



**Productivity Measurement of Call Centre Agents
using
A Multimodal Classification Approach**

Thesis

Author:

Abdelrahman M. AHMED

Supervisors:

Sergio Marin Toral

Khaled Shaalan

Submitted for The Degree Of:

Doctorate of Philosophy

Faculty of Electronic Engineering, University of Seville

2021

Abstract

Call centre channels play a cornerstone role in business communications and transactions, especially in challenging business situations. Operations' efficiency, service quality, and resource productivity are core aspects of call centres' competitive advantage in rapid market competition. Performance evaluation in call centres is challenging due to human subjective evaluation, manual assortment to massive calls, and inequality in evaluations because of different raters. These challenges impact these operations' efficiency and lead to frustrated customers. This study aims to automate performance evaluation in call centres using various deep learning approaches. Calls recorded in a call centre are modelled and classified into high- or low-performance evaluations categorised as productive or nonproductive calls.

The proposed conceptual model considers a deep learning network approach to model the recorded calls as text and speech. It is based on the following: 1) focus on the technical part of agent performance, 2) objective evaluation of the corpus, 3) extension of features for both text and speech, and 4) combination of the best accuracy from text and speech data using a multimodal structure. Accordingly, the diarisation algorithm extracts that part of the call where the agent is talking from which the customer is doing so. Manual annotation is also necessary to divide the modelling corpus into productive and nonproductive (supervised training). Krippendorff's alpha was applied to avoid subjectivity in the manual annotation. Arabic speech recognition is then developed to transcribe the speech into text. The text features are the words embedded using the embedding layer. The speech features make several attempts to use the Mel Frequency Cepstral Coefficient (MFCC) upgraded with Low-Level Descriptors (LLD) to improve classification accuracy. The data modelling architectures for speech and text are based on CNNs, BiLSTMs, and the attention layer. The multimodal approach follows the generated models to improve performance accuracy by concatenating the text and speech models using the joint representation methodology.

The main contributions of this thesis are:

- Developing an Arabic Speech recognition method for automatic transcription of speech into text.
- Drawing several DNN architectures to improve performance evaluation using speech features based on MFCC and LLD.

- Developing a Max Weight Similarity (MWS) function to outperform the SoftMax function used in the attention layer.
- Proposing a multimodal approach for combining the text and speech models for best performance evaluation.

The experiment goes through four stages: Data preparation, feature extraction, data modelling, and classification. The experiment was conducted on 7 hours of recorded calls from a real estate call centre in Egypt. The calls have been diarised to segregate the segments during which agents are talking from those in which customers or third parties are doing so. The data have been annotated manually and verified using Krippendorff's alpha for three raters, with 79.1% agreement among them. The text has been transcribed using lexicon-free Arabic speech recognition. The speech recognition acoustic model was trained using a 1200h Aljazeera corpus, and the corresponding language model was collected from the corpus and online web crawling. The speech transcription system achieved a 12% WER (Word Error Rate), which is outstanding compared to previous studies. The cascaded CNN-attention approach achieved the best classification accuracy in productivity measurement. This study enhanced the attention layer using the Max Weight Similarity (MWS) function instead of the SoftMax function. The experiment achieved accuracies of 91.4% for textual data, 92.88% for speech data, and 93.1% for the multimodal combination of text and speech. Findings also reveal some paralinguistic features associated with productive and nonproductive features, like Stuttering 'Umm Ahh' as a nonproductive feature and the tone level as a productive one. The experiment proves that productivity can be automatically detected and classified under each model type. The proposed approach is proven to outperform previous studies.

Acknowledgement

First of all, I would like to express my sincere gratitude to the University of Seville (Spain) for providing me with a learning opportunity in joining the doctorate program.

Further, I am grateful to my Supervisor, Professor Sergio Toral Marin, for his outstanding guidance, ultimate support, and highly appreciated efforts over five years to conduct the research project smoothly. His thoughtful comments and recommendations on this dissertation contributed to the core directions of the study. I am also grateful to my supervisor, Prof. Khaled Shaalan, for his guidance, comments, and motivation. Additionally, a special thanks to Dr Yasser Hifny for his help with machine learning techniques and resources.

To conclude, I cannot forget to thank my wife, Mona, for her continuous encouragement and motivation; and my dearest sister, Dr Hala, for her help and support.

Declaration

I hereby declare that this thesis represents my own work, which has been done after registration for the degree of PhD at the University of Seville and has not been previously included in a thesis or dissertation submitted to this or any other institution for any degrees or qualifications.

I have read the University's current research ethics guidelines and accept responsibility for the procedures' conduct in accordance with the University's ethics. I have attempted to identify all the risks that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the participants' rights.

Some of the material displayed herein has already been published in the form of the following publications:

Ahmed, A., Hifny, Y., Shaalan, K. and Toral, S., 2016, October. Lexicon free Arabic speech recognition recipe. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 147-159). Springer, Cham.

Ahmed, A., Toral, S. and Shaalan, K., 2016, October. Agent productivity measurement in a call centre using machine learning. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 160-169). Springer, Cham.

Ahmed, A., Hifny, Y., Toral, S. and Shaalan, K., 2018. A Call Center Agent Productivity Modelling Using Discriminative Approaches. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 501-520). Springer, Cham.

Ahmed, A., Hifny, Y., Shaalan, K. and Toral, S., 2019. End-to-end lexicon free Arabic speech recognition using recurrent neural networks. In *Computational Linguistics, Speech and Image Processing for Arabic Language* (pp. 231-248).

Ahmed, A., Toral, S., Shaalan, K. and Hifny, Y., 2020. Agent Productivity Modelling in a Call Center Domain Using Attentive Convolutional Neural Networks. *Sensors*, 20(19), p.5489.

Ahmed, A., Shaalan, K., Toral, S. and Hifny, Y., 2021. A Multimodal Approach to improve Performance Evaluation of Call Center Agent. *Sensors*, 21(8), p.2720.

Abdolrahman Ahmed Salem

Seville-SPAIN-JAN-2021


A handwritten signature in black ink, appearing to be 'Abdolrahman Ahmed Salem', written in a cursive style.

Table of Contents

1	Introduction.....	2
1.1	Background information about Call Centres.....	2
1.2	Performance Measurement Overview.....	3
1.3	Research Motivation	4
1.4	Challenges	5
1.5	Methods.....	6
1.6	Contribution	6
1.7	Thesis Outline	7
2	Background and Related Work.....	9
2.1	Introduction	9
2.2	Call Centre Technology Overview.....	9
2.3	Call Centres and Management Perceptions.....	12
2.4	Related Work.....	15
2.4.1	Generative versus Discriminative Machine Learning Approaches	16
2.4.2	Deep Learning Approaches.....	18
2.4.3	Multimodal Classification Approaches.....	21
2.5	Literature Review Gaps.....	27
2.6	Conclusion.....	30
3	Research Methodology and Framework	33
3.1	Introduction	33
3.2	Research Strategy	33
3.3	The Study Framework and Selected Methods.....	34
3.3.1	Call Diarisation	35
3.3.2	Speech-to-Text Transcription (Speech Recognition).....	36
3.3.3	Feature Extraction.....	41
3.3.4	Data Modelling and Classification.....	43
3.4	Conclusion.....	48
4	The Experiment and Findings.....	50
4.1	Introduction	50
4.2	Research Design and Plan	51
4.1	Data	53
4.2	Procedures	54
4.2.1	Stage 1: Data Preparation.....	54
4.2.2	Stage 2: Feature Extraction.....	59
4.2.3	Stage-3: Data Modelling.....	60
4.2.4	Stage-4: The Classification and Validation	62
4.3	Experimental Results.....	64
4.3.1	Speech Recognition Acoustic Modelling.....	64
4.3.2	Speech Productivity Modelling.....	65
4.3.3	Text Productivity Modelling.....	67
4.3.4	Multimodal Approach (Speech + Text)	68
4.4	The Attention Weight plot analysis.....	70
4.5	Modelling Performance.....	72
4.6	Experimental Assumptions and Considerations.....	74

4.7	Conclusion.....	75
5	Study Conclusion.....	77
5.1	Introduction.....	77
5.2	Meeting Research Aims and Objectives.....	78
5.3	Research Findings.....	81
5.3.1	Methodological Contribution.....	81
5.3.2	Practical Implications.....	84
5.4	Research Limitations.....	85
5.5	Lessons Learnt from the Study.....	86
5.6	Future Research.....	87
6	References.....	89

Tables of Figures

Figure 1: Contact Centre technology Portfolio	12
Figure 2: The single feature-based model.....	17
Figure 3: CNN structure – towardsdatascience.com.....	19
Figure 4: Subjectivity Elimination approach	21
Figure 5: Multimodal Approaches	24
Figure 6: Phones and hidden states	26
Figure 7: The Study framework.....	35
Figure 8: The RNN bidirectional layers.....	37
Figure 9: The study Framework.....	44
Figure 10: The Role of the Attention layer	45
Figure 11: Server Structure	52
Figure 12: Experiment Procedures.....	54
Figure 13: Text file combining the raters’ annotations.....	55
Figure 14: Segmentation boundary alignment between two sampling rates.	56
Figure 15: Detailed Neural Network Modelling Structure	61
Figure 16: The experiment parameters.	62
Figure 17: Model Accuracy per Approach	69
Figure 18: Attention weight graph.....	71
Figure 19: Attention weights for a sample segment	71
Figure 20: Time-Performance chart.....	73

List of Tables

Table 1: Contact Centre channels	11
Table 2: The multimodal approaches.....	24
Table 3: The performance measurement studies	30
Table 4: 65 provided Low-Level Descriptors (LLD)	42
Table 5: The Arabic letters and corresponding conversions.....	57
Table 6: MGB Challenge	65
Table 7: Speech Accuracy % per Model/Feature Type	66
Table 8: Max Weight Similarity (MWS) versus SoftMax functions	67
Table 9: Text Accuracy % per Model/Feature Type	67
Table 10: Multimodal Accuracy % per Speech and Text Models	69
Table 11: MWS vs SoftMax Accuracy Improvement	70
Table 12: Time-Performance for each Approach	73
Table 13: Study Objectives.....	78

List of Abbreviations

AHT	Average Handling Time
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
BDRNN	Bidirectional Recurrent Neural Network
BiLSTM	Bidirectional Long short-term memory
CNNs	Convolutional Neural Networks
CSR	Customer Service Representative
CTC	Connectionist Temporal Classification
DTMF	Dual Tone Multifrequency
DNN	Deep Neural Networks
DBM	Deep Boltzmann Machines
GMM	Gaussian Mixture Model
HRM	Human Resource Management
HMM	Hidden Markov Model
ICMI	International Customer Management Institute
LSTM	Long short-term memory
LSVM	Linear Support Vector Machine
LR	Logistic Regression
LLD	Low Level Descriptors
MFCC	Mel Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
MWS	Max Weight Similarity
OOV	Out of Vocabulary
RNN	Recurrent Neural Networks
SVM	Support Vector Machine
WER	Word Error Rate

Chapter 1: Introduction

1 Introduction

1.1 *Background information about Call Centres*

Call centres are the front doors where critical interactions with customers are handled. Efficient operations are crucial to an overall organisation's success, profitability, and reputation [1]. Call centres have become strategic assets of communication that encompass many potential channels and agents to enable customers to reach the information and services they need [2]. Furthermore, they are a 'cost-effective way of achieving increased service quality focus on reliability, responsiveness, assurance, tangibles, and empathy' [3].

The international customer management institute (ICMI) developed the following definition of contact centre¹ management: 'The art of having the right number of properly skilled people and supporting resources in place at the right time to handle an accurately forecasted workload, at service level with quality' [4]. This definition emphasises getting skilled people (staffing) to do the right thing (technical aspects) in an ultimate way (quality of service). Managing performance and monitoring the quality of service are defined as goal-oriented processes aiming to achieve the best performance from people, teams, resources, and the organisation as a whole. Call centres have grown over thirty decades to cover different communication channels, rather than just phone calls. Newly supported technologies like AI, speech recognition, chatbots, and business intelligence create significant challenges when evaluating overall call centre rather than individual performance [5]. Operations' performance efficiency and continuous improvement are critical factors in retaining the organisation's competitive advantage.

Measuring call centre agent performance is a fundamental and essential part of customer satisfaction and loyalty in outperforming rivals [6]. Managing performance

¹ The terms "call centre" and "contact centre" are interchangeable in the study.

and monitoring service quality is the ultimate goal of any call centre aiming for high profitability, reputation, skills retention, and customer loyalty.

1.2 **Performance Measurement Overview.**

Performance Measurement is ‘a goal-oriented process directed toward ensuring that organisational process is in place to maximise employees' productivity, teams, and the organisation’ [7]. The ultimate objective of performance management is to satisfy the customer by efficiently providing higher value than competitors [8]. Performance management in terms of quality and customer satisfaction draws on different methods and criteria to measure performance in the most accurate manner [9].

Performance measurement in call centres is performed using quantitative and qualitative methods [10]. The quantitative method considers the first call resolution, the average handling time of the call, the wrap-up time, and the adherence time² [11-13]. The qualitative method evaluates the recorded or live calls according to antecedent experience and subjective understanding [14]. For example, it judges the agent’s listening skills, communication skills, behaviour, and politeness, which differ from one evaluator to another.

Many research studies still aim to objectively evaluate call centres' overall performance using machine learning technology [15, 16]. Data mining targets a call centre's quantitative data and tries to draw a performance pattern [17]. The Paprzycki study tried to analyse the data collected quantitatively by collecting customer experience surveys and call centre Key Performance Indicators (KPIs). However, it still depends on human assessment. Carmel [18] proposed a more advanced approach to automatically analyse call contents, using speech recognition systems for conversation analysis. However, this approach lacks knowledge of the features embedded in the call

² First call resolution is the ability of the agent to resolve the customer’s need during the first call, with no need to follow up with a second call.

Average handling time (AHT) is the average talking time of the agent’s calls throughout their shift.

Wrap-up time is the time spent by an agent doing after-call work.

Adherence time is the total time that the agent spends behind the desk, ready to accept a new call. This time is calculated by excluding AHT and wrap-up time from the total working hours per shift [4] ICMI.

(2016). *ICMI | Call Center Training, Events, Certification, Resources, and Consulting*. Available: <http://www.icmi.com/>.

that determine productivity, making the analysis less useful. Ahmed, Hifny, Toral, and Shaalan [19] experimented with extracting productivity features using sentiment analysis, Naïve Bayes, logit regression, and support vector machine classifiers. The resulting accuracy ranged from 67% to 82%, which means that around 20%-30% of the recorded calls may be wrong, for many possible reasons, including incorrect annotations or incorrect classification methods to close the error gap. Human evaluation is still dominant compared to machine evaluation, which requires rigorous research to empower machine learning models to measure and predict performance. Productivity measurement automatization extends many applications, e.g. police communication radio recordings, recording analysis for airplane crash investigations, and recorded interviews. Therefore, further investigations into the evaluation process and the factors affecting it are still needed.

1.3 Research Motivation

Previous studies [18-21] examined the speech recognition output (text) to classify the performance and emotional phrases/words as either productive or nonproductive. One study evaluates agent performance using generative and discriminative approaches [19, 20]. This concept is based on transcribing the recorded calls to text and then binary classifying the output as productive or nonproductive. Other studies have been inspired by emotional recognition based on signal processing [21, 22] to classify the call centre agent's productivity. They evaluate agents' performance and the quality of service by determining the prominent emotions and speech analytics in the recorded calls.

The previous studies motivate the search for more sophisticated techniques for better classification accuracy. This study (thesis) aims to classify call centre agents' productivity using a multimodal approach that combines the speech processing and transcribed text models into one model for better classification. The study compares the productivity measurements resulting from the previous studies with those of the multimodal approach (speech-text). Furthermore, the study will enhance the two models (text/speech features) by comparing the proposed models with previous

studies. The next chapter will comprehensively explore the multimodal approaches and choose those that best fit the current study.

1.4 **Challenges**

This study aims to determine how to get a more objective measurement for performance evaluation in call centres. The advantage of the multimodal approach is that it can take the best parts of the speech and text approaches and combine them to improve classification accuracy.

The study's primary challenge is that productivity measurement differs from one person to another, so machine learning is subject to biased evaluation. In other words, the study requires the manual annotation of calls into the categories productive and nonproductive as an initial step before training the classification algorithm. The manual annotation depends on people's perceptions and antecedent experiences, so productivity measurement remains in square one. The second challenge has to do with the multimodal approach's data modelling, which depends on selecting speech signal features and text features, which are quite different from one another. Combining both models to improve performance accuracy is the core challenge in this study. There are other challenges, like speech transcription, data verification, and the reliability of the study. To overcome these challenges, the study objectives are to

- **Conduct a critical literature review** of operations efficiency and the quality of customer service in call centres, including the methods and technologies relevant to the study.
- **Build a multimodal conceptual model** for the call centre domain.
- **Demonstrate the capabilities of different machine learning classification approaches** to classify calls recorded by a call centre and their corresponding text into known classes (labels), which leads to fully automating the productivity measurement process.
- **Compare the multimodal model with previous related models** and make future research recommendations in different domains.

The following research questions guide the study objectives:

- **What are the best approaches to classify call centre agents' productivity?**
- **How can previous models be combined to improve accuracy?**
- **Can the proposed multimodal approach provide better classification performance than separate speech and text models?**

1.5 Methods

The study proposes using machine learning technology to classify agents' performance in call centres. State-of-the-art statistical models are typically based on discriminative models for the direct determination and prediction of performance. The proposed model will be based on Deep Neural Networks (DNN), using a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTMs) architectures. The data modelling is generated from the speech signal features and text features. More specifically, the data modelling for the speech signal is based on audio features, i.e. MFCC (Mel Frequency Cepstral Coefficient) features or the like.

In contrast, the text feature may depend on either a bag of words or word embedding. The study explores different combinations and tries to reach conclusions about the best accuracy. The details will be discussed in the literature review and the methodology chapters.

1.6 Contribution

The primary contribution of this study is to learn and evaluate agents' performance in call centres. Performance evaluation is a different paradigm than emotional recognition or sentiment analysis for several reasons. First, performance is related to the technical ability to answer the caller correctly. Second, productivity measurement should avoid emotional influence because it deviates from an objective evaluation, which is crucial. The study proposes several contributions to automate performance evaluation. The first is developing automatic Arabic speech recognition and transcription systems to convert recorded calls into text. The transcribed text is then

subjected to modelling into productive and nonproductive. The second contribution is speech modelling to classify agent productivity. The study attempts to investigate several DNN combinations based on MFCC and Low-Level Descriptors (LLD) speech features in order to achieve the best classification accuracy. The third contribution explores the best approaches for combining the individual models (text and speech). This remains a challenging task that provides an essential contribution for rating systems, speech recognition, forensic tracing, and text processing.

1.7 Thesis Outline

The thesis is organised into five chapters. Chapter 1 provides background about call centres and customer services, as well as the study's motivation and objectives. Chapter 2 reviews the machine learning literature within the call centre context, critically assessing and determining the research gaps in performance classification and the multimodal models. Chapter 3 creates the conceptual model and framework that the study follows to reach the proposed outcomes. Furthermore, it discusses the study methodology, the research design, and the methods used. Chapter 4 discusses the study setup and experiments according to the selected technologies and mathematical models. Chapter 5 addresses the final thesis conclusions and research contributions. It also proposes future extensions and study areas still under investigation for future research.

Chapter 2: Background & Related Work

2 Background and Related Work

2.1 Introduction

This chapter explores the literature on call centres and the new technologies applied in them during the last decades. It also discusses productivity measurement research in the call centre domain and the development of machine learning methods. The chapter should achieve the first study objective, which is to

- **Conduct a Critical literature review** of operations efficiency and the quality of customer service in call centres by determining factors and technologies relevant to the study.

The literature should also help find an answer to the research question,

- **What are the best approaches to classify call centre agents' productivity?**

Therefore, the next section gives an introduction to call centre technologies and challenges.

2.2 Call Centre Technology Overview

Call centres started in the 1950s at AT&T, the *Birmingham Press*, *Birmingham Mail*, and British Gas in Wales. They were based on legacy automatic call distributors (ACD) and public exchange switches (PBX). The peak of the call centre was in the 1990s [4] due to the evolution of computers/servers and technology convergence between IT and Telecom.

Call centre technology has developed exponentially in the last two decades, as summarised in Table 1:

#	Call Centre channels	Definition	Type of Communication
1	Telephony call centres	This is a legacy type of call centre, through a live voice call between the customer and agent.	One-to-one human interaction.
2	Interactive voice response (IVR) – Voice-enabled speech recognition	This is a self-service technology through which the caller selects choices using phone buttons (DTMF ³). Human interaction is minimal, only to seek help.	No human interaction (CSR), only phone button clicks or speech [2].
3	Voice recording system	This records a conversation between the CSR and the caller for quality monitoring and training purposes.	No human interaction [2].
4	Webchat	The agent receives a request for a chat through the company website or designated link.	A relaxed, one-to-one conversation that does not force a swift response from the agent, allowing them to search for answers in a knowledge base or something similar [2].
5	Emails/SMS	This is part of a call centre's collaboration channels; the email/GSM short message goes through a waiting queue till an agent is free to respond.	One-to-one conversation, conducted offline, does not force a swift response from the agent, allowing them to search for answers in a knowledge base or something similar [2, 4].
6	Social Media	This is a new call centre technology through which the customer comments on a social media post or ad, i.e., Facebook or Twitter; the post goes through a call centre queue for a response.	Many-to-one/many-to-many conversation. It is online and sometimes requires a quick response to inquiries or conversation threads.
7	Workforce management system	This tool is responsible for managing the agent's schedules/shifts, vacations, and resource prediction according to previous logs or calls.	A reporting or scheduling tool, which is irrelevant to customer calls or inquiries.
8	Automatic Speech Recognition	The machine takes the agent role to answer the call and respond with the appropriate action.	No human interaction[20, 22].

³ Dual Tone multifrequency (DTMF) is usually used in the call centre industry.

9	Chatbot	This is similar to web chatting, with a virtual agent that takes the real agent's place in responding to structured, predefined questions and answers.	No human interaction [23]
10	Emotional analysis and word spotting	Recorded calls are categorised to automatically report the customer's emotions or anger through signal processing and text processing (word spotting)	No direct interaction, but offline call analysis for quality assurance purposes [24].
11	Mobile application	This is the most booming and replacement to live calls as the customer inquiries are pushed directly	Human interaction can be present or automated, according to service complexity and sensitivity [25].

Table 1: Contact Centre channels

Call centre technologies are offered to match customer preferences and improve the customer experience for the appropriate channel [23, 26]. Following the terminology used throughout this study, the contact centre is a new definition of the legacy call centre, which extends calls to other communication channels. Many back-office systems link data together to grasp customer behaviour, loyalty, and product demand. Backend systems include customer relationship management (CRM), loyalty scoring systems, enterprise resource planning (ERP), etc. These systems intelligently link the data to build customers' profiles and predict their interests. Figure 1 illustrates the contact centre technology portfolio.

