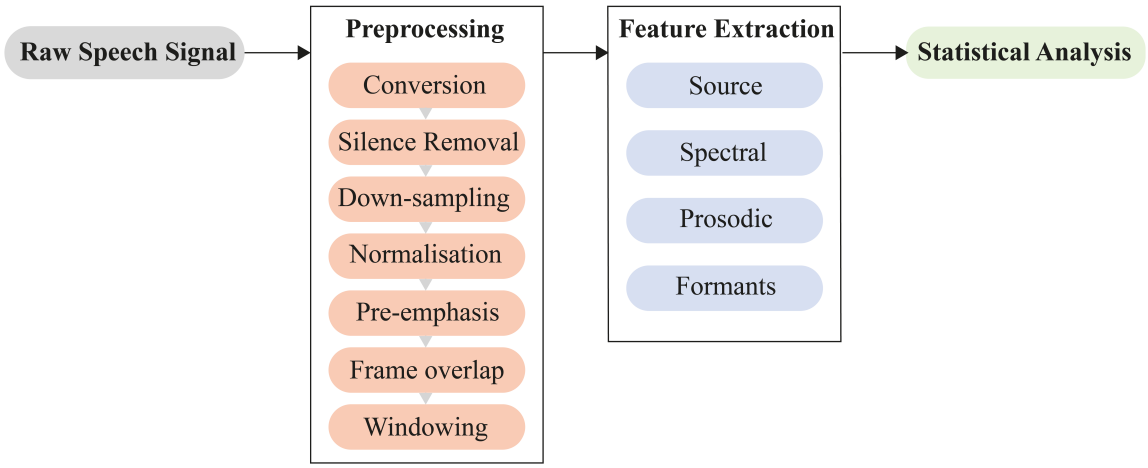

Fig. 3.1 A block diagram illustrating the steps to examine bio-acoustic features' reproducibility. These steps mainly including preprocessing steps, features extraction, and statistical analysis. Preprocessing steps comprise down-mixing signal, removing silent pauses, resampling speech signal (16 kHz), z-score normalisation, and signal pre-emphasis. Moreover, the features extraction step focuses on quantifying acoustic features. Statistical analysis using Intraclass Correlation Coefficient tests was applied to the quantified features.

speech boundaries using the MATLAB® detectSpeech function (The MathWorks, USA). The signals were then down-sampled to 16 kHz, commonly used for speech processing, to reduce the computational load [41], [93]. Samples were then normalised to eliminate differences from the recording environment using the z-score method that centres data to have a zero mean and unit variance [158]. Finally, a pre-emphasis filter was implemented with a coefficient value equal to 0.97, commonly used for speech applications, to enhance the signal-to-noise ratio, enhance higher frequency components (i.e., speech spectrum in the high-frequency region has a steep roll-off), and suppress some of the glottal effects from the vocal tract parameters [41], [159], [160]. The TF of pre-emphasis filter is given by the following:

$$H(z) = 1 - 0.97z^{-1}. \tag{3.1}$$

Since the speech signal is non-stationary, due to articulation effect, and considered stable only in short time intervals (typically 20–30 ms) [93], short-time analysis (framing) is required for analysis. The speech signal is segmented into frames of 20 ms duration, as recommended [93], [161]. The frames were overlapped by 50% to avoid introducing any spurious frequency components [93], [159] and track the temporal characteristics of individual speech. Afterwards, the Hamming window (Fig. 3.2), commonly used for speech processing, was applied to all frames to taper

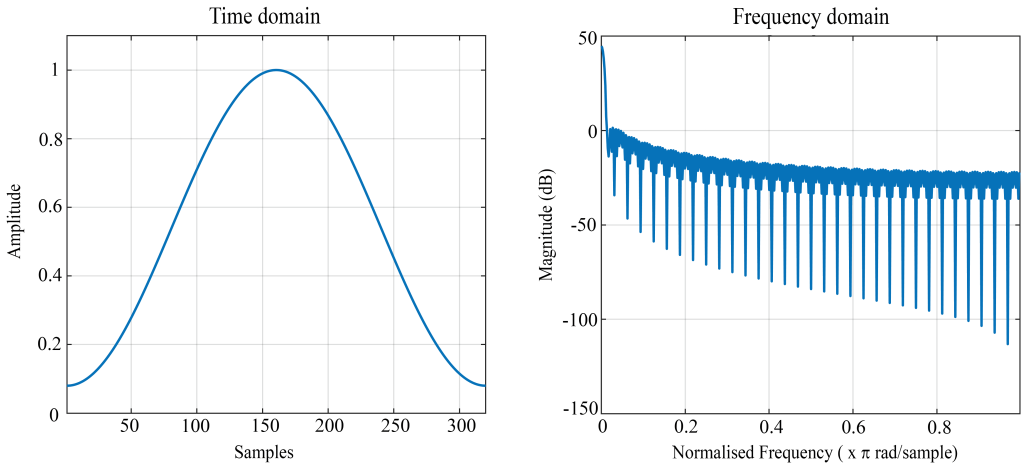

Fig. 3.2 Hamming window of 320 samples in time-domain and frequency-domain.

the signal in time-domain, reduce spectral leakage, and enhance harmonics [93], [159]. Owing to the windowing, information at the boundary of each frame is attenuated. Hence, overlapping helps in preserving information of the original signal.

3.2.2.2 Features Extraction

Speech feature extraction is at the core of the ability of speech processing systems to derive descriptive attributes of the signal [159]. Speech features can be categorised into two branches: acoustic and linguistic [162]. In this thesis, only acoustic features is considered, which can be divided into source, spectral, prosodic, and formants features [1]. Measuring these characteristics frame by frame is known as low level descriptors (LLD), while applying statistical functions over the LLD is known as statistical features [163]. A summary of the extracted features is provided in Table 3.2. The features were measured with the help of MATLAB®2021a (The MathWorks, USA) [164].

Source Features

The source features calculated over voiced regions included *jitter*, which quantifies the cycle-to-cycle variation in the glottal pulse timing period, and *shimmer*, which quantifies the cycle-to-cycle variation in the amplitude of the glottal pulse [31], [36]. They are defined by the following equations:

$$\text{Jitter}(\mu s) = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}|, \tag{3.2}$$

Table 3.2 SUMMARY OF THE EXTRACTED BIO-ACOUSTIC FEATURES.

Features Category [a]	Features	Statistical measurements
Source	Jitter	Mean, SD, percentile range (90–10%)
	Shimmer	Mean, SD, percentile range (90–10%)
Spectral	MFCC (1-13)	Mean, SD, percentile range (90–10%), skewness, kurtosis
	ΔMFCC	Mean
	ΔΔMFCC	Mean
	SR	Mean, skewness, kurtosis
	SC	Mean, percentile range (90–10%)
	SE	Mean, SD
	SF	Mean
	SS	Mean
	SK	Mean
Prosodic	Pitch	Mean, SD, percentile range (90–10%), skewness, kurtosis
	Loudness	Mean, SD
	VP	Mean
	ZCR	Mean, SD, skewness, kurtosis
Formants	F1	Mean, SD, percentile range (90–10%)
	F2	Mean, SD, percentile range (90–10%)

[a] MFCC: Mel-frequency cepstral coefficients; SR: Spectral roll-off; SC: Spectral centroid; SE: Spectral entropy; SF: Spectral flatness; SS: Spectral skewness; SK: Spectral kurtosis; VP: Voicing probability; ZCR: Zero-crossing rate; F1: First formant; F2: Second formant; SD: Standard deviation.

$$\text{Shimmer(dB)} = \frac{1}{N-1}\sum_{i=1}^{N-1} |20\log_{10}\frac{A_{i+1}}{A_i}|, \tag{3.3}$$

where T_i denotes the time period of the glottal pulse, N denotes the number of periods, and A_i represents the peak-to-peak amplitude [36], [165].

Both jitter and shimmer were determined by utilising the GCIs within each glottal cycle, which were detected automatically, over 60 ms frame duration, using the DYPSA algorithm built in the VoiceBox [58], [127]. Frame duration was chosen based on the closure onset time after glottal opening occurs in connected speech of approximately 60 ms.

Spectral Features

STFT is commonly used to analyse the frequency content of a non-stationary signal (i.e., speech signal). It is implemented to map a one-dimensional time-domain signal into a two-dimensional representation of time and frequency. In STFT, the DFT is applied to each windowed data of the speech signal, resulting in a matrix of complex numbers (magnitude and phase) for each point in time and frequency (Fig. 3.3). The kth element of STFT matrix can be expressed as follows:

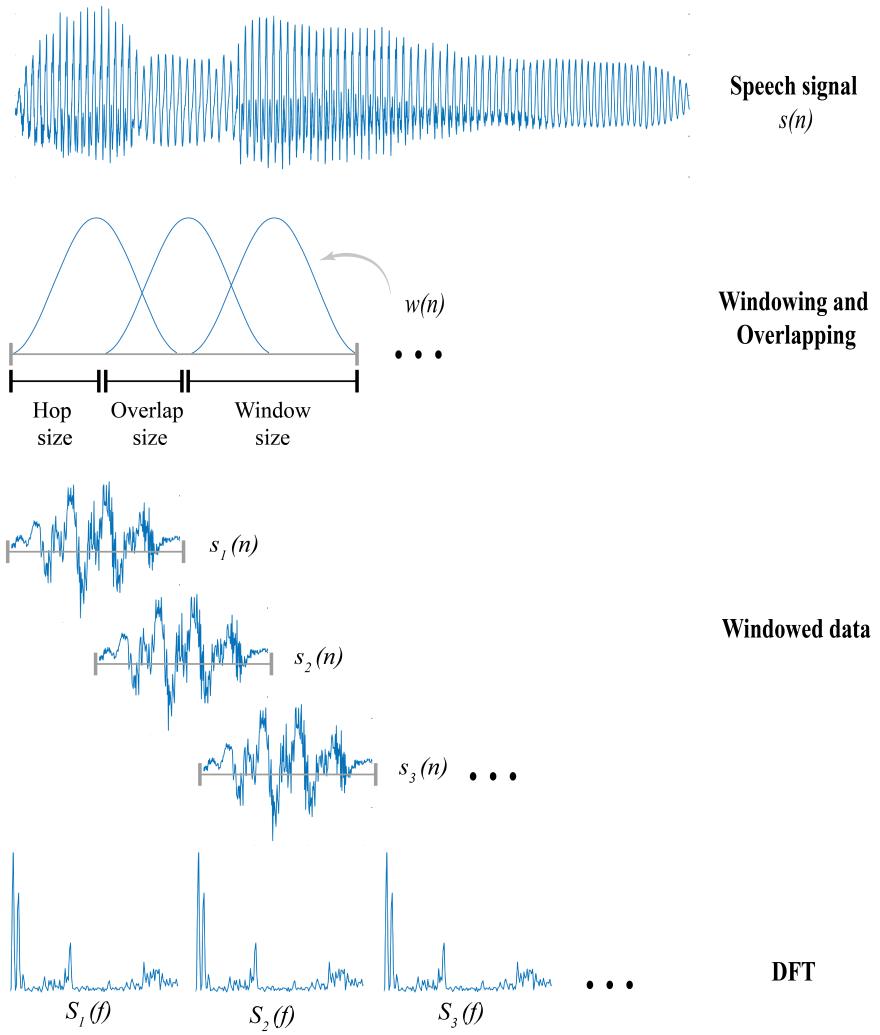

Fig. 3.3 An overview of Short-time Fourier transform (STFT) of a speech signal.

$$S_k(f) = \sum_{n=-\infty}^{\infty} s(n)w(n-kR)e^{-j\pi fn}, \tag{3.4}$$

where $S_k(f)$ is the DFT spectrum of the windowed data centred at kR time, $s(n)$ is the speech signal to be transformed, $w(n)$ is the window function; selectively

determine the analysed portion, and R is called the hop size; the number of samples between successive DFTs.

Since human hearing is relatively phase-insensitive [166], spectral shape descriptors are derived from the *amplitude spectrum* of the STFT, either a "magnitude" or "power" spectrum. The power spectrum is the magnitude squared of each frequency component. It characterises the signal's energy distribution over the frequency. One-sided power spectrum is used to estimate these descriptors.

For the power spectrum value S_k at bin k, the frequency f_k (in Hz) at bin k, and the band edges (bins) b_1 and b_2, the spectral shape descriptors are quantified as described below (i.e., STFT parameters: window size of 20 ms and overlap of 50%).

Spectral centroid, the first-order spectral moment, indicates the "centre of mass" of the spectrum. It is computed as the ratio between the sum of the weighted (by frequency) to the unweighted power spectrum [167]. This descriptor is defined in Hz as follows:

$$\mathrm{SC} = \frac{\sum_{k=b_1}^{b_2} f_k S_k}{\sum_{k=b_1}^{b_2} S_k}. \tag{3.5}$$

Spectral spread is known as spectral standard deviation. It is also considered the second-order spectral moment (denoted μ). This feature describes the spread of the power spectrum around its spectral centroid; closely related to the signal's bandwidth. Mathematically, spectral spread is defined as [167]:

$$\mu = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mathrm{SC})^2 S_k}{\sum_{k=b_1}^{b_2} S_k}}. \tag{3.6}$$

Spectral skewness is the third-order spectral moment. This descriptor characterises the asymmetry of spectrum distribution around its centroid value. For a given signal, SS can be calculated as follows [167]:

$$\mathrm{SS} = \frac{\sum_{k=b_1}^{b_2} (f_k - \mathrm{SC})^3 S_k}{\mu^3 \sum_{k=b_1}^{b_2} S_k}. \tag{3.7}$$

Spectral kurtosis is the fourth-order spectral moment. It is a measure of "non-Gaussianity"; indicating the *flatness* and *peakiness* of the spectrum around its

centroid. SK is defined as [167]:

$$\mathrm{SK} = \frac{\sum_{k=b_1}^{b_2} (f_k - \mathrm{SC})^4 S_k}{\mu^4 \sum_{k=b_1}^{b_2} S_k}. \tag{3.8}$$

Spectral entropy is a useful parameter for quantifying the regularity "peakiness" of power spectrum of speech signal [168]. It is calculated as:

$$\mathrm{SE} = \frac{-\sum_{k=b_1}^{b_2} S_k \log(S_k)}{\log(b_2 - b_1)}. \tag{3.9}$$

Spectral flatness corresponds to measuring the frequency distribution uniformity of a power spectrum [62]. It is obtained by taking the ratio of spectral distribution's geometric mean to its arithmetic mean. Spectral flatness, typically measured in dB, is determined according to Johnston [169] as follows:

$$\mathrm{SF} = \frac{(\prod_{k=b_1}^{b_2} S_k)^{\frac{1}{b_2 - b_1}}}{\frac{1}{b_2 - b_1} \sum_{k=b_1}^{b_2} S_k}. \tag{3.10}$$

Spectral roll-off points defines the frequency below which a certain percentage p (usually between 80% and 90%) of the total energy is concentrated [170]. For i roll-off point and p of 95%, SR (in Hz) is calculated by the following equation:

$$\sum_{k=b_1}^{i} |s_k| = p \sum_{k=b_1}^{b_2} S_k. \tag{3.11}$$

MFCC is a collection of coefficients used as features; they are constructed to represent the short-time power spectrum of a speech signal. These coefficients are obtained by computing DFT values of windowed speech data. These values are then grouped in critical bands and weighted by a bank of band-pass filters [39], [40]. Filters are designed to map the frequency range of human hearing. They are linearly spaced overlapped triangular filters in the Mel-scale and logarithmically spaced at frequencies higher than 1 kHz in linear frequency scale [14], [93]. The frequency axis is warped and evenly spaced with more weight (i.e., higher resolution) on lower frequency values according to the following equation [66]:

$$f_{\mathrm{mel}} = 1127.01048 \ln\left(\frac{f_l}{700} + 1\right), \tag{3.12}$$

where f_{mel} is the frequency in mel-scale, and f_l is the linear frequency in Hz.

The outputs of the filter bank are logarithmised and then decorrelated by means of the discrete cosine transform (DCT) [39], [40], instead of the IFT. Typically, the first 8 to 13 DCT components, identified by human hearing, represent the MFCC feature vector. Higher-order coefficients mainly contain the source signal because the vocal tract impulse response rapidly decays. Fig. 3.4 shows a typical block diagram representing the extraction steps of MFCC features. MFCCs omit temporal

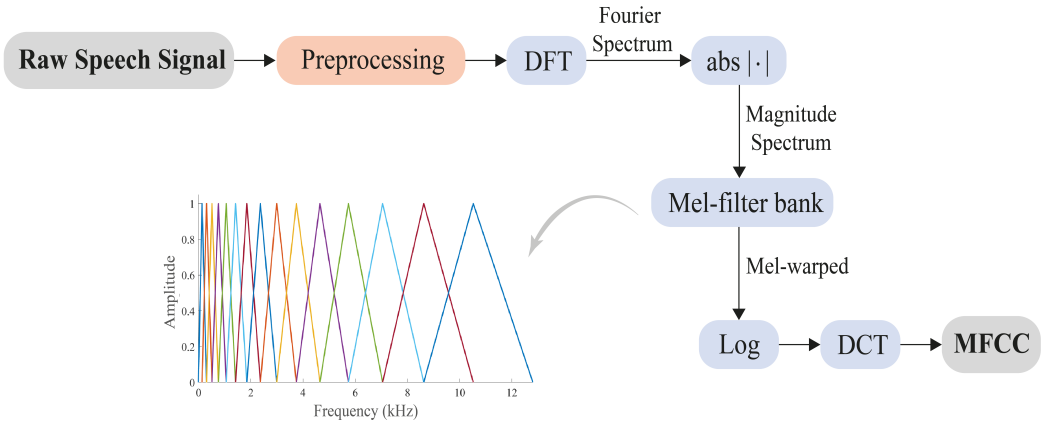

Fig. 3.4 Block diagram of mel-frequency-cepstrum coefficients (MFCCs) computation steps [75].

information of speech signals and are considered static features. ΔMFCC, the difference between the current and the previous coefficients, and ΔΔMFCC, the difference between the current and the previous delta values, are computed. In this thesis, I obtained the first 13 MFCCs, ΔMFCC, and ΔΔMFCC [40].

$$\Delta\text{MFCC}_m = \text{MFCC}_m(n) - \text{MFCC}_{m-1}(n). \tag{3.13}$$

$$\Delta\Delta\text{MFCC}_m = \Delta\text{MFCC}_m(n) - \Delta\text{MFCC}_{m-1}(n). \tag{3.14}$$

Prosodic Features

Pitch, expressed in Hz, is the auditory sensation of a sound frequency. It is closely related to F0, the lowest frequency component of quasi-periodic vibration of vocal folds, but they are not equivalent. It can be estimated by measuring the F0. A high pitch sound corresponds to a high-frequency sound wave, whereas a low pitch

sound corresponds to a low-frequency sound wave. Pitch is estimated in the short-time domain via the NACF method [171], defined as follows:

$$\mathrm{NACF}(\tau) = \frac{\sum_{n=0}^{N-1} x(n+\tau)x(n)}{\sqrt{\sum_{n=0}^{N-1} x^2(n+\tau) \sum_{n=0}^{N-1} x^2(n)}}, \tag{3.15}$$

where N is the frame length, τ is the time lags or delays, and $x(n)$ is the corresponded speech frame signal.

The NACF has the greatest value of one at zero time lag, at which the similarity of a signal with itself is 100%. An example NACF of a speech data frame is shown in Fig. 3.5. In this example, the next greatest peak after a lag of zero corresponded to the fundamental period.

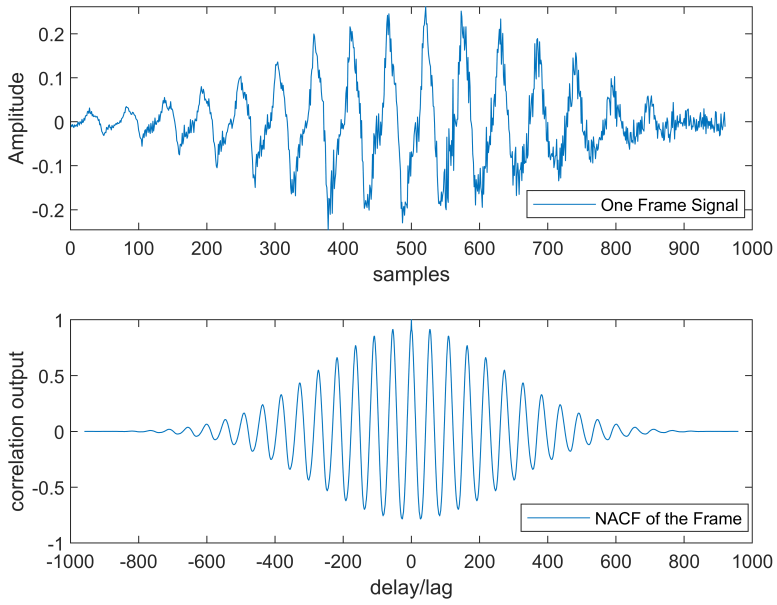

Fig. 3.5 Normalised correlation function (bottom) of a speech segment sampled at 16 kHz (top), resulting in a pitch value of 296.3 Hz.

