UNIVERSITY OF THE WESTERN CAPE

# A robust audio-based symbol recognition system using machine learning techniques

by

Qiming Wu

A thesis submitted in fulfilment for the
degree of Master of Science

in the
Faculty of Natural Sciences
Department of Computer Science

Supervisor: Dr Mehrdad Ghaziasgar
Co-supervisor: James Connan and Reg Dodds

February 2020

# Declaration of Authorship

I, QIMING WU, declare that this thesis "*Audio Recognition on Surface Alphabet*" is my own work, that it has not been submitted before for any degree or assessment at any other university, and that all the sources I have used or quoted have been indicated and acknowledged by means of complete references.

Signed: _____

Date: _____

*" I examined the poets, and I look on them as people whose talent overawes both themselves and others, people who present themselves as wise men and are taken as such, when they are nothing of the sort.*

*From poets, I moved to artists. No one was more ignorant about the arts than I; no one was more convinced that artists possessed really beautiful secrets. However, I noticed that their condition was no better than that of the poets and that both of them have the same misconceptions. Because the most skillful among them excel in their speciality, they look upon themselves as the wisest of men. In my eyes, this presumption completely tarnished their knowledge. As a result, putting myself in the place of the oracle and asking myself what I would prefer to be what I was or what they were, to know what they have learned or to know that I know nothing I replied to myself and to the gods: I wish to remain who I am.*

*We do not know neither the sophists, nor the orators, nor the artists, nor I what the True, the Good, and the Beautiful are. But there is this difference between us: although these people know nothing, they all believe they know something; whereas, I, if I know nothing, at least have no doubts about it. As a result, all this superiority in wisdom which the oracle has attributed to me reduces itself to the single point that I am strongly convinced that I am ignorant of what I do not know. "*

Socrates

# Abstract

This research investigates the creation of an audio-shape recognition system that is able to interpret a user's drawn audio shapes—fundamental shapes, digits and/or letters—on a given surface such as a table-top using a generic stylus such as the back of a pen. The system aims to make use of one, two or three Piezo microphones, as required, to capture the sound of the audio gestures, and a combination of the Mel-Frequency Cepstral Coefficients (MFCC) feature descriptor and Support Vector Machines (SVMs) to recognise audio shapes. The novelty of the system is in the use of piezo microphones which are low cost, light-weight and portable, and the main investigation is around determining whether these microphones are able to provide sufficiently rich information to recognise the audio shapes mentioned in such a framework.

# Acknowledgements

This thesis is a compilation of the highly regarded acknowledgements from many people that assisted me through the years. I would first like to thank my supervisor Dr Mehrdad Ghaziasgar for the central role of involvement and the weekly meetings during my study. My sincere appreciation to Reg Dodds for the dedicated enthusiasm and ideas.

I would like to extend a very special thanks to the Telkom/Cisco/Aria Technologies Centre-of-Excellence at the University of the Western Cape for their essential financial support without which this endeavour would not have been possible.

# Publications

- **Title:** Audio recognition of contact surface gestures

  **Authors:** Qiming Wu, Mehrdad Ghaziasgar, Reg Dodds, James Connan.

  Published [63] in the proceedings of the SATNAC Conference of 2017.

- **Title:** Robust audio-based digit recognition.

  **Authors:** Qiming Wu, Mehrdad Ghaziasgar, Reg Dodds, James Connan.

  Published [64] in the proceedings of the SATNAC Conference of 2017.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **ANN** | **A**rtificial **N**eural **N**etwork |
| **CC** | **C**epstral **C**oefficients |
| **CDA** | **C**opula **D**iscriminant **A**nalysis |
| **Cls** | **C**lassifier of **l**etter**s** |
| **CNN** | **C**onvolutional **N**eural **N**etwork |
| **CPU** | **C**entral **P**rocessing **U**nit |
| **DCT** | **D**iscrete **C**osine **T**ransform |
| **DS** | **D**igit **S**hapes |
| **DWT** | **D**ynamic **T**ime **W**arping |
| **FFT** | **F**ast **F**ourier **T**ransform |
| **FS** | **F**undamental **S**hapes |
| **GFCC** | **G**ammatone **F**requency **C**epstral **C**oefficients |
| **GPU** | **G**raphics **P**rocessing **U**nit |
| **HCI** | **H**uman **C**omputer **I**nteraction |
| **HMM** | **H**idden **M**arkov **M**odel |
| **IFFT** | **I**nverse **F**ast **F**ourier **T**ransform |
| **LLAP** | **L**ow **L**atency **A**coustic **P**hase |
| **LS** | **L**etter **S**hapes |
| **MFCC** | **M**el **F**requency **C**epstral **C**oefficients |
| **QDA** | **Q**uadratic **D**iscriminant **A**nalysis |
| **RAM** | **R**andom **A**ccess **M**emory |
| **RBF** | **R**adial **B**asis **F**unction |
| **SNR** | **S**peech to **N**oise **R**atio |
| **SSN** | **S**peech **S**haped **N**oise |
| **STFT** | **S**hort **T**erm **F**ourier **Transform** |
| **SVM** | **S**upport **V**ector **M**achine |