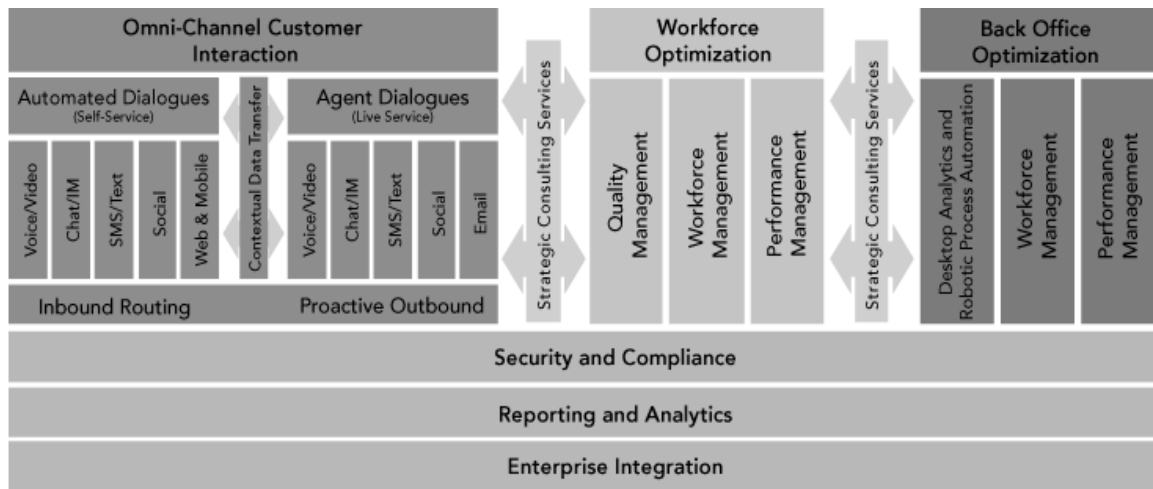


Figure 1: Contact Centre technology Portfolio⁴

However, technology development and high-performance computer processing still struggle to optimise and evaluate agents' performance and keep the customer satisfied [23]. It has been shown that 41% of customers share their customer service experience with others, while 66% of B2B⁵ and 52% of B2C⁶ customers have stopped buying after a bad customer experience [26]. Furthermore, customers sometimes confuse the quality of the product and the quality of service, which means that low quality of service may impact profitability [27].

Egypt entered the call centre industry at the beginning of 2000 due to an abundance of resources, language variety (nine languages), accent proficiency, and low operating costs [28, 29]. The key ingredient of call centres in Egypt is plentiful, well-educated resources. Each year, there are 250,000+ university graduates with high proficiency in Western languages (English, Spanish, German, and French) [28]. Call centres' operational costs in Egypt are the lowest in the region due to low worker wages and facility expenses [30]. Egypt achieved significant success in 2014; the value of such ICT exports and offshore services grew by 7%, according to the Information Technology Industry Development Agency (ITIDA) [31]. Asia (including the Gulf) was the primary destination for Egyptian ICT exports in 2013, accounting for 55% of the total value. It was followed by North and South America, with 27%; Europe, with 12%; and Africa, with 5% [32]. The current study focuses on call centres in Egypt, considering outsourcing services across different business lines.

2.3 *Call Centres and Management Perceptions*

The call centre is the road to glory for any organisation seeking to communicate and deliver business to customers. It is a routinised, restricted environment, or, as described in the literature, an 'Electronic Panopticon'. The Panopticon is a disciplinary concept where a guard in a tower monitors prison cells⁷. The phrase 'Electronic

⁴ Figure 1 was designed by aspect.com

⁵ Business to Business engagement business model.

⁶ Business to Consumer engagement business model

⁷ Wikipedia definition.

Panopticon' means that call centre agents are under the pressures of queues, a restricted environment, and comprehensive monitoring [33]. Fernie and Metcalf describe call centres as 'electronic sweatshops' due to their involving intensive activities with less autonomy than other jobs [34]. The management, including the quality assurance team, believe the stereotype that 'Call centres are neither complicated nor demanding and most of the interactions are basic, simple and scripted' [35]. On the other hand, the agents perceive their jobs as 'demanding and almost needing great attention through simultaneous subtasks,' such as listening and asking questions, operating the keyboard for data input, reading data on the screen, and answering the customer [35]. Furthermore, subjective evaluation opens the door for favouritism due to social ties [36]. It means that management may give a better evaluation to people with whom they have a close relationship, despite there being no difference in performance [36]. Different aspects of subjective evaluation have been studied from the management perspective. Abdelrahman Ahmed (2020) developed a study about the impact of subjective factors on performance evaluation. He concluded there were nine such factors: non-specific job skills, contextual performance, customer behaviour, standards contradiction, technology acceptance, channel development, stereotyping, cognitive bias, and self-serving⁸. Non-specific job skills are those that are irrelevant to the call's technical core, like communication skills, humour, listening skills, and accent fluency. Although these skills are essential and reflect on service quality, they are evaluated unequally by the quality team and are thus considered subjective. They are subtle evaluation criteria that depend on evaluator experience and self-evaluation [2]. Self-evaluation happens when the evaluator compares his/her performance with that of the agent through a social comparison process [37]. Contextual performance is when the loyalty and team spirit of the agent makes his/her evaluation stronger than those of others. The contextual performance is irrelevant to the agent's performance over the call, which leads to unfair judgment. Customer behaviour has a direct influence on subjective performance evaluation, which depends on the mood and communication methods of the customer. For example, it has been noticed that an angry customer reflects negatively on agent evaluation, even though it is irrelevant to

⁸ The Impact of Subjective Factors on Performance Evaluation: The Applied Case of Outsourced Call Centres in Egypt Based on Neural Networks Approach. Abdelrahman Ahmed, Thesis – 2020 (in Press).

both agent performance and the organisation's product. A long queuing time makes the customer frustrated, sending the message that 'your call is important to us, but your time is not' [2]. Evaluation issues could also arise due to a contradiction between call centre standards' quantitative and qualitative aspects. For example, the agent may be requested to shorten the calling time, to reach 22 calls per hour. Simultaneously, the same agent may be requested to elongate calls for better customer handling and intimacy. Both can bias evaluation when the evaluator has the right to choose one of these contradictory standards. Technology acceptance and channel development influence the evaluation due to the perception that the agent is capable of adapting to new technologies and doing much better than usual, regardless of the core technical competency. Technical core competency means following the call script and responding to the caller correctly [38]. Many other studies highlighted factors in different environments applied indirectly to call centres [39-44]. Subjective performance measures are limited by collusion [45], influence costs [46], bias [47], leniency in rating efficiency [48], and favouritism [36, 47]. A subjective evaluation impacts agents' performance and leads to high turnover, emotional exhaustion, a lack of well-being, and burnout [3, 35, 41, 49]. The conclusion is that performance measurement is not accessible or straightforward but a complicated activity that requires more research and solutions.

The second issue in call centre performance evaluation is that the evaluation process is very important when it depends on manual classification, considering extensive evaluations over time, e.g. one year. Hence, the evaluation is performed randomly on selected calls out of thousands of records. This leads to missing a more realistic performance during the majority of calls.

The third obstacle is the evaluators' diversity, in that they may rank the same agent's performance differently. A unified evaluation system can exert a significantly adverse impact on call centres' business when the baseline is overlooked. Avoiding subjectivity and automating performance evaluation is essential in reducing the time and effort associated with the manual evaluation process. This leads to the establishment of a performance baseline for the call centre with a unified evaluation system.

2.4 *Related Work*

Many studies have proposed productivity and performance evaluation constructs [50-52]. These studies try to understand the performance factors and how they reflect on employees' well-being from the perspectives of human resource management (HRM) or the efficiency of operations management (OM). However, they do not propose a mechanism to measure performance automatically. Automating performance evaluation is a dominating information technology strategy in accordance with the calls and heavy communications traffic back and forth to the call centre. For example, the agent has 20 calls per hour for 8 hours, which means 160 calls per day. Assuming a call centre has two shifts (in some cases, there can be three) with 100 agents per shift, the expected calls per day are $160 \text{ calls} \times 100 \text{ agents} \times 2 \text{ shifts} = 32000 \text{ calls/day}$. Therefore, automating the evaluation process is essential and comprehensive – essential because of the massive number of calls per day, making manual evaluation unaffordable; comprehensive because it holistically covers all the calls rather than only random samples as a general practice in call centres [2].

Many studies have tried to measure customer satisfaction from customer feedback [21, 53]. As we mentioned, customer satisfaction is subjective because it may depend on the product's quality rather than the agent's performance. The data mining approach [17] uses call centres' activity logs, like the average handling time and wrap-up time, as objective metrics. However, this misses essential parts of the call that are relevant to the technical core. The study's 'evaluation software' intends to use machine learning to classify the performance of agents in call centres using a Linear Support Vector Machine (LSVM) [54]. The software is based on predefined features like speech rate, voice intensity level, and emotional state. Yet, it is limited to these features and does not cover a broader range of other features embedded in the call. Call centres are a dynamic environment with characteristics that may vary depending on the business models, cultures, and regions. There are activity monitoring tools and systems for tracking employee activities in the workplace and call centres. There are many

commercial systems, like Teramind⁹, Staffcop¹⁰, ActivTrak¹¹, Veriato¹², and others. These systems monitor agents' activities by hosting spy software on PCs to monitor various activities like session time, active time, idle time, performance scoring, keystrokes, mouse movement, productive application usage, screen shots, and voice recording. These systems use a rule-based algorithm for drawing conclusions about performance behaviour and overall productivity measurement. The drawback of these systems is that they do not draw a baseline of performance for each environment or domain in which activities differ. Furthermore, these systems overlook deep statistical analysis using machine learning for performance evaluation.

Therefore, machine learning models should be extendable to cover various features rather than a predefined set of them. Moreover, the model should provide results that help determine additional features that were not counted initially. In other words, machine learning model results should help the researcher understand the classification problem and not be limited in their conclusions.

2.4.1 Generative versus Discriminative Machine Learning Approaches

There have been several studies of performance evaluation for call centres based on the text transcription of calls. A. Ahmed, Y. Hifny, K. Shaalan, and S. Toral proposed binary classification models for predicting productive/nonproductive evaluation [20]. Generally, these studies are based on data set annotation into a productive/nonproductive. The first study is built on a generative model using Naïve Bayes to classify calls using text features.

Mathematically, Naïve Bayes is a generative model that creates data \mathbf{x} given class \mathbf{c} . Accordingly, we are looking for the maximum value of both likelihood value $\mathbf{p}(\mathbf{x}|\mathbf{c})$

⁹ <https://www.teramind.co/>

¹⁰ <https://www.staffcop-enterprise.com/>

¹¹ <https://www.activtrak.com/>

¹² <https://www.veriato.com/>

and prior probability $\mathbf{p}(\mathbf{c})$ in order to predict class probability given input features. The predictive class is shown in equation (1)

$$\mathbf{p}(\mathbf{c}|\mathbf{x}) = \mathbf{argmax}[\mathbf{p}(\mathbf{x}|\mathbf{c})\mathbf{p}(\mathbf{c})] \quad (1)$$

We calculate the joint probability by multiplying the probability of words given class $\mathbf{p}(\mathbf{x}|\mathbf{c})$ with the class probability $\mathbf{P}(\mathbf{c})$ to get the highest features for each class (highest probability) [55], as in equation (2).

$$\mathbf{p}(\mathbf{C}_k|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \mathbf{p}(\mathbf{C}_k) \prod_{i=1}^n \mathbf{p}(\mathbf{x}_i|\mathbf{C}_k) \quad (2)$$

The text features are transcribed manually in the annotation processor. The study model is shown in Figure 2.

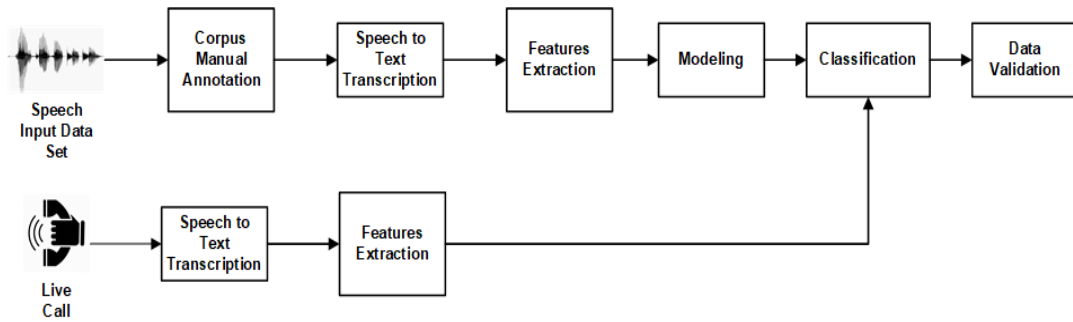


Figure 2: The single feature-based model

The modelling starts with corpus annotation, as mentioned before, into productive/nonproductive classes. The next step is speech-to-text transcription, performed manually or by an automatic transcription system [56]. The feature extraction step is concerned with generating a bag of words¹³ for each class. The study makes use of ‘word’ features for modelling and classification. Once the model is generated, the transcribed text from the live call is classified directly into the appropriate class. Another study followed the same framework for modelling and classification as shown in Figure 2, but using discriminative approaches [19]. This involves determining the probability of the productivity classes given the input

¹³ This is a modelling technique based on the frequency of word usage, regardless of grammar or of the order of the words.

features. As the input features are independent, the joint probability is as shown in equation (3).

$$\mathbf{p}(\mathbf{C}_k|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_{i=1}^n \mathbf{p}(\mathbf{C}_k|\mathbf{x}_i) \quad (3)$$

Several discriminative approaches have been used for performance evaluation, including Logistic Regression (LR) and the Linear Support Vector Machine (LSVM). Perera et al. developed software for the automatic handling of call centre agent performance [57]. They propose predefined factors like speech rate, voice intensity level, and emotional state to evaluate the performance of contact centre agents using Support Vector Machines (SVM). The classification was limited to these predefined features, which requires more investigation into other hidden factors. Ahmed et al. applied LR and an LSVM to the same data set of Naïve Bayes experiment [19]. Discriminative approaches outperformed the generative one, with an accuracy of 82.6% as compared to 66% (for Naïve Bayes). Previous studies demonstrate how machine learning can classify features with relatively high accuracy. However, the features are limited by the speech recognition system's accuracy, with an error rate of 22% for the Arabic transcription system [56]. Previous studies also dealt with agent performance as structured data. However, speech processing may provide better accuracy and thus improve classification accuracy. Speech data features are unstructured data types that require sophisticated neural network modelling structures, as discussed in the next section.

2.4.2 Deep Learning Approaches

There are many types of DNN, including recurrent neural networks (RNN) [58], Long Short-Term Memory (LSTM), Bidirectional LSTM [59], and Convolutional Neural Networks (CNN) [60]. CNN is a modified version of DNN for handling massive volumes of data, e.g. image processing, which requires a large network, vast parameters, many resources, and a great deal of time. The CNN structure divides layers into processing parts, e.g. image/signal dimension, and generates the corresponding parameters in terms of filters. The filter is a vector of trainable parameters concerned

with a smaller part of the whole image, for faster and more accurate processing (Figure 3).

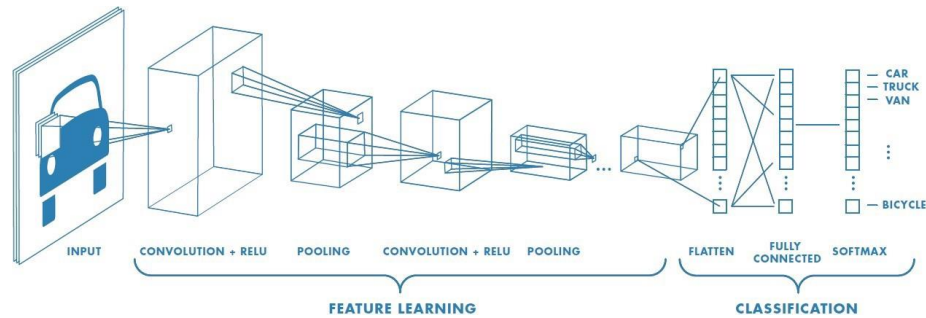


Figure 3: CNN structure – towardsdatascience.com

CNNs have proven to be significant in speech recognition and many other signal processing models [60]. CNNs help squash the frequencies' redundancy through their filters. Hence, they extract salient features through an efficient computational algorithm in a parallel mechanism. The LSTM is a form of recurrent neural network (RNN), which handles the gradient vanishing problem [61]. In RNNs, the gradient becomes very small for long sequences, which prevents the weights from being updated. LSTMs solve the gradient vanishing problem and can find longer temporal dependencies than simple RNNs. However, LSTM layers are slow to process the input sequence of speech frames. A variant of LSTMs known as bidirectional LSTMs (BiLSTM) [62] can integrate past and future information for better accuracy than legacy LSTMs. They combine two LSTMs in two directions: one operates forward, the other backward. Hence, each input frame at time t is aware of the past $t-1$ and future $t + 1$ contexts, which improves its accuracy.

DNN has additional layers on top of the previous networks to improve classification accuracy. The pooling layer squeezes the values to lower dimensions for better classification by either max pooling, which uses max values in the forwarded vector, or average pooling. The attention layer generates a context vector, which allows a greater focus on the significant hidden values of the previous layer. The CNN–Attention layer approach proves a significant improvement in speech recognition, image recognition, and emotional recognition [60, 63, 64]. The global max-pooling

layer is also commonly used in DNN structures to downsample the input vectors and reduce the dimensions to focus on the prominent features.

Speech studies use different approaches, like Mel Frequency Cepstral Coefficient (MFCC) or Filter Bank (FB), to extract speech features from the speech itself rather than the transcribed text [65, 66]. The MFCC represents the audio features in the frequency domain (non-linear spectrum) to be processed in numerical data to detect the vocal tract. It is based on the specific variation of the human ear's critical bandwidths with frequency and uses bandpass filters to capture the essential phonetic characteristics of speech [66].

Several studies applied speech processing to relatively large datasets to classify the acoustic features into targeted classes. Speech processing for productivity measurement was inspired by Emotional recognition studies [63, 67, 68]. Hifny and Ali developed an emotional recognition algorithm by extracting the MFCC features and classifying speech into seven emotional classes [63]. However, performance measurement requires much focus on the conversation aspects and the technical responses of parties, as mentioned before.

Abdelrahman Ahmed (2020) developed an approach to measure productivity by eliminating subjective factors ¹⁴. The general motive of the study was to classify the recorded calls as subjective or non-subjective. Subjective calls reflect the study variables: agent non-specific-job task and customer behaviour. Non-subjective calls are forwarded to productivity models to be classified as productive or nonproductive – see Figure 4. Abdelrahman's study was based on MFCC features and the DNN modelling structure. It draws a baseline of subjectivity with 82.5% classification accuracy. The main concern about the previous experiment is the imbalanced data set for each class. Imbalanced data sets may bias the accuracy, favouring the biggest class (non-subjective) [69]. F1 score has been applied, reducing the accuracy to 75%. Furthermore, the study could contribute indirectly to performance evaluation but fail to account for the productivity issues elaborated on in this study.

¹⁴ The Impact of Subjective Factors on Performance Evaluation: The Applied Case of Outsourced Call Centres in Egypt Based on Neural Networks Approach. Abdelrahman Ahmed, Thesis – University of Bradford, 2020 (in Press)

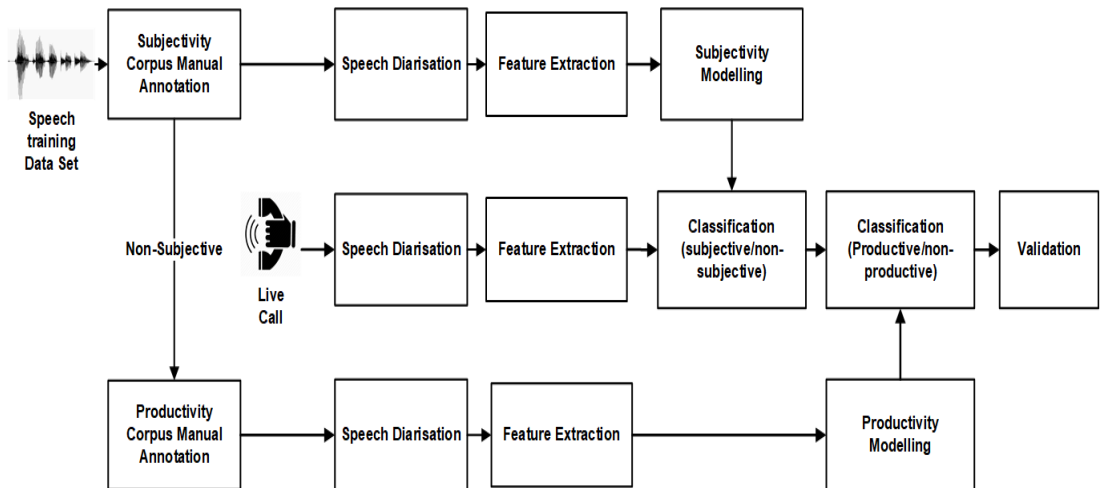


Figure 4: Subjectivity Elimination approach

2.4.3 Multimodal Classification Approaches

A supervised machine learning task is used to perform a single modality. For one data set, the training process generates the model and classifies it to check the output accuracy. Machine learning sometimes requires combining several types of data that are different in nature (text, image, or audio), performing extraction techniques, or doing both through multimodal machine learning. The multimodal approach aims to build models that can process and relate information from multiple modalities [70]. The multimodal approach can impact AI (artificial intelligence) on three levels: the application level, the performance level, and the methodological level. For example, the machine learning (AI) application-level tries to simulate human reactions to reading, hearing, feeling, smelling, and tasting. Each sensation is relevant to one model from a machine learning perspective. For example, ‘reading’ is relevant to language modelling, while ‘hearing’ is relevant to acoustic modelling. Another example is *Flickr*, a social website for sharing family and friends’ photos¹⁵. It provides the possibility to search for a photo or for its description. The photos always have tags that describe their contents. Hence, the multimodal approach can combine the image model with the text model to reach the content. Many applications have been proposed to combine text, images, speech, and videos, such as audio-visual speech recognition,

¹⁵ www.flickr.com

multimedia event detection, and media content description/indexing and retrieval [71, 72]. The second reason for using the multimodal approach is to improve the classification performance (accuracy) of the same task: combining models may improve the accuracy of the individual models. The third reason is concerned with the methodological level, which focuses on knowledge transfer among different models for better training. Co-learning explores how knowledge from one modality can help a computational model trained on a different modality, as exemplified by algorithms like co-training, conceptual grounding, and zero-shot learning [73].