Pitch floor and ceiling parameters are set to the interval [75–300 Hz] for male voices and [100–500 Hz] for female voices as recommended [92]. These boundaries are appropriate for analysing adult voices and avoiding pitch tracking errors that are generated due to inappropriate extreme values [172].

Sound pressure level is a physical value commonly used as indicator of the acoustic wave strength (i.e., sound intensity). It measures the acoustic pressure of a sound (denoted P) relative to a reference pressure (denoted P_0), where P_0 corresponds to the human hearing threshold 20 micro Pascal (µPa) at 1 kHz. This measure (i.e.,

SPL) is expressed as [40]:

$$\text{SPL(dB)} = 20\ \log_{10}(\frac{P}{P_0}). \tag{3.16}$$

Since the human ear is not equally sensitive to sounds with the same SPL and different frequencies, the perceived loudness of a sound is not directly equal to its SPL. Thus, the A-weighting curve filter has been adopted to weight SPL as a function of frequency, approximately in agreement with the frequency response properties of the human hearing for a pure tone [90]. The characteristic of A-weighting filter corresponds well with the perceived loudness and with the isophonic curve of 40 phons [90], [173]. The frequency response of this filter is shown in Fig. 3.6. On the A-weighted dB scale (dBA), low and high frequencies are given a relatively less weight compared to the mid-frequency range [88]. Consequently, A-weighted SPL is used as a proxy estimation of perceived loudness (in dBA).

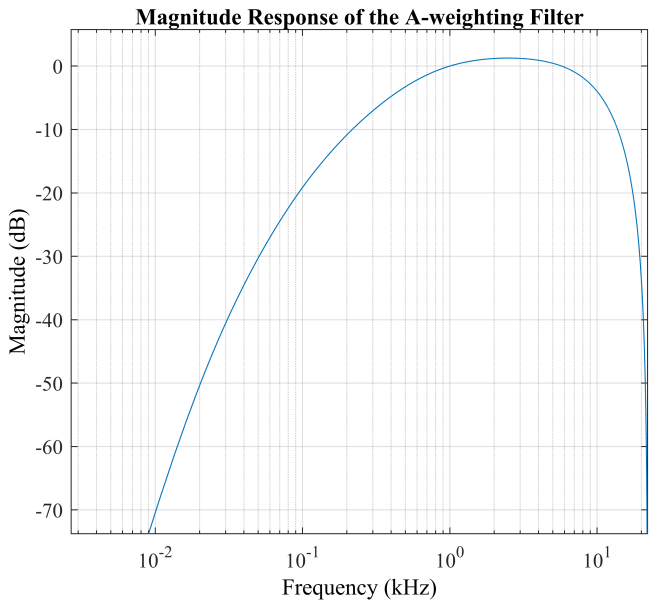

Fig. 3.6 The frequency response of A-weighting filter.

Zero-crossing rate, the number of times the speech signal passing the zero [93], is also extracted on a frame level. For the signal (s) and frame (i) of length N, the ZCR is defined by the following:

$$\text{ZCR}(i) = \frac{1}{2N}\sum_{n=1}^{N}|s_i(n) - s_i(n-1)|, \tag{3.17}$$

where $s_i(n)$=1 when the signal has a positive amplitude (>0) at n and 0 otherwise.

Voicing probability is obtained by applying a voice activity detector algorithm, introduced by Sohn *et al.* [174], over a windowed speech signal, in the frequency domain to detect speech-present. The probability threshold of transition from voiced to unvoiced frames is set to 0.2, while the transition from unvoiced to voiced frames is set to 0.1.

Formant Features

Formant frequencies (F1 and F2) are tracked frame-by-frame throughout a speech sample—after down-sampling, down-mixing, and removing silent pauses from the signal—using the automated formant tracking tool. This tool was built by Dissen *et al.* [96] to take in the raw speech signal and applied several preprocessing steps on it, including framing, overlapping, pre-emphasis filter, and windowing. Two sets of spectral features are then extracted from LPC analysis and pitch-synchronous spectra to parametrised the envelop of STFT. A recurrent neural network architecture is employed to consider temporal information of a signal's frames in the tracking process. The output layer of this network consists of the formant frequencies. This method was outperformed both Praat and WaveSurfer, as reported in [96].

Statistical Functions

Once bio-acoustic features are extracted across each a speech sample on a frame basis, several statistical functions are applied over these frames. These functions include the mean, standard deviation (SD), third- and fourth-order statistical moments (skewness and kurtosis), and percentile range (90–10%) value, leading to 125 bio-acoustic features per speech sample. These statistical functionals are relatively stable which helps in excluding the influence of tracking errors (e.g., pitch-halving or pitch-doubling errors). Because psychological investigations are concerned with the overall patterns of speech and classification results were not statistically significant between features measured at low-level, on a frame basis, and those measured by statistical functions [175], statistical function features was used in this thesis.

3.2.3 Statistical Analysis

Statistical analysis was carried out to determine how speech task length and speech task type affected bio-acoustic features reproducibility in men and women. Each speech sample was segmented three times at different percentages of speech

Table 3.3 DURATION OF SHORTENED SUB-SAMPLES OF TOTAL SAMPLE DURATION.

Sub-samples	75%	50%	25%
Duration (s)	93 ($\pm$19)	62 ($\pm$12)	31 ($\pm$6)

recording length (25%, 50%, and 75%) with a 25% sliding window, resulting in nine speech sub-samples. Table 3.3 summarises the duration in seconds of these sub-samples. Features calculated over 25% sub-samples were correlated with those obtained from 25%, 50%, and 75% non-overlapping randomly selected sub-samples. The correlation between features calculated over 25% randomly selected sub-samples and the full-length recording were also tested. Features extracted from a similar duration of 50% from the beginning and end of each recording were correlated. Therefore, correlation of five-pair sub-samples were examined: 25% vs. 25%, 25% vs. 50%, 25% vs. 75%, 25% vs. 100%, and 50% vs. 50%. Additionally, the speech characteristics calculated over the first ten seconds of the counting task were correlated to those extracted from the same duration of the reading-a-story task.

To assess the agreement level of the extracted features, intraclass correlation coefficients (ICC) [176] were calculated for individual features. ICC values of 0.00–0.39, 0.40–0.59, 0.60–0.74, and 0.75–1.00 were used to indicate poor, fair, good, and excellent agreement, respectively. Features with an excellent agreement level (ICC$\geq$0.75) were considered reproducible in this study. A two-way analysis of variance (ANOVA) test was performed over the ICC values to determine significant differences within the five correlation measurements and gender.

3.3 Results

3.3.1 Effect of Speech Task Duration

Fig. 3.7 summarises the number of reproducible features (ICC$\geq$0.75) at different lengths of speech data for the reading-a-story task. Comparing features obtained at 25% with 100% speech duration, 82 and 81 acoustic features were deemed reproducible in men and women, respectively. The number of reproducible features decreased to 53 in men and 57 in women when the same duration (25% sub-samples) were correlated. There was no statistical difference between men and women (P=0.52) in ICC values (out of 625) across five paired measurements (i.e., to consider effect of speech duration and genders on features reproducibility, in

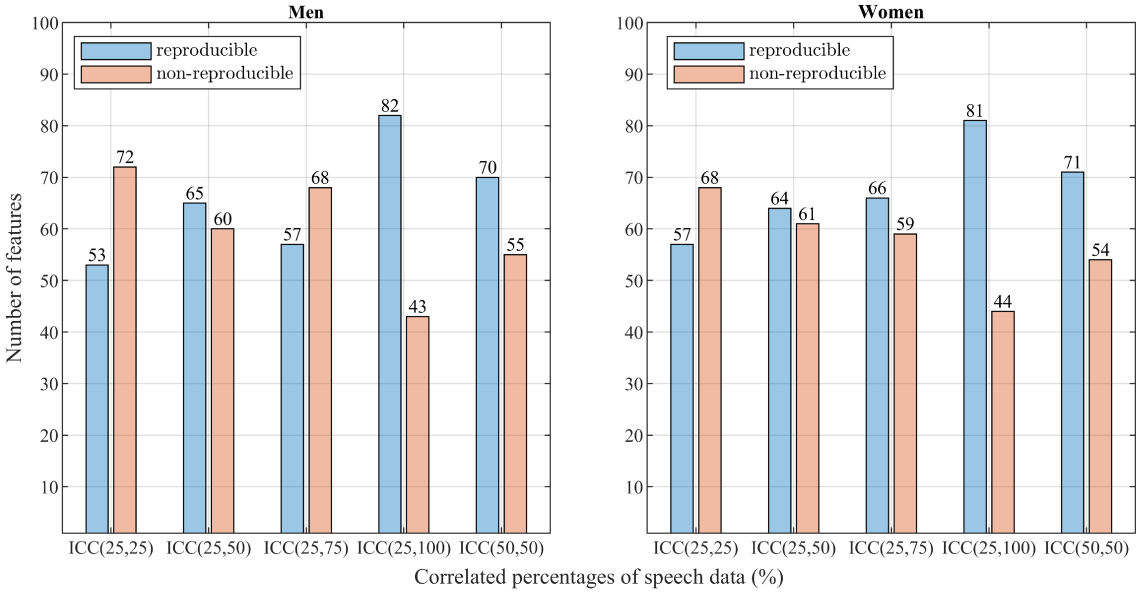

Fig. 3.7 Comparison of the number of reproducible bio-acoustic features as a function of correlated percentages speech data for men (left) and women (right). Data were extracted for different durations of the reading-a-story task.

general, two-way ANOVA test was applied on all measured ICCs across both genders).

The ICC values for feature categories are shown in Fig. 3.8. The duration had a considerable impact on source features' reproducibility. Jitter parameters (out of 3) achieved poor to fair reproducibility (ICC<0.59) across different speech durations (P=0.30; Fig. 3.8a); gender difference is significant (P=0.05). Shimmer parameters' agreement level (ICC<0.71) was similar (P=0.38) when considering different speech lengths; there was no significant difference between men and women (P=0.33; Fig. 3.8b).

MFCC coefficients, ΔMFCC, and ΔΔMFCC contributed about 73% to the total of measured features (91 out of 125). MFCC parameters were affected by speech duration (P<0.05). Gender has no effect on ICC values of these parameters (P=0.73). MFCC features also had fair-to-excellent reproducibility, with a mean ICC value around 0.75 in each measurement (Fig. 3.8c). Both ΔMFCC, and ΔΔMFCC attributes were influenced by reducing speech task length, resulting in a poor agreement level (ICC<0.32; Fig. 3.8d, Fig. 3.8e). Gender has a significant impact on ICCs of ΔMFCC (P=0.02), but it has no effect on ICCs of ΔΔMFCC (P=0.60).

Spectral shape characteristics showed high stability across reduction in speech task

lengths. SR parameters (out of 3) achieved excellent reproducibility (ICC>0.75) when speech duration decreased from full recording to 25%, with no significant gender difference was found (P=0.23; Fig. 3.8f). SC parameters (out of 2) had excellent and good-to-excellent agreement level in men and women, respectively, when speech duration is shortened (Fig. 3.8g). No significant difference was observed in SC ICCs between men and women (P=0.14). SE and SF were reproducible across different speech durations; no gender effect was found (P>0.05; Fig. 3.8h, Fig. 3.8i). SS and SK showed excellent reproducibility across duration reduction in men and women, as shown in Fig. 3.8j and Fig. 3.8k; only a statistical difference was found in SS between genders (P<0.05).

In terms of prosodic features, gender and speech duration reduction had a significant impact on ICCs of pitch parameters (P=0.0002 and P=0.001, respectively); however, pitch achieved an excellent agreement level (ICC>0.75) across all comparisons in both genders (Fig. 3.8l). A wide variation in loudness parameters was found, ranged from fair-to-excellent agreement; no gender effect was observed (P=0.75; Fig. 3.8m). ICC values of ZCR parameters showed fair-to-excellent agreement (>0.40) in both men and women, with no statistical difference was found between genders (P=0.20; Fig. 3.8n). VP attributes were varied and considered non-reproducible in men, while women maintained good-to-excellent ICC values (Fig. 3.8o).

ICC values of formants features' in men and women were statistically different at P<0.05. Although F1 and F2 parameters were considered reproducible across all duration comparisons (ICC>0.75), duration reduction impacted ICC values (P<0.05). At full sample duration, the ICCs of F1 parameters was around 0.95 for men and 0.93 for women. These values decreased gradually to nearly 0.90 and 0.83 for men and women, respectively, at the shortest speech length (Fig. 3.8p). Higher stability was observed in men than in women in ICCs of F2 parameters (Fig. 3.8q).

3.3.2 Effect of Speech Task Type

This experiment examined the reproducibility of bio-acoustic qualities calculated over the first ten seconds of two different tasks; reading-a-story and counting. Table 3.4 summarises the ICC values obtained by comparing these tasks for men and women. Most features showed high variability. Source features lost their reproducibility by changing speech tasks. The mean ICC values of jitter parameters were -0.05 in men and 0.03 in women. Shimmer ICC values were ranged between 0.002 and 0.07 and between -0.03 and -0.22 for men and women, respectively. For both genders, a poor agreement level was found in MFCC, ΔMFCC, and ΔΔMFCC

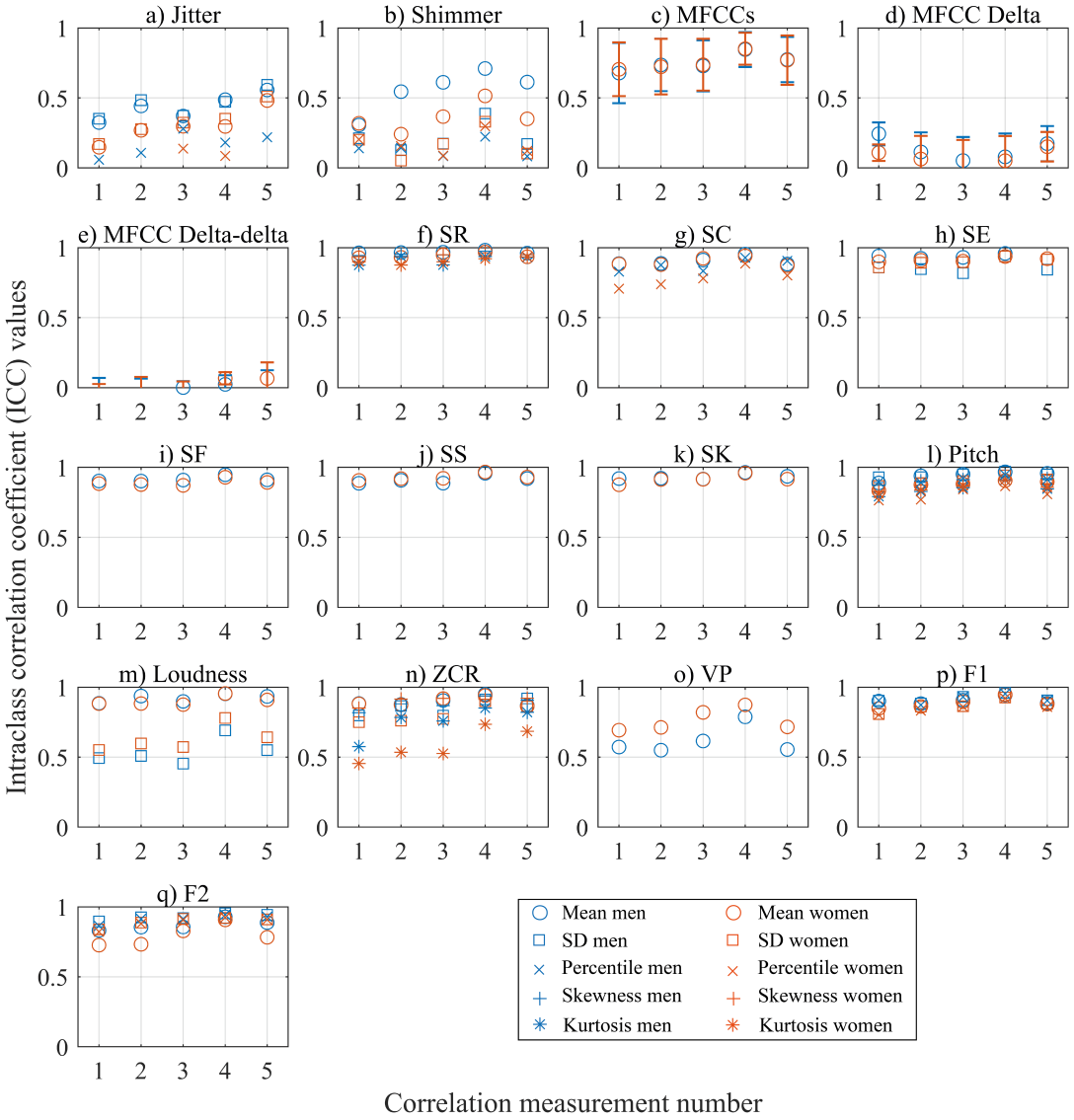

Fig. 3.8 The ICC values of bio-acoustic features for both men and women is shown in the scatter plot. The numbers on the x-axis can be interpreted as follows; 1: ICC(25% vs. 25%), 2: ICC(25% vs. 50%), 3: ICC(25% vs. 75%), 4: ICC(25% vs. 100%), 5: ICC(50% vs. 50%).

parameters. SR showed good-to-excellent stability in men (ICC>0.60) and fair-to-excellent stability in women (ICC>0.47). For both genders, poor reproducibility was found in SC and SE. Good and fair agreement levels were observed in SS and SK for men and women, respectively. The ICC of SF was around 0.5 for both genders. Speech task type impacted pitch reproducibility in men (ICCs: 0.36–0.67) and women (ICCs: 0.27–0.58). Variation in loudness and ZCR attributes was observed when comparing counting and reading tasks. VP was not a reproducible feature in both genders when speech task type is changed. In men and women, F1 parameters presented fair-to-good reproducibility, and F2

parameters showed fair reproducibility.

Table 3.4 ICC VALUES OF MEASURED BIO-ACOUSTIC FEATURES COMPARING TWO SPEECH TASKS: COUNTING AND READING-A-STORY.

Feature	Statistical measurements	ICC Value	
		Men	**Women**
Jitter	Mean	-0.12	-0.06
	SD	-0.13	-0.03
	Percentile range	0.09	0.18
Shimmer	Mean	0.07	-0.22
	SD	0.05	-0.07
	Percentile range	0.002	-0.03
MFCCs	Mean, SD, Percentile range, Skewness, Kurtosis	0.36 ($\pm$0.21)	0.35 ($\pm$0.20)
Δ**MFCC**	Mean	0.13 ($\pm$0.16)	0.09 ($\pm$0.16)
$\Delta\Delta$**MFCC**	Mean	0.03 ($\pm$0.08)	-0.03 ($\pm$0.08)
SR	Mean	0.79	0.78
	Skewness	0.80	0.72
	Kurtosis	0.60	0.47
SC	Mean	0.68	0.54
	Percentile range	0.48	0.37
SE	Mean	0.57	0.53
	SD	0.35	0.47
SF	Mean	0.55	0.54
SS	Mean	0.74	0.52
SK	Mean	0.62	0.45
Pitch	Mean	0.67	0.27
	SD	0.68	0.58
	Percentile range	0.60	0.41
	Skewness	0.44	0.33
	Kurtosis	0.36	0.33
Loudness	Mean	-0.18	-0.25
	SD	0.31	0.40
VP	Mean	-0.32	-0.30
~~**ZCR**~~	Mean	0.66	0.54

Continued on next page

Table 3.4 – continued from previous page

Feature	Statistical measurements	ICC Value	
		Men	**Women**
	SD	0.40	0.45
	Skewness	0.60	0.59
	Kurtosis	0.23	0.27
F1	Mean	0.46	0.56
	SD	0.67	0.65
	Percentile range	0.47	0.55
F2	Mean	0.51	0.47
	SD	0.44	0.50
	Percentile range	0.43	0.56

3.4 Discussion

In this chapter, the effects of speech task duration, speech task type, and gender on the reproducibility of bio-acoustic features in normal adults were examined. The main findings of this study are as follows: (i) the reproducibility of acoustic features steadily reduces proportional to speech duration down to about 30 s across gender; (ii) acoustic speech properties are less reproducible in less complex counting versus reading-a-story tasks; and, (iii) Some spectral (spectral shape descriptors), prosodic (pitch), and formants (F1, F2) features reached excellent reproducibility in both genders at different speech duration. Some spectral features (MFCC) and prosodic features (Loudness, ZCR) achieved excellent reproducibility at a longer duration. The reproducibility of source (Jitter, Shimmer), and other spectral (ΔMFCC, and ΔΔMFCC) features were lost when speech duration was changed. There were significant gender differences in jitter, ΔMFCC, SS, pitch, VP, and formants (F1, F2).