**VRAM** **V**ideo **R**andom **A**ccess **M**emory

*For my parents. . .*

# Chapter 1

# Introduction

## 1.1 Background and Motivation

One of the main research interests in the field of Human-Computer Interaction (HCI) is to provide the user with alternative and/or enhanced means of interacting with devices, as compared to traditional and ubiquitous input devices like the keyboard and mouse.

There are two broad areas of exploration in HCI research [24, 43, 61, 67]. The first is visual-based HCI which makes use of one or more cameras for input, coupled with computer vision techniques to process the input and effect a relevant reaction on a device. Examples of visual-based HCI include eye tracking for cursor control, facial recognition and hand gesture-based input such as sign language recognition systems [9, 15, 23, 25].

The second broad area of HCI research is audio-based HCI which makes use of one or more microphones for input, coupled with audio processing and recognition techniques to process the input and effect a relevant reaction on a device. Audio-based HCI can further be sub-divided into two main sub-areas, namely, speech-based HCI and non-speech-based HCI.

Speech-based HCI involves capturing and processing human speech and has recently re-gained significant traction and become a very popular research area with the advent and success of systems such as *Amazon Alexa*, *Apple Siri* and *Google Assistant*. As such, this research sub-area can be considered to have entered into a stage of relative maturity.

On the other hand, non-speech-based HCI can still be considered to be in its infancy, with research in this field being very limited in scale. Non-speech-based HCI involves capturing and processing generic sound signals such as stylus gestures, stylus-drawn

letters or numbers, taps [51], claps [38, 60] and other forms of non-speech audio in order to effect a relevant reaction on a device [19, 32, 33, 53]. Non-speech-based recognition has also been applied to a non-HCI context such as environmental sound classification, bird sound classification etc. [3, 8, 34, 58]. While the application context is different in these cases, the techniques used are very similar or the same, and may also be applied to HCI.

This research focuses on developing a novel non-speech-based HCI system. The intended system aims to capture and process the sound emitted by a generic stylus, such as the back of a pen, on a generic surface such as a table. The intent is to allow the user of the system to draw specific shapes on the surface, which can be taken as various information to the device.

The audio input will be captured by a contact microphone. A suitable choice of contact microphone is the Piezoelectric microphone, also known as "pickup" or "piezo" transducer microphone. Unlike conventional microphones, piezo microphones only detect surface or structure-borne vibrations and are insensitive to airborne pressure. This helps reduce the effects of air-borne noise on the system. Also, the piezo microphone has a low voltage requirement, is fabricated at very low cost, and it has a wide dynamic range [30]. The piezoelectric microphone in particular has a compact shape of a flat disk which makes the microphone exhibit an extremely low profile and convenience of usage. Furthermore, it requires no additional components like resistors for choking or capacitors for power filtering. It can be connected to any personal computer that has a standard 35mm audio-input jack. Therefore, the use of this input device will result in an input configuration that is low cost, low key, low complexity, sufficiently rich in acoustic information, and able to be connected to any standard personal computer without any hardware modifications whatsoever. This research therefore makes use of piezoelectric microphones which are henceforth referred to simply as "piezo microphones".

In terms of intended processing and output, the initial aim is to recognise a relatively limited set of fundamental abstract gesture shapes such as *Dot*, *Dash*, *Tick*, *Triangle*, etc. Note that these abstract shapes can be tied to specific tasks by the user at a later stage—the complexity here, and the focus of this research, is recognising these shapes.

Depending on the success of this initial system, the approach can systematically be re-packaged and re-applied to increasingly complex—and increasingly beneficial—tasks of digit recognition and, eventually, letter recognition. Digits and letters, respectively, represent an increasingly larger number of classes to be recognised, and therefore an increasing level of recognition complexity. Which brings us to—and to be clear—the final aim of this research; to arrive at a system that can recognise the uppercase letters

of the Roman alphabet, as drawn by a generic stylus on a given surface such as a table-top. The ability to eventually recognise letters will give the user access to text-based input at a very low cost and without any hardware complexity. Very importantly, a system of this kind will be very portable, requiring only a small low-key microphone to be moved around, as compared to moving around a physical mouse or, worse, a keyboard.