The question is, why can the model not be combined at the feature level? In other words, why do we not combine the text features with speech signals and generate one model? The answer is that the data structure is different for speech and text: speech uses MFCC features, and text uses a bag of words approach. The multimodal approach can be used on the same features but with different modelling approaches. Meinedo and Neto extracted the acoustic features using different methods, combining the acoustic models using a multilayer preceptor (MLP) for local probability and the Hidden Markov Model (HMM) for temporal modelling of the speech signals [74]. There are still challenges in multimodal approaches: representation, translation, alignment, and fusion [70]. Representation means the difference in a data structure, like text with a symbolic representation compared to audio presented as a signal. There is no straightforward answer to how best to represent features so as to empower models for best classification. Translation is another challenge where data mapping is highly subjective among modalities of the same data type. For instance, presenting audio using a vocal tract or prosodic representation does not mean correct or perfect translation. The third challenge is the alignment between modalities when determining a direct relationship between different data features or sub-elements. For example, in *Flickr*, matching photos with the corresponding description is a complex task in combination and validation. The fusion is about predicting when, for example, matching word-based speech recognition with visual lip motions.

There are several multimodal approaches in machine learning. They have been categorised into 1) feature-based representation, 2) weighted average or scoring representation, 3) joint representation, and 4) coordinated representation. Feature-based representation unifies the features before training the models. It requires the

features to be normalised and concatenated with a fixed width to create well-trained models [75, 76]. The problem with this approach is the possibility of undetermined collinearity between features, which may impact the model (overfitting) [77]. Another approach is the weighted average of the models' accuracy. This is similar to acoustic and language models' scoring function [56]. It requires balancing the power of each model as compared to others using scoring factors. The acoustic model is combined with the language model for the best probabilistic outcomes using a scoring function [56]. However, the weighted average method requires defining an external mechanism to balance the models' weights or exhaustive trials to reach the optimum scoring values.

The joint representation approach is based on training the models separately and combining them before or at the final layer (joined space) [73]. The advantage of joint representation is that it keeps the different features apart. First, the features are trained with corresponding weight enhancement; then, the training weights are forwarded to the final concatenated layer. However, the disadvantage of this approach is that it cannot handle missing data, e.g. Out of Vocabulary (OOV) words [78]. Probabilistic graphical models are another aspect of joint representation, using a latent random variable like Deep Boltzmann Machines (DBM) [79]. DBMs are similar to a neural network structure based on the joint probability distribution of the energy function [76]. They present data in a probabilistic manner that can later be forwarded to a neural network as initial (pre-trained) weights [75]. This is a powerful generative approach, but its major disadvantage is that it consumes high computational resources. Sequential representation is a common, well-known practice in multimodal training using Recurrent Neural Networks (RNN) and their corresponding variants (LSTMs/BiLSTM). The RNN is a unimodal approach for a time series training at time t . For instance, the DNN weights of the model at a later stage can be injected into an RNN for final classification.

The fourth approach is a coordinated representation that uses the similarity among representations rather than a joined space. This ensures a more structured representation for the resulting space [73]. This approach is concerned with the similarity between models to find less distance and distinguish the best classification [80]. For example, the bag of words and images linked to these words. This is a

straightforward mechanism to coordinate the models separately, but its disadvantage is subjectivity and poor ability in labelling for sophisticated featured data. Table 2 and Figure 5 illustrate the taxonomy of multimodal approaches.

Unified Features	Scoring Approach	Joint Representation	Coordinated Representation
Normalised, concatenated features	A weighted average of the models' respective accuracy	Shared Space	Similarity Approaches
	Scoring probabilities	Probabilistic graphical models	

Table 2: The multimodal approaches

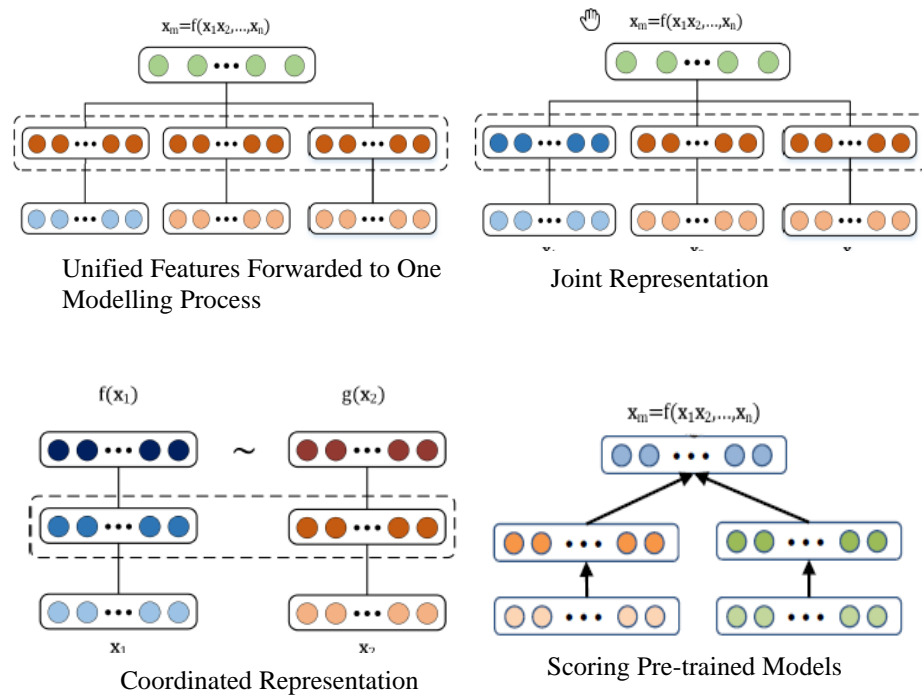


Figure 5: Multimodal Approaches

2.4.3.1 The Multimodal approach for Speech Recognition

It is important to shed light on speech recognition in this section for two reasons. First, speech recognition is a fundamental part of the transcription process required for performance evaluation. Automatic speech recognition transcribes speech into text to classify a performance as productive or nonproductive. Second, it demonstrates a generative multimodal approach, which is part of the scoring approach. The productivity study is conducted on Arabic speakers' agents in call centres located in Egypt. So, text transcription will highlight the speech techniques and challenges specifically related to the Arabic language.

Arabic is a challenging language and is considered one of the most morphologically complex [81-84]. Automatic Speech Recognition (ASR) for Arabic has been a research concern over the past decade [84, 85]. The Arabic language has limited resources in speech recognition because it requires a high level of experience in speech technology using a Hidden Markov Model (HMM)/Gaussian Mixture Model (GMM), as well as linguistic experts [86]. Most speech recognition systems are developed for Indo-European languages [85]. The Arabic language is structured from right to left, with a different pattern of vowelisation [87]. There have been many attempts to present Arabic speech recognition models [81]. However, it is still challenging to learn the methods used and get results competitive with those for other languages [88]. HMM/GMM acoustic models have significantly improved speech recognition in different languages, including Arabic [89]. Word alignment for acoustic modelling is performed by HMM state modelling, and more than one Gaussian model presents each state (phoneme local probability). HMM training is based on maximum likelihood. The lexicon and language model are prepared before the decoding process so the decoder will match the best score for each word defined in the dictionary [90]. HMM/GMM modelling still has some drawbacks: (1) it requires a deep knowledge of HMM; (2) it is developed under the assumption that observations are independent, which does not comply with the vocal tract. In HMM, each phone call is jointly presented by three emitting states, as shown in Figure 6.

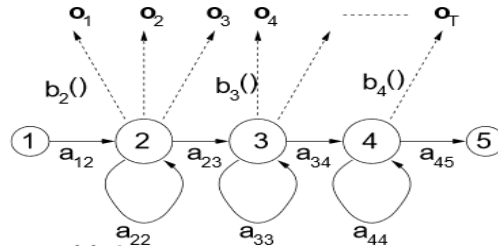


Figure 6:Phones and hidden states

Each phone is characterised by three states: a state at the beginning of the phone, the state when it is in the middle, and at the state at the end of the phone, i.e. a_{12} , a_{22} , a_{23} . Each state is represented by more than one gaussian probability distribution. The observations, i.e. o_1, o_2, o_3 in Figure 6, are speech observations according to hidden states [91]. The neural networks have achieved great jumps in the fields of speech recognition, handwriting, visual recognition, and others [92]. By mixing neural networks (NN) and HMM/GMM, speech recognition outperforms a tandem HMM/GMM method. The tandem method [93] is a feature extraction approach that uses a deep neural network (DNN) to obtain features complimentary to the MFCC [94]. This setup outperforms the HMM model. The hybrid approach, which is not included in this study, combines the HMM and Artificial Neural Network (ANN). It uses HMMs for sequential modelling and ANN models as flexible discriminant classifiers to estimate a scaled likelihood [95]. This approach, also combined with DNN and BDRNN, provides a performance improvement [96]. The hybrid method can be generalised using deep conditional random field methods [97]. Yet data preparation and integration for an HMM/GMM tandem or hybrid are very complicated and require much processing time.

Speech recognition can also be treated from the perspective of language models. A language model is a probability distribution of hypothesised words generated by scoring the words' probability in a previous word sequence [98]. Scores are estimated before the training process when the domain is defined [98]. The language model is used to resolve ambiguous utterances [82, 99]. For example, two sentences, 'it takes two' or 'it takes too', could not be decoded by an acoustic model until it is combined with the language model. The n-gram language model is the number of words count

(frequency) proceeding the n-words in sequence. The following equation determines the word probabilities:

$$P(\mathbf{c}_1, \dots, \mathbf{c}_m) = \prod_{i=1}^m P(\mathbf{c}_i | \mathbf{c}_1, \dots, \mathbf{c}_{i-1}) \quad (4)$$

Where \mathbf{c}_i is the position of the word in the stream of words (sentence). Most speech recognition systems use 3-grams, 4-grams, up to 7-grams. The higher the gram order, the higher the certainty (and the lower the entropy).

The previous discussion was about multimodal acoustic models using hybrid probability models (GMM/HMM/NN) in terms of multimodal approaches. However, there should be a multimodal alternative that combines an acoustic and language model for the best classification. The following equation presents the relationship between an acoustic and language model:

$$\hat{\mathbf{c}} = \mathbf{arg\ max}[\log P(x|\mathbf{c}) + \alpha \log p(\mathbf{c})] \quad (5)$$

Where $\hat{\mathbf{c}}$ is the hypothesised word.

Chapter 3 describes the methodology of Arabic speech recognition and the approaches used to improve both acoustic and language models. The approaches will use a deep neural network for automatic transcription and the corresponding productivity measurement framework.

2.5 Literature Review Gaps

The broadest gap in existing studies is how to evaluate a call centre's agent performance from the recorded calls using text and speech processing. Many studies have tried to measure customer satisfaction from feedback [21, 53], but this perception is about product quality rather than agent performance. The data mining approach [17] uses activity logs, like the average handling time and wrap-up time, that look like

objective evaluation. However, it measures agent efficiency but misses essential parts of the call, namely the technicalities that the agent should fulfil. Many commercial systems use rule-based algorithms to monitor users' activities on their PCs. Yet, these systems overlook comprehensive statistical analysis and the building of feature relations using machine learning approaches.

Perera et al. developed software to automatically handle a call centre agent's performance [57]. They propose predefined factors like speech rate, voice intensity level, and emotional state to evaluate the performance of contact centre agents using a Support Vector Machine (SVM). However, their classification was limited to these predefined features, requiring more investigation into other hidden factors. Sudarsan et al. examined several systems to evaluate performance in call centres based on prohibited words, emotional recognition, and greeting words [100]. Their framework was based on pre-trained machine learning platforms like Google, Wit, and Sphinx for transcription, word analytics, and emotional detection. A big data analytics application has also been applied to recorded calls to detect the quality of service delivered to the customer. It was based on the Hadoop Map-Reduce framework and utilised text similarity algorithms such as Cosine and n-gram [101]. It also integrated slang word lists into the monitoring system; however, the study was limited to text processing, and speech processing was overlooked.

Other studies were based on text classification using words relevant to productivity measurement [102, 103]. Ahmed et al. transcribed calls to text using a speech recognition engine [56] and then classified them according to a pre-annotated corpus with a binary classifier (productive/nonproductive). The productivity classification was based on a Naïve Bayes generative model and determined the posterior probability of a productivity class conditioned on observations [20]. A similar study was conducted based on a discriminative approach [19]. Logistic Regression and a Linear Support Vector Machine (LSVM) were used to improve the classification accuracy. However, the text processing was based on a bag of words, and it did not take advantage of the deep learning approach using word embeddings for better classification. A speech processing study based on MFCC 13 features and DNN was also conducted for non-subjective classification as part of the broader performance

evaluation study. The study intended to eliminate subjective factors like non-specific tasks and customer behaviour. However, it requires a performance evaluation of the non-subjective calls. Finally, it is worth mentioning that the previous studies have not considered a multimodal approach for both text and speech processing, which is one of the main contributions of this work. Table 3 summarises the previous studies and the features used.

SN	The Study Approach	Features	Reference
1	Statistical Analysis	Structured data	[21] Rychalski and A. Palmer, 'Customer Satisfaction and Emotion in the Call Centre Context,' in <i>The Customer is NOT Always Right?</i> 2017 [50] D. Chicu, M. del Mar Pàmies, G. Ryan, and C. Cross, 'Exploring the influence of the human factor on customer satisfaction in call centres,' 2019
2	Data Mining approach	Data mining for structured data	[17] M. Paprzycki, A. Abraham, R. Guo, and S. Mukkamala, 'Data mining approach for analysing call centre performance,' 2004
3	Discriminative approach	Unstructured data-speech processing	[104] K. Perera, Y. Priyadarshana, K. Gunathunga, L. Ranathunga, P. Karunaratne, and T. J. I. J. S. R. P. Thanthriwatta, 'Automatic evaluation software for contact centre agents' voice handling performance,' 2019 [105] V. Sudarsan and G. Kumar, 'Voice call analytics using natural language processing,' 2019
		Unstructured Data – language processing	[20] A. Ahmed, Y. Hifny, S. Toral, and K. Shaalan, 'A call center agent productivity modelling using discriminative approaches,' 2018
4	Deep learning and generative approach	Speech recognition, n-gram	[106] B. Karakus and G. Aydin, 'Call centre performance evaluation using big data analytics,' 2016

	Generative approach	Unstructured data – generative approaches – language processing	[19] A. Ahmed, S. Toral, and K. Shaalan, ‘Agent productivity measurement in a call centre using machine learning,’ 2016 [107] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, ‘Automatic analysis of call-centre conversations,’ 2005 [108] M. A. Valle, S. Varas, and G. A. J. E. S. w. A. Ruz, ‘Job performance prediction in a call centre using a naive Bayes classifier,’ 2012
5	Deep learning	Unstructured data – speech processing	A. Ahmed, ‘The impact of subjective factors on performance evaluation: The applied case of outsourced call centres in Egypt based on neural networks approach’ (Thesis, University of Bradford), 2020.

Table 3: The performance measurement studies

2.6 Conclusion

The chapter reviews the literature on call centres, productivity measurement, and their relevancy to machine learning approaches. There are text modelling methods using generative and discriminative approaches. The generative approach has been used for text classification via Naïve Bayes to determine the word sequence probability given the annotated classes. The discriminative approaches determine the joint probability of the targeted classes given a set of features under the assumption that the features are independent. The discriminative approaches for productivity measurement include logit regression and the linear support vector machine (LSVM). The logit regression uses a sigmoid classifier to classify the data within the range 0 to 1. The LSVM classifies the data using a hyperplane surrounded by a margin to categorise the text as productive or nonproductive. Speech has been used for subjectivity detection and classification into agent non-specific-job tasks, customer behaviour, and non-subjective classes.

This chapter discusses the four main categories of the multimodal approach: unified features, weighted average representation, joint representation, and coordinated

representation. The multimodal approaches show significant classification accuracy when the natures of data, structure, and features are different. Features can be unified before training models. This requires normalising the features and concatenating them with a fixed width for well-trained models. Another approach is the weighted average of the model accuracies. This is similar to acoustic and language models' scoring function. It requires balancing the power of each model as compared to others using so-called scoring factors. Joint representation is based on a shared space for data modelling and propagates them into the classification layer. The coordinated representation is the fourth category, determining model similarities to find the best matches. Arabic speech recognition is an important part of the productivity measurement process as it transcribes speech into text (the text studies). The Arabic language is one of the most challenging languages to process using AI technologies due to its morphological complexity. Arabic speech recognition applies a joint representation by combining a GMM for local probability and an HMM for sequence probability. The generative multimodal approach is used in speech recognition when combining the acoustic and language models into a scoring equation to improve classification accuracy. The first study question, 'What are the best approaches to machine learning for productivity measurement?', is partially answered by highlighting the strength of using deep learning for text and speech modelling, combined with a multimodal approach. The next chapter discusses the proposed methodology and framework by going through the steps and procedures for measuring agents' performance in call centres.

Chapter 3: Research Methodology and Framework

3 Research Methodology and Framework

3.1 Introduction

Chapter 3 digs deep into the methodology of the machine learning approach to productivity measurement. The literature reviews in Chapter 2 highlighted the challenges facing performance evaluation in call centres: first, the quality team's evaluation is subjective, depending on their previous experience [2]; second, manual evaluation is unaffordable due to the massive number of calls; and third, external factors impact evaluation, such as customer behaviour and the management bias in favour of frustrated customers and thus against agents [53, 104]. Previous studies tried to automate evaluation using AI and machine learning with various classification models. The next sections discuss this research strategy and design. The study framework proposes different neural network combinations for speech and text to achieve the best classification accuracy. Then, this study proposes combining the network structures with the best results, following the multimodal approach, to improve classification performance. The study selected a joint representation multimodal approach with different data features and performed classification using a shared (concatenated) layer. The methods used in the experiment stages and the modelling approaches will be discussed.

This chapter should achieve the first study objective:

- **Build a multimodal conceptual model for the call centre domain.**

The conceptual model should help answer to the following research question:

- **How can previous models be combined to improve accuracy?**

3.2 Research Strategy

The research strategy delineates the guidelines of the study from the beginning to the end. It is based on deductive and inductive approaches (abductive approach) [105]. The deductive approach proves the theoretical perspective using empirical validation. Quantitative methods will validate the characteristics of the data set and its eligibility to be modelled in a way that removes human subjectivity. So, the accuracy validated using comprehensive statistical models via machine learning will be considered proof

of the success or failure of the experiment. The inductive approach will be followed to explore additional characteristics of the calls and understand questions of productivity. In other words, the study results will be analysed to suggest additional characteristics of the performance. The strategy is fulfilled by collecting an appropriate data set that generalises the experiment for the same domain and context. The models should consistently simulate evaluators' behaviour to avoid subjective effects. Also, it should have the capability to automate classification smoothly and rapidly. The most important strategy is to explore the best methods based on combining previous knowledge to achieve the study objectives and answer the research questions. Therefore, the framework of the study exploits the most appropriate design to increase classification accuracy. The next section discusses methods taken from previous studies to establish the study framework and related research activities.

3.3 *The Study Framework and Selected Methods*

Performance (productivity) in call centres is related to individual proficiency and the relevant activities that contribute to the organisational 'technical core' [38]. The technical core refers to call centres' standards, which an agent should follow with relevant technical knowledge to respond to customer inquiries. The technical core standards involve following the call scripts and predefined scenarios: starting with a greeting, verifying the caller's account, and responding to customer inquiries correctly [35]. The recorded call comprises streams of speech and text features that present the sequence of the conversation and the corresponding productivity annotation. As shown in Figure 7, the study framework is split into five stages: 1) call diarisation, 2) speech-to-text transcription, 3) feature extraction, 4) data modelling, and 5) classification. Several modelling approaches will be followed to achieve the best accuracy.

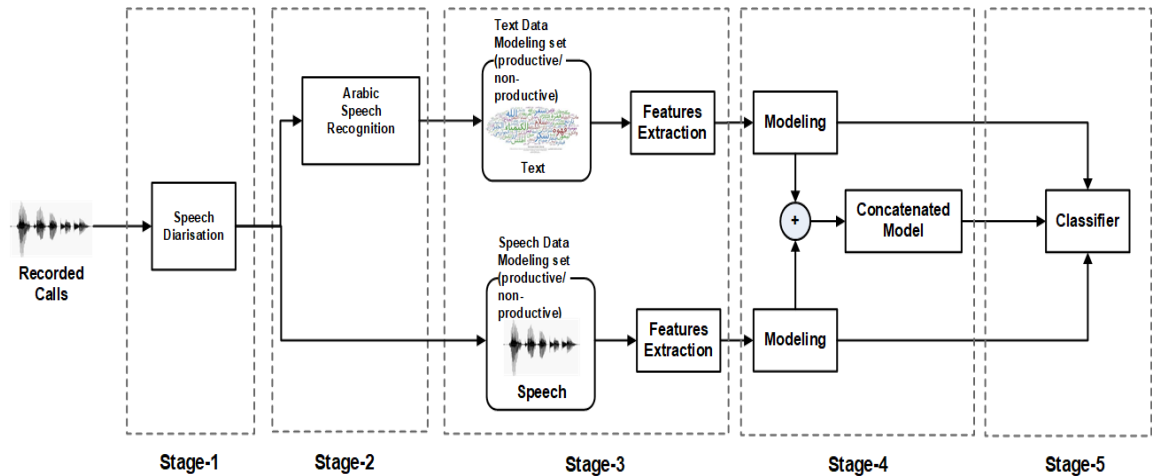


Figure 7: The Study framework.

3.3.1 Call Diarisation

Generally, recorded or live calls in call centres include two parties: the agent and the customer¹⁶. Productivity measurement is mainly concerned with evaluating the agent’s part of the call. Diarisation is a signal processing algorithm to split the voice stream into smaller chunks. More specifically, speaker diarisation is the process of splitting the speakers’ utterances into separate segments [106].

Hence, the diarisation step should extract the agent’s part of the call for evaluation and exclude the customer or any third party. The challenging part of speaker diarisation is diarisation accuracy because such algorithms are subject to error rates for similar segments. Overlapping utterances, when the agent and customer speak simultaneously without considering the conversation switching points, are critical issues. Many algorithms and studies are dedicated to speaker diarisation, but they are outside the scope of this study [107, 108]. This study will use the SIDEKIT toolkit (S4D), which supports the Bayesian Information Criterion (BIC), Hierarchical Agglomerative Clustering (HAC), and Viterbi decoding algorithms [109].

¹⁶ The call may include a third party, such as an auxiliary call to a team leader or supervisor for some inquiries.

[12] J. C. Abbott, *The executive guide to call center metrics*. Robert Houston Smith Publishers, 2004.