Interview based diagnostic and prognostic assessments for common psychiatric illnesses, such as major depression, have limited reliability and predictive accuracy [177]. Examining reproducibility of acoustic features at different speech durations has become of clinical interest to improve the accuracy of the assessment and provide valuable insights that can drive the assessment. Few studies have explored the impact of voice sample length on speech characteristics [150]–[152]. Previous work has largely focused on evaluating only one type of acoustic property against time. Scherer *et al.* suggested that at least 3 s of recording are required for accurate reading of speech perturbations [150]. To the best of my knowledge, no study has

systematically investigated the influence of decreasing the length of a speech signal on the reproducibility of bio-acoustic features in healthy individuals.

The source features' measurements were not reproducible when fewer voice samples were considered. Perturbation measurement stability is dependent on the components of speech in the location of the selected segments, for instance, there is high variability between different vowels [150]. Selecting a more stable speech segment, periodic (repetitive) or nearly periodic (nearly-repetitive) waves, leads to a more consistent result [44]. Based on the measurements of logMel and MFCC features of cropped signals from about 8 s to about 1 s, Neumann and Vu reported that a system for emotion detection performed sufficiently, despite a slight loss in accuracy compared to the use of full samples [151]. The current study found that the reproducibility of MFCC features reduces as duration shortens, which might cause a loss of prediction accuracy in such a system [151].

The results showed that the pitch parameters were reproducible with reduced sample duration in both women and men. A study on German speakers showed a substantial effect of utterance length on the variability of F0 measurements [152]. Zraick *et al.* also showed that the estimated pitch value of White women varied for different speech durations [153]. Several factors may have contributed to the differences between this study and prior investigations, including speech task type, differences of speakers' language, and method used to compute pitch. In this study, the analysis was limited to English speakers who read a story, and the NACF method was used to extract pitch. Nishinuma *et al.* report an effect of shorter sample duration on the loudness measurement [178]. Similarly, the current study showed a wide variation in loudness parameters across all duration comparisons. We demonstrated the considerable impact of duration on VP feature reproducibility in men and women [179], [180].

A study of French and German speakers has shown that decreasing the speech duration influences both F1 and F2 as a function of vowel duration [181]. Although a similar pattern is observed in this study, where short speech data impacted formants' qualities, the ICC values remained high ($>$0.75). For men and women, formants measures were reproducible across durations. Results showed higher stability in men's formants than women's.

Several studies have investigated the impact of the speech task on acoustic parameters [143], [182], [183]. This study tested the reproducibility of a wide range of voice parameters during counting and reading-a-story tasks. Results demonstrate that changing speech tasks impacted at least 96% and 98% of the

measured acoustic qualities for men and women, respectively, even if the duration was identical (10 s).

Several studies suggest that vowel type impacts shimmer parameters [114], [184]. Similarly, this study found that the shimmer feature was not reproducible between tasks. Results indicate that although the task type had a significant effect on the measurements of pitch features in women, it achieved good reproducibility in men. This finding that is in line with Sandage's *et al.* and Zraick *et al.* [144], [185]. Hence, some spectral shape descriptors (i.e., SR and SC) are relatively stable across speaking tasks.

Gender differences in speech arising from difference in vocal cord anatomy lead to dissimilarities in some acoustic features such as F0 and jitter [84], [165]. In this study, when men and women were analysed separately, significant differences in the correlation analysis of some speech properties is found, including jitter, ΔMFCC, SS, pitch, VP, and formants, suggesting that the pattern of reliable markers may be different across gender.

This study has several limitations. First, this study assessed the reproducibility of bio-acoustic features derived from native English speakers only and in a dataset with an identical recording setup; findings did not validate on voice samples across different datasets, languages or environments. Hence this study may have overestimated real-world generalised reproducibility. Second, the reproducibility of acoustic parameter examined in individuals with healthy voices. Furthermore, conducted experiments in the current study are restricted to a sample of participants aged 18 to 44 years (mean = 27 years), limiting generalisability to older or younger individuals. Additionally, scripted speech tasks, while allowing standardised comparisons, do not elicit natural speech [186]. Results need to be validated with future studies on natural speech samples across more diverse age groups, languages and environments, compared between clinical and healthy control samples. Finally, given that the content of the speech data impact the acoustic features, thus when 25% of full speech duration was correlated with the full recording duration, speech content of 25% is included within the full recording, which may affect the correlation results.

3.5 Conclusion

This study has examined the effect of speech duration and speech task on the reproducibility of bio-acoustic qualities. Shortening the speech duration from full duration to 25% of total speech duration reduces the reproducibility of measured

speech features (out of 125) from 82 and 81 to 53 and 57 in men and women, respectively. Spectral shape, pitch, and formants reached excellent reproducibility. MFCCs, loudness, and ZCR achieved excellent reproducibility only at a longer duration. Reproducibility of source, MFCC derivatives, and VP was poor. Clinicians may have to collect a minimum of speech data to achieve a high number of reproducible bio-acoustic features (at least one minute and a half in the case of the reading-a-story task). In addition, changing the speech task has a significant effect on the measurements of features; around 97% of features in both genders lost reproducibility, in part due to the short counting task duration. Therefore, researchers may have to build and train speech-task specific models (classifier/regressor). Gender factor has a significant impact on the reproducibility of jitter, ΔMFCC, SS, pitch, VP, and formants qualities. Bio-acoustic features are less reproducible in shorter samples and are affected by gender.

Chapter 4

Reproducibility of Bio-acoustic Features in Non-depressed and Depressed Speakers

4.1 Introduction

Bio-acoustic measures have been increasingly used in clinical and research pursuits in an effort to analyse the voice [187]. Reflections of depression in bio-acoustic characteristics of the patients' speech have previously been explored [1], [111]. Assessment of these characteristics to identify depression requires a practical, methodological, and statistical framework. Experimental and computational approaches for the application of speech-based depression detection studies are varied, challenging the comparability of their outcomes.

Previous studies on speech-based depression prediction task used different speech samples length [10], [128], [136], [142]. Espinola *et al.* recorded interview speech samples without setting a duration limit on their recordings, resulting in a mean duration of control group 8.8 (± 2.31) minutes, while that of depressed group 7.8 (± 3.55) minutes [136]. Helfer *et al.* conducted their analysis on speech samples with a duration ranged between 3 and 6 minutes per session [10]. Other researchers used speech data of clinical interviews with a mean duration of 16 (± 5) minutes [128], [142]. These differences make it imperative to understand the impact of speech task duration on the stability of bio-acoustic measurements in depressed and non-depressed voices.

Few studies have explored duration effects on commonly used bio-acoustic qualities (e.g., pitch, jitter, and shimmer) of individuals with normal voices [150], [153]. However, to my knowledge, no study has systematically investigated the influence of speech length on reproducibility of bio-acoustic properties in depressed patients. Analysis of these qualities might be affected by speech tasks duration. Alghowinem *et al.* used an equal amount of speech data (92 s) from each subject to avoid the potential effects of duration differences on the analysis results, comparing different classifiers and using spontaneous speech [131]. They showed a best classification performance when a hybrid classifier, combing GMM and SVM, was employed. Another study reported a better classification result when smaller parts of speech data were used, from the beginning of the reading task, compared to the whole speech sample [163].

As the duration of the reading-a-story task affects bio-acoustic metrics' reproducibility of normal speakers (as shown in Chapter 3), efforts need to be made to examine features' reproducibility on depressed and non-depressed individuals. Therefore, this chapter aims to assess the impact of speech durations on the reproducibility of bio-acoustic features obtained from *depressed* and *non-depressed* men and women during a spontaneous speech task.

4.2 Methods

4.2.1 Dataset

The database consists of a subset of DAIC and called the Wizard-of-Oz interviews (DAIC-WOZ) [129]. It contains audio recordings from 189 clinical interviews of English speakers. Most speakers (n = 133; 77 men, 56 women) were classified as "non-depressed", whereas 56 speakers (25 men, 31 women) were classified as "depressed". A virtual psychologist, operated by a human interviewer outside the room, was used to interview the participants. A binary label and depression severity level based on the PHQ-8 self-reported depression scale was assigned to every participant. Transcription of each interview, including annotation of each sentence's start- and end-time, is also provided for verbal and non-verbal indicators. Duration of speech data ranged from 7 to 33 minutes, with a mean duration of about 16 (± 5) minutes. All voice samples were recorded at a 16 kHz sampling rate and saved as uncompressed WAV format.

Only participants who spoke for at least two minutes, after the interviewer speech was removed, were included in this study. This allows for a comparison with the

Table 4.1 CHARACTERISTICS OF THE PARTICIPANTS ENROLLED IN THE STUDY.

Class	Gender	n	Mood score (PHQ-8)	p-value [a]
Depressed ($n = 56$)	Men	25	14.28 ($\pm$3.09)	0.96
	Women	31	14.52 ($\pm$4.19)	
Non-depressed ($n = 131$)	Men	76	3.52 ($\pm$2.90)	0.84
	Women	55	3.49 ($\pm$3.08)	

[a] p-values refer to the comparison of mood score between men and women in each group.

results in Chapter 3. Consequently, two non-depressed participants (one from each gender) were removed. Table 4.1 shows the basic information of the included participants.

4.2.2 Biomedical Speech Signal Processing

Speech samples were passed through several pre-processing steps before the feature extraction algorithm was performed. Segmentation was applied to isolate participants' speech segments by utilising the time points provided in the transcript files. Once the pure participant speech had been extracted, speech data from the first two minutes of each participant's speech was extracted. Silent pauses were removed from these sub-samples, and then z-score data normalisation was applied. Filtering, framing, windowing, and overlapping were also applied to the speech data. Pre-processing steps are detailed in Section 3.2.2.1.

A set of bio-acoustic features comprising 53 low-level descriptors were extracted from short-time analysis windows. These windows span over the whole speech sample and being processed at regular time steps (i.e., equal to half of the window size). These features include source features (jitter and shimmer), cepstral features (the first 13 MFCC, ΔMFCC, and $\Delta\Delta$MFCC), spectral shape features (SR, SC, SE, SF, SS, and SK), prosodic features (pitch, loudness, ZCR, and VP), and formants features (F1 and F2). Once these descriptors had been extracted, five statistical functions were applied over a feature's frames to provide a representation of the feature distribution across a speech sample, resulting in 125 features characterised a speech sample. Section 3.2.2.2 includes a detailed methodology for bio-acoustic features extraction.

Table 4.2 INTERPRETATION OF INTRACLASS CORRELATION COEFFICIENT VALUES.

ICC values	Agreement Level
0.00 – 0.39	Poor
0.40 – 0.59	Fair
0.60 – 0.74	Good
0.75 – 1.00	Excellent

4.2.3 Statistical Analysis

Statistical analysis was performed to determine the effect of speech task duration on bio-acoustic feature reproducibility in depressed and non-depressed men and women. Bio-acoustic features calculated at 30 s sample duration were compared to those obtained from non-overlapping randomly selected sub-samples shortened to 90 s, 60 s, and 30 s. Features quantified from 120 s speech sample length were also compared to those measured at 30 s duration. Additionally, features measured from a similar duration of 60 s from the beginning and end of each sample (i.e., a total length of 120 s) were correlated. The ICC statistical test was performed to calculate the agreement level of the extracted features at different speech durations. Interpretation of ICC values is summarised in Table 4.2. Features are considered reproducible at ICC$\geq$0.75. A two-way ANOVA test was carried out over the ICCs to determine significant differences between the correlation measurements and gender. A similar statistical analysis is performed in Section 3.2.3.

4.3 Results

4.3.1 Effect of Speech Duration in Non-depressed Participants

Fig. 4.1 summarises the number of reproducible features at different speech data lengths. Comparing acoustic features measured at 30 s with 120 s speech duration, 71 and 70 features were reproducible in men and women, respectively. These numbers decreased to 38 in both genders when similar durations of 30 s were correlated.

Fig. 4.2 shows the ICC values for feature categories. Jitters' ICCs varied between poor-to-good agreement in women and between poor and fair agreement in men across different speech durations (Fig. 4.2a). A wide variation observed on ICCs of shimmer parameters. In women, shimmer parameters achieved good agreement,

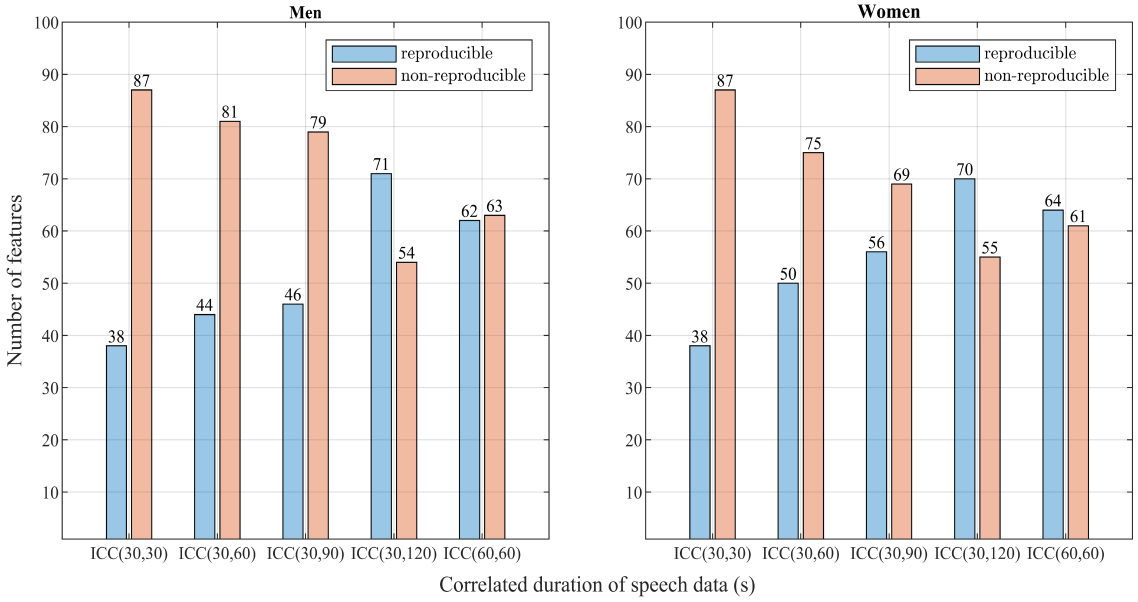

Fig. 4.1 Comparison of the number of reproducible bio-acoustic features as a function of correlated speech data duration for men (left) and women (right). Data were extracted for different durations of the spontaneous speech task.

while in men achieved excellent agreement (Fig. 4.2b). No significant gender difference was observed on ICCs of jitter and shimmer parameters at P=0.90 and P=0.66, respectively.

MFCC parameters affected by duration reduction (P<0.05), showing fair-to-excellent reproducibility; gender had no effect on ICCs of these parameters (P=0.90; Fig. 4.2c). When speech duration is shortened, ΔMFCC reached poor and fair ICC values in women and men, respectively, and ΔΔMFCC showed poor ICC values, with no gender impact was found (P=0.15 and P=0.09, respectively; Fig. 4.2d, Fig. 4.2e).

SR parameters showed good-to-excellent agreement in women and fair-to-excellent agreement in men across reduction of speech data (Fig. 4.2f); a significant gender difference was observed (P=0.01). Duration reduction also had a significant impact on ICCs of SC parameters (P=0.03; Fig. 4.2g); agreement level reduced from excellent-to-fair in men and from good-to-fair in women. Both duration reduction and gender had no significant impact on the agreement level of SE (P=0.85 and P=0.51, respectively); ICCs of men and women >0.54 (Fig. 4.2i). SF and SS were reproducible in men and lost their reproducibility in women across duration reduction; no gender effect was found (P>0.05; Fig. 4.2h, Fig. 4.2j). SK was not reproducible across all comparisons (Fig. 4.2k); no statistical difference was observed between genders (P=0.20).

Duration reduction and gender significantly impacted F0 parameters ($P<0.05$). F0 achieved good-to-excellent agreement in men and fair-to-excellent agreement in women when duration is reduced (Fig. 4.2l). A wide variation in loudness parameters was found and no gender effect was observed ($P=0.14$; Fig. 4.2m). ICCs of ZCR parameters showed poor-to-excellent agreement (ICCs>0.25) in men and fair-to-good agreement in women (ICCs>0.40), with no statistical difference was found between men and women ($P=0.60$; Fig. 4.2n). VP attributes maintained excellent ICCs in women while showed good-to-excellent ICCs in men; gender differences significant ($P=0.01$; Fig. 4.2o). Gender difference had no significant impact on ICCs of F1 and F2 at $P=0.90$ and $P=0.80$, respectively. During duration reduction, in both men and women, ICC values of F1 parameters achieved good-to-excellent agreement, while those of F2 achieved fair-to-excellent agreement (Fig. 4.2p, Fig. 4.2q). Table 4.3 summarised *p*-values of the two-way ANOVA test for comparing features' (i.e., obtained from non-depressed speakers) reproducibility at different speech duration and across men and women.

4.3.2 Effect of Speech Duration in Depressed Participants

Fig. 4.3 summarises the number of reproducible features across different speech lengths. The number of reproducible features (out of 125) decreased from 70 to 31 in men and from 59 to 42 in women with a shorter sample length.

Fig. 4.4 shows the ICC values for feature categories. Jitter parameters achieved poor-to-good agreement (ICC=0.06–0.64) at different speech durations (Fig. 4.4a); gender difference was significant ($P<0.05$). Shimmer parameters' agreement level was similar ($P=0.94$) when different speech lengths were considered; gender had no significant effect ($P=0.30$; Fig. 4.4b). Sample duration had a significant impact on the reproducibility of MFCC parameters, with a mean ICC value around 0.8 in each measurement (Fig. 4.4c); no statistical difference was found between men and women at $P=0.73$. ΔMFCC attributes influenced by reducing speech task length, resulting in poor-to-fair agreement (ICC<0.50; Fig. 4.4d); a significant difference between men and women was observed ($P=0.01$). Although duration reduction and gender had no significant impact on ICCs of $\Delta\Delta$MFCC attributes ($P=0.43$ and $P=25$, respectively; Fig. 4.4e), $\Delta\Delta$MFCC is a non-reproducible feature.

SR parameters achieved fair-to-excellent agreement (ICC>0.43) when speech duration is decreased, with no significant gender difference was found ($P=0.96$; Fig. 4.4f). ICC values of SC parameters varied between poor-to-excellent in men and women when speech duration is shortened (Fig. 4.4g); no genders difference was found on ICCs ($P=0.10$). Duration reduction and gender had no significant

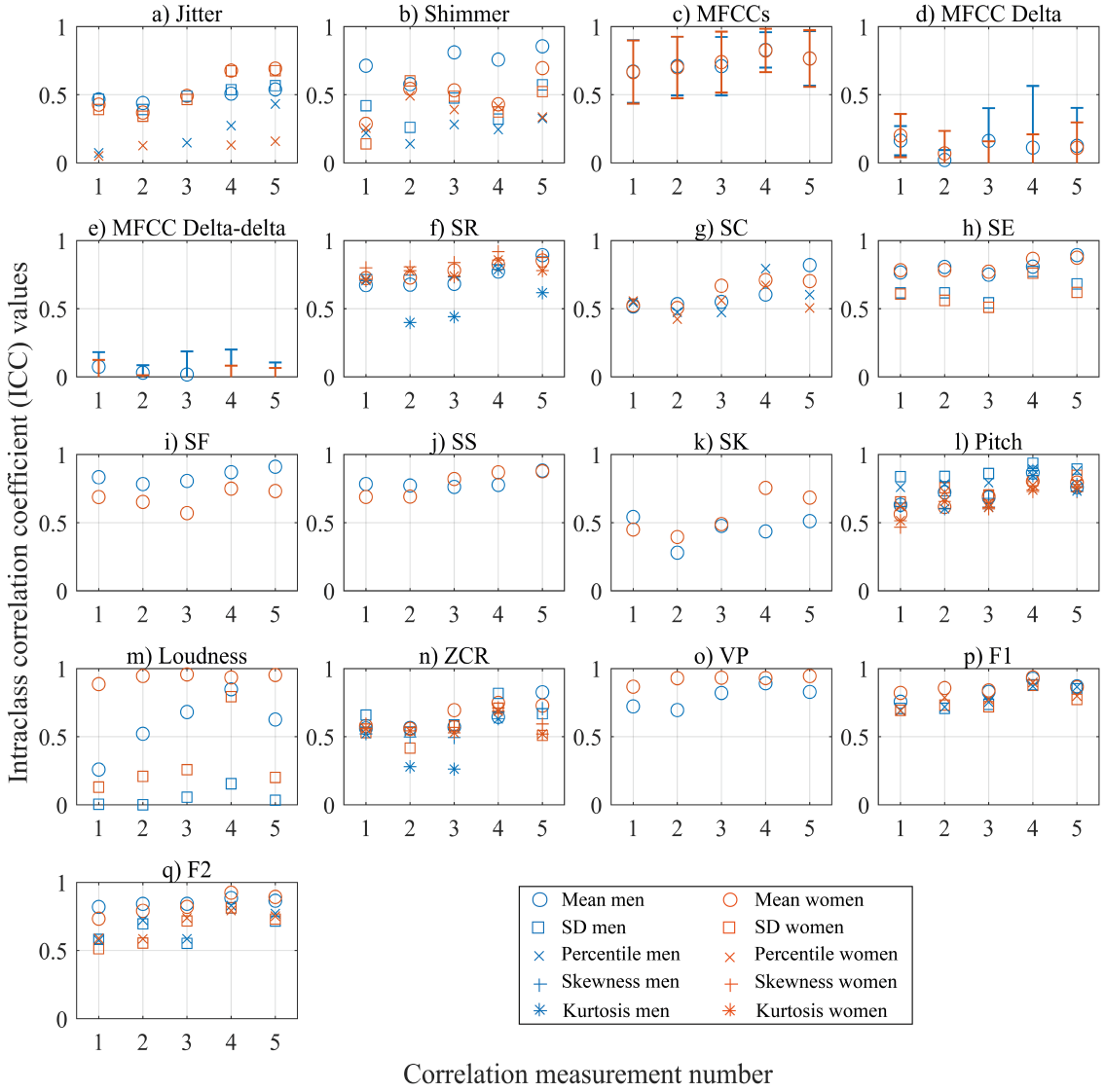

Fig. 4.2 The ICC values of bio-acoustic features for both non-depressed men and women is shown in the scatter plot. The numbers on the x-axis can be interpreted as follows; 1: ICC(30 s vs. 30 s), 2: ICC(30 s vs. 60 s), 3: ICC(30 s vs. 90 s), 4: ICC(30 s vs. 120 s), 5: ICC(60 s vs. 60 s).

impact on ICCs of SE, SF, and SS ($P>0.05$; Fig. 4.4h, Fig. 4.4i, Fig. 4.4j). SK reached excellent reproducibility in women and fair reproducibility in men when considering shorter speech samples; gender differences was significant (Fig. 4.4k).