There are a limited number of research projects that have attempted to create similar systems to the one proposed in this research [32, 33, 53, 62, 69, 70], although these systems make use of conventional condenser microphones, such as those found in modern mobile devices. The configuration proposed in this research, i.e. using piezo microphones is a novel alternative that is of a much lower cost compared to condenser microphoes. Therefore, one of the main goals of this research is to determine whether these low-cost microphones can provide sufficient sound definition and quality to enable the audio recognition problems mentioned previously to be carried out.

Another point of investigation is the *number of* piezo microphones that will be required to enable a sufficiently-high recognition accuracy. The most desirable configuration for the proposed system is with a single microphone. However, the complexity of the recognition problems that are being investigated—especially considering that the complexity increases from fundamental shapes, to digits, and finally to letters—may require more than one source of audio information. A single microphone provides a single dimension of information and can only be used to distinguish between audio shapes based on the number and length of strokes in the shapes. For example, the two letters "A" and "H" may sound exactly the same from a single sound source given they may both be drawn with three distinct strokes. Adding a second microphone into the configuration makes it possible to infer the two-dimensional spatial distribution of audio shapes, and in this case it is expected that "A" and "H", for example, would no longer sound similar since the three strokes are drawn in different relative locations. A third microphone may serve as a confirmation source that can limit ambiguities and sources of noise.

The processing component of the proposed system consists of two fundamental elements: a feature descriptor that transforms the audio signals of the relevant audio shapes into standardised and learnable features; and a classifier that is trained on these features to recognise the audio shapes. The feature descriptors and classifiers used in speech recognition may be suitable in this regard, since it is possible to apply them to non-speech-based applications due to their robustness.

According to a number of researchers, one suitable feature descriptor that represents audio very effectively with dynamic features, and is very robust and effective under

various environmental conditions is the Mel-Frequency Cepstral Coefficients (MFCC) feature descriptor [39, 44, 55, 71]. The MFCC extracts both linear and non-linear attributes of an audio signal, and is well-suited to classification problems. It has been applied to a wide range of domains, both speech and non-speech, with success. Therefore, the MFCC is used in this research.

With a suitable audio descriptor in place, the recognition or classification of specific audio shape classes is carried out with a classifier or statistical model. Trained on an existing set of audio shape data, such a model must be able to accurately predict the audio shape of previously unseen or new audio samples. There are a variety of choices when selecting a classifier or statistical model. These include a variety of "traditional" techniques such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Bayesian classifiers, Random Forests, to name a few, as well as hybrid approaches proposed by various researchers that mostly combine two traditional techniques [1, 27, 36]. These techniques have all been applied to a variety of classification problems, with differing but arguably comparable performance.

Of these techniques, SVM-based classification appears to be a very good choice of technique, since it is a very popular and effective choice in a variety of applications, including data mining, electrical engineering and other fields [28, 47, 65, 66]. It has been repeatedly shown to be accurate and robust, and its tendency to always generalise across classes makes training easier than some other techniques such as ANNs [5]. SVM-based classification makes use of statistical learning theory which is able to handle practical problems like pattern recognition [59]. It is also well-suited to high-dimensional data. Therefore, SVMs are used in this research.

In summary, this research aims to create an audio-shape—fundamental shapes, digits and/or letters—recognition system that allows the user to draw audio shapes on a given surface such as a table-top using a generic stylus such as the back of a pen. It then makes use of one, two or three piezo microphones to capture the sound of the audio gestures, and a combination of the MFCC feature descriptor and SVMs to recognise audio shapes. This framework will initially be applied to generic fundamental shapes and, depending on the success, will be re-applied to written digits, and finally to uppercase alphabet characters.

## 1.2 Research Question

Based on the discussion in Section 1.1, the following research question can be formulated: "How accurately can the proposed configuration, i.e. a combination of one or more piezo

microphones, the MFCC feature descriptor and SVMs, recognise audio shapes drawn by a generic stylus on a given surface?" This can be further broken down into the following sub-questions, to provide an answer to the main research question:

1. How accurately can the proposed configuration recognise a limited number of *fundamental shapes*?
2. How accurately can the proposed configuration recognise written *digits*?
3. How accurately can the proposed configuration recognise written *upper case alphabet characters*?
4. How many microphones are sufficient to ensure high-accuracy recognition of the audio shapes?

## 1.3 Research Progression and Research Objectives

The research will be conducted in the following order and with the following objectives which, once met, will lead to answers to the questions in the previous section.