3.3.2 Speech-to-Text Transcription (Speech Recognition)

The Arabic transcription system is a DNN modelling approach based on the Stanford CTC (Connectionist Temporal Classification) source code¹⁷ [110]. The engine is based on RNNs and the CTC objective function. The acoustic model will be trained using 1200 hours of the Aljazeera broadcast news TV corpus, collected by QCRI [111]. The speech-to-text transcription system consists of three components: a BDRNN acoustic model, a language model, and a character-based decoder. The character-based decoder does not need a lexicon or word dictionary in the decoding process, unlike word-level decoders. In addition, the training and decoding process is based on Arabic graphemes. The objective function used to train BDRNNs is CTC, which removes the need for pre-segmented acoustic observations. The evaluation for the test set will be performed on both word and character levels to validate the results against other word-based models. The next subsections discuss in more detail the acoustic model using BDRNNs/CTC, the n-gram language model and the lexicon-free character-based decoder.

3.3.2.1 Acoustic Model

The Bidirectional Recurrent Neural Network (BDRNN) will be applied to train the acoustic model to score each character, given the input data. Moreover, the CTC objective function (loss function) maximises the probabilities of predicting the correct characters.

A Bidirectional RNN computes the probability of the output character c appearing at a given time input x_t . It consists of a few hidden layers followed by a SoftMax output layer (illustrated in the next sections). The scoring at each layer depends on the current input x_t and the previous hidden state s_{t-1} . Hence, it does not model information based on the future acoustic context [96]. BDRNN has separate hidden layers for scoring, based on the past and future context, to overcome this limitation. Each layer is computed separately by going forward from $t - 1$, t and $t + 1$ parallel with a

¹⁷ <https://github.com/amaas/stanford-ctc>

backward computation from $t + 1$, t and $t - 1$. Then, the hidden layers are summed together as in Figure 8.

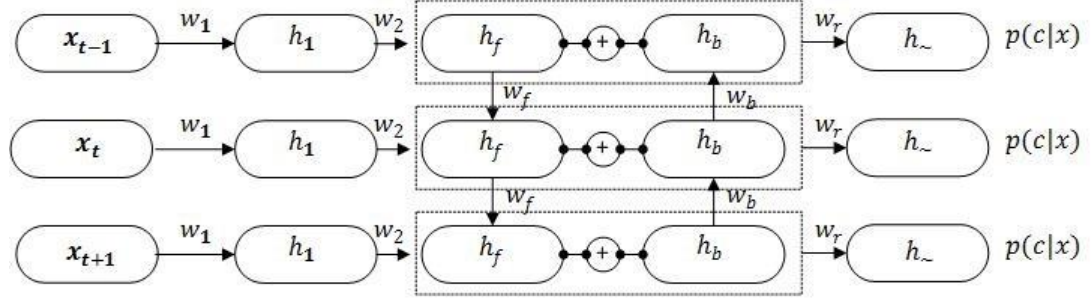


Figure 8: The RNN bidirectional layers

The first layer of scoring is based on equation (6).

$$h_t^{(1)} = f(W^{(1)T}x_t + b_1) \quad (6)$$

The second layer in Figure 8 is the BDRNN hidden layer j , comprising the partial sum of a forward and backward layer (temporal layer) at time t .

$$h_t^{(j)} = h_t^{(f)} + h_t^{(b)} \quad (7)$$

The hidden forward and backward layers are computed independently, with weight matrices $W^{(f)}$ and $W^{(b)}$. The partial hidden layer takes input from the previous hidden layer $h_t^{(j-1)}$. Therefore, the hidden layers $h_t^{(f)}$ and $h_t^{(b)}$ at time t are computed as shown in equation (8).

$$\begin{aligned} h_t^{(f)} &= f\left(W^{(j)T}h_t^{(j-1)} + W^{(f)T}h_{t-1}^{(f)} + b^{(j)}\right), \\ h_t^{(b)} &= f\left(W^{(j)T}h_t^{(j-1)} + W^{(b)T}h_{t+1}^{(b)} + b^{(j)}\right), \end{aligned} \quad (8)$$

where $f(z) = \min(\max(z, 0), \mu)$ is a rectified linear activation function clipped to a maximum possible activation μ to prevent overflow [112]. The final layer of the BDRNN computes the output distribution $p(c|x_t)$ using a SoftMax function:

$$p(c_k|x_t) = \frac{e^{-(w_k^{(s)T}h^{(\cdot)}+b_k^{(s)})}}{\sum_{j=1}^K e^{-(w_j^{(s)T}h^{(\cdot)}+b_j^{(s)})}}, \quad (9)$$

where $W_k^{(s)}$ is the k th column of the output weight matrix $W^{(s)}$ and $b_k^{(s)}$ is a scalar bias term. The vector $h^{(\cdot)}$ is the hidden layer representation of the final hidden layer in our BDRNN. The set of all expected characters K includes the blank symbol ($_$).

The objective function used to train the BDRNN is the Connectionist Temporal Classification (CTC). CTC removes the need for pre-segmented training data. Given an input sequence X of length T , CTC assumes the probability of a character sequence C of length T is computed as follows:

$$P(C|X) = \prod_{t=1}^T p(c_t|X) \quad (10)$$

The network output at different times is conditionally independent, given the input [113]. Afterwards, the total probability of any label sequence can be found by summing the probabilities of its different alignments. In particular, the CTC objective function $CTC(X, W)$ is the likelihood of the correct final transcription W , which requires integrating over the probabilities of all character sequences C of length T [110].

$$\begin{aligned} CTC(X, W) &= \sum_{C_W} P(C|X) \\ CTC(X, W) &= \sum_{C_W} \prod_{t=1}^T p(c_t|X) \end{aligned} \quad (11)$$

A CTC collapse function constructs possible shorter output sequences from our T -length sequence of output characters. It collapses any repeated characters in the original sequence of length T . For example, the word ‘so’ in English is equivalent to any of the following [sso, soo, _so, s_o, so_]. Like HMM, a dynamic programming algorithm is used to compute this loss function and its gradient for the BDRNN parameters.

3.3.2.2 Language Model

As mentioned in Chapter 2, the language model computes the probability of a word sequence. It is used to resolve ambiguous utterances during the decoding [99]. For example, two sentences, ‘it takes two’ and ‘it takes too’, are acoustically confusable. When language model scores are combined with acoustic scores, the ambiguity may be resolved. As we used characters instead of words during decoding, equation (12) is the $n - gram$ character-based prior probability.

$$p(c_1 \dots c_m) = \prod_{i=1}^m p(c_i | c_1 \dots c_{i-1}), \quad (12)$$

where c_i is the character position in the stream of characters. Most speech recognition systems use 3-grams, 4-grams, up to 5-grams. The higher-order models imply higher certainty (low entropy). We need to extend the n-grams to the highest possible order to increase the certainty per word and for the preceding words for the character-based approach. For example, assuming that we have a 4-gram word-based decoder and the words on the average count of letters about four letters, we may need a 16-gram order in a character-based language model (4 words \times 4 letters). Limitations on computational resources may hinder the possibility of achieving such an order (16-grams). Furthermore, it consumes a great deal of time in the decoding process, which diminishes the prominent advantage of the lexicon-free decoders discussed in the next section.

3.3.2.3 Decoding

The beam search decoder is used in the transcription engine to decode the sentences at a character level. The beam search calculates the likelihood of a character sequence of a specific length (beam length). The beam length is user-defined, so the longer the beam, the more accurate the decoding process. This method gives more advantages than word-level decoding for two reasons. First, the decoding speed at the character level is much higher than at the word level because of the lexicon. The search time for a lexicon-based decoder is a function of the number of words to be searched. For the

Arabic language, the lexicon may contain up to 2M words. Hence, lexicon-based decoders may be very slow. On the other hand, character-based decoders depend on the number of characters (e.g., 35) used to train a BDRNN. Hence, they are faster than a lexicon-based decoder [114]. Second, character-based decoding overcomes the Out of Vocabulary (OOV) problem exhibited by word decoding [110].

The collapse function ignores non-blank symbols due to the time shift of character alignment, which produces the same character again. It also controls the hypothesized characters to be repeated or the characters residing between two blanks. Furthermore, the sum of the acoustic model and the language model probability logs, equation (13), avoids underflow (a problem in processing very small values) and increases the algorithm's speed. α is the scaling factor of the language model to balance the weight of the language model probability. For example, when α is adjusted by a small value (fraction), the effect of the language model on the probability of the predicted character given the input, or vice versa, is small. The probability of the predicted character is shown in equation (13).

$$c^\psi = \operatorname{argmax}[\log p(x|c) + \alpha \log p(c)] \quad (13)$$

where c^ψ is the hypothesized character with a given sequence length (beam length). β is the insertion bonus, which is the scaling factor of the final insertion of the character string. This is the exponent value of the length of the generated string multiplied by the probability of the hypothesized string. If $\beta < 1$, a reduction factor is applied, which reduces the opportunity of the decoder to insert the hypothesized string (conservative decoder). The beam length is the length of the hypothesized character sequence to be processed through probability calculation. By increasing the beam length, the decoder accuracy increases, and the decoding process consumes much more time than for a shorter beam length.

3.3.3 Feature Extraction

Feature extraction is an important stage in machine learning, transforming the input data set into a trainable form. The next stage consists of extracting the speech and text features available from previous steps. The general framework intends to model both the speech and the text and concatenate them to improve the accuracy of the final performance evaluation. As the nature of speech data is quite different from that of text, the features extraction methods are also different. Feature extraction is a wide research area, so this study focuses on the approaches and toolkits available without going too deeply into a methodological perspective.

3.3.3.1 Speech Features Extraction

Many studies have used MFCC (Mel Frequency cepstral coefficient) to extract the signal features in the frequency domain [63, 65, 66]. MFCC conversion changes audio signals into numbers (frequencies) to identify the salient features (coefficients) from the audio file and to ignore unimportant features (noise). MFCC goes through multiple stages of windowing and a Discrete Fourier Transform (DFT) to extract 13 voice coefficients. MFCC feature extraction is performed by going through the audio files with a 25 millisecond (ms) window size and 10 ms steps for each window (overlapped) [66]. Each frame is presented as 13 MFCC features and forwarded to the model input layer for training. As MFCC focuses on the vocal tract, it works efficiently with speech recognition [56].

However, providing MFCCs with 13 features may limit classification performance; so, the study will investigate the effect of extended features on accuracy. An additional 65 features will be used, based on the INTERSPEECH 2016 Computational Paralinguistics Challenge (2016 COMPARE) [115]. That study was concerned with different speech feature problems, like the classification of deceptive vs non-deceptive speech, estimation of the degree of sincerity, and identification of the native language. It includes energy-related, spectral-related, and voicing-related Low-Level Descriptors (LLDs), comprising logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. The features are presented in Table 4.

54 spectral LLD
RASTA-style auditory spectrum
MFCC 1–14
Spectral energy
Spectral Roll Off Point
Entropy, Spectral Flux, Skewness, Variance, Kurtosis, Slope, Harmonicity, Psychoacoustic Sharpness
7 voicing-related LLD
Probability of voicing, F0 by SHS - Viterbi smoothing
Jitter, logarithmic HNR, Shimmer
PCM fftMag spectral Centroid SMA numeric
4 energy-related LLD
Sum of the auditory spectrum
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate

Table 4: 65 provided Low-Level Descriptors (LLD)

3.3.3.2 Text Features Extraction

In previous performance evaluation studies [19, 20], the features extracted were based on a bag of words approach. The bag of words counts the occurrences of each word in the sentence, considering the words as independent of each other. Therefore, the first drawback is that it does not consider the statement's context [116]. The second drawback is the large vector size arising from the vocabulary size, which is not practical. Third, this representation leads to highly sparse vectors, as words rarely occur in many documents. Accordingly, word embedding representation is preferred for contextual extraction. The Word2Vec approach computes continuous vector representations of words from very large data sets [117]. Google's BERT (Bidirectional Encoder Representations by Transformers) developed a sophisticated technique for generating vectors representing the contexts of statements, based on pre-trained deep bidirectional representations from an unlabelled text, by jointly conditioning on both left and right contexts in all layers [118]. Word embedding approaches achieved significant improvement in natural language processing and deep learning modelling [117, 119]. However, the pre-trained models do not support the Arabic language in the

call centre domain¹⁸. Therefore, the embedding layer is used instead to get a representable conversion in call centres. The words are indexed based on the whole vocabulary and passed to the neural network to train the text model.

3.3.4 Data Modelling and Classification

Data modelling is the architecture of the neural networks used for the modelling process. There are several alternatives for modelling speech and text. This study proposes two main schemes for modelling text and speech, as shown in Figure 9. Once the speech is diarised, two separate branches can be distinguished in Figure 9. The first branch is for modelling speech using CNNs, cascaded CNN-LSTMs, and an attention layer. Each speech modelling subbranch presents one or more deep learning architecture combinations, e.g. CNN, CNN-attention, CNN-LSTM, and CNN-LSTM-Attention layers. The frames are forwarded to the speech modelling subbranches to obtain the speech model with the best accuracy. The text is transcribed using an automatic speech recognition system, and the features extracted using a word embedding layer. The text branch follows the same speech neural network structure to attain the best accuracy for the text branches. The models with the best accuracy for speech and text are then merged (concatenated) at the last neural network layer for sigmoid binary classification.

3.3.4.1 CNNs and BiLSTMs

CNNs are widely used for signal processing and speech recognition tasks [60]. They help scan the extracted features' frames to obtain the best classification accuracy through the filters. This study considers two main branches, as shown in Figure 9: one for the text features and another for the speech features. Each branch is in turn divided into four subbranches that follow a similar scheme: two make use of 1D-CNN layers with *tanh* activation functions, followed by either a global-max-pooling layer or an

¹⁸ There are several pre-trained models that support Arabic, including QARiB, mBERT, and ArabicBERT. They are trained using the OSCAR corpus (web dump), social medial and tweets [120] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. J. a. p. a. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations," 2021.

attention layer; the other two make use of a 1D-CNN-BiLSTM, also followed by either a global-max-pooling layer or an attention layer.

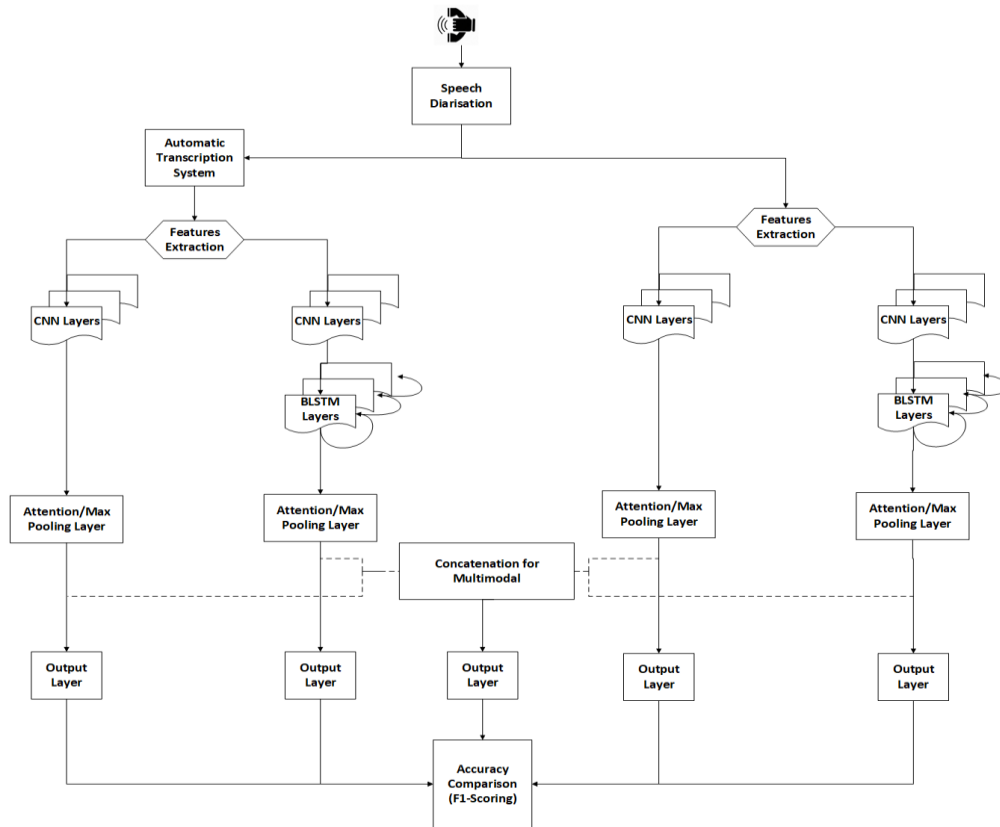


Figure 9: The study Framework

Finally, a logit sigmoid output layer performs binary classification into productive or nonproductive calls. Combinations of different models are proposed in this study to identify the best classification performance.

3.3.4.2 Attention layer

The sequence of vectors (frames) produced by a CNN or LSTM are forwarded to the attention layer to be converted into a context vector [64, 121, 122]. The attention weights are propagated to the SoftMax function at time t to generate the probability of the frame out of one to the remaining frames in the same speech segment. Figure 10 illustrates the role of the attention layer in our approach:

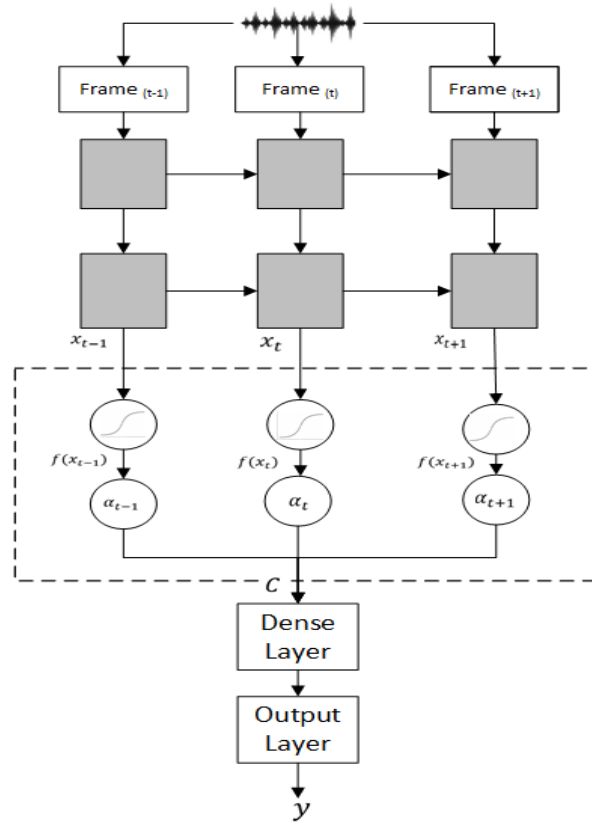


Figure 10: The Role of the Attention layer

- The modelling layers are shown in solid grey.
- The attention layer is the dotted box, including the circles representing the calculation of the attention weights. The SoftMax function calculates the attention weights and generates the context vector C .
- The context vector is fed into a dense layer with a *tanh* activation function.

The output layer is the logit regression (sigmoid) function for the segment classification. Then, the context vector is generated from the weighted average of the frames' probabilities. For each vector, \mathbf{x}_t , in a sequence of inputs, i.e. $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$, the attention weights, α_t , are calculated as follows:

$$\alpha_t = \frac{\exp(f(x_t))}{\sum_{j=1}^T \exp(f(x_j))} \quad (14)$$

where $f(x_t)$ is defined by the parameter w as follows:

$$f(x_t) = \tanh(w^T x_t) \quad (15)$$

The weighted average of the SoftMax generated weights, and the input vector are summed to get the context vector C .

$$C = \sum_{t=1}^T \alpha_t x_t \quad (16)$$

The Dense layer D uses a *tanh* activation function:

$$D = \tanh(W^T C + b) \quad (17)$$

where W comprises the weights of the hidden layer, and b is the bias. The Logit function is the output layer for the two classes (productive/nonproductive).

$$y = \text{Logit}(D) \quad (18)$$

3.3.4.3 Max Weights Similarity (MWS)

The attention layer uses the SoftMax function to determine the probability of the hidden layer weights [64]. The SoftMax function converts a vector of real values into probability values that sum up to one [123]. The SoftMax function is called either multi-class logistic regression or the Softargmax function. The wide variety of speech features ($n \times \text{features} \times 25\text{ms frame}, 10\text{ms frameshift}$) means that SoftMax can efficiently perform speech processing. However, it may be less efficient in text than in speech because the generated context vectors for text features have values quite close to each other. The attention layer, therefore, has insufficient variability to reach a value with significant accuracy. This study proposes the Max Weight Similarity (MWS) function instead of SoftMax to overcome the limited variability in the feature set. MWS aims to collapse the training weights around a reference value, which is the maximum value of the vector. The MWS function determines the similarity between the maximum value in the vector and the remaining values. For each vector x_t in a sequence of inputs x_1, x_2, \dots, x_T , and for $f(x_t)$ in equation (18), the attention weights α_t and the maximum value β_t of the vector are given by

$$\beta = \max(\exp(f(x_1)), \exp(f(x_2)), \dots, \exp(f(x_T))) \quad (19)$$

The cosine similarity equation for vectors \mathbf{a} and \mathbf{b} is as follows:

$$\text{Cosine_Similarity} = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (20)$$

The weights α_t of the context vector \mathbf{C} in equation (19) are as follows:

$$\alpha_t = \frac{\exp(f(x_t)) \cdot \beta}{\|\exp(f(x_t))\| \|\beta\|} \quad (21)$$

where $\|\exp(f(x_t))\|$ is the normalized value of the vector of weights.

If the frame is one-dimensional, the max value is scalar. In this case, the similarity is as follows:

$$\alpha_t = |\exp(f(x_t)) - \beta| \quad (22)$$

Then, the maximum value of the vector is chosen to give significant attention to the values of the vector compared to others. Finally, MWS will be applied to text and on the speech branches to compare its efficiency with that of the SoftMax function.

3.3.4.4 Multimodal Approach

Many studies have been developed for deep learning multimodal approaches [71, 72, 124]. We chose to merge the trained models for text and speech models in this study by forwarding the concatenated layer to the output layer for classification. This is a joint representation multimodal approach that makes use of the shared space in a neural network¹⁹. In Figure 9, the dotted lines indicate the merging layer that combines

¹⁹ More details are discussed in Chapter-2.

the two speech and text models from the different branches. Different combinations of speech and text models are merged until we achieve the best accuracy. In equation (17), a merged dense layer concatenates the dense activation output from text and speech branches, as shown in equation (23).

$$D_{Merged_Dense} = Concatenate(D_{Speech}, D_{Text}) \quad (23)$$

Finally, the merged dense layer is forwarded to the classifier shown in equation (18). The dense layer size is the total number of units for both speech and text layers.