Lengths of speech data and gender had a significant impact on ICCs of F0 parameters ($P<0.05$). Across all comparisons, F0 showed good-to-excellent agreement in men and poor-to-excellent agreement in women (Fig. 4.4l); higher stability was observed in men than in women. ICC values of loudness parameters varied from poor-to-excellent in women and between poor and fair in men; no

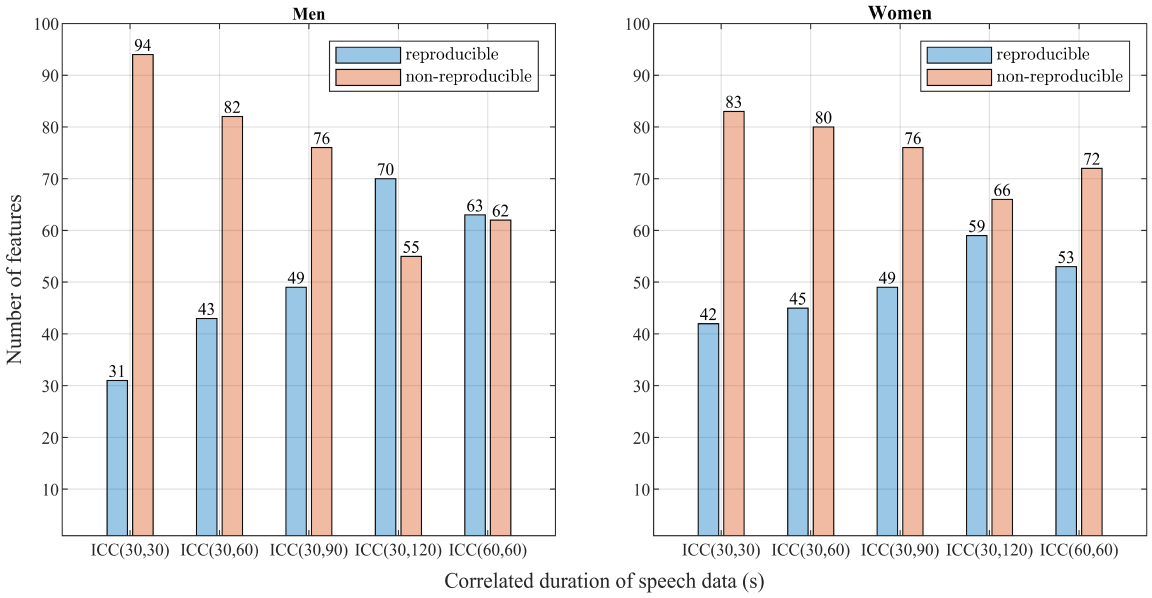

Fig. 4.3 Comparison of the number of reproducible bio-acoustic features as a function of correlated duration of speech data for men (left) and women (right). Data were extracted for different durations of the spontaneous speech task.

gender effect was observed (P=0.61; Fig. 4.4m). ZCR parameters showed poor-to-excellent ICCs (>0.40) in men and poor-to-good ICCs (>0.30) in women; a statistical difference was found between genders (P=0.01; Fig. 4.4n). VP attribute was reproducible in women and lost its reproducibility in men across duration reduction (Fig. 4.4o).

Gender had a significant impact on ICCs of F1 parameters at P<0.05, but had no significant impact on those of F2 at P=0.17. Although F1 parameters were reproducible across all durations (ICC>0.75), a significant effect of duration reduction was observed on F1 ICCs (P=0.04; Fig. 4.4p). Shortening speech samples significantly affect F2 parameters (P<0.05), where some parameters lost their reproducibility in both genders (Fig. 4.4q). Table 4.3 summarised *p*-values of the two-way ANOVA test for comparing features' (i.e., obtained from depressed speakers) reproducibility at different speech duration and across men and women.

4.3.3 Bio-acoustic Features' Reproducibility Comparison between Depressed and Non-depressed Participants

Table 4.3 also shows *p*-values of the two-way ANOVA test, comparing the effects of group factor (i.e., depressed and non-depressed) and speech task duration factor on ICC values of bio-acoustic features. Group and duration differences were not

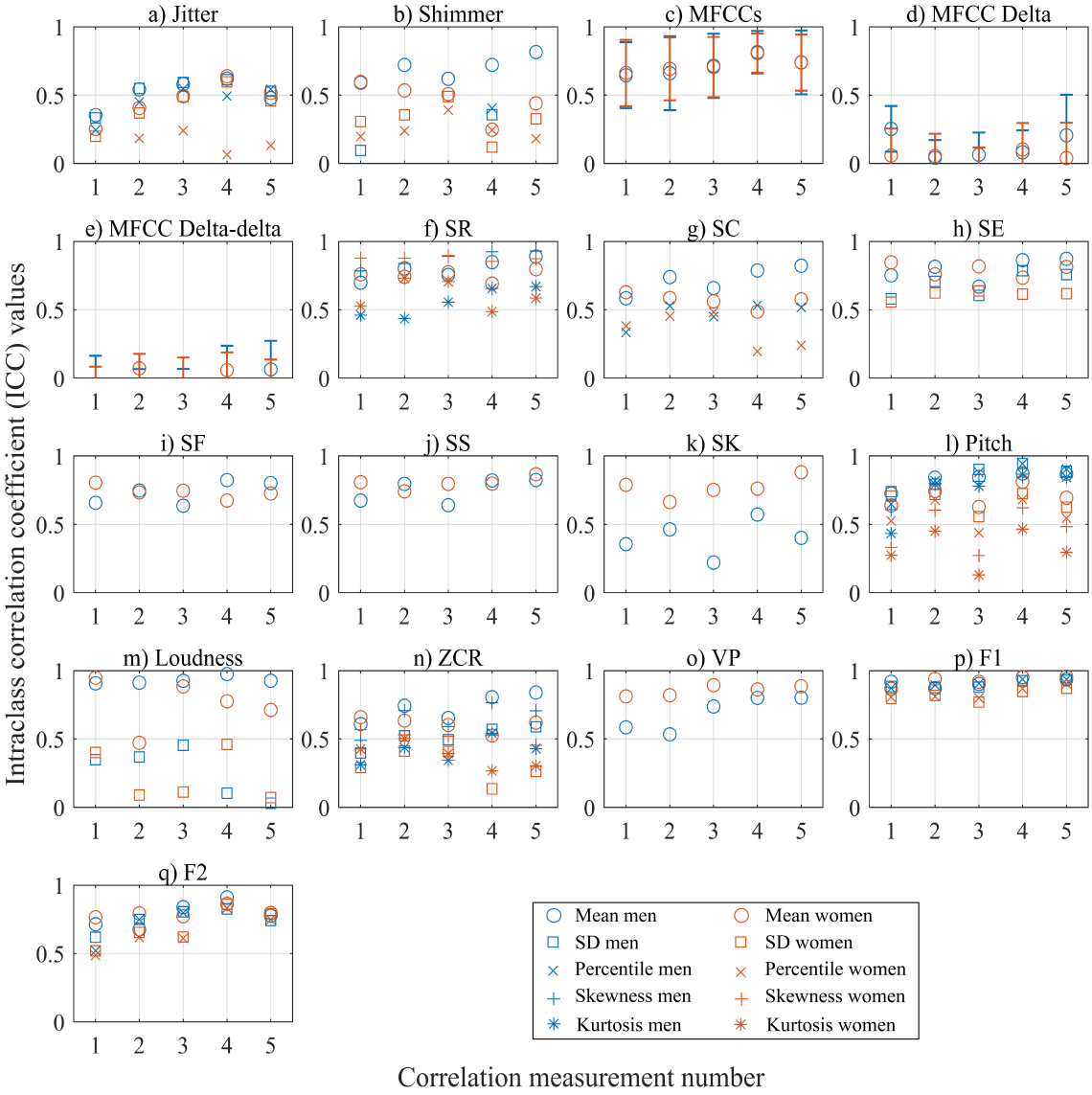

Fig. 4.4 The ICC values of bio-acoustic features for both depressed men and women is shown in the scatter plot. The numbers on the x-axis can be interpreted as follows; 1: ICC(30 s vs. 30 s), 2: ICC(30 s vs. 60 s), 3: ICC(30 s vs. 90 s), 4: ICC(30 s vs. 120 s), 5: ICC(60 s vs. 60 s).

significant in ICCs of jitter parameters at P=0.30 and P=1.00, respectively. Shimmers' agreement levels were impacted by participants group (P=0.05); no effect of speech duration was found on ICCs of these parameters (P=0.95). MFCC had similar ICC values when comparing speech with and without depression. Similarly, ΔMFCC ICCs were similar when depressed and non-depressed participants were compared. Task duration had a significant impact on ICCs of MFCC ($P< 0.05$), but it did not affect ICCs of ΔMFCC (P=0.14). Agreement level of ΔΔMFCC descriptors impacted by groups and shortening speech durations. A similar level of reproducibility in spectral shape features was found when depressed and non-depressed participants

and different speech durations were compared. No differences observed in ICC values of pitch, loudness, and ZCR parameters at group and duration levels. VP ICCs affected by duration changes ($P > 0.05$); no significant difference was found on ICCs of both depressed and non-depressed groups. While groups differences were significant on the agreement level of F1 parameters, it was not significant on the agreement level of F2 parameters. Duration influenced ICC values of both F1 and F2 parameters at $P < 0.05$.

4.4 Discussion

This study investigated the reproducibility of bio-acoustic measures in depressed and non-depressed individuals across different speech durations of a spontaneous speech task. Reproducibility of bio-acoustic features decreases with speech duration reduction among depressed and non-depressed men and women. In depressed and non-depressed individuals, source (jitter and shimmer), some cepstral (ΔMFCC and ΔΔMFCC), some spectral shape (SR, SC, SE, SF, and SK), some prosodic (loudness and ZCR) features are less reproducible when speech duration is changed. MFCC, SS, VP, and F2 achieved excellent reproducibility at a longer duration in both groups (i.e., depressed and non-depressed) and among both genders. Pitch parameters are less reproducible at shorter speech duration, with higher stability is observed in men than in women. F1 had excellent reproducibility in both depressed and non-depressed voices and across men and women. Furthermore, gender differences are statistically significant in SR, pitch, and VP for non-depressed voices, while it is significant in jitter, ΔMFCC, SK, pitch, ZCR, VP, and F1 for depressed voices.

In the clinical evaluation of patients' voices, it is essential to understand the stability of acoustical voice parameters. Previous studies investigated the influence of speech duration on some acoustic measurements of normal speakers. Still, a validation and generalisation of reproducibility findings of features measured from a reading-a-story task (Section 3.3.1) have to be explored on a different dataset and/or a different speech task type (i.e., spontaneous task).

Generally, the effect of speech sample length on bio-acoustic measures' reproducibility is consistent with the study in Chapter 3. In this study, the reproducibility analysis indicates a similar reduction pattern in the number of reproducible features when speech duration is shortened among depressed and non-depressed men and women. This reduction is higher in spontaneous speech compared to the reading task (Section 3.3.1). A potential reason behind this could be the nature of the spontaneous speech task, which is highly variable (i.e.,

acoustically and linguistically) and difficult to control [163], [188]. In a reading task, a lower articulation rate, more F0 variation, and less shimmer were reported compared to a spontaneous speech [189].

The results clearly show that shortening speech duration substantially impacts the stability of source parameters measured from the spontaneous speech; a similar finding was observed on those of reading-a-story (shown in Section 3.3.1). Usually, perturbation analysis is preferably performed on sustained vowels as it is expected to be steady and quasi-periodic [44]. However, continuous speech tasks present essential information about the coordination of the respiratory and laryngeal subsystems, which is important to evaluate a speaker's voice [182]. Jitter and shimmer measurements are significantly influenced by extraneous variables such as loudness of voice signal and speech content (e.g., vowel type) [183], which might be related to the instability observed in this study.

Results revealed that the reproducibility of spectral shape features quantified from the spontaneous speech task among depressed and non-depressed participants is sensitive to the changes of speech duration. However, this result was not consistently observed on those extracted from normal speakers and read a pre-defined story (Section 3.3.1). Furui *et al.,* characterised a spontaneous speech by a reduction in the spectral distribution compared to that of reading task [188]. Such a factor might be contributed to this difference in results. Although Nakamura *et al.* found an acoustic reduction of the MFCC spectrum as the speech becomes more spontaneous [190], the reproducibility of MFCC parameters was similar between participants' speech samples, regardless of speech task type, gender, and whether participants are depressed or not.

In terms of prosodic features, the reproducibility of pitch parameters gradually reduced as spontaneous speech sample is shortended (non-depressed participants); in contrast to the reading task, they remained reproducible across all duration comparisons (normal voices; as presented in Section 3.3.1). Hudson and Holbrook identified differences in F0 measures between reading and spontaneous speech samples [191], which could be a possible explanation for differences in the reproducibility findings. Differences in recording environments may also affect this result. Draxler *et al.* reported the dependency of F0 variability on both speaking task and duration [152], which is in line with my results (i.e., pitch is affected by speech duration). Further, stability of pitch parameters was higher for men than for women speakers, regardless of speaking tasks and whether participants were in depression or not. Gender differences in F0 are reported in a previous study [84] and have to be considered in future predictive speaking tasks. An earlier study

characterised continuous speaking contexts by frequent fluctuations in pitch and loudness to reflect intonation patterns, and emphatic stress [182]. My results showed a relatively strong effect of duration on the reproducibility of loudness attributes in depressed and non-depressed spontaneous speech; the same pattern was obtained from individuals with healthy voices and who read a story (Section 3.3.1).

My results also found that variation in duration had significant influences on both F1 and F2 qualities in depressed and non-depressed speech. F1 parameters maintained good to excellent reproducibility, while F2 was not reproducible in shorter speech duration. Although a similar pattern was observed in reading-a-story, where short speech data impacted formants' qualities, these parameters remained reproducible (Section 3.3.1). Vocal tract resonances are affected by changes in phonation [183], which might lead to a substantial difference in formants' reproducibility among different speech tasks; conversational speech is spoken more rapidly and less carefully articulated.

This study has several limitations. First, bio-acoustic features are derived from a restricted sample of English speakers; my results have to be investigated on voice samples of different challenges (languages and/or environments) and different contexts. Second, knowing that age impact acoustic characteristics of voice, I conducted this experiment without considering the age factor, suggesting a replication on matched age and another for a more diverse age. Furthermore, reproducibility assessment was performed using only the first two minutes of each recording, allowing standardised comparisons between speech samples. Hence, my findings remain to be confirmed by further research on longer speech durations and at different parts of speech samples. Finally, given that the content of the speech data impact the acoustic features, thus when 30 s speech duration was correlated with 120 s duration, speech content of 30 s is included within 120 s, which may affect the correlation results.

4.5 Conclusion

This study shows the impact of spontaneous speech duration on bio-acoustic features' reproducibility in depressed and non-depressed men and women. Results indicate that the reproducibility of bio-acoustic measures is largely affected by speech duration and speaker gender. For non-depressed participants, the number of reproducible features (out of 125 features) decreased from 71 to 38 in men and from 70 to 38 in women when speech sample length is decreased from 120 s to 30 s.

Similarly, in depressed voices, this number is reduced from 70 and 59 to 31 and 42 for men and women, respectively. Understanding the stability of acoustic measures is essential. Speech duration and gender have to be considered when acoustical measurements are used in the clinical setting.

Table 4.3 P-VALUES OF ICCs OF BIO-ACOUSTIC PARAMETERS COMPARING FEATURES' REPRODUCIBILITY IN DEPRESSED AND NON-DEPRESSED SPEAKERS SEPARATELY AND GROUPED AT DIFFERENT SPEECH DURATIONS AND ACROSS BOTH GENDERS.

Category	Features	*p*-value for comparing features reproducibility [a]					
		Non-depressed		Depressed		Non- and depressed	
		Duration	Gender	Duration	Gender	Group	Duration
Source	Jitter	0.34	0.90	**0.03**	**0.01**	0.30	1.00
	Shimmer	0.42	0.66	0.95	0.27	**0.05**	0.95
Cepstral	MFCC	**0.00**	0.90	**0.00**	0.73	0.12	**0.00**
	ΔMFCC	0.15	0.20	0.10	0.18	0.50	0.14
	ΔΔMFCC	0.11	0.09	0.43	0.26	**0.09**	**0.02**
Spectral	SR	0.05	**0.01**	0.83	0.96	0.40	0.65
	SC	**0.03**	0.84	0.94	0.11	0.25	0.52
	SE	0.51	0.85	0.77	0.46	0.94	0.48
	SF	1.00	1.00	1.00	1.00	1.00	1.00
	SS	0.23	0.90	0.42	0.29	0.50	0.10
	SK	0.28	0.20	0.60	**0.01**	0.32	0.40
Prosodic	Pitch	**0.00**	**0.00**	**0.003**	**0.00**	0.12	0.75
	Loudness	0.81	0.14	0.94	0.61	0.97	0.80
	VP	0.25	**0.01**	0.17	**0.02**	0.90	**0.004**
	ZCR	**0.00**	0.60	0.80	**0.01**	0.09	0.13
Formants	F1	**0.00**	0.90	**0.04**	**0.01**	**0.003**	**0.00**
	F2	**0.02**	0.80	**0.00**	0.18	0.54	**0.01**

[a] Bold *p*-value represents a significant difference.

Chapter 5

Automated Depression Detection System

5.1 Introduction

Human speech is sensitive to slight changes in the speaker's mental state. Neurophysiological changes associated with depression influence the laryngeal dynamics and then controlling ability of the vocal folds vibration [31]. Cognitive and physiological changes in a depressed individual's affect the speech production process [192]. Such changes potentially impact the acoustic qualities of the produced speech in a measurable way.

Computerised acoustic analysis could objectively evaluate speech with depression [187]; however, this analysis is affected by acoustic content and speech duration. Studies involving bio-acoustic characteristics, addressing depressed and non-depressed speakers, used different speaking task and different durations [10], [128], [136], [142]. Although many previous studies investigated the impact of speech task type on depression prediction [2], [163], [193], a relatively few studies has investigated the influence of short speech segments on the classification performance of acoustic measurements of depressed and non-depressed individuals [163], [194], [195].

Low *et al.* optimised the length of test data (i.e., 0.5, 1, 2, and 3 minutes) to maximise the classification accuracy of their model. They extracted acoustic descriptors based on Teager energy operator. The highest accuracy was reported for utterance length of one minute [118]. Alghowinem *et al.* studied the discriminative power of a classifier when smaller parts of speech are used; selected

from the beginning of a reading task. Their results showed a better classification performance for those parts compared to the whole speech sample [163]. Another study reported a good predicting accuracy (reached 73%) of voice parameters (26 LLD) extracted form short voice recordings (10 s) [59].

Afshan *et al.* trained a classifier on voice quality features and i-vectors derived from the full lengths of interview recordings (1.8 minutes). They then examined the predictive power of their classifier on a test data with shorter speech durations (40 s, 30 s, 20 s, and 10 s). Their findings revealed that the classification accuracy decreases when the speech length is shortened [194]. More recently, Huang *et al.* found that the speech file length influences the performance of depression detection, showing a higher F1 score at longer speech files. Their evaluation was based on convolutional neural networks and using vocal tract coordination features (e.g., MFCC and formants) [195].

Speech duration could affect the acoustic measurement outcomes, and therefore, classification results. Analysing acoustic parameters as a function of speech duration and its relation to depression identification may ultimately yield to a better understanding about the behaviour of these features and then the most informative speech segments.