1. Implement the framework that includes audio capture from one, two or three microphones.
2. Implement the MFCC feature descriptor for the audio processing procedure and incorporate it into the framework.
3. Incorporate SVM machine learning into the framework for training and testing.
4. Collect a data set of generic fundamental shapes. The fundamental shapes used will be: "Dot", "Dash", "Tick", "Triangle", "Square", "Cross", "Circle".
5. Collect a data set of written digits—"0" to "9".
6. Collect a data set of written uppercase letters—"A" to "Z".
7. Train and test the framework on the fundamental shapes data set using one microphone. If the accuracy obtained is sufficiently high, then move on to the next objective, otherwise re-train with two or three microphones as required.
8. Train and test the framework on the digit shapes data set using the smallest number of microphones that was found to be capable of providing a sufficiently high accuracy in Objective 7. If the accuracy obtained is sufficiently high, then

move on to the next objective, otherwise re-train with two or three microphones—depending on whether the objective started with one, two or three microphones—as required.

9. Train and test the framework on the letter shapes data set using the smallest number of microphones that was found to be capable of providing a sufficiently high accuracy in Objective 8. If the accuracy obtained is sufficiently high, then all objectives will be considered to have been successfully met. If not, re-train with two or three microphones—depending on whether the objective started with one, two or three microphones—as required.

Note that, although Objective 7 states that an increasing number of microphones—up to three—will be used if one or two microphones do not provide a sufficiently high accuracy, it is unlikely that this will take place. It is expected that one microphone will be a sufficiently rich source of information for that recognition problem.

Also, Objective 8 makes provision for the case where one and two microphones are unable to provide for high accuracy recognition, in which case three microphones will be employed. The Objective does not explicitly cater for the case in which neither of the three microphone configurations can provide for a sufficiently high accuracy. For the sake of completeness only, it is stated that, in such a case, this research will conclude that the proposed configurations can not support high-accuracy recognition of letters, which is the eventual goal of this research, and no further objectives beyond Objective 8 will be undertaken. This, although possible, is not expected to happen. The two-dimensional information from two microphones is expected to be sufficiently rich as to provide for high accuracy recognition of written digits.

## 1.4 Premises

The following strategic premises and assumptions are made for this research in order to limit the work to a feasible scope, and all of which can serve as the basis for future work:

1. Recognition will be carried out on single characters i.e. no word or sentence segmentation strategy will be incorporated into the system at this stage.

2. Recognition will be done on shapes, digits and upper case Roman alphabet letters only.

3. For the purposes of this study, a single generic surface, namely a table-top, will be selected and used for the duration of experiments.

## 1.5 Thesis Outline

This chapter is where the research questions and objectives are stated. The remainder of the thesis is arranged as follows.

**Chapter 2:** *Research Methodology.* This chapter provides a detailed discussion around an appropriate philosophical stance that this research takes and the methodological protocol that will be followed in this research, that will help lead to a prototype of the proposed system and an answer to the research question set out in this chapter.

**Chapter 3:** *Related Work.* This chapter looks at related work in the literature that focus on the recognition of various audio signatures, with recognition carried out using various feature descriptors and classifiers. The approaches and results of related research projects are examined and the performance of various feature descriptors and classifiers are examined. Ultimately, the chapter demonstrates the novelty of the proposed approach as described in this chapter relative to other research.

**Chapter 4:** *Techniques of Audio Recognition.* This chapter focuses on the MFCC feature descriptor and SVM classification, the techniques that make up the components of the proposed system's audio recognition process. Their processes and functionality is discussed in detail. Together with Chapter 3 formulates the interrelationships between components and presents an educated solution.

**Chapter 5:** *Design and Implementation.* This chapter entails the experimental set up of the system, hardware usage, processes and data in order to address objectives in listed in Section 1.3. This chapter is a constructed test solution to help gain insight into the system's practicality.

**Chapter 6:** *Experimental Results and Analysis.* This chapter describes the results produced from experiments carried out in order to respond to research questions imposed in Section 1.2. This chapter consolidates the knowledge gained from experiments and observations which shows that the hypothesis has not been falsified.

**Chapter 7:** *Conclusion.* This chapter concludes that the research questions have been successfully answered and indicates directions for future work.

# Chapter 2

# Research Methodology

In order to justify research, a methodological protocol needs to be described and followed that leads to a legitimate "proof" for hypotheses. In other words, evidence, i.e an observed artefact is necessary to support such "proof". It is mandatory for scientific practice and research to involve processes, methods and tools within such protocols. Such practice will serve its purpose when the body of knowledge is expanded upon in the research domain. Figure 2.1 shows an illustration of a hypothetical body of knowledge which is contributed to, and expanded via, research practice within a given research domain. Note that this body of knowledge refers to the knowledge within a specific research domain and not the general knowledge body which overrules other research domains [45].