3.4 **Conclusion**

This chapter discussed the study's methodology and the guidelines of the proposed experiment to achieve the study objectives. The research strategy focuses on empirical observation and inductive conclusions from the experimental results. The framework proposes several neural network combinations, to compare the results and reach the best accuracy among them. The neural network configuration follows the final results in similar studies, and is expected to have high classification accuracy for productivity measurement. This study uses five stages for speech and text processing: 1) speech diarisation, 2) speech recognition, 3) feature extraction for speech and text, 4) data modelling and 5) classification. Arabic language support is a critical part of the study, to ensure the possibility to measure performance based on the Arabic corpus. In machine learning, data validation is the bottom line of any study, verifying and validating the outcomes. The next chapter reveals the experiment results and discusses the study findings and conclusions.

Chapter 4: Research Experiment and Findings

4 The Experiment and Findings

4.1 Introduction

Chapter 3 discussed the research strategy and the methodology of productivity measurement. It proposed several modelling approaches, considering CNNs, BiLSTMs, and attention and global-max-pooling layers. The study comprises five stages: speech diarisation, speech-to-text transcription, feature extraction, data modelling, and classification. Chapter 4 illustrates the research design and setup to achieve the study objectives. Also, it elaborates upon the experiment with more detailed procedures, considering different data sets for speech recognition and productivity modelling. The first data set is used for acoustic modelling, which is required to transcribe the speech into text (speech recognition). The second uses the productivity data set, which divides the call centre calls subject to evaluation into productive and nonproductive ones. Data preparation consists of the annotation process, which is a critical part of the experiment because it is performed manually. Manual annotation can be biased due to raters' perceptions, and this can compromise the aim of reaching an objective evaluation. Diarisation is also challenging due to overlapping speech and the low sampling rates (8KHz) of the phone calls, which can mislead the algorithm into split the speakers. Speech recognition shows similar issues because of the Word Error Rate (WER), which impacts the accuracy of transcribed text and, in turn, affects productivity modelling because of Out-Of-Vocabulary (OOV) words.

Several feature extractions methods (MFCC/LLD) will be used in the study to test classification performance. The modelling is performed over different neural network structures to determine the best accuracy among different networks configurations. Considering that, the study keeps the network configuration fixed in order to smoothly and accurately compare the results with previous studies. Hence, the experiment aims to outperform previous studies through its extended features, network structure and multimodal approach.

This chapter tries to achieve the following study objectives:

- **Demonstrate different machine learning classification approaches** to classify the phone calls recorded in call centres and the corresponding text, leading to a fully automated productivity measurement process.
- **Compare the multimodal model with previous models** mentioned earlier and make future research recommendations in different domains.

The chapter should provide answers to the study question,

- **Can the proposed multimodal approach provide better classification performance than separate speech and text processing?**

4.2 *Research Design and Plan*

This research aims to classify the phone calls recorded in call centres as productive or nonproductive to evaluate agent performance. The research design is built on the conceptual model described in Chapter 3 and requires an appropriate data set to validate the outcomes. The sample data will be collected from a real call centre environment in Egypt to validate the model and generalise the outcomes. The ethical considerations are challenging because of the privacy of the data: the consent of the participants should be granted. Thus, the experiment will use a private data set from Luminous technologies²⁰ collected for research purposes. The annotation process is subjective, as mentioned in Chapter 3, and statistical analysis is required to ensure the consistency of annotation among raters. Kappa and Krippendorff's methods are usually used to determine the agreement between two or more raters through the annotation process [125]. Cross-validation and F1-scoring measures are mandatory to avoid data annotation bias and validate model accuracy. For the inductive approach, plotting graphs are required to imply the productivity aspects that contribute further to performance evaluation.

²⁰ info@luminous-technologies.com

Luminous technologies provides a data set and a machine learning environment for training the models. It requires high computational power in servers and graphical processing units (GPUs) to build the DNNs and attempt different NN approaches in a shorter time. The training is performed over Luminous servers with various types of GPUs: dual-slot Tesla k80, two NVidia GPUs, Quadro M4000, and M5000. The programming language is Python version 3.8 in a Linux environment²¹. The classification models were built using Keras 2.4.3 and TensorFlow 2.3, based on the Conda virtual environment. The script uses the Keras library, which is an open-source neural network library written in Python. It can easily run on top of different machine learning backends, like TensorFlow, Microsoft Cognitive Toolkit, R, and Theano [126]. The setup structure is illustrated in Figure 11.

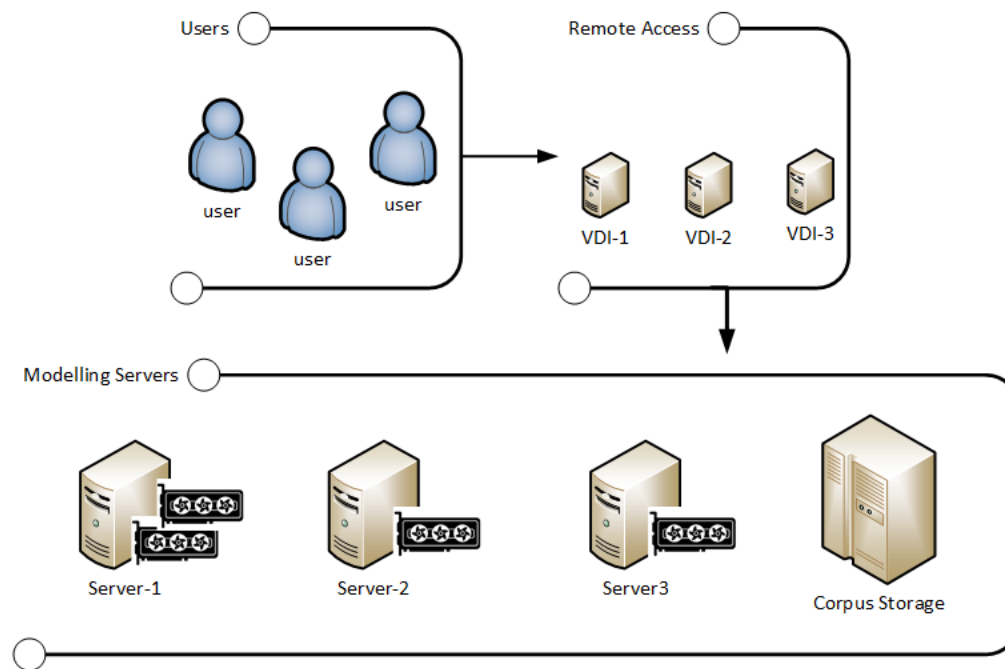


Figure 11: Server Structure

The modelling servers are connected via a unified storage server. The storage server hosts the corpus of speech recognition and productivity data modelling sets²². This is necessary because the corpus requires a great deal of storage for the recorded calls. Furthermore, the extracted feature files and modelled data files require storage as well.

²¹ Linux operating systems Ubuntu 15, 18, 20, Mint 20.

²² The data sets are discussed in next section

The users have a virtual remote desktop (VDI) to access the servers remotely and securely. Each server has a shared folder on the storage server for training the data and posting the generated models and classification output.

4.1 *Data*

Two data sets will be used in the study. The first is for the acoustic modelling of speech recognition. It comprises 1200 hours of the Aljazeera broadcast news TV corpus collected by QCRI [111]. The data were collected and transcribed by QCRI [127]. The duration of an episode is typically 20–50 minutes, which can be split into three broad categories: conversation (63%), interview (19%), and reportage (18%). Conversational speech includes multiple dialects and overlapping talkers, in typical political debate and talk show programs. The recordings come from TV programs using the fluent form of Arabic called Modern Standard Arabic (MSA)²³. It has been roughly estimated that more than 70% of the speech is MSA. The rest is spoken in different Dialectal Arabic (DA), such as Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR). The corpus contains 12 domain classes: politics, economy, society, culture, media, law, science, religion, education, sports, medical, and military.

The second data set is for productivity modelling. Securing the availability and accessibility of the data set is crucial tasks that requires exposing the contact centre's internal and confidential data for model training. Hence, ethical approval has been granted for collecting an experimental corpus for research purposes from a real estate call centre located in Egypt. A VoIP call centre with a built-in call recording system was used between 2014 and 2015 to collect real calls over landline phones, with a sampling rate of 8 kHz. The selected random calls consist of seven hours and over 30 calls (14 min per call on average), which is considered adequate compared to similar studies [63, 128]. The corpus comprises six different agents between 25 and 35 years old: two females and four males. The calls were previously diarised in [128], so we know the talking time is 40% for females and 60% for males. The naming convention for the recorded calls is built from the metadata in the form Date, Time, Agent ID,

²³ This was developed in the Arab world in the late 19th and early 20th centuries. It is the language used in academia, print and mass media, law and legislation, though it is generally not spoken as a first language, similar to Classical Latin – Wikipedia.

Speaker ID (by the diariser), the call direction, Inbound or Outbound, so the wave file name takes the form 'DATE-TIME_AGENT-ID_SPK-ID_CALL-DIRECTION(INBOUND-OUTBOUND).wav'.

4.2 Procedures

For simplicity, the experiment will be divided into stages, and each stage will discuss the activities in individual sections denoted a capital letter. This section is concerned with the tools, toolkits, and algorithms used for each step in the proposed methodology. All the algorithms described in the previous chapter are mandatory for conducting the experiment. However, not all of them are part of the study mentioned with the corresponding references without details. Figure 12 illustrates the experiment stages, indicated in dotted boundaries.

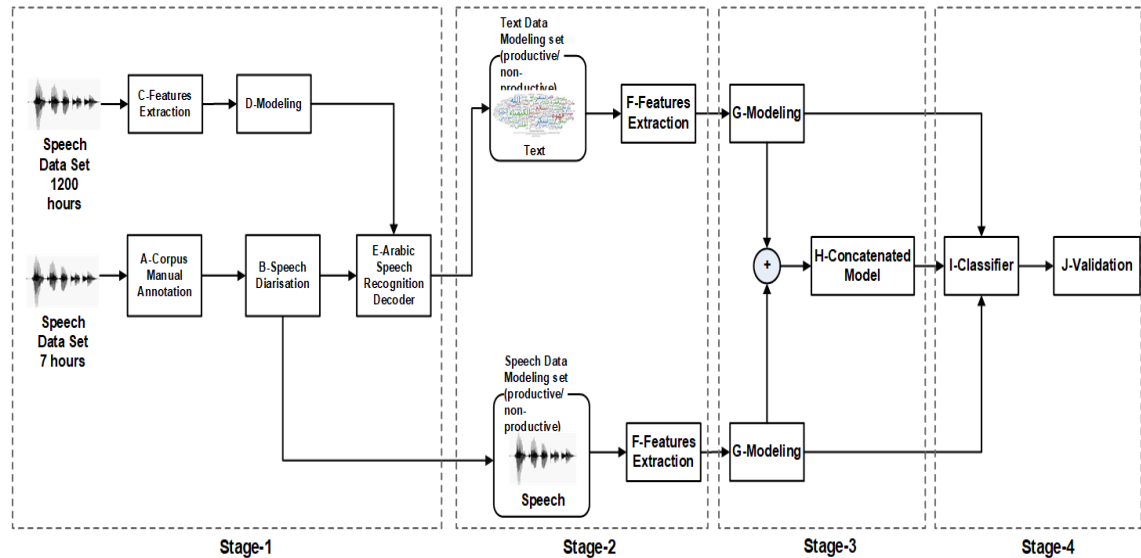


Figure 12: Experiment Procedures

4.2.1 Stage 1: Data Preparation

It was mentioned in the Data section that there are two types of data: (1) the 1200 hours of data for speech recognition and modelling, used to transcribe the 7 hours of speech data into text, (2) the experimental seven hours of data for productivity measurement. Data preparation includes manual annotation, speech diarisation, acoustic model feature extraction, speech modelling, and text transcription (decoding).

A- Manual Annotation of Recorded Calls

The recorded calls should be annotated as productive/nonproductive for training. The annotation is performed manually. The annotators must focus on the technical core of the call and ignore other subjective factors, like communication skills, the speaker's tone, etc. This step requires an orientation session for the raters to understand the concept and avoid misjudgement. After the manual annotation, the calls identified by unique numbers will be manipulated using Kappa Alpha for two raters or Krippendorff's alpha for more than two. The consistency of the raters should be around 80% for this to be considered a valid annotation [125].

Three raters are selected, respectively with two, five, and eight years of experience in call centres. They are full-time supervisors with experience in the real estate domain. They had an orientation session on the technical core that can be identified by listening to recorded calls. They started rating the recorded calls as nonproductive (labelled with number 0) or productive (labelled with number 1). They were requested to move the audio file to a folder belonging to the designated class (productive folder, nonproductive folder). The file labels (0,1) have been listed in a text file and processed using a Python script based on the Natural Language Toolkit (NLTK) and its library 'Agreement'. The text file is comma-separated, indicating the file name, rater-1, rater-2, and rater-3, as shown in Figure-13.

```
Filename, rater-1, rater-2, rater3
02-098.wav,0,0,0
02-111.wav,0,0,0
04-075.wav,1,1,1
04-137.wav,1,1,1
02-028.wav,1,1,1
02-076.wav,1,1,1
02-008.wav,1,1,1
04-029.wav,1,0,1
01-078.wav,1,1,1
01-121.wav,1,1,1
02-066.wav,1,1,1
04-042.wav,1,1,1
03-030.wav,1,1,1
03-050.wav,0,1,0
```

Figure 13: Text file combining the raters' annotations

The calls have been sliced and diarised, as discussed in the Speech Diarisation section. The rater agreement (Krippendorff's alpha) over all 503 files is 79.1%, which is considered acceptable. The annotation can thus be accepted as objective.

B- Speech Diarisation

The diarisation step extracts the agent part for evaluation and excludes the customer or third party. The challenging part of speaker diarisation is its accuracy, because the algorithms are subject to error rates for similar segments. Hence, the calls are sliced using the Sox toolkit²⁴ into smaller segments of 20 seconds each and adjusted on the silences when possible. This also helps achieve better diarisation and speech recognition [81, 95, 98]. The data is diarised using the SIDEKIT tool (S4D for short) [109]. The diarisation is conducted using cascaded methods that involve the following steps: Bayesian Information Criterion (BIC), Hierarchical Agglomerative Clustering (HAC), and Viterbi decoding. The segmentation process sometimes fails to detect the speakers because of the low sampling rate (8 kHz). Hence, the diarisation is performed in two steps. The first step is to upgrade the sampling rates to 16 kHz, using Sox to extract the segmentation boundaries. The second step applies the segment information extracted from the previous step to the 8 kHz corpus, as shown in Figure 14.

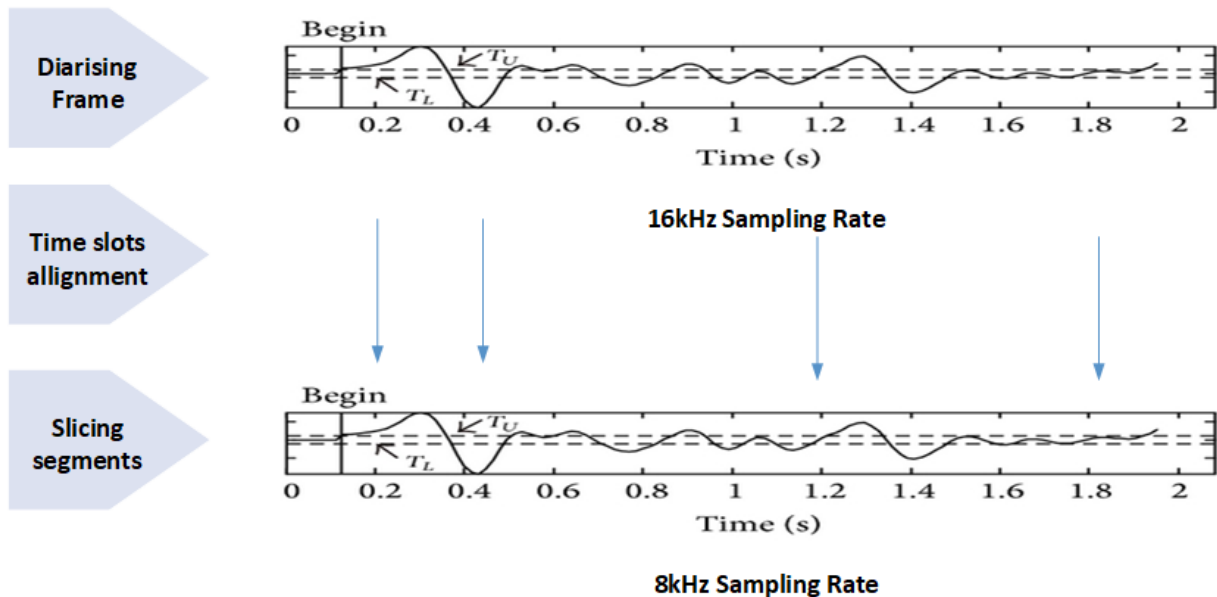


Figure 14: Segmentation boundary alignment between two sampling rates.

²⁴ The upgrading and slicing of segments is performed using the SOX toolkit, <http://sox.sourceforge.net/>.

C- Feature Extraction for Speech Recognition

The front-end preparation consists of the transliteration of the input data from Arabic to Latin to prepare for speech recognition training. The Latin characters are transcribed into numerical values to prepare for neural network processing. Then, the audio features are extracted to build the input data matrix. A sample of the character set is shown in Table 5.

Arabic Letter	أ	د	ش	ل	ك	#	*
Transliteration	ga	d	sh	l	k	#	*
English equivalent letter	A	D	SH	L	K	#	*
Alias	2	10	22	39	31	0	20

Table 5: The Arabic letters and corresponding conversions

The transliteration process maps each letter from Arabic to the corresponding Latin character. We added spaces between each character because the character length can be different (Arabic characters can be transliterated into either one or two Latin characters, as shown in example (1)). We use the hash symbol (#) to indicate the start of a sentence, the star symbol (*) for the end of a sentence, and the separator symbol (|) for spaces between the words. These special characters help the decoder detect sentence and word boundaries.

Example 1. Example of the transliteration of an Arabic statement into Latin characters:

حيث تقوم الحياة السياسية على التعددية الحزبية

hh y th | t q w m | a l h h y a t | a l s y a s y t | a a l a | a l t a a d d y t | a l h h
z b y t *

The transliteration process transforms the right-to-left statement (Arabic writing direction) into a left-to-right one (Latin). Buckwalter²⁵ is a powerful open-source tool for Arabic to Latin transliteration. However, we built our look-up list to easily edit the character set, which contains 39 characters. Also, the experiment uses different character combinations to determine the best transcription results. This is because different Arabic letters can represent the same English letter, e.g. ‘A’ ~ ‘أ, ا, or ء’. Then, the transcription is converted into the corresponding number in the mapping list, as shown in Table 5.

Example 2. Numerical transformation

hh y th | t q w m | a l hh y a t *
 1 14 37 12 39 11 30 36 33 39 7 32 14 37 7 11 20

The feature extraction in this study (speech recognition part) is based on a filter bank (FB) instead of the Mel Frequency Cepstral Coefficient (MFCC). The empirical results in previous work [56] show that an FB outperforms an MFCC in speech and speaker recognition technology [129]. FB acts as a bandpass filter for the audio signal in the frequency domain. It works by projecting features into a higher-dimensional space in which classification can be easier [98]. The speech representation uses a Log-Fourier-transform-based filter bank with 40 coefficients (plus energy) distributed on a Mel-scale, together with their first and second temporal derivatives, resulting in a 123-element feature vector. The features are pre-processed, with a zero mean, unit variance, and acoustic context information. The context window is ± 10 frames before and after the current frame (21 frames). Hence, the feature dimensions are 123x21.

D- Acoustic and Language models

The training is performed over 1200 hours, with a 15-gram language model and 20 iterations (epochs). The high gram order (15) is because the transcription is based on

²⁵ <http://www.qamus.org-/transliteration.htm>

a character level; 15-grams is thus mostly equivalent to 4-grams in a word-based model. The feature extraction of the training and testing data sets is carried out into three files, which are required for the training and decoding process: 1) Alias is the text transcription into numbers, 2) Key is the number of frames extracted per audio file, and 3) Feat is the audio features extracted. The experiment proposes two types of language model, the Pseudo Language Model (PLM) and the Real Language Model (RLM). The PLM is collected from the data set to reduce language model perplexity and adjust the decoder parameters. Afterwards, the PLM is replaced by an RLM of 980k unique words from Aljazeera broadcasts, supported by conversational text collected from Twitter in the same domain, to minimize Out-Of-Vocabulary words (OOV). The RLM will be tested based on 7-, 9-, 14- and 15-gram orders for both the PLM and RLM, with modified Kneser-Ney smoothing using the KenLM toolkit²⁶.

E- The Arabic Speech Recognition Decoder

The beam search decoder is used in the transcription engine to decode the sentences at a character level. The optimum values of the parameters α and β of the acoustic and language models should be adjusted for the best transcription accuracy. As per previous studies [56, 58], the estimated parameters of scoring equations (5),(13) are: $\alpha = 5$, $\beta = 3.8$, and beam length =150.

4.2.2 Stage 2: Feature Extraction

F- Feature Extractions for Text and Speech

The features are extracted for speech productivity modelling using MFCC and LLD. The experiment is performed over the two feature types and compares the effect of feature extension accordingly. The Essentia toolkit has been used to extract 13 MFCCs [130] segmented into 25 ms frames with a 10 ms shift. The OpenSMILE toolkit for speech feature extraction generated 65 features [131]. This is based on the INTERSPEECH 2016 Computational Paralinguistics Challenge (2016 COMPARE) [115]. It includes energy-related, spectral-related, and Low-Level Descriptors (LLDs),

²⁶ <https://kheafield.com/code/kenlm>

including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness.

The text features use a word embedding layer for the indexed words resulting from the transcribed Arabic text. The indices will then be handed to the embedding layer (the second layer following the input layer) to generate the corresponding training weights and forwarded to the neural network.

4.2.3 Stage-3: Data Modelling

G- Data Modelling for Text and Speech

The neural network configuration and structure and the corresponding hyperparameters are significant for accuracy of the final classification. The text and speech modelling structure has been discussed earlier and shown in detail in Figure 15. The most important neural network hyperparameters are the number of layers and nodes, the learning rates, and the dropout rates [76]. Optimization techniques can find the optimum hyperparameters to reach the best classification performance [132]. However, fixed hyperparameters have been selected and applied in this experiment, for several reasons. First, the proposed hyperparameters follow the configuration defined in the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [133], which various previous studies have used for emotional recognition and performance measurement [63, 128].