Hence, this chapter aims to evaluate the association between depression and acoustic vocal qualities measured at different lengths of speech data. It also aims to explore the predictive ability of those qualities (i.e., quantified at different speech durations) with application to depression detection.

5.2 Methods

Fig. 5.1 illustrates the employed approach to identify MDD by utilising bio-acoustic features in this thesis. It mainly consists of features engineering and machine learning algorithm. Feature engineering is a process applied to extract the most discriminative descriptors from the raw audio signal, resulting in a numeric feature vector that characterise the input sample. This vector is then provided as an input to the machine learning algorithms (e.g., SVM) to construct and validate the classification model [81].

Speech samples from the DAIC-WOZ dataset (explained in Section 4.2.1) were also used in this study. Four sub-samples were obtained from the beginning of each spontaneous speech sample with durations of 120 s, 90 s, 60 s, and 30 s. Several pre-processing steps were applied on these sub-samples (detailed in

Section 3.2.2.1). Bio-acoustic characteristics were then computed, based on 53 LLD, for further analysis. Descriptive statistics were applied over features' frames, leading to a feature vector with 125 features per a speech sample (see Table 3.2). Most of these characteristics have been verified to be useful for depression detection [1], [118], [163] and quantified in the same way as described in Section 3.2.2.2.

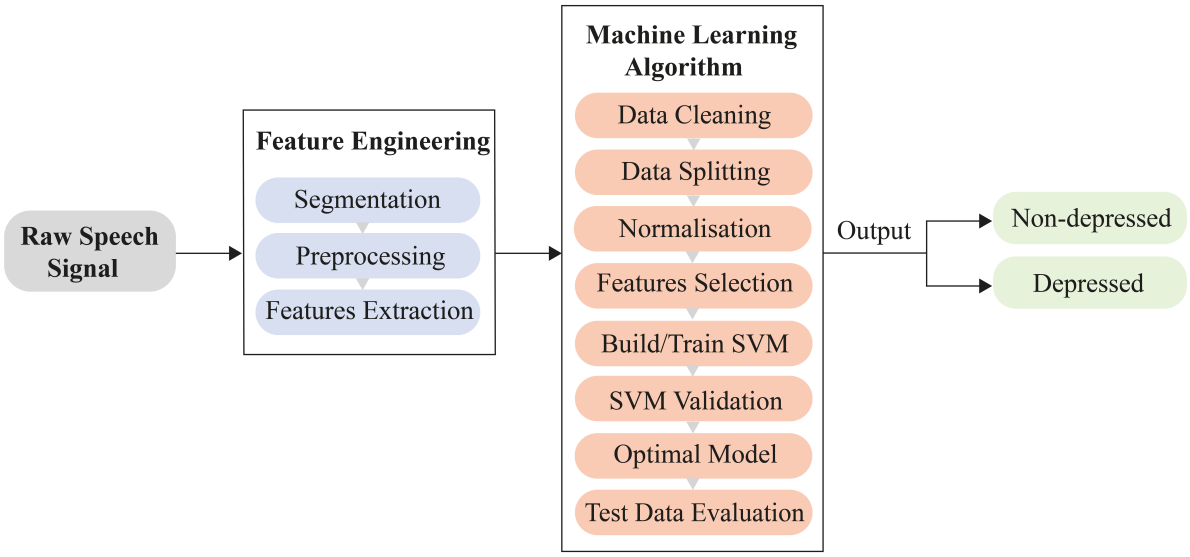

Fig. 5.1 A framework illustrates the employed approach for Major Depressive Disorders detection, considering bio-acoustic speech features extraction and machine algorithm. Output is a binary classification of speech characteristics of depressed or non-depressed participants.

5.2.1 Statistical Analysis

Two-way ANOVA followed by Tukey-Kramer *post-hoc* test (i.e., pairwise comparisons) was carried out on each voice descriptor obtained at four different speech durations. Two-way ANOVA test was used to determine whether the differences in bio-acoustic descriptors of participants groups (i.e., depressed and non-depressed) and different speech durations were statistically significant. *Post-hoc* was applied to identify where the differences in time occurred. The significance level was defined at $\alpha = 0.05$.

5.2.2 Machine Learning Algorithm

5.2.2.1 Data Preprocessing

Data preprocessing, which is a crucial step in machine learning analysis, removes irrelevant and redundant information and enhances the generalisation performance of learning models [196]. Several preprocessing steps were performed on the dataset, including data preparation, splitting the dataset, feature scaling, and feature selection.

Data Preparation

Spending time preparing and cleaning a dataset enhances the quality of the data and improves model performance [197]. Dataset exploration, outlier identification, and checking for missing values were performed.

Dataset Split

Dataset is commonly divided into two random and non-overlapping sets, which are referred to as *training set* and *test set*. The training set is used to train the models, while the test set is used to evaluate the generalisation performance of these models. A sample of the training set is held out (i.e., the validation set) to tune the model's hyperparameters. Using cross-validation techniques, a model is iteratively trained on the reduced training set (i.e., the full training set minus the validation set), after which it is validated on the validation set while tuning the hyperparameters. The purpose of this process is to estimate how the model's results will generalise to an independent dataset. Lastly, the model with the best performance on the validation set is selected and subjected to a generalisation assessment using the test set [197].

In this study, 75% of the dataset samples was used for the training set and 25% was held out for the test set. To mitigate the impact of the limited amount of data, as well as maximise the use of the data [163], [198], a leave-one-out (LOO) cross-validation strategy was used. LOO is an exhaustive cross-validation approach that uses a single observation from the training set as the validated data, while the rest are used as the training data. Besides, this strategy also capable of overcoming overfitting of the training data (i.e., the model "well explained" the training data), while improving the generalisation ability [199].

Feature Scaling

Feature scaling is a common requirement for many machine learning algorithms. Scaling feature values ensures that all numerical attributes are set within the same range and reduces bias towards the higher values. The most widely adopted

scaling techniques are normalisation and standardisation. In *normalisation*, data values are mapped to a range from 0 to 1, whereas in *standardisation*, data values are mapped to a distribution with a zero mean and unit variance [197]. In SVM, as an example, the objective function (i.e., the optimal hyperplane with the widest margin) is influenced by the scale of the input features [198]; as described in Section 5.2.2.2. In this study, the feature *standardisation* approach was applied. Scaling parameters were calculated on the training set and then applied to the test set and validation set, where one observation per experiment was left out (i.e., LOO), which helped to avoid data leakage during the model testing process.

Feature Selection

Feature selection techniques are primarily performed to eliminate irrelevant, redundant, and noisy features. This results in a compact subset of the most promising and informative qualities [196], [197]. Employing these techniques improves the generalisation performance of learning algorithms [59], [196]. Feature selection methods are divided into two categories: subset feature selection and feature transformation [200].

Subset feature selection aims to select a subset of features that minimise redundancy and maximise relevance to the class label. The three main categories of subset feature selection method are filters (i.e., a ranking-based statistical test), wrappers (i.e., a machine learning-based classifier), and embedded (i.e., also a machine learning-based classifier) [201]. Filtering techniques are more suitable for small datasets due to their ability to avoid overfitting problems [202]. Using this method, features are ranked based on certain criteria (e.g., Fisher Score), where the highest ranked features are chosen for further processing [201].

Transformation-based dimensionality reduction, also known as feature extraction, is used to remove noisy and redundant features. It transforms the original features into a new feature set with a lower-dimensional space [201]; knowledge of the class labels is not required. *Principal component analysis* (PCA) is the most popular dimensionality-reduction algorithm [197]. Statistically, PCA utilises orthogonal transformation to transform correlated features data into *principal components* (PCs) through an eigen-analysis [203], the aim of which to create uncorrelated components.

In this study, ANOVA and PCA method were performed on the standardised (i.e., scaled) dataset in each LOO cross-validation; their parameters are determined on the training set and then applied on the test set. PCA was applied to reduce feature space dimensionality, while preserving 98% of the original dataset variance.

5.2.2.2 Support Vector Machines

The SVM algorithm, introduced in the late 1970s by Vapnik and his collaborators, is considered a state-of-the-art classifier [163]. Evidence indicates that SVM classifier can robustly handle small datasets and achieve favourable generalisation properties [163], [192]. Compared to the most widely used machine learning approaches [2], [131], SVM has been effectively shown a great performance in the mental state prediction and classification tasks on the association of speech [192].

SVM establishes an optimal decision boundary in feature space to separate the training data points (or instances) into two discrete classes, while maintaining the widest distance (or maximum margin) between them. This boundary, known as a hyperplane, is represented as a function of *support vectors*; training points lie on the margin [204] (shown in Fig. 5.2). Boundary violations and training errors permitted by slack variables (denoted ζ) are penalised by the regularisation parameter (denoted C). This parameter controls the trade-off between the margin maximisation and training error minimisation [2]. Relaxing the margin constraints, softly penalising training points, and permitting a certain amount of training errors are useful for dealing with complex and overlapping real-world data [204]. Mathematically, such an optimisation problem can be represented as follows [205]:

$$\begin{aligned} \text{minimise}_{\mathbf{w},b,\zeta} : \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\zeta_i, \\ \text{subject to} : \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \zeta_i, \\ & \zeta_i \geq 0\,, \text{ for all } i = 1, \ldots, n, \end{aligned}$$

where $\mathbf{w}$ is the vector normal to the hyperplane, C is the regularisation parameter, ζ_i is the slack variable, $y_i \in \{1, -1\}$ is the target value, $\phi(\mathbf{x_i})$ is the mapping function, $\mathbf{x}_i \in R^n$ is the training vector, and b is a scaler bias.

Training data points of the two classes are not always linearly separable. Kernel techniques have been introduced to address the limitation of linear SVM and makes it applicable to complex real-world problems. These kernel functions, such as *radial basis function* (RBF), map feature space into a higher-dimensional space to find a non-linear decision boundary [204], where separability between the classes is achieved. SVM with RBF (Gaussian) kernel is employed in this study in a binary (i.e., depressed vs. non-depressed) gender independent modelling using LibSVM toolbox. Mathematically, RBF kernel can be represented as follows [205]:

$$\phi(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2\Big), \tag{5.1}$$

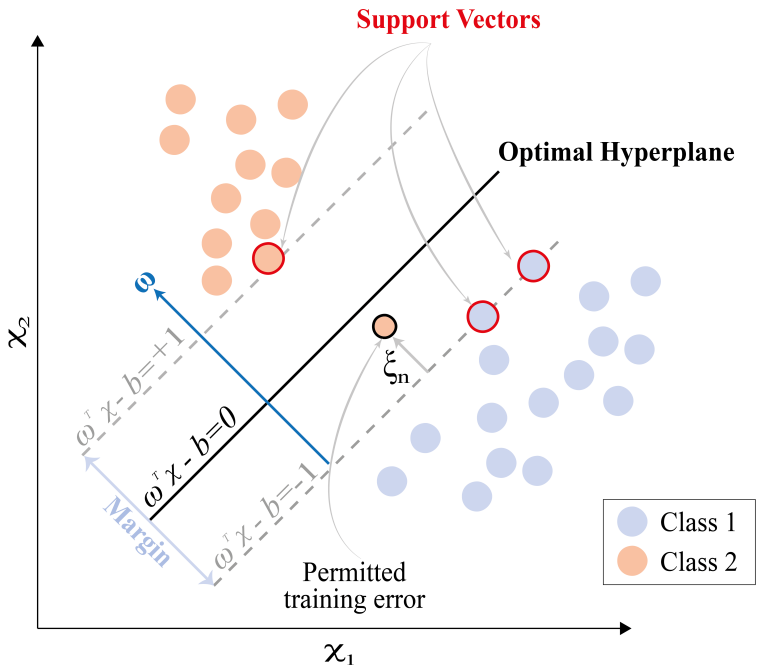

Fig. 5.2 Basic Support Vector Machine classifier of linearly separable data with soft margin hyperplane, permitting a number of training errors. An optimal hyperplane separating data points of two classes (i.e., class 1 and class 2), and support vectors lies closest to the hyperplane.

where $||\mathbf{x}_i - \mathbf{x}_j||$ is the distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ and γ is the gamma hyperparameter.

Optimisation of SVM can also be achieved by manipulating γ parameter. It measures the range of influence of training points at which decision boundary is defined. Increasing γ value causes a smaller influence range of training points, resulting in an irregular decision boundary and wiggling around individual point. Conversely, decreasing γ value causes a wider range of influence of the training points, resulting in a smoother decision boundary [197]. Fig. 5.3 presents implementation of SVM classifiers using a RBF kernel, showing the impact of increasing and decreasing both γ and C hyperparameters.

SVM training involves tuning a number of hyperparameters (γ and C). Finding the optimal hyperparameters could enhances the classification performance and predicting independent data accurately (i.e., the test data). C hyperparameter (tested values: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 10000]) was optimised using a grid search method with LOO cross-validation, while γ hyperparameter

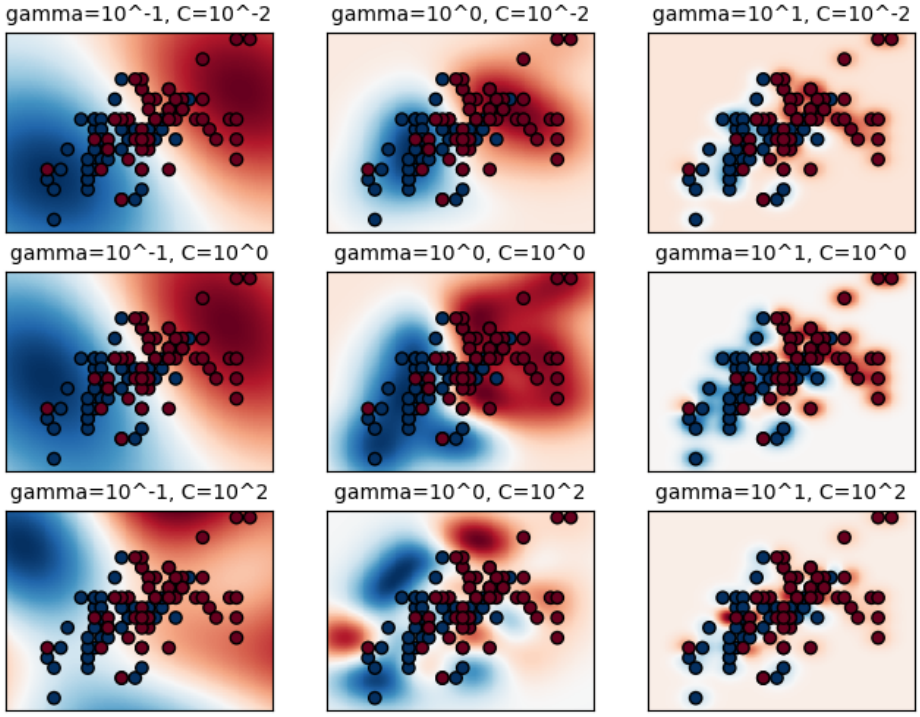

Fig. 5.3 SVM classifiers using a radial basis function kernel [206].

was kept in a default setting "scale", based on the following equation:

$$\gamma = \frac{1}{n_{\text{features}} \times X \cdot \text{var}()}, \tag{5.2}$$

where n_{features} is the number of features and $X \cdot \text{var}()$ is features variances.

In this study, since the used dataset is unbalanced (i.e., 131 non-depressed and 56 depressed), the class-weight parameter was set to be in "balanced" mode to adjust the weights of each class automatically. Although gender independent classifier was employed, the gender of each participant was added as a feature to the extracted set of bio-acoustic features, resulting in 126 features. Discriminative power of these features in a binary (i.e., depressed/non-depressed) manner using RBF SVM was investigated multiple times for different lengths of speech data, abbreviated at 30 s, 60 s, 90 s, and 120 s from the beginning of each speech recording.

Additionally, four representations from every speakers with 30 s speech duration and 30 s sliding window were extracted from each speech recording, resulting in a

Table 5.1 CONFUSION MATRIX FOR BINARY CLASSIFICATION.

Actual/ Predicted	Positive class	Negative class
Positive class	True positive (t_p)	False negative (f_n)
Negative class	False positive (f_p)	True negative (t_n)

748 (187×4) speech samples. For these samples, RBF SVM calssifier was employed and 5-fold cross-validation technique was used to fit the large amount of data.

5.2.2.3 Evaluation metrics

Several statistical methods based on the confusion matrix (shown in Table 5.1) were used to evaluate the classification performance [197]. In this study, *accuracy*, *precision*, *recall*, and weighted-averaged *F1 score* were computed.

Accuracy measures the ratio of correctly predicted instances to the total number of instances evaluated. It can be calculated as follows:

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + f_p + t_n + f_n}. \tag{5.3}$$

Precision is used to measure the ratio of correctly predicted positive instances over the positive class predictions. It is defined as follows [197]:

$$\text{Precision} = \frac{t_p}{t_p + f_p}. \tag{5.4}$$

Recall, also called sensitivity, is used to measure the ratio of positive instances that are correctly classified. Recall can be computed as follows [197]:

$$\text{Recall} = \frac{t_p}{t_p + f_n}. \tag{5.5}$$

Weighted F1 score is a weighted *harmonic mean* of the precision and recall. In this study, each label was given a weight equal to its support (i.e., the number of true instances for every label) [197]. This metrics can be measured by the following equation [207]:

$$F_\beta = (1 + \beta^2)\frac{Precision \times Recall}{\beta^2 Precision + Recall}, \tag{5.6}$$

where β is the relative importance given to recall over precision.

5.3 Results

5.3.1 Association of bio-acoustic features with depression and duration

Table 5.2 summarises the statistically significant bio-acoustic characteristics associated with depression and which of these were significantly affected by duration changes. It also shows the specific source of significance among the four investigated speech durations if task duration is significant. Of 125 bio-acoustic features, only 37 were significantly different between depressed and non-depressed individuals. Among those 37 features, seven qualities showed a significant change over speech duration. Fig. 5.4 shows the mean and standard error of these qualities at different speech durations.

Mean jitter feature in depressed patients was significantly (P=0.007) lower than that of non-depressed participants. Mean pitch values increased in depressed compared to non-depressed individuals; no significant effect of speech duration was observed (P=0.543). Both group and duration had a significant ($P<0.05$) impact on mean SF descriptor, with a slightly higher mean value for non-depressed group in comparison with depressed group (as shown in Fig. 5.4a). A significant difference, indicated by *post-hoc* test, found between mean SF feature measured at 30 s speech duration and those measured at 60 s, 90 s, and 120 s. Eleven out of 13 mean MFCCs affected by depression ($P< 0.05$). Most of these coefficients did not impact by changes in duration. Only mean MFCC1 values were significantly influenced by speech task duration (P=0.040); differences in feature values were observed between 30 s and 120 s speech segments (Fig. 5.4d). In depressed group, mean VP feature values showed a significant reduction (P=0.011) compared to non-depressed group. Mean F1 parameter was significantly higher in depressed than in non-depressed individuals; changes in task duration were not significant.

Depression significantly affected SD jitter quality (P=0.007), showing a lower value in depressed group; no significant speech duration differences was found (P=0.596). Depressed participants' voices had a significant (P=0.024) reduction in SD ZCR feature as compared to those of non-depressed speakers (Fig. 5.4c). Speech sample length influenced the value of this feature (P=0.009); differences were observed between the samples of 30 s duration and those with 90 s and 120 s durations. SD MFCC7, MFCC9, MFCC10, and MFCC13 values increased in depressed than non-depressed speakers; duration changes had no significant impact on these measurements.

Higher values of skewness pitch parameter observed in depressed participants compared to non-depressed participants (Fig. 5.4d); changes of value of this parameter across different durations was significant (P=0.001). Skewness of MFCC8 and MFCC13 was significantly lower in speech with depression (P$<$ 0.020) and was not affected by sample length. Like skewness pitch, kurtosis pitch was increased with depression; duration factor affected this measure where differences found between 30 s duration and those at 90 s and 120 s (P=0.014; Fig. 5.4e). Depression and speech duration influenced kurtosis ZCR parameter (P=0.032 and P=0.004, respectively; Fig. 5.4f). Some kurtosis MFCCs properties were affected by depression (MFCC4, MFCC5, MFCC9, and MFCC10), while others did not show a statistically significant difference.

Percentile range shimmer quality was significantly higher in speech with depression; no duration effect was observed (P=0.626). Percentile range MFCC7, MFCC10, and MFCC13 values increased significantly with depression (P$<$ 0.05), with no significant duration impact was found. Both depression and speech length impacted percentile range SC feature at P=0.045 and P=0.0004, respectively. This measure at 30 s speech duration was statistically different from the ones at longer duration, defined at 60 s, 90 s, and 120 s (Fig. 5.4g).