Numerous research methodologies exist and an appropriate strategy to apply to this research would be *System Development*. System development comprises of five phases: (1) concept design; (2) constructing the architecture of the system; (3) prototyping; (4) product development; and (5) technology transfer [13]. Although this methodology does not perfectly align with the nature of this research, its general direction and phases are in agreement with this research. As such, minor adjustments to the phases have been made in order to tailor this methodology to be compatible with this research.

The adapted system development phases are as follows: (1) conceptual framework; (2) system design; (3) prototyping; and (4) evaluation. This serves as the scientific method in order to complete the research objectives set out in Section 1.3 and arriving at an answer to the research question and sub-questions detailed in Section 1.2. Figure 2.2 shows an overview of the phases from the system development method and the adapted version of phases for this research.

The following section opens with a discussion on Philosophy, and the sections that follow

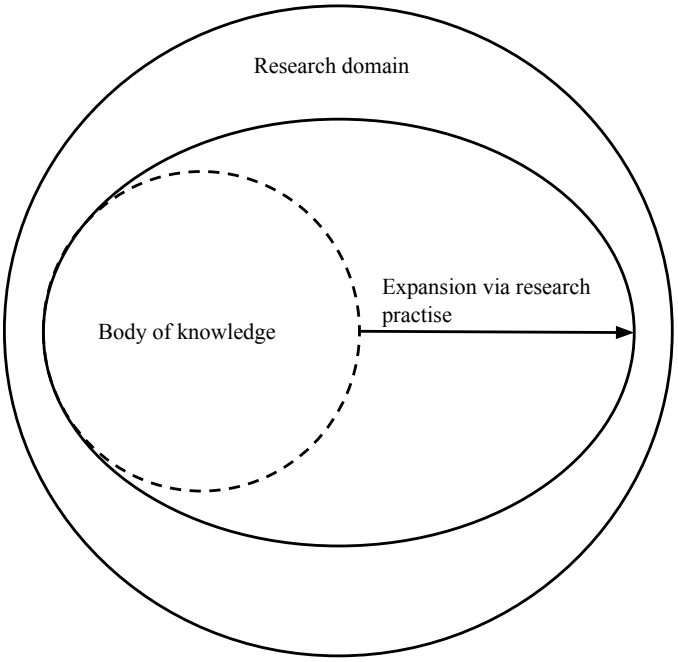


FIGURE 2.1: Illustration of a hypothetical body of knowledge in a specific research domain, and its expansion.

discuss each of the phases of the adapted system development methodology illustrated in Figure 2.2 and the content is then summarised.

## 2.1 Philosophy

Philosophy, the fundamental *of* fundamentals; it is the shining seed that is the root of all academic branches, i.e it is the very nature of thought. As this could arguably be the grand origin of all possible topics that any paper may discuss, its subjects are complex and intertwined, and thus any remote in-depth exploration could result in derailing the focus of this thesis. Therefore, discussion is restricted to the nature of the research conducted in this paper.

Gaining research insight is to take a stance. A stance consists of an approach and an assumption. An approach is a spectrum of either subjectivist or objectivist artefacts. An assumption is the questioning of a certain fundamental aspect. Figure 2.3 shows the philosophical assumptions positioned on a spectrum of approaches. Ontology is the questioning of reality and Epistemology is the questioning of knowledge [7].

The approach spectrum is explained by using *Truth* as an example. In the subjectivism side of the spectrum, truth only takes hold in the mind, which is to say that one statement holds true for one person but might not hold true for another person. This purely

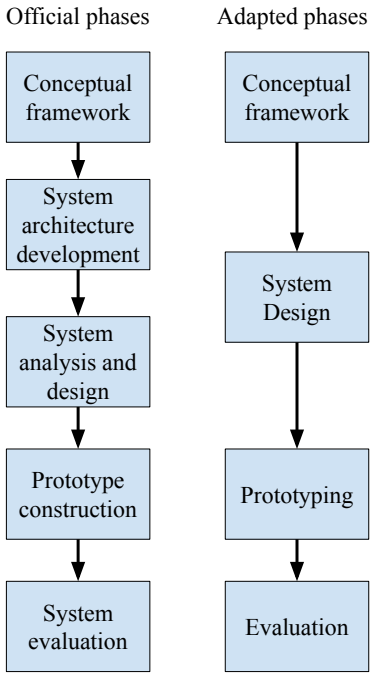


FIGURE 2.2: Research methodology: a) Overview of the official system development method; b) the adapted version of the methodology for this research.