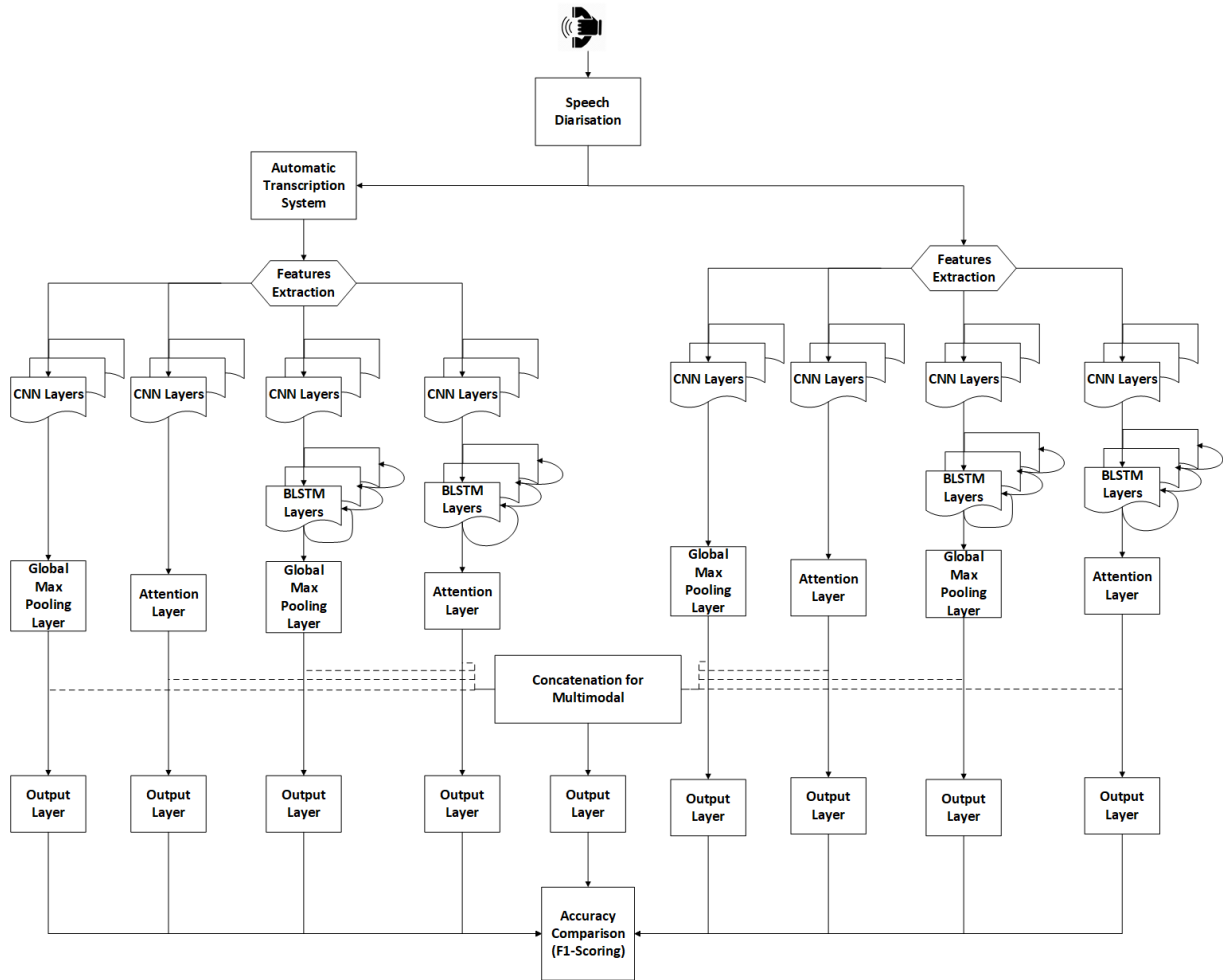


Figure 15: Detailed Neural Network Modelling Structure

The second reason is to compare results to emphasise the efficiency of the multimodal approach as compared to previous studies. The proposed experiment hyperparameters are defined in Figure 16. The upper branch is for speech processing (CNNs), with 13 or 65 input features. The lower branch is for text (CNNs-LSTMs) with an input layer of 128 words (max length). Several combinations of neural network structure are used to find the best accuracy.

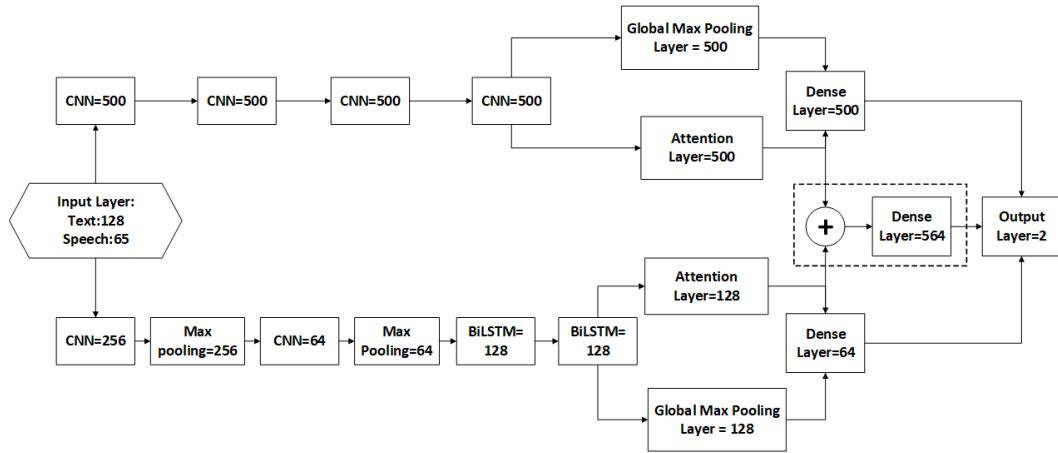


Figure 16: The experiment parameters.

H- The Multimodal Data Modelling

This step is required to increase the classification accuracy by combining (merging) the models at the final layer in Figure 16. The dotted box is the merged dense layer of batch size=32 with the following hyperparameters:

$$\begin{aligned} \text{Merged_Dense}(\text{Batch_size} = 32, \text{Param} = 564) \\ = \text{Conc}(\text{DSpeech}(32,500), \text{DText}(32,64)) \end{aligned}$$

The merged dense layer is forwarded to the output layer for the binary classification.

4.2.4 Stage-4: The Classification and Validation

I- Classification

The final layer for binary classification. Its size is two units, for productive/nonproductive classes. It gives the values 0 (nonproductive) and 1 (productive).

J- The Validation.

Data modelling starts by splitting the data randomly into training (80%) and test (20%) sets²⁷ to determine the optimum training parameters w . The test set is smaller because it is required only for model verification. Every iteration (epoch) is a one-time training of the whole data set to update the model parameters. The data set is split into batches

²⁷ It can be 90% for training and 10% for validation, depending on the data set size and the task.

that help train and update the parameters on smaller chunks of data. The training keeps iterating until it reaches the optimum weights of the model. The model created is called ‘Fold’. At this point, part of the data had not been trained (test data), so the data was shuffled and re-split into training and test sets. This process is called cross-validation and serves to recompose the training and test sets to the corresponding folds. The models created are denoted by ‘Fold_Number_iterations_accuracy’. The experiment has five folds due to the corpus size (7 hours) [134].

The annotation process may lead to an imbalanced data set when the size of the annotated classes is quite different from a class to another. An imbalanced data set may bias accuracy, favouring the largest class [69]. Imbalanced data should be re-adjusted with a bigger corpus and re-annotated. Because the experiment is limited to only seven hours, other statistical techniques are used to measure classification accuracy, such as the F1 score [135], a measurement of model accuracy on a data set. This score is based on the average precision and recall of the resulting label as compared to the annotated class [136]. Referring to hypothesis test cases, TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative. Accuracy is the ratio of the correctly predicted observations to total observations, while Precision is the ratio of the correctly predicted positive observations to the total predicted positive observations, as shown in equation (24). Recall is the ratio of correctly predicted positive observations to all observations on the same class, as shown in equation (25). The F1-score is the weighted average of the precision and recall, as shown in equation (26).

$$Precision = \frac{TP}{TP+FP} \quad (24)$$

$$Recall = \frac{TP}{TP+FN} \quad (25)$$

$$F1 - Score = 2x \frac{(Precision \times Recall)}{Precision + Recall} \quad (26)$$

4.3 **Experimental Results**

The experiment followed the procedures mentioned previously. The next subsections detail the experiment's milestones: speech recognition modelling, speech productivity modelling, text productivity modelling, and the multimodal approach.

4.3.1 **Speech Recognition Acoustic Modelling**

The 1200-hour data set was trained and tested using a development set provided by QCRI [111]. QCRI provided overlapping and non-overlapping data sets for testing. Several studies contributed to the QCRI challenge: QCRI (1200 hours), LIUM (650 hours), MIT (1200 hours), NDSC (680 hours), and Seville (1200h). The lexicon-free speech recognition system achieved a WER of 12.03% for non-overlapping recordings and a WER of 22.89% for the overlapping data set. The gap between the overlapping and non-overlapping data sets is around 11%, which is quite high compared to other studies (3%–6%) [111]. This implies that while the RNN-CTC lexicon-free system achieves competitive results for non-overlapping files, it shows poor immunity to cross-talking speech as well as to noise, as shown in Table 6. The 15-gram language model consumes a high amount of computational resources and time due to longer gram probability calculations. It is therefore recommended to find other alternatives for language model generation, like RNNs.

Furthermore, the WER of 12% raises an issue in productivity modelling because of OOV. For instance, the incorrectly transcribed words may be misinterpreted in the productivity text modelling and multimodal accuracy. The text model accuracy is equal to $(\text{Text_model} \times 88\% \text{ WER})$. The automated transcription system should be improved in the production environment by modelling adaptation to the acoustic and language models. Also, the data set should be collected from a call centre rather than a TV news data set. The 7 hours of productivity modelling were too small a corpus for speech acoustic modelling, and must therefore be extended for speech recognition to a minimum of 100 hours. However, to verify the productivity modelling and isolate the WER issue, the transcription was manually revised to correct the mistranscribed text (WER=0).

Acoustic Model	Language Model	System/Affiliation	WER Overlapped	WER non-Overlapped
TDNN²⁸, LSTM, BiLSTM	Tri-grams-LMRNN	QCRI	17.3	14.7
DNN-TDNN	4-grams	LIUM	19.2	16.7
CNN-TDNN-GLSTM-HLSTM	4-grams-LMRNN	MIT	19.9	17.3
LSTM-TDNN	LMRNN	NDSC	23.8	18.2
RNN-CTC	15-grams	Seville	22.89	12.03

Table 6: MGB Challenge

4.3.2 Speech Productivity Modelling

The experiment includes two types of feature for speech modelling: MFCC and LLD. The MFCC is extracted using the Essentia toolkit for 13 features in the frequency domain. The OpenSMILE toolkit for feature extraction [131] has been used to collect 65 features. It used the speech configurations of INTERSPEECH 2016 Computational Paralinguistics Challenge (2016 COMPARE) [115]. The 2016 COMPARE includes MFCC, energy-related, spectral-related, and Low-Level Descriptors (LLDs), including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. The features are given in Chapter 3, Table 4. The resulting accuracy from training and validating the models is detailed in Table 7.

There is a significant improvement in speech classification when using LLD rather than MFCC, with the highest improvement being 8.4%. The resulting accuracy of CNN-BiLSTM is higher than that of the CNN model, as expected. This proves that

²⁸ (TDNN) Time Delay Neural Networks, Guided LSTM(GLSTM), Hierarchal LSTM (HLSTM), Large Margin RNN (LMRNN)

the BiLSTM can handle a more extensive sequence of data for processing than CNNs. However, the processing time of CNN is one-fourth that of BiLSTM, on the same setup. BiLSTMs consume much more time because of their recurrent units. The accuracy improvement when using BiLSTMs as compared to CNNs is less than 1%, making it questionable if the additional time and resources are worth such a minor improvement.

Classification Method	Features Type	Accuracy
CNN	MFCC	82.7%
CNN-Attention	MFCC	84.27%
CNN-BiLSTM	MFCC	83.55%
CNN-BiLSTM-Attention	MFCC	83.54%
CNN	LLD	90.1%
CNN-Attention	LLD	92.48%
CNN-BiLSTM	LLD	92.67%
CNN-BiLSTM-Attention	LLD	92.68%

Table 7: Speech Accuracy % per Model/Feature Type

The attention layer increased CNN's performance to a value very close to that provided by CNN-BiLSTM. Nevertheless, for CNN-BiLSTM, the attention layer does not affect, as the results before and after adding it are almost the same. This indicates that the attention layer provides higher performance with CNNs than with BiLSTMs. Moreover, the accuracy of CNNs with an attention layer is higher than that of the CNN-BiLSTM-attention approach with MFCC, but not with LLD. This means that the feature type affects the performance of the attention layer, especially with CNNs.

The Max Weight Similarity (MWS) in Table-8 indicates that MWS replaces the SoftMax function to show the difference in performance. The attention layer with MWS gives a slight improvement of 0.2% for CNN compared with SoftMax but about 0.18% less accuracy for CNN-LSTM.

Classification Method	Features Type	Accuracy
CNNs-Attention	LLD	92.48%
CNNs-Attention+MWS	LLD	92.88%
CNNs-BiLSTMs	LLD	92.67%
CNNs-BiLSTMs-Attention	LLD	92.68%
CNNs-BiLSTMs-Attention+MWS	LLD	92.25%

Table 8: Max Weight Similarity (MWS) versus SoftMax functions

4.3.3 Text Productivity Modelling

The word embedding layer has been applied to the transcribed Arabic indexed words. The generated dictionary is around 4k words, with a max stream length of 128 words. The same deep learning structure as in Figure 15 was applied, with the attention layer using SoftMax and MWS. The results were compared with previous text classification results [20] using Logit and SVM based on a bag of words. The results are reported in Table 9.

Classification Method	Features Type	Accuracy
Naïve Bayes	Bag of words	67.3%
Logistic Regression	Bag of words	80.76%
Linear Support Vector Machine (LSVM)	Bag of words	82.69%
CNN	Word Embedding	90.73%
CNN-Attention	Word Embedding	90.98%
CNN-Attention+MWS	Word Embedding	91.4%
CNN-BiLSTM	Word Embedding	89.87%
CNN-BiLSTM-Attention	Word Embedding	91.19%
CNN-BiLSTM-Attention+MWS	Word Embedding	91.12%

Table 9: Text Accuracy % per Model/Feature Type

Deep learning text classification using the embedding of words shows a significant improvement of 8.7% over the generative and discriminative approaches (bag of words). MWS is more accurate than SoftMax only for the CNN approach (0.42%). The same applies to the speech approach. CNN-BiLSTM is less accurate than the CNN-Attention model, matching the results from the previous section. This is because BiLSTM is more efficient for long data streams, which is not the case in short conversations in a call centre. Accordingly, the attention layer does not provide a significant classification improvement in CNN-BiLSTMs, as compared with CNNs.

4.3.4 Multimodal Approach (Speech + Text)

This step is required to increase classification accuracy by combining (merging) the models at the final layer. The dotted box in Figure 15 is the merged dense layer of the text and speech training branches. The highest and comparable accuracy models are combined where the lower accuracies have been excluded (Speech or Text). The results are reported in Table 10.

The multimodal approach provides better classification accuracy by combining the CNN-attention model for speech features and the CNN-attention model for text features; both are implemented with the MWS function instead of the SoftMax function. The Multimodal MWS approach allows for a 0.22% improvement in speech modelling and a 1.7% improvement in text modelling. The accuracy of the multimodal classification using MWS was slightly better than that using the SoftMax layer, by 1.34% for the same model. Findings reveal that the multimodal approach improves upon previous approaches and the uncombined models, as shown in Figure 17.

Text Model	Speech Model	Multimodal Accuracy
CNN	CNN	90.44%
CNN-Attention	CNN	90.10%
CNN	CNN- Attention	92.63%
CNN	CNN- Attention+MWS	92.90%
CNN-Attention	CNN- Attention	91.76%
CNN-Attention+MWS	CNN- Attention+MWS	93.10%
CNN	CNN-BiLSTM-Attention	91.80%
CNN	CNN-BiLSTM-Attention+MWS	91.90%
CNN-Attention	CNN-BiLSTM	90.36%
CNN-Attention+MWS	CNN-BiLSTM	91.10%
CNN-Attention	CNN-BiLSTM-Attention	91.00%
CNN-Attention+MWS	CNN-BiLSTM-Attention+MWS	91.10%

Table 10: Multimodal Accuracy % per Speech and Text Models

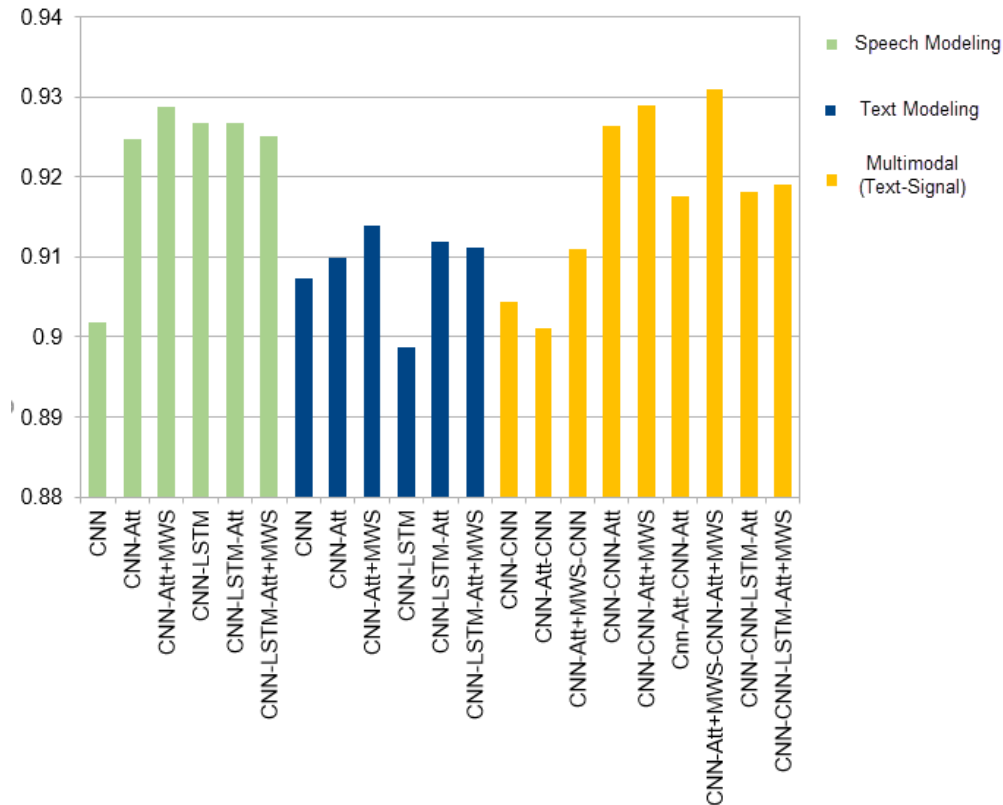


Figure 17: Model Accuracy per Approach

Table-11 summarizes the results using MWS and SoftMax functions for the models with the highest accuracy.

Methods	Speech Model	Text Model	Multimodal
SoftMax	92.68%	90.98%	91.76%
MWS	92.88%	91.4%	93.1%
Delta	0.2%	0.42%	1.34%

Table 11: MWS vs SoftMax Accuracy Improvement

4.4 *The Attention Weight plot analysis*

The attention layer focuses on the significant frame of the frame sequence for classification. The evolution of the attention weights at time t is illustrated in Figure 18. The attention layer generates the context vector of the frames computed by the input and the attention weights. By applying equation (14) to the attention weights, a graphical representation is generated for the attention weights versus the training segment frames. The attention weights may help illuminate linguistic and paralinguistic features in the call to glean some intuition about the performance behind the conversation. A sample of four selected graphs appears in Figure 18. This figure shows the attention weight curve for a sample segment of the call. The x axis represents frames per call, and the y axis represents attention weights as an output of the SoftMax function. High weight means the classifier pays attention to a significant frame in the remaining sequence. By analysing the peaks of the graphs of all segments, better knowledge of the features that impact productivity during a call can be obtained. The peaks at time t are annotated and matched with their corresponding wave segments. By listening to those wave files, the first observation is irrelevant for determining productivity and corresponds to situations such as drop call cadence, customer talk, and crowd noise. Customer talk is a small portion of such conversation that was not excluded accurately through the diarisation process. Crowd noise is the background noise of other agents in the call centres. The graph in Figure 19 illustrates two peaks, at frames 70 and 155, of the cross-talk between the customer and the agent.

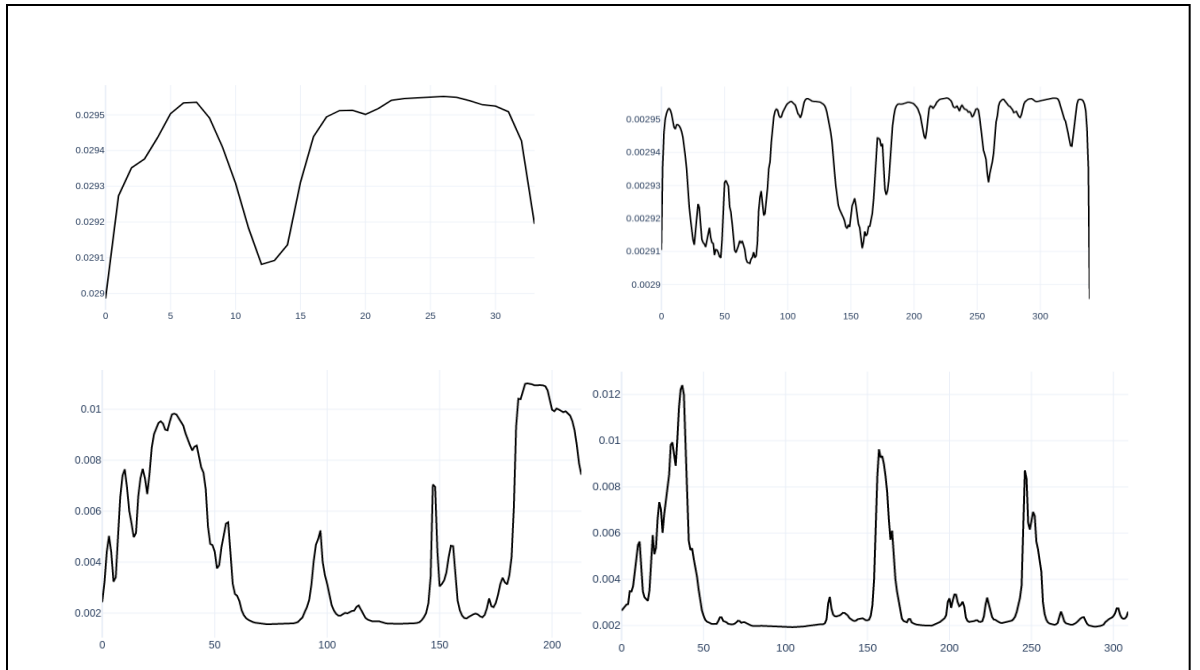


Figure 18: Attention weight graph²⁹

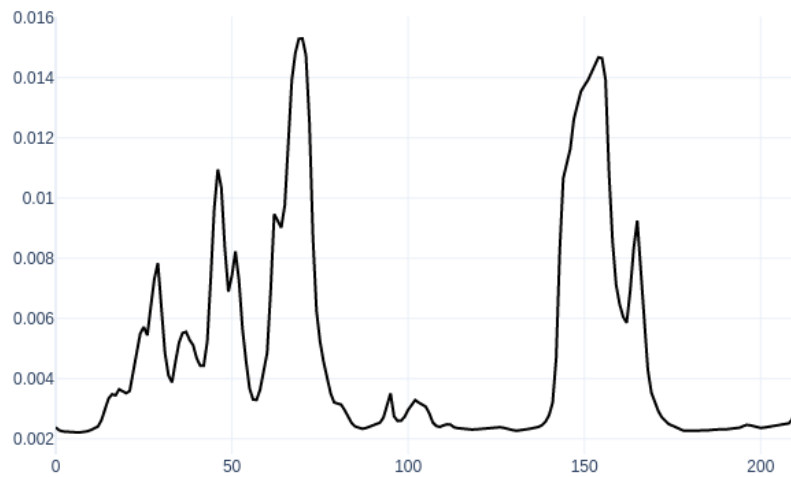


Figure 19: Attention weights for a sample segment

The second observation is that there are significant features, such as **paralinguistic** features, in the call. These features are summarized in the following points:

- Stuttering; ‘Umm Ahh’: This is common in nonproductive calls, repeated during the call, with an average duration from one to two seconds.