Table 5.2 BIO-ACOUSTIC FEATURES ASSOCIATED WITH DEPRESSION AND SPEECH DURATION

Feature	**Effect** [a]	***p*-value**			**Post-hoc**[c]
		Group	**Duration**[b]	**Interaction**	
Mean Jitter	D<ND	0.007	0.582	0.967	–
Mean Pitch	D>ND	0.003	> 0.999	> 0.999	–
Mean SF	D<ND	0.009	**0.0005**	0.999	Seg. 2,3,4
Mean MFCC1	D<ND	0.007	**0.040**	0.989	Seg. 4
Mean MFCC2	D<ND	0.0045	0.267	0.989	–
Mean MFCC3	D>ND	0.044	0.197	> 0.999	–
Mean MFCC4	D>ND	0.013	0.460	0.999	–
Mean MFCC5	D<ND	0.0003	0.912	0.973	–
Mean MFCC7	D>ND	0.002	> 0.999	0.999	–
Mean MFCC8	D<ND	< 0.0001	0.981	0.994	–
Mean MFCC9	D>ND	0.0008	0.971	0.989	–
Mean MFCC10	D<ND	0.0087	0.992	0.995	–
Mean MFCC12	D>ND	0.024	0.936	0.998	–
Mean MFCC13	D<ND	0.046	0.950	0.996	–

Continued on next page

Table 5.2 – continued from previous page

Feature	Effect [a]	*p*-value			Post-hoc[c]
		Group	Duration [b]	Interaction	
Mean VP	D<ND	0.011	0.994	0.864	–
Mean F1	D>ND	0.011	0.998	0.997	–
SD Jitter	D<ND	0.007	0.596	0.942	–
SD ZCR	D<ND	0.024	**0.009**	0.992	Seg. 3,4
SD MFCC7	D>ND	0.023	0.593	0.979	–
SD MFCC9	D>ND	0.031	0.835	0.999	–
SD MFCC10	D>ND	0.048	0.902	0.999	–
SD MFCC13	D>ND	0.019	0.905	0.924	–
Skewness Pitch	D>ND	0.008	**0.0013**	0.999	Seg. 2,3,4
Skewness MFCC8	D<ND	0.020	0.909	0.913	–
Skewness MFCC13	D<ND	0.0005	0.990	0.999	–
Kurtosis Pitch	D>ND	0.0001	**0.014**	0.999	Seg. 3,4
Kurtosis SR	D>ND	0.0004	0.437	0.954	–
Kurtosis ZCR	D>ND	0.032	**0.0004**	0.835	Seg. 2,3,4
Kurtosis MFCC4	D>ND	0.005	0.630	0.921	–
Kurtosis MFCC5	D>ND	0.0001	0.230	0.961	–
Kurtosis MFCC9	D>ND	0.016	0.933	0.607	–
Kurtosis MFCC10	D<ND	0.012	0.938	0.945	–
Percentile range Shimmer	D>ND	0.0001	0.626	0.509	–
Percentile range MFCC7	D>ND	0.041	0.617	0.973	–
Percentile range MFCC10	D>ND	0.035	0.941	0.999	–
Percentile range MFCC13	D>ND	0.022	0.887	0.938	–
Percentile range SC	D<ND	0.045	**0.0004**	0.894	Seg. 2,3,4

[a] D, Depressed participants; ND, non-depressed participants.

[b] Bold *p*-values represent significant differences of speech duration.

[c] Seg.2, speech segment at 60s; Seg.3, speech segment at 90s; Seg.4, speech segment at 120s.

5.3.2 Effect of speech duration on classification performance

Fig. 5.5 illustrates the depression classification results of bio-acoustic features obtained at different durations of spontaneous speech task and analysed using RBF SVM classifier. Generally, the classification performance of the speech segments with 30 s duration outperformed the ones at longer speech durations. Classification accuracy was around 70% for the features measured at 30 s speech duration. This percentage decreased to around 60% at 60 s, and then it stayed relatively stable

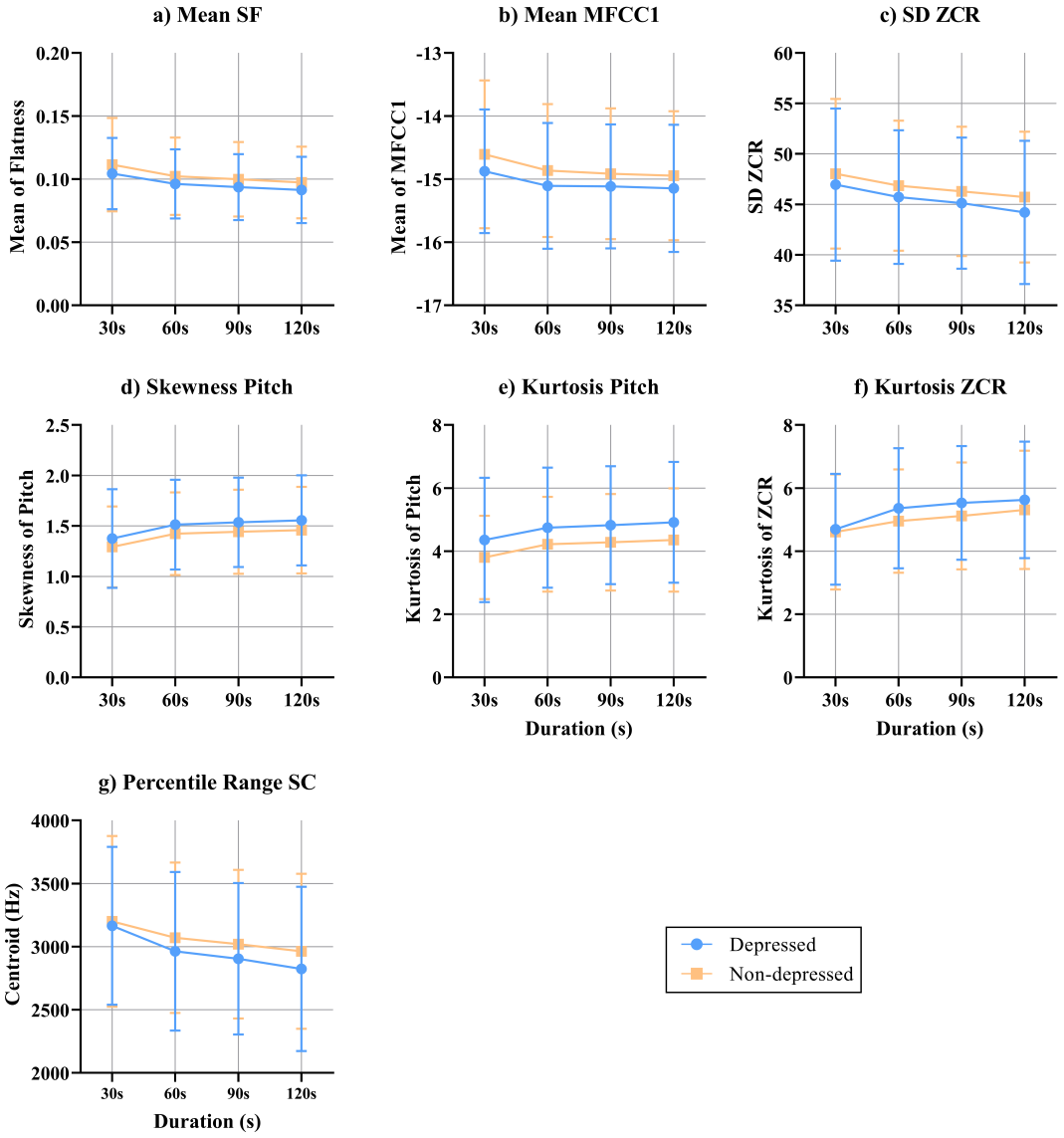

Fig. 5.4 Mean and standard error of bio-acoustic features that are significantly affected by speech duration in depressed and non-depressed participants.

when 90 s and 120 s of speech durations were considered. Similarly, the weighted F1 score achieved 70% at 30 s before it decreased to around 60% at 120 s.

Recall measures were stable (=40%) when task duration was increased from 30 s to 90 s. At 120 s, it decreased to around 25%. Speech duration also impacted the precision values. Increasing speech sample length worsened the precision measure. It decreased from around 59% to around 39% when features measured at 30 s and 120 s were analysed, respectively.

Table 5.3 showed the performance of depression identification by utilising multiple

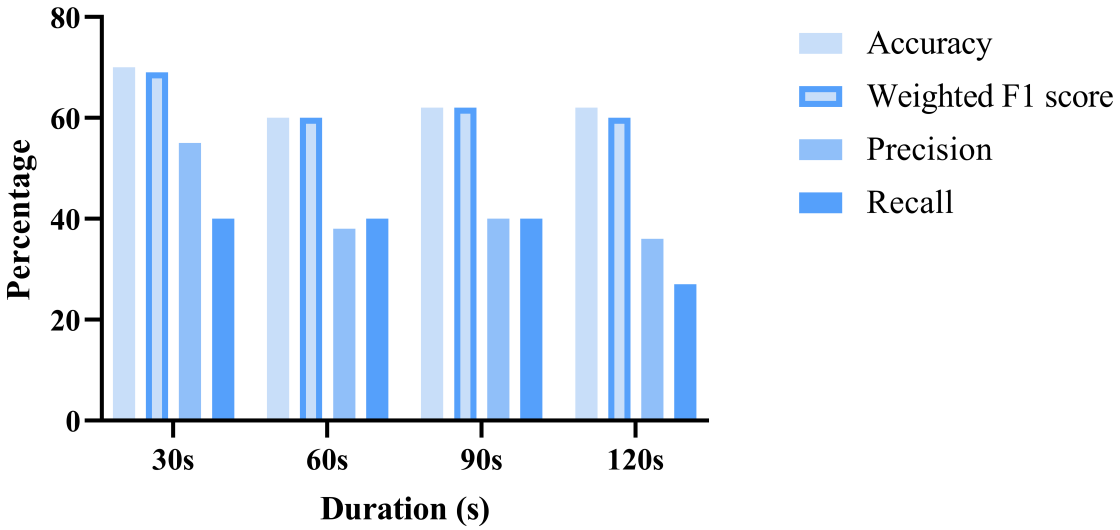

Fig. 5.5 Classification results using RBF SVM at different speech sample lengths.

Table 5.3 CLASSIFICATION PERFORMANCE USING MULTIPLE REPRESENTATIONS FROM EACH SPEAKER.

Evaluation metrics	Percentage
Accuracy	73.26%
Weighted F1 score	72.30%
Recall	49.20%
Precision	63.62%

representations (i.e., four representations from each speaker) with 30 s speech duration. For this experiment, accuracy was around 73%, weighted F1 score was around 72%, recall was around 49% , and precision achieved 63%.

5.4 Discussion

This study explored the effects of depression and speech duration on the bio-acoustic characteristics obtained from spontaneous speech task. It also examined the classification performance of acoustic measures, quantified at different lengths of speech samples, using RBF SVM classifier. Study findings can be summarised as follows: (i) depression significantly impacted some of the bio-acoustic features; (ii) length of speech sample affected the values of mean SF and MFCC1 features, SD ZCR feature, skewness pitch feature, kurtosis pitch and

ZCR features, and percentile range SC feature. Significant differences observed on those quantified at 30 s duration compared to the ones measured at 60 s, 90 s, and 120 s; (iii) SVM predictive ability of depression slightly improved when speech duration is shortened, with accuracy achieved more than 70% at 30 s.

Previous studies characterised speech with depression by higher jitter and shimmer values [31], [111]. Similarly, percentile range shimmer feature in this study was higher in depressed individuals. However, this study showed a lower mean jitter value in depressed compared to non-depressed groups, which is not consistent with the findings in [31], [111]. A potential reason behind this inconsistency is that Kiss and Vicsi and Quatieri and Malyska, performed their experiments on vowel voices and voiced parts of reading-a-story [31], [111], whereas, in this study, spontaneous speech task was used (i.e., highly variable). Additionally, this study found that the changes of mean jitter and percentile range shimmer over time were not statistically significant. The potential reason behind this could be the speech length used in the current study (i.e., the shortest length was 30 s), which is much longer than the suggested duration of at least 3 s to get accurate measures of perturbations qualities [150].

In this study, most of the spectral shape features did not significantly impact by depression. This could be the reason behind the poor performance of depression prediction reported by Lopez-Otero *et al.* when only spectral features were used [17]. Therefore, spectral shape descriptors could not be good indicators of depression. Results showed that depression only affected the values of some of SF, SC, and SR parameters. These parameters could be promising spectral shape features for depression analyses. This study also found that mean SF and percentile range SC qualities contain significant information to discriminate between speech with and without depression, presenting lower values in depression. Their changes across different speech lengths were significant. Hussenbocus *et al.* found higher SC values in non-depressed speech [69]; a similar finding observed in this study. A flat spectrum was observed in speech with depression [60], this implies patients experiencing depression have a high SF value. However, in the current study, lower values of mean SF feature were observed in depressed participants compared to non-depressed participants. In line with the current study, results presented by Stolar *et al.* showed SR values were higher for depressed than non-depressed groups when cut-off points were higher than 55%, [28]. Meaning that energy is concentrated in the higher frequencies range in depressed participants, while it is concentrated in the lower frequencies range in non-depressed participants. Alghowinem *et al.* observed that utilising MFCCs

features for identifying individuals with depression gives relatively good classification performance; however, MFCCs with its first- and second-derivatives performed slightly better than MFCCs alone [163]. Although this study showed that the differences in mean MFCCs parameters between the two groups are significant, ΔMFCC and ΔΔMFCC were not.

In terms of prosodic measurements, pitch parameters (mean, skewness, and kurtosis) showed higher values in depressed group, which is consistent with findings of Quatieri and Malyska [31]. Speech sample length impacted skewness and kurtosis pitch features. Duration of speech task affects the F0 stability, as reported in [152], [153]. Even though loudness quality was observed to be significantly lower in depressed participants [175], it was not a significant feature in this study. Similar to my finding, Alpert *et al.* found no statistical significant differences in loudness measure, comparing depressed and non-depressed participants' voices [208]. Depression and speech duration significantly affected ZCR (SD and kurtosis) characteristics. Notably, ZCR might be considered an informative feature when discriminating between speech with and without depression. Similar to spectral shape features, using only prosodic features to capture depression level resulted in a poor performance of learning model [17].

By exploring the effect of depression on formants' properties, only differences on mean F1 was statistically significant between the groups, showing higher values in depressed patients. A similar increasing pattern was observed by Mundt *et al.*, but no significant differences between the control and depressed group were reported [187]. This study also showed that the duration changes were insignificant in F1 mean. Although this result contradicts that found by Gendrot and Adda-Decker, who found the speech duration affected both F1 and F2 [181], a different speech task is used in the current study. They utilised more stable speech samples (i.e., vowels), while this study used a highly variable spontaneous speech.

Some bio-acoustic metrics showed instability across time. This instability is mainly observed at the speech segments with 30 s duration (i.e., selected from the beginning of the speech task) compared to those with longer durations (i.e., 60 s, 90 s, and 120 s). Instability of the beginning (vocal onset) and end (vocal offset) of the speech sample, caused by changes in aerodynamic and muscular adjustment factors, was reported [44]. Due to these changes, an increase in the speech perturbations values, jitter and shimmer, as an example, was found. Consequently, acoustic analyses were usually performed on stable regions (i.e., with minimum jitter, shimmer, and nonlinear dynamic parameter correlation dimension with the maximum signal-to-noise ratio) of the speech sample [44]. This instability could

explain the contradictory results in classification performance found in previous studies [163], [194], [195].

This study also showed that as the speech samples duration decreases from 120 s down to 30 s, the classification performance slightly improved. The RBF SVM classifier gave relatively better classification accuracy results (by around 10%) when the bio-acoustic features extracted from the first 30 s of a speech sample was utilised. Supporting this finding, Alghowinem *et al.* pointed out using a smaller amount of speech data, starts from the beginning of speech recordings, gives better results than using the full speech length [163]. Although several published studies found that shortening speech duration resulted in a reduction in depression classification performance [194], [195], direct comparison between current study results and their results is not straightforward. Their results were found using different acoustic features and/or different learning algorithms. Also, Afshan *et al.* performed their study on Mandarin speakers [194], while this study used speech samples of English speakers. Huang *et al.* utilised speech data with durations shorter than 35 s [195]. This study examined classification performance at 30 s duration and longer.

This study indicates that the beginning of speech provides essential discriminative information that might be considered while analysing speech for depression. Comparing the results of bio-acoustic features analyses with depression classification performance at different durations, the instability of these features at the beginning of speech signals could be a factor affecting the depression identification, yielding a relatively better discriminative power, which could augment the conventional evaluation methods in the clinic.

There are some limitations in this study. First, both men and women were analysed together. Some bio-acoustic features are well known to be gender-dependent (e.g., pitch). Gender-based learning models are outperformed gender-independent models [59]. Due to the limited number of depressed participants, gender-independent model was employed in this study and only added gender as a feature to the feature set. Hence, a further investigation is required to analyse men's and women's voices separately. Second, speech samples of different class labels used in this study are unbalanced, suggesting a replication on a balanced dataset to avoid computational bias. Furthermore, this study includes only two minutes selected from the beginning of each recording and analyses bio-acoustic features at four different speech durations. Therefore, analysing speech data using a shorter speech sample length (>30 s) with more time steps is necessary. Finally, this study used RBF SVM classifier, limiting generalisation on other learning

algorithms. Future studies have to be conducted using different classifiers to validate and generalise findings of depression detection.

5.5 Conclusion

This study evaluated the association between bio-acoustic features obtained from different speech durations and depression. It also investigated the impact of spontaneous speech task length on the performance of RBF SVM classifier. Only 37 acoustic characteristics (out of 125) provided important information to differentiate between depressed and non-depressed participants. The most informative bio-acoustic features were: jitter, shimmer, pitch, VP, ZCR, SF, SR, SC, and MFCC. Of 37 measures, only seven qualities were influenced by duration changes. Instability of these features is observed at the beginning of each speech recording (i.e., measured at 30 s) compared to longer speech durations (60 s, 90 s, and 120 s). Further, this study found a slight improvement in the classifier predictive power when speech segments with 30 s duration were utilised.

Speech sample length affected the stability of bio-acoustic descriptors, and therefore, their analyses through SVM classifier. Differences in the prediction performances for the speech at different durations may underscore the need for more research to link variations of acoustic features quantified from the beginning of a speech to the performance of depression prediction. Therefore, vocal acoustic measures quantified from the beginning of speech are unstable and seem to be more effective in clinical depression prediction.

Chapter 6

Conclusions

Speech signal carries essential information about the physiological condition and pathophysiological state of a speaker. Bio-acoustic qualities of voice show evolving value in analysing psychiatric illnesses. Obtaining a sufficient speech sample length to quantify these qualities is essential. However, the impact of sample duration on the reproducibility of bio-acoustic features of speech with and without depression has not been systematically explored. Still, discriminative power of acoustic features obtained from depressed and non-depressed individuals across different sample lengths has to be explored. This thesis delineated the effect of speech task duration on the reproducibility, stability, and classification of depressed and non-depressed bio-acoustic characteristics. This chapter summarises the main findings of the studies conducted for this thesis and potential future studies extended form this thesis.

6.1 Thesis Summary

Examining reproducibility of bio-acoustic characteristics against changes in speech task durations and speech task types is the most significant contribution of this thesis. The evaluation was conducted over features quantified from normal English-speaking adults ($n = 185$; 87 m, 98 w) for reading-a-story and counting tasks. It was also carried out on features obtained from a spontaneous speech task of depressed ($n = 56$; 25 m, 31 w) and non-depressed ($n = 133$; 77 m, 56 w) English speakers. The extracted bio-acoustic features are source, spectral, cepstral, formant, and prosodic features. The intraclass correlation coefficients (ICCs) test was used to assess the degree of agreement between features measured at different durations. A two-way analysis of variance (ANOVA) test was applied to evaluate the

influence of duration and gender on the agreement level of feature parameters. Results showed that the number of reproducible features (out of 125) in reading-a-story and spontaneous speech tasks decreased stepwise with duration reduction. Gender differences on ICC values of some acoustic measures were significant. Changing speech task type from reading-a-story to counting tasks, keeping speech duration at 10 s, significantly impacted the reproducibility of the most measured qualities, in part due to the short counting task duration. Therefore, bio-acoustic features are less reproducible in shorter samples and are affected by gender.

Major depressive disorder (MDD) altered several bio-acoustic features of speech. The identification of MDD by analysing these characteristics might be considered an objective biomarker. However, the stability of bio-acoustic features across different durations may affect the classification performance of depressed and non-depressed acoustic features. Therefore, this thesis explored the effect of depression on bio-acoustic measures across different lengths of spontaneous speech task, determined at 30 s, 60 s, 90 s, and 120 s from the beginning of each recording. A two-way ANOVA test was conducted to evaluate the differences between depressed and non-depressed acoustic features and among the four sample lengths. The *post-hoc* test was applied, if the duration factor is significant, to determine where the differences between speech durations came from. Experimental results found that only 37 features were highly sensitive to depression. Of these measures, only seven features were affected by duration changes; differences were mainly found between speech segments with 30 s duration and other durations. Finally, this thesis assessed the predictive ability of depression against changes in length of speech data utilising a support vector machine algorithm with a radial basis function (RBF SVM). Results indicated that classification performance of bio-acoustic qualities was affected by task duration. Shortening speech duration down to 30 s slightly improved classification metrics.