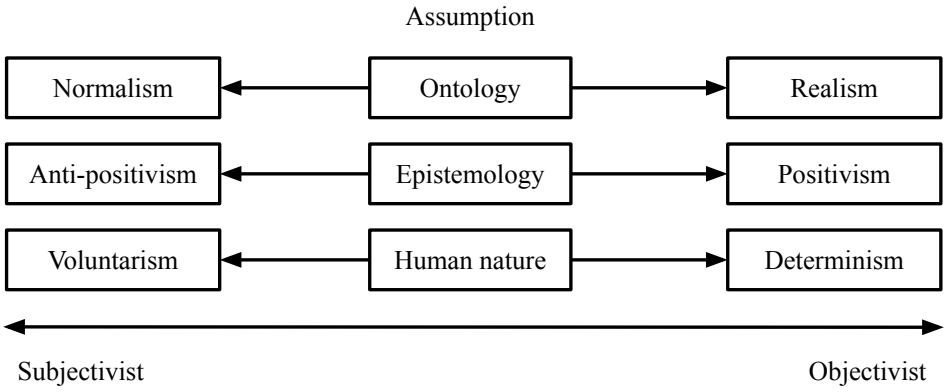


FIGURE 2.3: Research philosophy approach and assumption.

depends on the perception of one's mind and there is no universal "truth" or "knowledge" or even "reality", but only one's mental experience.

The objectivism side of the spectrum, on the other hand, identifies truth and falsehood as aspects of both conceptual and perceptual knowledge. Thus, truth *is* reality, which is processed or perceived by a conscious, logical mind, such that processing or perception is done with reasoning. Objectivism does allow for statements to be true for one person and false for another, only if there are perceptual evidences for the statement that contradicts and different levels of evidence will result in different levels of determinacy.

For example: the statement is probable, certain, or arbitrarily true [49].

The relevant assumption for this research is targeted at epistemology—the questioning of knowledge. Taking a subjectivist approach to this assumption would imply that the objective reality does not exist outside of the human mind. Thus "reality" then holds true only in one's mental experience. This stance is anti-positivism, also known as interpretivism. The main cause of concern with the anti-positivist stance is that "reality" is considered to only be our perception which is imagined. Thus, knowledge is then only value-laden and research—the pursuit of knowledge or expanding the body of knowledge—becomes subjective and cannot be objectively consistent [21, 31].

Taking an anti-positivist stance renders scientific research endeavour inconsistent, and the proper option to consider is the positivist stance. The positivist stance implies that the world exists outside of the human consciousness—the individual man—as an empirical entity and with tangible structures. Therefore knowledge of reality—the world—can be obtained through logical and sensible observation as well as measurement. This philosophical stance justifies the endeavour of scientific research well and this research will follow suit [17, 18, 31, 40].

## 2.2 Conceptual Framework

Before any actual design or construction can take place, fundamental reasoning is required. The research itself should be justified by answering research queries posed by questions formulated from the research problem. This sets the foundation of the research. In order to answer the research questions, the results and performance are studied to gain insight. A target focus is provided throughout the research development cycle by having a defined problem statement. The problem statement discussed in Section 1.1 provides a conceptual framework to form theory building. Theory building has different categories that differ within various disciplines. The appropriate category for this research to adapt would be a framework of guidance and management of ideas. This provides recommended actions to take [41].

## 2.3 System Design

The design phase usually comprises of two parts: (1) architecture design and (2) system design. In this adaptation, both architecture and system design are considered as a single phase. However, the combined phase may be broken down into two sub-phases. Although the two form a single phase, each has its own speciality and function.

Architecture design within system development provides a guide for planning construction. Chapter 3 dives into the research for such guides—obtained from other previous existing works—to plan the construction for the system's design. It lays out the functionality and relations between system components, as well as the interactions between them. Researchers evaluate the constraints which the environment imposes on the system, thus influencing the specification for system requirements, objectives and functionality. This evaluation may be an assumption, however, an educated one, so it may be adjusted empirically without difficulty. Under the results from assumptions, system design and implementation are done according such results.

System design is done according to system architecture which drives the system requirements and functionality. Chapter 4 holds the relevant design application for knowledge base at this stage. This involves an understanding of the research domain, application of scientific and technical knowledge, formulating a proposed solution, as well as an alternative solution. Design elements act as the layout for the implementation stage. All data structures, data, knowledge bases, modules and functions should be specified during this phase [14].