²⁹ The graph is drawn using plotly python library.

- Tone level: a high tone triggers a high level of attention for productive calls. The proper tone level is an important factor in call centres, indicating the wakefulness and enthusiasm of the agent. The primary reason for a customer to become frustrated is an insincere tone of voice from the person handling their query [137].
- Tone level also indicates a nonproductive call on the customer side. It indicates that customers are frustrated when they cannot get answers to their inquiries or complaints.

Therefore, the attention weights provide additional factors that influence the productivity factors, based on the inductive approach (mentioned in Chapter 3), and some intuition about how neural networks operate. These factors appear to be subjective, which is not recommended for objective evaluation of the technical core. However, they provide a practical approach to linguistic and paralinguistic analysis and implications for using the attention weight plots. The paralinguistic analysis can be extended to the technical part.

4.5 ***Modelling Performance***

Modelling time and prediction time are two basic aspects of machine learning. Modelling time is the time required to extract features and to train and validate the model. Prediction time is the time required to classify the data based on the pre-trained models. Big data and limited resources are challenges in performance modelling, for which it is important to figure out the required resources, effort, and time. Prediction performance is critical for real-time classification, where a few seconds' delays may affect, e.g. the response times of an online service. The performance evaluation framework is proposed as an offline prediction, so a delay of a few minutes is not an issue. Yet, modelling time is a critical factor due to the data size and the newly added calls that require continuous adoption for model enhancement.

When it comes to modelling performance based on processing time, modelling text is much faster than speech. The size and the numerical complexity of the speech features require much training time to reach the optimum weights (optimisation). Nevertheless,

it is essential to mention that speech recognition is a mandatory requirement before text classification, which consumes much modelling time. Automatic call transcription is a comprehensive process that consumes around 30 h for acoustic modelling. It depends on the approaches used (HMM/GMM, DNN, RNN, CNN, LSTM, etc.) and the computational resources. This study considered the minimum time as representing the maximum modelling performance. For instance, the shortest modelling time is that of Speech-CNN (~0.75 hours), ranked as 100% performance. Relative scaling is required to rank performance based on time, not accuracy. Table 12 summarises the time and performance of each modelling approach.

	Text Modelling					Speech Modelling				
	LSVM	Naïve Bayes	CNN-BiLSTM-Attention	CNN-BiLSTM	CNN-Attention	CNNs	CNN-BiLSTM-Attention	CNN-BiLSTM	CNN-Attention	CNNs
Time (Hour)	31	30	35	34	31	30.5	25.5	22	2.75	0.75
Performance %	2.42	2.50	2.14	2.21	2.42	2.46	2.94	3.41	27.27	100

Table 12: Time-Performance for each Approach

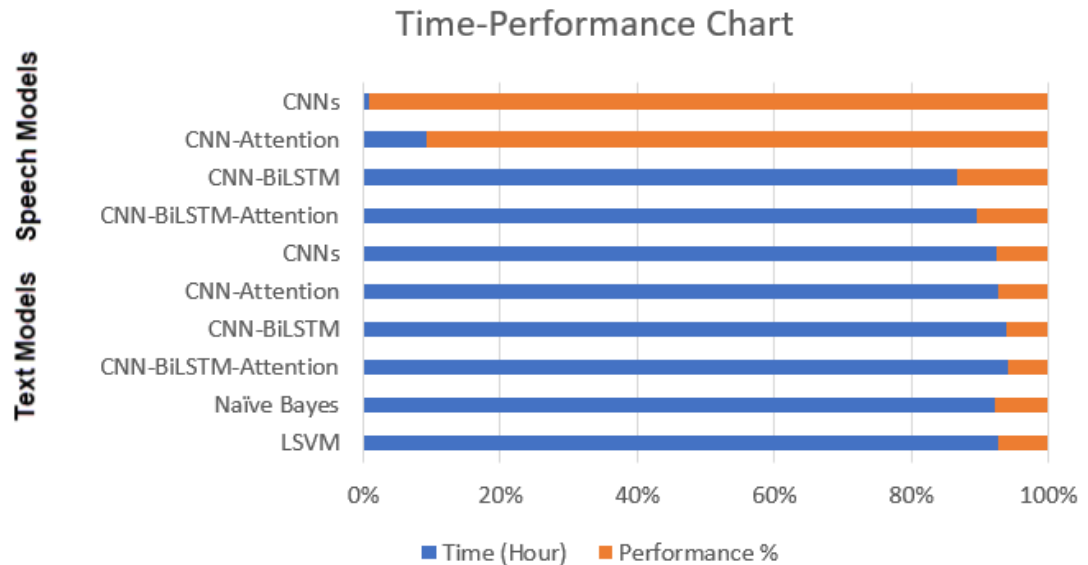


Figure 20: Time-Performance chart

Figure 20 illustrates performance vs time for each model. A longer time (in blue) indicates weaker performance (in orange) and vice versa. The multimodal approach depends on both speech and text, so the modelling time is the longest in text modelling, as mentioned previously.

4.6 *Experimental Assumptions and Considerations*

The experiment depicts several classification models based on text, speech, and multimodality. Many factors affect the performance of the models, such as feature type, the number of features, the relevance of features, e.g. prosodic features, and the network combinations. The experiment has been performed under several constraints and considerations:

- 1- The experiment used predefined hyperparameters that follow previous studies. These configurations were fixed to compare the influence of the features and the network alternatives on modelling performance.
- 2- The speech recognition achieved 12% WER, which outperforms other systems on the same corpus. However, manual transcription was required to get the absolute network classification performance and avoid OOV words.
- 3- The experiment has been performed offline, which means the diarisation, speech transcription, and feature extraction are ready at once. In production, a well-defined mechanism must synchronise all activities, prioritise tasks, and get the classification done on the spot. The Keras-TensorFlow handles text and speech training simultaneously and is appropriate for the final classification layer for the modelling part.
- 4- The manual annotation is a challenging part of the experiment, intended to avoid subjective evaluation. It is a showcase of the process that ensures objective evaluation. However, it is an exhaustive process that requires the definition of more sophisticated algorithms for annotation, like unsupervised clustering using, e.g. K-Means, or automatic annotation. The models generated by the experiment can be reused to annotate a newly added corpus and extend the experiment outcomes. However, this process is critical for keeping the models' baseline within a considerable limit and avoiding deviation by the model.

Therefore, the previous assumptions should be considered when evaluating and applying this experiment for any further investigation. The experiment's limitations and future research avenues will be discussed in detail in the next chapter.

4.7 Conclusion

Chapter 4 discusses the procedures and experimental results of the productivity measurement. The experiment's procedures go through the following stages: 1) data preparation, 2) feature extraction, 3) data modelling, and 4) classification. Data preparation involves the annotation process carried out by three trained raters. Krippendorff's alpha has been applied to verify the agreement between raters (79.1%), which is deemed acceptable as an objective evaluation when the alpha is close to 80% (the percentage required in social science). The speech diarisation uses the S4D toolkit to split the speakers and focus solely on agent performance. Arabic speech recognition has been trained using a 1200h Aljazeera corpus from TV news and talk shows. The Speech model achieved 12% WER and outperformed other systems using the same corpus with different sizes. The second stage is feature extraction, where the text features are extracted using embedding layers and speech features are extracted, with 13 MFCC and 65 LLD features. The Essentia toolkit has been used for MFCC feature extraction. The OpenSMILE toolkit has been used to extend the speech features to include 65 LLD features based on INTERSPEECH 2016 configurations. The third stage is data modelling, where several combinations have been used for text, speech, and multimodal approaches. Finally, the fourth stage is data classification and validation using five-fold cross-validation with F1-scoring. The experiment conducted outperforms other systems. It shows that a DNN based on word embedding outperforms bag of words modelling based on generative and discriminative approaches by around (8.7%). The second finding is that the LLD speech features outperformed the MFCC features given the same network structure and configurations (8.4%). The accuracy of the final multimodal approach is 93.1%, an improvement of approximately 1.7% over text classification and 0.22% over speech processing. The Max weight similarity (MWS) method slightly improved accuracy as compared to the SoftMax function and is recommended for further investigation over different domains.

The next chapter concludes the study and highlights the study results, limitations, recommendations, and future recommended studies.

Chapter 5: Study Conclusion

5 Study Conclusion

5.1 Introduction

Call centres are one of the essential access gates allowing organisations to manage customers' needs. They have improved over five decades to better handle and support the customers. Even though contact centres started operating a long time ago, they remain a unique channel for communicating with customers smoothly and efficiently. As it is the critical touch point of any organisation to interact with customers, this research studied call centre performance because it has the following notable characteristics:

Energetic environment: Call centres are an energetic and densely packed environment where calls go back and forth through different channels based on time constraints.

Resources varieties: Call centres mix different resources, including human resources, technology, and processes.

Similarity: Unlike many industries, all call centres have an analogous structure, even with different businesses and technologies.

Great Improvement: Many studies over the past three decades were concerned with call centres from a different perspective; however, continuous development in this area from technological and business perspectives opens a door for further research.

Generalisable workplace: Call centres cover many aspects of business and technology, which helps researchers scale studies in the field to cover many other business challenges.

The precise performance evaluation linked to the call centre’s competitive advantage is significant. Therefore, this study has chosen to critically review the literature on call centres, define the gaps, and search for study contributions. The next sections conclude the journey of four chapters by exploring the following themes:

- 1- Research aims and objectives.
- 2- Research findings and outcomes
- 3- The methodological contribution
- 4- The practical contribution
- 5- Research limitations
- 6- Lessons learnt from the study
- 7- Future Research.

5.2 Meeting Research Aims and Objectives

The study's core objectives drive the direction of the research activities and guide the study outcomes. The study objectives are distributed, as shown in table 13:

#	The Objective	Chapter	Title
1	Objective 1	Chapter Two	Literature Review
2	Objective 2	Chapter Three,	Research conceptual Model Research Methodology
3	Objective 3	Chapter Four	Study findings
4	Objective 4	Chapter Four Chapter Five	Study findings

Table 13: Study Objectives

The study aims to better perform evaluation at the help desk and call centres to achieve a more objective measurement. The study objectives are

- **Objective 1: Conduct a critical literature review** in operations efficiency and quality of customer service in call centres, including the methods and technologies relevant to the study.

- **Objective 2: Build the multimodal conceptual model** for the call centre domain, considering relevant components like automatic transcription and performance evaluation for text and speech features.
- **Objective 3: Demonstrate the different classification approaches in machine learning technology** to classify the calls recorded at a call centre and the corresponding text, leading to the full automation of the productivity measurement process.
- **Objective 4: Compare the multimodal model with previous models** mentioned earlier and make future research recommendations in different domains.

Objective 1: Reviewing Literature

The literature has been reviewed to determine the research gaps and define the theoretical model. Several gaps have been determined from previous studies, which help draw the conceptual model and study methodology as follows:

- 1- **Study Gap 1:** Rule-based systems lack the ability to perform deep statistical analysis of performance measurement, as compared to machine learning algorithms.
- 2- **Study Gap 2:** The machine learning models that depend on structured data, data mining, or predefined factors overlook essential productivity measurement factors and limit their ability to measure performance objectively.
- 3- **Study Gap 3:** The machine learning approaches based on generative and discriminative approaches achieved low accuracy due to using a ‘bag of words’ approach or classification methodology.
- 4- **Study Gap 4:** The DNN performance modelling approach could outperform the legacy ML approaches. However, more investigation is required for performance evaluation using MFCC, as well as the extended features.
- 5- **Study Gap 5:** The multimodal approach has not been examined adequately in the performance evaluation and call centre domains, based on the literature reviewed.

- 6- **Study Gap 6:** Previous studies processed a pre-transcribed text but have not proposed automation of call centre performance measurement based on an automatic transcription system embedded within the solution.

Objective 2: Build a multimodal conceptual model for the call centre domain.

The second objective is to identify ways to overcome the research gaps by building a conceptual model of the DNN that combines several neural networks architectures and proposing a multimodal approach.

The objective has been achieved through the following points:

- Development of a full-fledged automatic Arabic transcription system to transcribe the recorded calls into text for evaluation.
- Building a DNN structure that deals with text and speech features.
- Extension of the Speech features from MFCC to LLD.
- Extraction of the Text features using embedding layers instead of a bag of words.
- Development of the Max Weight Similarity (MWS) function to improve the attention weights.
- Development of Multimodal neural networks based on a joint representation approach.

Objective 3: Demonstrate the different classification approaches in machine learning to fully automating the productivity measurement process.

The third objective has been achieved by automating productivity measurement using machine learning via the following activities:

- Agent performance evaluation is performed on the agent talking time for the recorded call.
- As per the literature review, state-of-the-art machine learning approaches are used for the best classification accuracy. The machine learning structures are the CNN,

BiLSTM, and attention layer. The speech and text models are compared to previous studies based on the new features, and are shown to give higher accuracy.

- The modelling process is performed through the manual annotation of the recorded calls, with statistical proof of consistency among raters.
- The feature extraction process is carried out to train the machine for parameter prediction based on the text and speech inputs.
- The speech recognition used for text transcription outperformed the other studies, based on the QCRI 1200h TV news/talk shows corpus.

Objective 4: Compare the multimodal model with the previous models mentioned earlier and make future research recommendations in different domains.

Several comparisons with previous studies have been conducted. Finally, the multimodal proposed structure achieved better accuracy than both speech and text models and previous studies using legacy ML.

5.3 *Research Findings*

5.3.1 **Methodological Contribution**

The conceptual model is built on the following best-of-breed algorithms and structures from previous studies:

1. Focusing on the technical part of agent performance,
2. Evaluating the modelling corpus objectively,
3. Transcribing speech to text,
4. Extending the features for both text and speech,
5. Modelling the data based on the best possible neural network structure,
6. Combining the best accuracy from text and speech using a multimodal structure.

The diarisation algorithm was proposed to split the agent's part of a call recording from the customer's. It is necessary to objectively evaluate the agent apart from any other subjective aspects, e.g. a frustrated customer. The manual annotation was

proposed under statistical constraints, using Krippendorff's alpha to avoid subjective annotation. It is important to measure the raters' agreement and avoid biased classification. Arabic speech recognition is presented to transcribe the speech to text. Then, word embedding, MFCC, and Low-Level speech Descriptors (LLD) are used as input features for the text and speech branches, respectively. The data modelling for speech and text was based on CNN, BiLSTM, and the attention layer. The multimodal approach was joint representation, using a shared space (model concatenation) for best classification performance. The Max Weight Similarity (MWS) function was deployed, and the results were compared to those achieved with the SoftMax function. It has been concluded that the CNN-Attention structure provides the best accuracy among the speech (MFCC-LLD), text (embedded words), and multimodal approaches. The attention layer performs better when combined with the CNN layers. Also, BiLSTM does not provide the best performance for short utterances, which often occur in the call centres. Furthermore, a slight improvement has been achieved using MWS rather than the SoftMax function, something worthy of more investigation in future research.

Data has been annotated manually and verified using Krippendorff's alpha for three raters with an agreement of 79.1%. The text has been transcribed using lexicon-free Arabic speech recognition. The acoustic model was based on a 1200h Aljazeera corpus and a corresponding language model collected from the corpus and from online crawling. The speech transcription system achieved 12% WER, which is an outstanding performance compared to previous studies. The features were extracted from text using a word embedding layer and speech from using Low-Level Descriptors (LLD). Following previous studies, the modelling was carried out on a cascaded CNN-attention for best classification accuracy. The experiment achieved accuracies of 91.4% for text, 92.88% for speech, and 93.1% for the multimodal approach. Some paralinguistic features derived from the calls are relevant to productive and nonproductive features, e.g. Stuttering 'Umm Ahh' is a nonproductive feature, while the tone level is productive. The experiment proves that machine learning modelling can automatically detect and classify the recorded calls into performance scale. The results have been compared to previous studies and shown to outperform them.

The study's main contribution is productivity measurement/classification on the pre-annotated corpus (productive and nonproductive). This opens a wide door for exploring the productivity factors that can be detected from calls rather than predefined factors that restrict the modelling systems to determine beyond. It assesses the annotation process as a critical challenge because it reflects the quality of the experiment outcomes. The annotation process based on the raters' agreement is essential to avoid subjective evaluation. Furthermore, the multimodal approach to best evaluate performance covers various aspect of call centre agent performance. Performance evaluation should conceptually cover many different aspects of performance rather than individual factors. The study would propose that including several performance models and combining them in a joint representation approach is most likely to achieve better accuracy, fair judgment, and objective evaluation. The study has proved that extended speech features (LLD) perform better than fewer ones (MFCC). The expansion of features from the vocal tract into the prosodic level highlights the conversational context adequately, where the models are much more robust in dealing with the calls.

This study contributes to improving the performance of the attention layer by using Max Weight Similarity (MWS). MWS re-adjusts the hidden weights from previous layers around the max vector(s). It helps the content vector pay more attention to similar values around the max vector and less attention to values at a longer distance from it. The MWS is proposed as a replacement for the SoftMax function; more details about their respective performance will be discussed in the next section.

There are complementary contributions worth mentioning. First, the lexicon-free speech recognition performs impressively, indicating the power of RNN-CTC in acoustic modelling. The character-based approach avoids common OOV issues and increases the beam search of the decoder (no dictionary). Second, upgrading sampling rates from 8kHz to 16kHz for phone calls gives better diarisation for splitting the speakers.

5.3.2 Practical Implications

The practical application of this study would see call centres adopt a form of performance evaluation based on the conceptual model presented here. Call Centres embrace various data sources, like recorded calls, chatting, emails, management meetings, evaluation forms, screenshots, CRM/ERP back-end systems, etc. Each of these sources of data can serve as input for one or more of the study models. Recorded calls are among the most common channels for evaluation in call centres [2, 10]. However, inputting various data source into the multimodal setup may lead to better evaluation than using recorded calls as a single data source.

Furthermore, the application can be smoothly extended to cover several evaluation levels, such as excellent, average, and poor performance. This is accomplished by extending the annotation classes and replacing the logit function with the SoftMax function at the classification layer. The rest of the structure remains the same, in terms of the text and speech features.

The bottom line of the practical application is to automate performance evaluation in a manner most likely to be fair in a high subjective environment (call centres). The application goes deep into the recorded calls (unstructured data) to determine performance for a huge number of calls. The application resolves critical problems in call centre operations arising from unfair judgment about agents' performance. Underestimating agents' performance, frivolous orientations, and low annual raises lead to emotional exhaustion [138], burnout [139], and high turnover [137], which impact both the business and the quality of customer service. Performance modelling keeps workers confident, loyal, and in positive well-being, with every piece of work monitored and evaluated based on a unified and robust baseline [35, 49].

From the customer side, the baseline also helps illuminate complaints by considering the attention weights and corresponding performance attributes. Spotting a specific part of the call instantly, along with the corresponding performance level, reduces the time and effort needed to determine the weak points that should be covered through vocational training. It also gives a clear picture of the customers' complaints, if they

are relevant to the agent's performance or to other factors like the company's product, processes, or policies [140].

Nevertheless, monitoring performance tightly has several drawbacks. As mentioned in Chapter 2, performance monitoring and evaluation apply the concept of the panopticon to the agents, like prisoners monitored by jail guards [33]. Unfortunately, this leads to exhaustion, burnout, and high turnover, as mentioned earlier. Agents' participation in the evaluation process is thus highly recommended, and evaluation results should be shared with the agent to provide feedback and justify the results. This helps the agents overcome their fears and proactively improve their performance, and will brilliantly help to enhance the ML modelling based on their experience and feedback.

5.4 **Research Limitations**

The study illustrates several classification models based on text, speech, and multimodality. Many factors affected the performance of the models, like feature type, feature size, the relevance of features, i.e., prosodic features, and the network structure. However, several limitations are highlighted for further research:

- 1- Machine learning based on supervised learning is limited to human subjective evaluation through annotation. This limitation is overcome by using many raters with a good reliability check (Krippendorff's alpha). However, rater agreement does not guarantee that the annotation is objective. So, the study can be assumed to measure performance based on the evaluation baseline of the raters.
- 2- The experiment used predefined hyperparameters taken from previous studies. These configurations were kept the same to compare the influence of the features and the networks alternatives on modelling performance. However, it does not show the effect on the results of optimising hyperparameters.
- 3- The study does not provide a mechanism to handle the effect of OOV on performance measurement due to speech recognition achieving 12% WER.

- 4- The experiment has been performed offline, so the diarisation, speech transcription, and feature extraction were ready at once. The study does not provide a well-defined mechanism to synchronise activities, prioritise tasks, and complete classification on the spot.
- 5- The annotation process is an exhaustive one that requires the definition of more sophisticated algorithms for automatic annotation.

5.5 Lessons Learnt from the Study

Through the research journey, there are lessons as an outcome of the current study. The lessons are categorised on the organisation level and research level. Organisations should have a secure and comprehensive system to evaluate performance over different aspects of workers, technology, and processes. Most organisations are concerned with abstract aspects of performance, but it is necessary to define a procedure that also includes organisational resources and the integrity of these resources.

Evaluation measurement is not a luxury but an essential procedure for fair judgement of organisational performance and competitive advantage. Building text or speech models for performance evaluation is the ultimate goal. However, building a comprehensive form of performance evaluation using different aspects of data is more important than focusing on one aspect, like recorded calls.