6.2 Future Research

The findings of this thesis were based on speech samples of reading-a-story task (i.e., mean duration=124.4 s) and counting task (i.e., mean duration=26 s) for English speakers with healthy voices (as shown in Chapter 3). They were also based on spontaneous speech task (duration=120 s) for depressed and non-depressed English speakers (as shown in Chapter 4 and 5). Therefore, verifying the main findings

presented in this thesis using different speech corpora, speech task types, speech task lengths, and languages could be valuable to obtain more reliable results.

Bio-acoustic features have shown remarkable success in the depression detection field. This thesis continues to support that depression exhibit some changes in bio-acoustic characteristics (as shown in Chapter 5). Therefore, it would be beneficial to build a fully automated system integrated into the clinical settings for depression detection, including speaker diarisation, features extraction, and feature analysis. It would also be valuable to investigate more the relation between features instability along the speech sample (i.e., during vocal onset) and the classification performance of depressed and non-depressed bio-acoustic features. Employing a similar method (i.e., features extraction from short speech duration) using different SVM kernels, different learning algorithms and a large and balanced depression corpora would be helpful. Selecting the story to be read could be further refined in future through a process of optimisation (e.g. the story that maximises the reproducibility and results of classification), thus improving standardisation. Further studies on monitoring mental health remotely using voice-controlled interfaces (e.g., Amazon Alexa and Apple Siri) is required, especially when Alexa device became Health Insurance Portability and Accountability Act (HIPAA) compliant by setting standards for protection-sensitive data of patients and avoiding user privacy violations [209].

Finally, this thesis only focused on analysing acoustic characteristics of voice as a function of speech duration. Analyses of speech linguistic features at different speech durations could be a rich source to identify depression.

Appendix A

Supplementary Material

SMRC Self-Management Resource Center

Personal Health Questionnaire Depression Scale (PHQ-8)

Over the **last 2 weeks**, how often have you been bothered by any of the following problems?
*(circle **one** number on each line)*

How often during the past 2 weeks were you bothered by...	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

Scoring

If two consecutive numbers are circled, score the higher (more distress) number. If the numbers are not consecutive, do not score the item. Score is the sum of the 8 items. If more than 1 item missing, set the value of the scale to missing. A score of 10 or greater is considered major depression, 20 or more is severe major depression.

Bibliography

[1] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Invest. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, Feb. 2020, ISSN: 2378-8038. DOI: 10.1002/lio2.354.

[2] H. Jiang, B. Hu, Z. Liu, *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Commun.*, vol. 90, pp. 39–46, Jun. 2017, ISSN: 0167-6393. DOI: 10.1016/j.specom.2017.04.001.

[3] H. Jiang, B. Hu, Z. Liu, *et al.*, "Detecting depression using an ensemble logistic regression model based on multiple speech features," *Comput. Math. Methods Med.*, vol. 2018, Sep. 2018, ISSN: 1748-6718. DOI: 10.1155/2018/6508319.

[4] M. H. Farouk, "Speech production and perception," in *Application of Wavelets in Speech Processing*. Springer, 2018, pp. 5–10.

[5] J. Kreiman and D. Sidtis, "Producing a voice and controlling its sound," in *Foundations of Voice Studies*. Oxford, UK: Wiley-Blackwell, Apr. 2011, ch. 2, pp. 25–71, ISBN: 0631222979, 9781444395068. DOI: 10.1002/9781444395068.ch2.

[6] P. Song, "Assessment of vocal cord function and voice disorders," in *Principles and Practice of Interventional Pulmonology*. New York: Springer, New York, NY, Nov. 2012, pp. 137–149, ISBN: 978-1-4614-4291-2,978-1-4614-4292-9. DOI: 10.1007/978-1-4614-4292-9_14.

[7] R. T. Sataloff and S. W. Beet, "Clinical anatomy and physiology of the voice," in *Voice Science*. San Diego: Plural Publishing, Apr. 2017, ch. 6, pp. 67–106, ISBN: 9781597568623.

[8] E. L. Stegem´oller, "The neuroscience of speech and language," *Music therapy perspectives*, vol. 35, no. 2, pp. 107–112, 2017, ISSN: 0734-6875.

[9] B. Diop, R. Collignon, M. Guèye, and T. Harding, "Diagnosis and symptoms of mental disorder in a rural area of senegal," *Afr. J. Med. Med. Sci.*, vol. 11, no. 3, pp. 95–103, Sep. 1982, ISSN: 0309-3913.

[10] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), 2013, pp. 2172–2176.

[11] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, Dec. 2018, ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2018.07.007.

[12] Y. Tahir, Z. Yang, D. Chakraborty, *et al.*, "Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia," *PloS one*, vol. 14, no. 4, e0214314–e0214314, Apr. 2019, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0214314.

[13] C. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, no. 4, pp. 452–460, Apr. 1976. DOI: 10.1109/PROC.1976.10154.

[14] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker dependency of spectral features and speech production cues for automatic emotion classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, Apr. 2009, pp. 4693–4696. DOI: 10.1109/ICASSP.2009.4960678.

[15] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), Aug. 2011, pp. 2997–3000. DOI: 10.21437/Interspeech.2011-750.

[16] T. Boonla and T. Yingthawornsuk, "Assessment of vocal correlates of clinical depression in female subjects with probabilistic mixture modeling of speech cepstrum," in *Proc. 11th Int. Conf. Contr., Autom., Syst.*, Oct. 2011, pp. 387–391.

[17] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," in *Proc. 2nd Int. Workshop Biom. Forensics (IWBF)*, IEEE, Mar. 2014, pp. 1–6, ISBN: 978-1-4799-4370-8. DOI: 10.1109/IWBF.2014.6914245.

[18] M. Senoussaoui, M. Sarria-Paja, J. Santos, and T. Falk, "Model fusion for multimodal depression classification and level detection," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, ser. AVEC '14, ACM, Nov. 2014, pp. 57–63, ISBN: 145033119X.

[19] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *J. Biomed. Inf.*, vol. 83, pp. 103–111, Jul. 2018, ISSN: 1532–0464. DOI: DOI:10.1016/j.jbi.2018.05.007.

[20] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proc. Audio/Vis. Emotion Challenge Workshop*, ser. AVEC '18, New York, NY, U. S., Oct. 2018, pp. 23–30. DOI: 10.1145/3266302.3268997.

[21] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Vocal acoustic analysis and machine learning for the identification of schizophrenia," *Res. Biomed. Eng.*, vol. 37, no. 1, pp. 33–46, Mar. 2021, ISSN: 2446-4732. DOI: 10.1007/s42600-020-00097-1.

[22] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 272–281, Nov. 2020. DOI: 10.1109/JSTSP.2019.2955022.

[23] N. Cummins, "Automatic assessment of depression from speech: Paralinguistic analysis, modelling and machine learning," Ph.D. dissertation, School of Elect. Telecommun. Eng., UNSW Australia, Sydney, Australia, 2016.

[24] F. Scibelli, G. Roffo, M. Tayarani, *et al.*, "Depression speaks: Automatic discrimination between depressed and non-depressed speakers based on

nonverbal speech features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, IEEE, Apr. 2018, pp. 6842–6846, ISBN: 9781538646588. DOI: 10.1109/ICASSP.2018.8461858.

[25] WHO, *Depression fact sheets*, https://www.who.int/news-room/fact-sheets/detail/depression, 2020.

[26] Who, *Depression: A global crisis*, World Mental Health Day, 2012.

[27] C. D. Mathers, E. T. Vos, C. E. Stevenson, and S. J. Begg, "The burden of disease and injury in australia," *Bull. W. H. O.*, vol. 79, no. 11, pp. 1076–1084, 2001, ISSN: 0042-9686. DOI: 10.1590/S0042-96862001001100013.

[28] M. N. Stolar, M. Lech, S. J. Stolar, and N. B. Allen, "Detection of adolescent depression from speech using optimised spectral roll-off parameters," *Biomed. J. Sci. Technol. Res.*, vol. 5, no. 1, pp. 1–10, Jun. 2018.

[29] V. Mitra and E. Shriberg, "Effects of feature type, learning algorithm and speaking style for depression detection from speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, Apr. 2015, pp. 4774–4778, ISBN: 1467369977. DOI: 10.1109/ICASSP.2015.7178877.

[30] M. Valstar, B. Schuller, K. Smith, *et al.*, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, ser. AVEC '14, ACM, Nov. 2014, pp. 3–10, ISBN: 145033119X. DOI: 10.1145/2661806.2661807.

[31] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), 2012, pp. 1059–1062.

[32] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 142–150, Apr. 2013, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2012.38.

[33] J. F. Cohn, T. S. Kruez, I. Matthews, *et al.*, "Detecting depression from facial actions and vocal prosody," in *3rd Int. Conf. Affective Comput. Intell. Interact. (ACII)*, IEEE, Sep. 2009, pp. 1–7, ISBN: 9781424448005. DOI: 10.1109/ACII.2009.5349358.

[34] A. A. König, N. Linz, R. Zeghari, *et al.*, "Detecting apathy in older adults with cognitive disorders using automatic speech analysis," *J. Alzheimer's Dis.*, vol. 69, no. 4, pp. 1183–1193, 2019.

[35] C. Pérez, Y. Campos-Roca, L. Naranjo, and J. Martín, "Diagnosis and tracking of parkinson's disease by using automatically extracted acoustic features," *J. Alzheimer's Dis. Parkinsonism*, vol. 6, no. 5, Sep. 2016, ISSN: 2161-0460. DOI: 10.4172/2161-0460.1000260.

[36] R. Singh, *Profiling Humans from their Voice*, 1st ed. 2019. Singapore: Springer Singapore, 2019, ISBN: 981-13-8403-7.

[37] C. Henley, "External brain anatomy," in *Foundations of Neuroscience*. Michigan State University, 2021.

[38] *Larynx*, https://www.getbodysmart.com/larynx.

[39] M. M. Sondhi, Y. Huang, and J. Benesty, *Springer handbook of speech processing*, ser. Springer Handbooks. Springer, 2007, ISBN: 3540491287.

[40] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

[41] F. Eyben, *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*. Cham: Springer International Publishing AG, 2016, ISBN: 3319272985; 9783319272986;

[42] *Vocaltract*, https://commons.wikimedia.org/wiki/File:VocalTract.svg.

[43] J. Koreman, "The effects of stress and f0 on the voice source," *Phonus*, vol. 1, pp. 105–120, 1995.

[44] A. E. Olszewski, L. Shen, and J. J. Jiang, "Objective methods of sample selection in acoustic analysis of voice," *Ann. Otol., Rhinol., Laryngol.*, vol. 120, no. 3, pp. 155–161, Mar. 2011, ISSN: 0003-4894. DOI: 10.1177/000348941112000303.

[45] W. Jianglin, J. An, and M. T. Johnson, "Features for phoneme independent speaker identification," in *Int. Conf. Audio, Lang., Image Process.*, IEEE, Jul. 2012, pp. 1141–1145, ISBN: 1467301736. DOI: 10.1109/ICALIP.2012.6376788.

[46] J. P. Teixeira and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Comput. Sci.*, vol. 100, pp. 271–279, 2016, ISSN: 1877-0509. DOI: DOI:10.1016/j.procs.2016.09.155.

[47] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and hnr parameters," *Procedia Technol.*, vol. 9, pp. 1112–1122, 2013, ISSN: 2212-0173.

[48] T. Haji, S. Horiguchi, T. Baer, and W. J. Gould, "Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation," *J. Acoust. Soc. Am.*, vol. 80, no. 1, pp. 58–62, 1986, ISSN: 1520-8524. DOI: 10.1121/1.394083.

[49] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–9, 2009, ISSN: 1687-6180.

[50] K. Waghmare, S. Kayte, and B. Gawali, "Analysis of pitch and duration in speech synthesis using psola," *Commun. Appl. Electron.*, vol. 4, no. 4, pp. 10–18, Feb. 2016, ISSN: 2394-4714.

[51] I. R. Titze and H. Liang, "Comparison of fo extraction methods for high-precision voice perturbation measurements," *J. Speech, Lang., Hear. Res.*, vol. 36, no. 6, pp. 1120–1133, Dec. 1993.

[52] Y. Maryn, P. Corthals, M. D. Bodt, P. V. Cauwenberge, and D. Deliyski, "Perturbation measures of voice: A comparative study between multi-dimensional voice program and praat," *Folia phoniatrica et logopaedica*, vol. 61, no. 4, pp. 217–226, Sep. 2009, ISSN: 1021-7762. DOI: 10.1159/000227999.

[53] P. Boersma, "Should jitter be measured by peak picking or by waveform matching?" *Folia Phoniatr. logopaedica*, vol. 61, no. 5, pp. 305–308, Nov. 2009, ISSN: 1021-7762. DOI: 10.1159/000245159.

[54] K. Daoudi and A. J. Kumar, "Pitch-based speech perturbation measures using a novel gci detection algorithm: Application to pathological voice classification," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sep. 2015.

[55] Y. Chien, M. Borsky, and J. Guðnason, "F0 variability measures based on glottal closure instants," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*

(Interspeech), Sep. 2019, pp. 1986–1989. DOI: 10.21437/Interspeech.2019-1326.

[56] J. Pérez and A. Bonafonte, "Automatic voice-source parameterization of natural speech," in *9th Eur. Conf. Speech Commun. Technol.*, 2005.

[57] T. Ewender and B. Pfister, "Accurate pitch marking for prosodic modification of speech segments," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Citeseer, Sep. 2010, pp. 178–181.

[58] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007. DOI: 10.1109/TASL.2006.876878.

[59] W. Pan, J. Flint, L. Shenhav, *et al.*, "Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders," *PloS one*, vol. 14, no. 6, e0218172–e0218172, 2019, ISSN: 1932-6203.

[60] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000, ISSN: 0018-9294. DOI: 10.1109/10.846676.

[61] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proc. 10th Int. Conf. Digital Audio Eff.*, Bordeaux, vol. 237, Sep. 2007, DAFX–1–DAFX–8.

[62] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, Jan. 2020, ISSN: 0003-682X.

[63] A. Lerch, *An introduction to audio content analysis : applications in signal processing and music informatics*. Hoboken, New Jersey: Wiley, 2012, ISBN: 1-283-80405-0.

[64] A. Tursunov, S. Kwon, and H.-S. Pang, "Discriminating emotions in the valence dimension from speech using timbre features," *Appl. Sci.*, vol. 9, no. 12, Jun. 2019, ISSN: 2076-3417.

[65] M. Sonn, *Psychoacoustical terminology*. Raytheon Company, Submarine Signal Division, 1969.

[66] T. Giannakopoulos, *Introduction to audio analysis : a MATLAB approach*, First edition. Oxford: Academic Press, 2014, ISBN: 0-08-099389-3.

[67] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7 : multimedia content description interface*. Berlin: Wiley, 2002, ISBN: 0471486787.

[68] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *J Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1252–1263, 2000.

[69] A. Y. Hussenbocus, M. Lech, and N. B. Allen, "Statistical differences in speech acoustics of major depressed and non-depressed adolescents," in *9th Int. Conf. Signal Process. and Commun. Syst. (ICSPCS)*, IEEE, Dec. 2015, pp. 1–7. DOI: 10.1109/ICSPCS.2015.7391781.

[70] G. Chenghui, Z. Heming, T. Zhi, Y. Zongyue, and G. Xiaojiang, "Feature analysis on emotional chinese whispered speech," in *Int. Conf. Inf., Networking, Autom. (ICINA)*, vol. 2, 2010, pp. V2–137–V2–141. DOI: 10.1109/ICINA.2010.5636965.

[71] B. P. Bogert, M. J. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and

saphe cracking," in *Proc. Symp. on Time Series Analysis*, vol. 15, New York: John Wiley and Sons, Inc., Jun. 1963, pp. 209–243.

[72] A. Oppenheim and R. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 95–106, 2004. DOI: 10.1109/MSP.2004.1328092.

[73] J. Williamson, T. Quatieri, B. Helfer, R. Horwitz, B. Yu, and D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proc. 3rd Int. Workshop Audio/Vis. Emotion Challenge*, ser. AVEC '13, ACM, Oct. 2013, pp. 41–48, ISBN: 9781450323956. DOI: 10.1145/2512530.2512531.

[74] M. Morvidone, B. L. Sturm, and L. Daudet, "Incorporating scale information with cepstral features: Experiments on musical instrument recognition," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1489–1497, 2010, ISSN: 0167-8655.

[75] T. B́ackstŕom, *Cepstrum and mfcc*, https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC, 2019.

[76] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, Dec. 2017, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2017.08.015.

[77] A. Dev and P. Bansal, "Robust features for noisy speech recognition using mfcc computation from magnitude spectrum of higher order autocorrelation coefficients," *Int. J. Comput. Appl.*, vol. 10, no. 8, pp. 36–38, 2010.

[78] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1998, pp. 617–620. DOI: 10.1109/ICASSP.1998.675340.

[79] "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986. DOI: 10.1109/TASSP.1986.1164788.

[80] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, 2017, ISSN: 0957-4174.

[81] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114 591, 2021, ISSN: 0957-4174.

[82] L. Baghai-Ravary, "Speech production and perception," in *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*, 1st ed. 2013. New York, NY: Springer New York, 2013, ch. 2, pp. 7–11, ISBN: 1-283-62432-X. DOI: 10.1007/978-1-4614-4574-6_2.

[83] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *J. Nonverbal Behav.*, vol. 20, no. 2, pp. 83–110, 1996, ISSN: 1573-3653.

[84] A. P. Simpson, "Phonetic differences between male and female speech," *Lang. Linguist. compass*, vol. 3, no. 2, pp. 621–640, Mar. 2009, ISSN: 1749-818X. DOI: 10.1111/j.1749-818X.2009.00125.x.

[85] M. Hasan and T. Shimamura, "An efficient pitch estimation method using windowless and normalized autocorrelation functions in noisy

environments," *Int. J. Circuits Syst. Signal Process.*, vol. 6, no. 3, pp. 197–204, 2012.

[86] Q. Wang, X. Zhao, and J. Xu, "Pitch detection algorithm based on normalized correlation function and central bias function," in *10th Int. Conf. Commun. Networking China (ChinaCom)*, 2015, pp. 617–620. DOI: 10.1109/CHINACOM.2015.7498011.

[87] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *J. Acoust. Soc. Am*, vol. 126, no. 4, pp. 2061–2071, 2009. DOI: 10.1121/1.3203668.

[88] D. C. Giancoli, "Sound," in *Physics: principles with applications*. Pearson Education, 2016, ch. 12, pp. 328–358, ISBN: 9781292057125.

[89] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Bell Syst. Tech. J.*, vol. 12, no. 4, pp. 377–430, 1933.

[90] S. C. on Emerging and N. I. H. Risks, *Potential health risks of exposure to noise from personal music players and mobile phones including a music playing function: Preliminary report*, 2008.

[91] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, 2020, ISSN: 1381-2416.

[92] P. Boersma and D Weenink, *Praat: Doing phonetics by computer*, https://www.fon.hum.uva.nl/praat/, 2019.

[93] T. ´Ozseven and M. Düğenci, "Speech acoustic (spac): A novel tool for speech feature extraction and classification," *Appl. Acoust.*, vol. 136, pp. 1–8, Jul. 2018, ISSN: 0003-682X. DOI: 10.1016/j.apacoust.2018.02.009.

[94] M. Stanek and L. Polak, "Algorithms for vowel recognition in fluent speech based on formant positions," in *36th Int. Conf. Telecommun. Signal Process. (TSP)*, 2013, pp. 521–525. DOI: 10.1109/TSP.2013.6613987.

[95] X. E. Sun and J. Y. Luo, "Matlab-based formant estimation," *Appl. Mech. Mater.*, vol. 577, pp. 798–801, Jul. 2014, ISSN: 1660-9336. DOI: 10.4028/www.scientific.net/AMM.577.798.

[96] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *J. Acoust. Soc. Am*, vol. 145, no. 2, pp. 642–653, Jan. 2019, ISSN: 0001-4966. DOI: 10.1121/1.5088048.

[97] F. Torres, *What is depression?* https://www.psychiatry.org/patients-families/depression/what-is-depression, 2020.

[98] *Depressive disorders*, ser. DSM-5 selections. Arlington, Virginia: American Psychiatric Publishing, 2016, ISBN: 1-61537-037-4.

[99] S. D Østergaard, S. O. W. Jensen, and P. Bech, "The heterogeneity of the depressive syndrome: When numbers get serious," *Acta psychiatrica Scandinavica*, vol. 124, no. 6, pp. 495–496, 2011, ISSN: 0001-690X.

[100] K. Singer, "Depressive disorders from a transcultural perspective," *Soc. Sci. Med.*, vol. 9, no. 6, pp. 289–301, 1975, ISSN: 0037-7856. DOI: https://doi.org/10.1016/0037-7856(75)90001-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0037785675900013.

[101] R. Freedman, D. A. Lewis, R. Michels, *et al.*, "The initial field trials of dsm-5: New blooms and old thorns," *Am. J. Psychiatry*, vol. 170, no. 1, pp. 1–5, 2013, ISSN: 0002-953X.