The design phase is an element of both architecture and system. During the application of this phase, architecture and system design are concurrently executed as the development cycle progresses.

## 2.4 Prototyping

Within system development, a research venture would certainly build a prototype. This is the creation of an artefact, which is a possible solution to address the problem. Chapter 5 holds the relevant design application for prototyping at this stage. It is an engineering-oriented practice to test the proposed system in a practical setting [41]. The purpose of prototyping is to achieve proof of concept specified in Section 2.2 and to fulfil requirements specified in Section 2.3. This is in order to demonstrate the practical aspect of the proposed system and provide insight into any unsatisfactory elements. Stressing the prototype may provide accumulated experience and knowledge so that any re-design and adjustments can be made without difficulty.

## 2.5 Evaluation

When the final prototype has been completed, it is then tested for its performance as well as evaluated to observe its impact on the target audience. Chapter 6 holds the relevant

result readings and information for evaluation at this stage. Evaluation of results from testing infers answers for queries set in the conceptual phase discussed in Section 2.2. Development takes on an evolutionary process; knowledge and experiences gained from developing one system frequently lead to further research in developing new systems or extension onto existing ones.

## 2.6 Summary

This chapter described in detail a methodological protocol that will be followed in this research, that will help lead to a prototype of the proposed system and an answer to the research question set out in the previous chapter. An adaptation of the system design methodology was proposed.

Section 2.1 discussed the appropriate philosophical stance this research has taken. A positivist stance is considered to be the proper stance for this research, which is an enquiry of knowledge that is based on observation, logical thinking and measurement on the tangible reality.

Section 2.2–2.5 described the adapted research methodology that will be used to guide the development life cycle adapted for this research.

The next chapter conducts a review of relevant literature in the field associated with this research.

# Chapter 3

# Related Work

Previous research on audio recognition develops a theme of using specific feature extraction techniques along with template matching or machine learning techniques to recognise non-speech audio shapes or environmental sounds. While research into non-speech-based audio recognition currently still hasn't reached maturity, the techniques used are found to either be the same as those used in speech-based audio recognition applications, or at least based on such origins. The fact that these techniques have successfully been applied to non-speech audio recognition applications demonstrates their capabilities as regards being able to adapt them to other non-speech-based audio applications.[4]

This chapter reviews existing research studies done on non-speech audio processing applications, with emphasis on the usage of audio feature extraction techniques together with classification techniques. Where possible, the hardware used to collect audio data, the feature extraction and classification technique used, the data set used and the approach used to test their system will be described for each study. This will all collectively be used to demonstrate the novelty of the proposed approach.

This chapter is structured as follows: Section 3.1 looks into relevant studies from previous work that have applications that are closely related to the system proposed in this research, i.e. those that are more generally focused on recognising shapes, digits or letters from the audio generated while writing on a surface. Section 3.2 discusses more general audio recognition research studies that use feature descriptors and classifiers. The chapter is then summarised with a table that summarises all of the related studies described.

## 3.1 Shape, Digit or Letter Recognition Research

This section looks into and discusses studies that are closely related to the task of this research i.e. make use of audio to recognise characters. Different approaches of how the hardware and technique are used on characters to create an audio context is examined.

The following subsections will each discuss and detail a relevant research study related to this research area.

### 3.1.1 Stroke Recognition Using a Mobile Device Microphone

Zhang et al. [70] produced a text input system named *SoundWrite*. The system captures the audio signal of handwritten text using the integrated microphone in a mobile phone. The writing is done using fingers on a wooden table. The system is built for mobile devices running Android and has a predefined set of inputs. The system aims to recognise seven unique strokes as primary inputs which, in combination, produce various written characters. Figure 3.1 shows the primary strokes: a dot, horizontal dash, vertical dash, left diagonal dash, right diagonal dash, left arc and right arc. This is likely inspired by the structure of Wubi Chinese input method [68].


FIGURE 3.1: Zhang et al.'s primary strokes for written text entry, adapted from [70].

Combinations of primary strokes are then used to infer characters. For example, referring to the labels of the primary strokes in Figure 3.1: the combination of S4, S5 and S2 forms the letter "A"; the combination S3, S7 and S7 forms the letter "B". A $k$-Nearest-Neighbours classifier with Euclidean distance was used to recognise strokes. The features used were extracted from the amplitude spectrum density function which is obtained by applying the Fast Fourier Transform (FFT) to the audio signal.

Figure 3.2 shows Zhang et al.'s results for the stroke's recognition accuracy. The overall accuracy achieved was 90.3%. According to the paper, it is possible that with a significantly higher microphone quality in mobile devices, a higher accuracy result might be achieved. Table 3.1 shows the accuracy results over three different mobile devices.