Technologies like machine learning open the door to improvement in learning about and detecting performance; however, human interaction and judgement are still the dominant factors in revising and weighing the results. Complementing the hypotheses and practices is essential to reach the truth.

Data collection and analysis are quite challenging for researchers, since the annotation and modelling processes are costly in terms of time, effort, and money. Statistical methods are tools to process data, but the interpretation, analysis, and implications of

the results rest on the researchers' shoulders. The most challenging factor in the research, from my perspective, is the researcher's motivation, which may lead to predefined thoughts about the results and inadvertent trial(s) in jumping to conclusions.

5.6 ***Future Research***

The study discussed performance evaluation through theoretical and practical models. The recommended future studies can be summarised as follows:

- 1- Applying hyperparameter optimisation and comparing the results with the current study. It is important to draw out the contribution of the network configurations in achieving the optimum results.
- 2- Reducing OOV. The transcription system achieved 12% WER. Generally, machine learning studies, and specifically speech recognition systems, never achieve 100% accuracy. Hence, OOV is still an issue affecting performance evaluation. It is recommended that the speech recognition acoustic model be trained on recorded calls from the call centre to achieve better accuracy and reduce OOV. Furthermore, using Natural Language Processing (NLP) algorithms, like words' synonyms (WordNet) or similarity based on a large vocabulary, would help reduce the OOV to a minimum.
- 3- Defining a mechanism to synchronise the experimental tasks of diarisation, speech transcription, and feature extraction, to complete classification on the spot.
- 4- Improving automatic annotation. Manual annotation is an exhausting process that requires automation. So, it is important to define more sophisticated algorithms for this task, like unsupervised clustering using, e.g. K-Means, or automatic annotation based on pre-trained models (one-time manual annotation). The experiment generated models that can be reused to annotate a newly added corpus and to extend the experiment outcomes. However, this process is critical to keeping the models' baseline within a considerable limit and avoiding deviation.
- 5- Extending the features of the text and speech. There are pre-trained models, like BERT and Word2Vec, that offer extended features, e.g. 768 features/vector.

There are also Arabic BERT models like QARiB, mBERT, and ArabicBERT [120]. These models support the Arabic language based on web corpora, but not the call centre domain. Therefore, it is worth testing these models and generating custom models based on call centre corpora for comparison.

- 6- Improving recognition accuracy. Speech recognition technology provides a wide area of research for this area, like self-training [141] for acoustic models with autogenerated labels and no prior transcription.
- 7- Investigating what is called performance behaviour, which is relevant to performance evaluation. This can be determined by applying machine learning classifiers to data collected from the employee's PC, describing their activities: browsing websites, active time, session time, and idle time.

6 References

- [1] A. Cosseboom, *Enable Better Service: A Customer Service Contact Center Story of Breaking Away from the Norm Through Creativity, Technology and Innovation*. Amazon Digital Services LLC - Kdp Print Us, 2019.
- [2] B. Cleveland, *Call Center Management on Fast Forward: Succeeding in the New Era of Customer Relationships*. ICMI Press, 2012.
- [3] C. Dormann and F. Zijlstra, "Call centres: High on technology—high on emotions," *European Journal of Work and Organizational Psychology*, vol. 12, no. 4, pp. 305-310, 2003.
- [4] ICMI. (2016). *ICMI / Call Center Training, Events, Certification, Resources, and Consulting*. Available: <http://www.icmi.com/>
- [5] R. Andrade, S. Moazeni, and J. Ramirez-Marquez, "Contact Center Operations Management Systems Architecture and Reliability," *Available at SSRN 3320821*, 2018.
- [6] J. Paul, A. Mittal, and G. Srivastav, "Impact of service quality on customer satisfaction in private and public sector banks," *International Journal of Bank Marketing*, vol. 34, no. 5, pp. 606-622, 2016.
- [7] R. L. Cardy and B. Leonard, *Performance management: Concepts, skills, and exercises*. ME Sharpe, 2011.
- [8] J. P. Liyanage and U. Kumar, "Towards a value-based view on operations and maintenance performance management," *Journal of Quality in Maintenance Engineering*, vol. 9, no. 4, pp. 333-350, 2003.
- [9] S. Brignall and J. J. I. J. o. S. I. M. Ballantine, "Performance measurement in service businesses revisited," 1996.
- [10] R. Rubingh, *Call Center Rocket Science: 110 Tips to Creating a World Class Customer Service Organization*. CreateSpace Independent Publishing Platform, 2013.
- [11] P. Reynolds, "Call center metrics: Best practices in performance measurement and management to maximize quitline efficiency and quality," *North American Quitline Consortium*, 2010.
- [12] J. C. Abbott, *The executive guide to call center metrics*. Robert Houston Smith Publishers, 2004.
- [13] L. Gil, G. Iddo, and Y. Dana, "Spending more time with the customer: service-providers' behavioral discretion and call-center operations," *Service Business*, vol. 9, no. 3, pp. 427-443, 2015.
- [14] J. A. Judkins, M. Shelton, and D. Peterson, "System and method for evaluating agents in call center," ed: Google Patents, 2003.
- [15] M. Wöllmer, "Context-Sensitive Machine Learning for Intelligent Human Behavior Analysis," *Universitätsbibliothek der TU München*, 2013.
- [16] C. C. Helper, "White Paper: Quality Management Automation – ROI Calculation Guide," ed, 2019.
- [17] M. Paprzycki, A. Abraham, R. Guo, and S. Mukkamala, "Data mining approach for analyzing call center performance," in *Innovations in Applied Artificial Intelligence*: Springer, 2004, pp. 1092-1101.

- [18] D. Carmel, "Automatic analysis of call-center conversations," vol. DOI: 10.1145/1099554.1099684 · Source: DBLP, A. R. Ron Hoory, Ed., ed. <https://www.researchgate.net/publication/221614459>, 2005, p. 8.
- [19] A. Ahmed, Y. Hifny, S. Toral, and K. Shaalan, "A Call Center Agent Productivity Modeling Using Discriminative Approaches," in *Intelligent Natural Language Processing: Trends and Applications*: Springer, 2018, pp. 501-520.
- [20] A. Ahmed, S. Toral, and K. Shaalan, "Agent productivity measurement in call center using machine learning," in *International Conference on Advanced Intelligent Systems and Informatics*, 2016, pp. 160-169: Springer.
- [21] A. Rychalski and A. Palmer, "Customer Satisfaction and Emotion in the Call Centre Context," in *The Customer is NOT Always Right? Marketing Orientations in a Dynamic Business World*: Springer, 2017, pp. 67-70.
- [22] S. Scheidt and Q. Chung, "Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation," *International Journal of Information Management*, 2018.
- [23] C. C. Helper, "White Paper: 2019 Contact Centre Trends You Need to Know," ed, 2019.
- [24] 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," *arXiv preprint arXiv:1706.00612*, 2017.
- [25] A. De Keyser, S. Köcher, L. Alkire, C. Verbeeck, and J. Kandampully, "Frontline Service Technology infusion: conceptual archetypes and future research directions," *Journal of Service Management*, vol. 30, no. 1, pp. 156-183, 2019.
- [26] C. Group, "Contact Center Satisfaction (CCSI) 2018 - CFI Group," ed, 2019.
- [27] Z. W. Lee, T. K. Chan, A. Y.-L. Chong, and D. R. Thadani, "Customer engagement through omnichannel retailing: The effects of channel integration quality," *Industrial Marketing Management*, vol. 77, pp. 90-101, 2019.
- [28] S. Kamel and M. Hussein, "Xceed: pioneering the contact center industry in Egypt," *Journal of Cases on Information Technology (JCIT)*, vol. 10, no. 1, pp. 67-91, 2008.
- [29] M. Boussebaa, S. Sinha, and Y. Gabriel, "Englishization in offshore call centers: A postcolonial perspective," *Journal of International Business Studies*, vol. 45, no. 9, pp. 1152-1169, 2014.
- [30] Oxford, "The Report: Egypt 2019," B. Group, Ed., ed, 2020.
- [31] Oxford, "The Report: Egypt 2016," R. Group, Ed., ed, 2017.
- [32] A. N. H. Zaied, A. H. Ali, and H. A. El-Ghareeb, "E-government Adoption in Egypt: Analysis, Challenges and Prospects," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 52, no. 2, pp. 70-79, 2017.
- [33] P. Bain and P. Taylor, "Entrapped by the 'electronic panopticon'? Worker resistance in the call centre," *New technology, work and employment*, vol. 15, no. 1, pp. 2-18, 2000.
- [34] S. Fernie and D. Metcalf, *(Not) hanging on the telephone: payment systems in the new sweatshops*. Centre for Economic Performance, London School of Economics and Political Science, 1998.
- [35] J. Wegge, R. Van Dick, G. K. Fisher, C. Wecking, and K. Moltzen, "Work motivation, organisational identification, and well-being in call centre work," *Work & Stress*, vol. 20, no. 1, pp. 60-83, 2006.

- [36] K. Breuer, P. Nieken, and D. Sliwka, "Social ties and subjective performance evaluations: an empirical investigation," *Review of managerial Science*, vol. 7, no. 2, pp. 141-157, 2013.
- [37] J. Suls and L. Wheeler, *Handbook of social comparison: Theory and research*. Springer Science & Business Media, 2013.
- [38] S. Sonnentag and M. Frese, "Performance concepts and performance theory," *Psychological Management of Individual Performance*, p. 1, 2003.
- [39] A. Frederiksen, F. Lange, and B. Kriechel, "Subjective performance evaluations and employee careers," *Journal of Economic Behavior & Organization*, vol. 134, pp. 408-429, 2017.
- [40] M. Ketokivi, "Point-counterpoint: Resource heterogeneity, performance, and competitive advantage," *Journal of Operations Management*, vol. 41, pp. 75-76, 2016/01/01/ 2016.
- [41] A. S. DeNisi and R. D. Pritchard, "Performance appraisal, performance management and improving individual performance: A motivational framework," *management and Organization Review*, vol. 2, no. 2, pp. 253-277, 2006.
- [42] P. E. Levy and J. R. Williams, "The social context of performance appraisal: A review and framework for the future," *Journal of management*, vol. 30, no. 6, pp. 881-905, 2004.
- [43] J. P. Campbell, R. A. McCloy, S. H. Oppler, and C. E. Sager, "A theory of performance," *Personnel selection in organizations*, vol. 3570, 1993.
- [44] G. T. Milkovich, A. K. Wigdor, and I. ebrary, *Pay for performance: evaluating performance appraisal and merit pay* (no. Book, Whole). Washington, D.C: National Academy Press, 1991.
- [45] J. Tirole, "Hierarchies and bureaucracies: On the role of collusion in organizations," *Journal of Law, Economics, & Organization*, vol. 2, no. 2, pp. 181-214, 1986.
- [46] P. R. Milgrom, "Employment contracts, influence activities, and efficient organization design," *Journal of political economy*, vol. 96, no. 1, pp. 42-60, 1988.
- [47] C. Prendergast and R. Topel, "Discretion and bias in performance evaluation," *European Economic Review*, vol. 37, no. 2-3, pp. 355-365, 1993.
- [48] G. C. Kane, "Are You Part of the Email Problem?," (in English), *MIT Sloan Management Review*, vol. 56, no. 4, p. 0, Summer Summer 2015 2015-07-20 2015.
- [49] S. Grebner, N. Semmer, L. L. Faso, S. Gut, W. Kälin, and A. Elfering, "Working conditions, well-being, and job-related attitudes among call centre agents," *European Journal of Work and Organizational Psychology*, vol. 12, no. 4, pp. 341-365, 2003.
- [50] B. Marr and A. Neely, "Managing and measuring for value: the case of call centre performance," 2004.
- [51] D. Holman, C. Chissick, and P. Totterdell, "The effects of performance monitoring on emotional labor and well-being in call centers," *Motivation and Emotion*, vol. 26, no. 1, pp. 57-81, 2002.
- [52] J. Anton, V. Bapat, and B. Hall, *Call center performance enhancement using simulation and modeling*. Purdue University Press, 1999.
- [53] D. Chicu, M. del Mar Pàmies, G. Ryan, and C. Cross, "Exploring the influence of the human factor on customer satisfaction in call centres," *BRQ Business Research Quarterly*, vol. 22, no. 2, pp. 83-95, 2019.

- [54] K. N. N. Perera, Y. Priyadarshana, K. Gunathunga, L. Ranathunga, P. Karunarathne, and T. Thanthriwatta, "Automatic Evaluation Software for Contact Centre Agents' voice Handling Performance," 2019.
- [55] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, 2006.
- [56] A. Ahmed, Y. Hifny, K. Shaalan, and S. Toral, "Lexicon free Arabic speech recognition recipe," in *International Conference on Advanced Intelligent Systems and Informatics*, 2016, pp. 147-159: Springer.
- [57] K. Perera, Y. Priyadarshana, K. Gunathunga, L. Ranathunga, P. Karunarathne, and T. J. I. J. S. R. P. Thanthriwatta, "Automatic Evaluation Software for Contact Centre Agents' voice Handling Performance," vol. 5, pp. 1-8, 2019.
- [58] A. Ahmed, Y. Hifny, K. Shaalan, and S. Toral, "End-to-End Lexicon Free Arabic Speech Recognition Using Recurrent Neural Networks," *Computational Linguistics, Speech And Image Processing For Arabic Language*, vol. 4, p. 231, 2018.
- [59] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 2013, pp. 273-278: IEEE.
- [60] D. Palaz and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," *Idiap2015*.
- [61] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107-116, 1998.
- [62] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602-610, 2005.
- [63] Y. Hifny and A. Ali, "Efficient Arabic Emotion Recognition Using Deep Neural Networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6710-6714: IEEE.
- [64] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.
- [65] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 2006, p. 4 pp.: IEEE.
- [66] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer science and Technology*, vol. 16, no. 6, pp. 582-589, 2001.
- [67] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *arXiv preprint arXiv:1911.00432*, 2019.
- [68] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312-323, 2019.
- [69] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth international conference on natural computation*, 2008, vol. 4, pp. 192-201: IEEE.

- [70] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Challenges and applications in multimodal machine learning," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, 2018, pp. 17-48.
- [71] S. E. Kahou *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," vol. 10, no. 2, pp. 99-111, 2016.
- [72] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.
- [73] T. Baltrušaitis, C. Ahuja, L.-P. J. I. t. o. p. a. Morency, and m. intelligence, "Multimodal machine learning: A survey and taxonomy," vol. 41, no. 2, pp. 423-443, 2018.
- [74] H. Meinedo and J. P. Neto, "Combination of acoustic models in continuous speech recognition hybrid systems," in *Sixth International Conference on Spoken Language Processing*, 2000.
- [75] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [76] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning* (no. 2). MIT press Cambridge, 2016.
- [77] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage, 2013.
- [78] S. K. D'mello and J. J. A. C. S. Kory, "A review and meta-analysis of multimodal affect detection systems," vol. 47, no. 3, pp. 1-36, 2015.
- [79] Y. Bengio, A. Courville, P. J. I. t. o. p. a. Vincent, and m. intelligence, "Representation learning: A review and new perspectives," vol. 35, no. 8, pp. 1798-1828, 2013.
- [80] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 49-56.
- [81] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A complete KALDI recipe for building Arabic speech recognition systems," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 2014, pp. 525-529: IEEE.
- [82] E. Othman, K. Shaalan, and A. Rafea, "Towards resolving ambiguity in understanding arabic sentence," in *International Conference on Arabic Language Resources and Tools, NEMLAR*, 2004, pp. 118-122: Citeseer.
- [83] K. Shaalan, M. Magdy, and A. Fahmy, "Analysis and feedback of erroneous Arabic verbs," *Natural Language Engineering*, vol. 21, no. 02, pp. 271-323, 2015.
- [84] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, p. 14, 2009.
- [85] F. Diehl, M. J. Gales, M. Tomalin, and P. C. Woodland, "Morphological decomposition in Arabic ASR systems," *Computer Speech & Language*, vol. 26, no. 4, pp. 229-243, 2012.
- [86] K. Shaalan, "A survey of arabic named entity recognition and classification," *Computational Linguistics*, vol. 40, no. 2, pp. 469-510, 2014.
- [87] K. Shaalan, H. M. Abo Bakr, and I. Ziedan, "A hybrid approach for building Arabic diacritizer," in *Proceedings of the EACL 2009 workshop on computational approaches to semitic languages*, 2009, pp. 27-35: Association for Computational Linguistics.

- [88] V. Radha and C. Vimala, "A review on speech recognition challenges and approaches," *doaj.org*, vol. 2, no. 1, pp. 1-7, 2012.
- [89] P. Motlicek, D. Imseng, B. Potard, P. N. Garner, and I. Himawan, "Exploiting foreign resources for DNN-based ASR," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1-10, 2015.
- [90] M. Attia, Y. Samih, K. F. Shaalan, and J. van Genabith, "The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database through the Detection and Lemmatization of Unknown Words," in *COLING*, 2012, pp. 83-96.
- [91] S. Young *et al.*, "The HTK book (for HTK version 3.5)," *Cambridge University Engineering Department, Cambridge, UK*, 2015.
- [92] S. Raschka, "Python Machine Learning," ed: Packt Publishing, 2015.
- [93] H. Hermansky, D. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 3, pp. 1635-1638: IEEE.
- [94] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82-97, 2012.
- [95] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *Adaptive Processing of Sequences and Data Structures*: Springer, 1998, pp. 389-417.
- [96] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 6645-6649: IEEE.
- [97] Y. Hifny, "Unified Acoustic Modeling using Deep Conditional Random Fields," *Transactions on Machine Learning and Artificial Intelligence*, vol. 3, no. 2, p. 65, 2015.
- [98] D. Yu and L. Deng, *Automatic Speech Recognition*. Springer, 2012.
- [99] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 1996, pp. 310-318: Association for Computational Linguistics.
- [100] V. Sudarsan and G. Kumar, "Voice call analytics using natural language processing," 2019.
- [101] B. Karakus and G. Aydin, "Call center performance evaluation using big data analytics," in *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, 2016, pp. 1-6: IEEE.
- [102] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, "Automatic analysis of call-center conversations," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 453-459.
- [103] M. A. Valle, S. Varas, and G. A. J. E. S. w. A. Ruz, "Job performance prediction in a call center using a naive Bayes classifier," vol. 39, no. 11, pp. 9939-9945, 2012.
- [104] S. Hudson, H. V. González-Gómez, and A. Rychalski, "Call centers: is there an upside to the dissatisfied customer experience?," *Journal of Business Strategy*, vol. 38, no. 1, pp. 39-46, 2017.

- [105] J. W. Creswell, *Research design: qualitative, quantitative, and mixed methods approaches* (no. Book, Whole). Thousand Oaks, CA;London,;: Sage, 2009.
- [106] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557-1565, 2006.
- [107] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4353-4356: IEEE.
- [108] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010, vol. 2010.
- [109] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier, "S4D: Speaker Diarization Toolkit in Python," in *Interspeech*, 2018.
- [110] A. L. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng, "Lexicon-free conversational speech recognition with neural networks," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.
- [111] A. Ali *et al.*, "The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition," *arXiv preprint arXiv:1609.05625*, 2016.
- [112] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315-323.
- [113] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369-376.
- [114] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764-1772: PMLR.
- [115] B. Schuller *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, 2016, pp. 2001-2005.
- [116] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson, 2014.
- [117] T. Mikolov, K. Chen, G. Corrado, and J. J. a. p. a. Dean, "Efficient estimation of word representations in vector space," 2013.
- [118] J. Devlin, M.-W. Chang, K. Lee, and K. J. a. p. a. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.
- [119] T. Mikolov, E. Grave, P. Bojanowski, C. Puhersch, and A. J. a. p. a. Joulin, "Advances in pre-training distributed word representations," 2017.
- [120] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. J. a. p. a. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations," 2021.
- [121] A. Norouzian, B. Mazoure, D. Connolly, and D. Willett, "Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed," in *ICASSP 2019-2019 IEEE International*

- Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7310-7314: IEEE.
- [122] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *International Conference on Machine Learning*, 2017, pp. 1243-1252: PMLR.
- [123] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*: Springer, 1990, pp. 227-236.
- [124] D. Ramachandram and G. W. J. I. S. P. M. Taylor, "Deep multimodal learning: A survey on recent advances and trends," vol. 34, no. 6, pp. 96-108, 2017.
- [125] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77-89, 2007.
- [126] A. Gulli and S. Pal, *Deep Learning with Keras*. Packt Publishing Ltd, 2017.
- [127] A. Ali, Y. Zhang, and S. Vogel, "Qcri advanced transcription system (qats)," in *Spoken Language Technology Workshop (SLT)*, 2014.
- [128] A. T. Ahmed, S.; Shaalan, K.; Hifny, Y., "Agent Productivity Modeling in a Call Center Domain Using Attentive Convolutional Neural Networks," (in English), *Sensors Journal* vol. 20, no. 19, p. 11, 2020/9/25 2020.
- [129] A. Mertins and D. A. Mertins, *Signal analysis: wavelets, filter banks, time-frequency transforms and applications*. John Wiley & Sons, Inc., 1999.
- [130] D. Bogdanov *et al.*, "Essentia: An audio analysis library for music information retrieval," in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.*, 2013: International Society for Music Information Retrieval (ISMIR).
- [131] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462.
- [132] G. I. Diaz, A. Fokoue-Nkoutche, G. Nannicini, H. J. I. J. o. R. Samulowitz, and Development, "An effective algorithm for hyperparameter optimization of neural networks," vol. 61, no. 4/5, pp. 9: 1-9: 11, 2017.
- [133] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," vol. 42, no. 4, pp. 335-359, 2008.
- [134] F. Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.
- [135] S. Kotsiantis, D. Kanellopoulos, P. J. G. I. T. o. C. S. Pintelas, and Engineering, "Handling imbalanced datasets: A review," vol. 30, no. 1, pp. 25-36, 2006.
- [136] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data," in *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2019, pp. 14-18: IEEE.
- [137] S. Deery, R. Iverson, and J. Walsh, "Work relationships in telephone call centres: Understanding emotional exhaustion and employee withdrawal," *Journal of Management studies*, vol. 39, no. 4, pp. 471-496, 2002.

- [138] S. Echchakoui and D. Baakil, "Emotional Exhaustion in Offshore Call Centers: A Comparative Study," *Journal of Global Marketing*, vol. 32, no. 1, pp. 17-36, 2019.
- [139] P. Wang, T. A. Wagner, S. L. Boyar, S. A. Corman, and R. B. McKinley, "The Relationship Between Organizational Family Support and Burnout Among Women in the Healthcare Industry: Core Self-Evaluation as Moderator," in *Handbook on Well-Being of Working Women*: Springer, 2016, pp. 283-296.
- [140] J. P. Wilson, *The Call Centre Training Handbook: A Complete Guide to Learning & Development in Contact Centres*. Kogan Page, 2009.
- [141] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7084-7088: IEEE.