[102] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *J. Affective Disord.*, vol. 114, no. 1, pp. 163–173, 2008, ISSN: 0165-0327.

[103] M. Hamilton, "A rating scale for depression," *J. Neurol., Neurosurg. Psychiatry*, vol. 23, no. 1, pp. 56–62, Feb. 1960, ISSN: 0022-3050. DOI: 10.1136/jnnp.23.1.56.

[104] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, *et al.*, "The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): A psychometric evaluation in patients with chronic major depression," *Biol. Psychiatry*, vol. 54, no. 5, pp. 573–583, 2003, ISSN: 0006-3223. DOI: 10.1016/S0006-3223(02)01866-8.

[105] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," *J. Pers. Assess.*, vol. 67, no. 3, pp. 588–597, 1996, ISSN: 0022-3891. DOI: 10.1207/s15327752jpa6703_13.

[106] S. L. Ruyak, N. K. Lowe, E. J. Corwin, M. Neu, and B. Boursaw, "Prepregnancy obesity and a biobehavioral predictive model for postpartum depression," *J. Obstet., Gynecol., Neonat. Nurs.*, vol. 45, no. 3, p. 326, 2016.

[107] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. A., and D. P. Subha, "Automated eeg-based screening of depression using deep convolutional neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, 2018, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2018.04.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260718301494.

[108] J. K. Darby and H. Hollien, "Vocal and speech patterns of depressive patients," *Folia Phoniatr. et Logopaedica*, vol. 29, no. 4, pp. 279–291, 1977.

[109] E. Kraepelin, "Manic depressive insanity and paranoia," *J. Nerv. Ment. Dis.*, vol. 53, no. 4, p. 350, 1921.

[110] F. H́onig, A. Batliner, E. Ńoth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: Relevant features and relevance of gender," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Singapore: International Speech Communication Association (ISCA), Sep. 2014, pp. 1248–1252. DOI: 10.21437/Interspeech.2014-313.

[111] G. Kiss and K. Vicsi, "Mono- and multi-lingual depression prediction based on speech processing," *Int. J. Speech Technol.*, vol. 20, no. 4, pp. 919–935, Dec. 2017, ISSN: 1381-2416. DOI: 10.1007/s10772-017-9455-8.

[112] Y. Jia, Y. Liang, and T. Zhu, "An analysis of voice quality of chinese patients with depression," in *22nd Conf. Orient. COCOSDA Int. Comm. Co-ord. Stand. Speech Databases Assess. Tech. (O-COCOSDA)*, 2019, pp. 1–6. DOI: 10.1109/O-COCOSDA46868.2019.9060848.

[113] W. J. Silva, L. Lopes, M. K. C. G., and A. A. Almeida, "Voice acoustic parameters as predictors of depression," *J. Voice*, 2021, ISSN: 0892–1997. DOI: https://doi.org/10.1016/j.jvoice.2021.06.018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199721002058.

[114] M. C. Franca, "Acoustic comparison of vowel sounds among adult females," *J. Voice*, vol. 26, no. 5, 671.e9–671.e17, Sep. 2012, ISSN: 0892-1997. DOI: 10.1016/j.jvoice.2011.11.010.

[115] T. Yingthawornsuk, H. K. Keskinpala, D. France, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, "Objective estimation of suicidal risk using vocal output characteristics," in *Ninth International Conference on Spoken Language Processing*, 2006.

[116] N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France, "Evaluation of voice acoustics as predictors of clinical depression scores," *J. Voice*, vol. 31, no. 2, 256.e1–256.e6, 2017, ISSN: 0892-1997.

[117] B. Schuller, A. Batliner, D. Seppi, *et al.*, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2007, pp. 2253–2256.

[118] M. A. L.-S. Low, C. Namunu, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Jul. 2011, ISSN: 0018-9294. DOI: 10.1111/ecc.13033.

[119] M. Yağanoğlu, "Real time wearable speech recognition system for deaf persons," *Comput., Electr. Eng.*, vol. 91, p. 107 026, 2021.

[120] T. Taguchi, H. Tachikawa, K. Nemoto, *et al.*, "Major depressive disorder discrimination using vocal acoustic features," *J. Affective Disord.*, vol. 225, pp. 214–220, 2018, ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2017.08.038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032717300344.

[121] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed people: A cross-situation study," *BMC Psychiatry*, vol. 19, no. 1, pp. 300–300, Oct. 2019, ISSN: 1471-244X. DOI: 10.1186/s12888-019-2300-7.

[122] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychol. Bull.*, vol. 99, no. 2, pp. 143–165, 1986, ISSN: 0033-2909.

[123] H. Hollien, "Vocal indicators of psychological stress," *Ann. N. Y. Acad. Sci.*, vol. 347, no. 1, pp. 47–72, Jun. 1980, ISSN: 0077-8923. DOI: 10.1111/j.1749-6632.1980.tb21255.x.

[124] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *J. Psychiatr. Res.*, vol. 27, no. 3, pp. 309–319, 1993, ISSN: 0022-3956. DOI: 10.1016/0022-3956(93)90041-Y.

[125] K. Vicsi, D. Sztaho, and G. Kiss, "Examination of the sensitivity of acoustic-phonetic parameters of speech to depression," in *IEEE 3rd Int. Conf. Cognitive Infocommunications (CogInfoCom)*, IEEE, 2012, pp. 511–515, ISBN: 1467351873. DOI: 10.1109/CogInfoCom.2012.6422035.

[126] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An investigation of emotional speech in depression classification.," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2016, pp. 485–489.

[127] M. Brookes, *Voicebox: Speech processing toolbox for matlab*, www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[128] S. Scherer, G. Stratou, J. Gratch, and L. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), Aug. 2013, pp. 847–851. DOI: 10.21437/Interspeech.2013-240.

[129] J. Gratch, R. Artstein, G. M. Lucas, *et al.*, "The distress analysis interview corpus of human and computer interviews.," in *LREC*, May 2014, pp. 3123–3128.

[130] F. Eyben, M. W′ollmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, ser. MM '10, ACM, Oct. 2010, pp. 1459–1462, ISBN: 1605589330. DOI: 10.1145/1873951.1874246.

[131] S. Alghowinem, R. Goecke, M. Wagner, *et al.*, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, May 2013, pp. 8022–8026, ISBN: 1479903566. DOI: 10.1109/ICASSP.2013.6639227.

[132] M. Valstar, J. Gratch, B. Schuller, *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 3–10.

[133] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, 2014, pp. 960–964, ISBN: 9781479928934.

[134] H. Long, Z. Guo, X. Wu, B. Hu, Z. Liu, and H. Cai, "Detecting depression in speech: Comparison and combination between different speech types," in *IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, IEEE, 2017, pp. 1052–1058. DOI: 10.1109/BIBM.2017.8217802.

[135] E. W. McGinnis, S. P. Anderau, J. Hruschak, *et al.*, "Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood," *IEEE J. Biomed. Health Inf.*, vol. 23, no. 6, pp. 2294–2301, 2019, ISSN: 2168-2194.

[136] C. W. Espinola, J. C. Gomes, J. M. Pereira, and W. P. dos Santos, "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study," *Res. Biomed. Eng.*, vol. 37, no. 1, pp. 53–64, 2020, ISSN: 2446-4732.

[137] A. Saidi, S. Othman, and S. Saoud, "Hybrid cnn-svm classifier for efficient depression detection system," in *4th Int. Conf. Adv. Syst. Emergent Technolog.*, 2020, pp. 229–234. DOI: 10.1109/IC_ASET49463.2020.9318302.

[138] V. Aharonson, A. Nooy, S. Bulkin, and G. Sessel, "Automated classification of depression severity using speech - a comparison of two machine learning architectures," in *IEEE Int. Conf. Healthcare Inf. (ICHI)*, 2020, pp. 1–4. DOI: 10.1109/ICHI48887.2020.9374335.

[139] S. Lee, S. W. Suh, T. Kim, *et al.*, "Screening major depressive disorder using vocal acoustic features in the elderly by sex," *J. Affective Disord.*, vol. 291, pp. 15–23, 2021, ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2021.04.098. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032721004304.

[140] M. Patil and V. Wadhai, "Selection of classifiers for depression detection using acoustic features," in *2021 Int. Conf. Comput. Intell. Comput. Appl.*, 2021, pp. 1–4. DOI: 10.1109/ICCICA52458.2021.9697240.

[141] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomed. Signal Process. Control*, vol. 71, p. 103 107, 2022, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103107. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421007047.

[142] E. A. Stepanov, S. Lathuiliére, S. A. Chowdhury, *et al.*, "Depression severity estimation from multiple modalities," in *Proc. 20th IEEE Int. Conf. e-Health Netw., Appl. Serv. (Healthcom)*, IEEE, Sep. 2018, pp. 1–6. DOI: 10.1109/HealthCom.2018.8531119.

[143] S. H. Choi and C.-H. Choi, "The effect of gender and speech task on cepstral-and spectral-measures of korean normal speakers," *Audiol. Speech Res.*, vol. 12, no. 3, pp. 157–163, 2016. DOI: http://doi.org/10.21848/asr.2016.12.3.157.

[144] R. I. Zraick, S. D. Skaggs, and J. C. Montague, "The effect of task on determination of habitual pitch," *J. Voice*, vol. 14, no. 4, pp. 484–489, Dec. 2000, ISSN: 0892-1997. DOI: 10.1016/S0892-1997(00)80005-3.

[145] A. P. Vogel and A. T. Morgan, "Factors affecting the quality of sound recording for speech and voice analysis," *Int. J. Speech Lang. Pathol.*, vol. 11, no. 6, pp. 431–437, Oct. 2009, ISSN: 1754-9507. DOI: 10.3109/17549500902822189.

[146] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Commun.*, vol. 121, pp. 16–28, Aug. 2020, ISSN: 0167-6393. DOI: 10.1016/j.specom.2020.03.008.

[147] J. Weiner, M. Angrick, S. Umesh, and T. Schultz, "Investigating the effect of audio duration on dementia detection using acoustic features," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), Sep. 2018, pp. 2324–2328. DOI: 10.21437/Interspeech.2018-57.

[148] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "A deep neural network based end to end model for joint height and age estimation from short duration speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, May 2019, pp. 6580–6584. DOI: 10.1109/ICASSP.2019.8683397.

[149] M. Sigmund, "Gender distinction using short segments of speech signal," *Int. J. Comput. Sci. Network Secur. (IJCSNS)*, vol. 8, no. 10, pp. 159–162, Oct. 2008.

[150] R. C. Scherer, V. J. Vail, and C. G. Guo, "Required number of tokens to determine representative voice perturbation values," *J. Speech, Lang., Hear. Res.*, vol. 38, no. 6, pp. 1260–1269, Dec. 1995, ISSN: 0022-4685.

[151] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *CoRR*, vol. abs/1706.00612, Jun. 2017. arXiv: 1706.00612.

[152] C. Draxler, F. Schiel, and T. Ellbogen, "F0 of adolescent speakers-first results for the german ph@ ttsessionz database," in *Proc. 6th Int. Conf. Lang. Resour. Eval. (LREC'08)*, May 2008, pp. 2275–2278.

[153] R. I. Zraick, K. Y. Birdwell, and L. Smith-Olinde, "The effect of speaking sample duration on determination of habitual pitch," *J. Voice*, vol. 19, no. 2, pp. 197–201, Jun. 2005, ISSN: 0892-1997. DOI: 10.1016/j.jvoice.2004.01.010.

[154] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *15th Annu. Conf. Int. Speech Commun. Assoc. (interspeech)*, International Speech Communication Association (ISCA), 2014.

[155] D. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *J. Acoust. Soc. Am.*, vol. 118, no. 5, p. 3177, Nov. 2005.

[156] C. Dong and K. Jingming, "A robust voice activity detector applied for amr," in *Proc. 5th Int. Conf. Signal Process. 16th World Comput. Congr. (WCC- ICSP)*, vol. 2, IEEE, Aug. 2000, pp. 687–692.

[157] J. A. Morales-Cordovilla, N. Ma, V. Sánchez, J. L. Carmona, A. M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 4808–4811. DOI: 10.1109/ICASSP.2011.5947431.

[158] B. Schuller, B. Vlasenko, F. Eyben, *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affective Comput.*, vol. 1, no. 2, pp. 119–131, Jul. 2010, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2010.8.

[159] I. McLoughlin, *Applied Speech and Audio Processing: With Matlab Examples*. Cambridge: Cambridge University Press, 2009, ISBN: 0521519543; 9780521519540. DOI: 10.1017/CBO9780511609640.

[160] K. S. Rao, *Speech Recognition Using Articulatory and Excitation Source Features*, 1st ed. 2017. Cham: Springer International Publishing, 2017, ISBN: 3-319-49220-9.

[161] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, 2. Prentice-Hall, Englewood Cliffs. New Jersey: Elsevier Ltd, 1980, vol. 12, ISBN: 0031-3203. DOI: 10.1016/0031-3203(80)90010-2.

[162] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Commun.*, vol. 53, no. 9, pp. 1198–1209, 2011, ISSN: 0167-6393. DOI: 10.1016/j.specom.2011.05.002.

[163] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, 2013, pp. 7547–7551.

[164] I. The MathWorks, *Audio toolbox*, https://au.mathworks.com/help/audio/index.html, U. S., 2021.

[165] J. P. Teixeira and P. O. Fernandes, "Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices," *Procedia Technol.*, vol. 16, pp. 1228–1237, 2014. DOI: 10.1016/j.protcy.2014.10.138.

[166] G. S. Ohm, "Noch ein paar worte über die definition des tones," *Annalen der Physik*, vol. 138, no. 5, pp. 1–18, 1844.

[167] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *Tech. Rep.; IRCAM*, vol. 54, no. 0, pp. 1–25, 2004.

[168] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust asr," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, IEEE, 2004, pp. I–193.

[169] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 314–323, 1988.

[170] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, IEEE, 1997, pp. 1331–1334.

[171] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Am.*, vol. 52, no. 6B, pp. 1687–1697, 1972, ISSN: 0001-4966.

[172] C. D. Looze, A. Ghio, S. Scherer, G. Pouchoulin, and F. Viallet, "Automatic analysis of the prosodic variations in parkinsonian read and semi-spontaneous speech," in *Speech Prosody 6th Int. Conf.*, 2012, p. 4.

[173] S. Namba and S. Kuwano, "Psychological study on leq as a measure of loudness of various kinds of noises," *J. Acoust. Soc. Jpn.*, vol. 5, no. 3, pp. 135–148, 1984.

[174] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999, ISSN: 1070-9908. DOI: 10.1109/97.736233.

[175] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, M. Breakspear, *et al.*, "Characterising depressed speech for classification," in *14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Lyon, France: International Speech Communication Association (ISCA), Aug. 2013, pp. 2534–2538.

[176] S. Arash, *Intraclass correlation coefficient (icc)*, https://au.mathworks.com/, U. S., 2016.

[177] M. Cearns, N. Opel, S. Clark, *et al.*, "Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: A multimodal machine learning approach," *Transl. Psychiatry*, vol. 9, no. 1, pp. 285–9, Nov. 2019, ISSN: 2158-3188. DOI: 10.1038/s41398-019-0615-2.

[178] Y. Nishinuma, A. Di Cristo, and R. Espesser, "How does vowel duration affect loudness in a cv syllable?" *Speech Communi.*, vol. 3, no. 1, pp. 39–47, 1984, ISSN: 0167-6393. DOI: 10.1016/0167-6393(84)90007-4.

[179] C. Shih, B. Möbius, and B. Narasimhan, "Contextual effects on consonant voicing profiles: A cross–linguistic study," in *Proc. 14th Int. Congr. Phonetic Sci. (ICPhS99)*, vol. 2, Aug. 1999, pp. 989–992.

[180] B. Möbius, "Corpus-based investigations on the phonetics of consonant voicing," *Societas Linguistica Europaea*, vol. 38, no. 1-2, pp. 5–26, 2004. DOI: 10.1515/flin.2004.38.1-2.5.

[181] C. Gendrot and M. Adda-Decker, "Impact of duration on f1/f2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in french and german," in *Proc. 9th Eur. Conf. Speech Commun.*

Technol. (Interspeech), Lisbon, Portugal: International Speech Communication Association (ISCA), Sep. 2005, pp. 2453–2456.

[182] S. Y. Lowell and J. A. Hylkema, "The effect of speaking context on spectral- and cepstral-based acoustic features of normal voice," *Clin. Linguist. phonetics*, vol. 30, no. 1, pp. 1–11, Jan. 2016, ISSN: 0269-9206. DOI: 10.3109/02699206.2015.1087049.

[183] S. N. Awan, A. Giovinco, and J. Owens, "Effects of vocal intensity and vowel type on cepstral analysis of voice," *J. Voice*, vol. 26, no. 5, 670.e15–670.e20, Sep. 2012, ISSN: 0892-1997. DOI: 10.1016/j.jvoice.2011.12.001.

[184] S.-H. Choi and C.-H. Choi, "The stability and variability based on vowels in voice quality analysis," *Phonetics Speech Sci.*, vol. 7, no. 1, pp. 79–86, Mar. 2015.

[185] M. J. Sandage, L. W. Plexico, and A. Schiwitz, "Clinical utility of cape-v sentences for determination of speaking fundamental frequency," *J. Voice*, vol. 29, no. 4, pp. 441–445, Jul. 2015, ISSN: 0892-1997. DOI: 10.1016/j.jvoice.2014.09.005.

[186] A. Romana, J. Bandon, M. Perez, *et al.*, "Automatically detecting errors and disfluencies in read speech to predict cognitive impairment in people with parkinson's disease," in *Proc. 22nd Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), 2021, pp. 1907–1911. DOI: 10.21437/Interspeech.2021-1694.

[187] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biol. Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012, ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2012.03.015.

[188] S. Furui, M Nakamura, T. Ichiba, and K. Iwano, "Analysis and recognition of spontaneous speech using corpus of spontaneous japanese," *Speech Commun.*, vol. 47, no. 1, pp. 208–219, 2005, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2005.02.010.

[189] G. P. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Commun.*, vol. 22, no. 1, pp. 43–65, 1997, ISSN: 0167-6393.

[190] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Comput. Speech Lang.*, vol. 22, no. 2, pp. 171–184, 2008, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2007.07.003.

[191] A. I. Hudson and A. Holbrook, "Fundamental frequency characteristics of young black adults: Spontaneous speaking and oral reading," *J. Speech, Lang., and Hear. Res.*, vol. 25, no. 1, pp. 25–28, 1982.

[192] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, May 2015, ISSN: 0167-6393. DOI: 10.1016/j.specom.2015.03.004.

[193] H. Long, X. Wu, Z. Guo, J. Liu, and B. Hu, "Detecting depression in speech: A multi-classifier system with ensemble pruning on kappa-error diagram," *J. Health Med. Inf.*, vol. 8, no. 5, pp. 1–8, 2017, ISSN: 2157-7420. DOI: 10.4172/2157-7420.1000293.

[194] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Sep. 2018, pp. 1676–1680. DOI: 10.21437/Interspeech.2018-1399.

[195] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, IEEE, 2020, pp. 6549–6553, ISBN: 9781509066315.

[196] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.

[197] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol: O'Reilly Media, Incorporated, 2019, ISBN: 9781492032649.

[198] M. H. Sanchez, D. Vergyri, L. Ferrer, *et al.*, "Using prosodic and spectral features in detecting depression in elderly males," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, International Speech Communication Association (ISCA), Aug. 2011, pp. 3001–3004.

[199] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers," Sep. 2016, pp. 1943–1947. DOI: 10.21437/Interspeech.2016-1339.

[200] M. Masaeli, G. Fung, and J. G. Dy, "From transformation-based dimensionality reduction to feature selection," in *ICML*, 2010.

[201] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[202] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in *IEEE Int. Conf. Comput. Intell. Comput. Res.*, 2014, pp. 1–6. DOI: 10.1109/ICCIC.2014.7238499.

[203] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1, pp. 37–52, 1987, ISSN: 0169-7439.

[204] B. Sch´olkopf, A. J. Smola, and B. Sch´olkopf, ser. Adaptive computation and machine learning series. Cambridge, Massachusetts: MIT Press, 2001, ISBN: 0-262-25693-2.

[205] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology*, vol. 2, no. 3, pp. 1–27, 2011, ISSN: 2157-6904.

[206] scikit learn, *Rbf svm parameters*, https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html.

[207] N. Chinchor, "Muc-4 evaluation metrics," in *Proc. 4th Conf. Message Understanding*, ser. MUC4 '92, McLean, Virginia: Association for Computational Linguistics, 1992, 22–29, ISBN: 1558602739.

[208] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *J. Affective Disord.*, vol. 66, no. 1, pp. 59–69, 2001, ISSN: 0165-0327. DOI: https://doi.org/10.1016/S0165-0327(00)00335-9.

[209] A. Chen, *Amazon's alexa now handles patient health information*, https://www.theverge.com/2019/4/4/18295260/amazon-hipaa-alexa-echo-patient-health-information-privacy-voice-assistant, 2019.