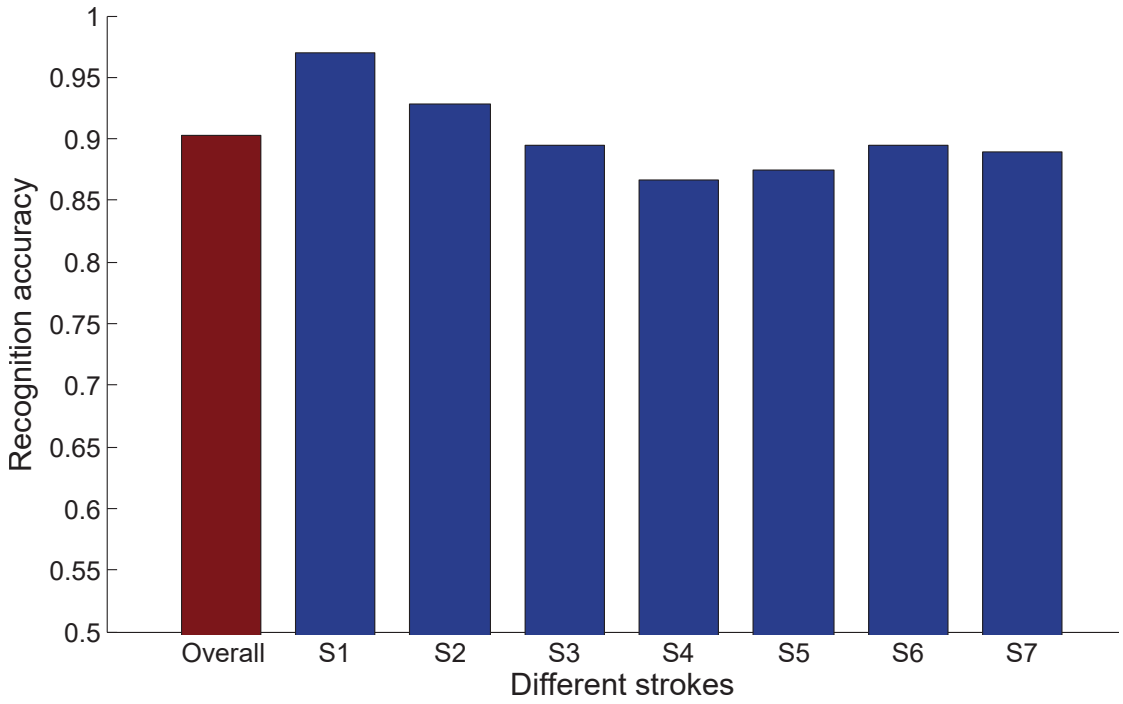


FIGURE 3.2: Stroke recognition accuracy of Zhang et al.'s system, adapted from [70].

| **Device** | Huawei U9508 | Motorola MT887 | Samsung G3568V |
|---|---|---|---|
| **Accuracy(%)** | 90.3 | 88.5 | 90.1 |

TABLE 3.1: Stroke recognition accuracies of Zhang et al.'s *SoundWrite* system on different mobile devices.

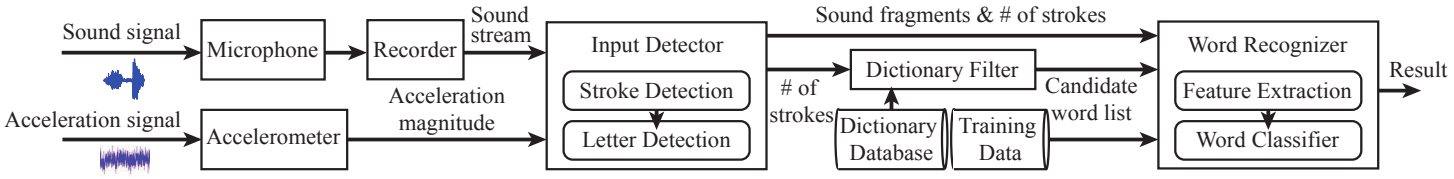


FIGURE 3.3: High-level overview of Yu et al.'s system [69].

### 3.1.2 Letter Recognition For Information Interception

Yu et al. [69] produced a similar system named *WritingHacker*, which also makes use of mobile devices to capture gestures. However, the system was designed from the perspective of hackers to investigate whether it is possible to intercept information such as passwords in this way. The system uses a two-stage approach in which it first determines the letter(s) drawn and then groups them up using a dictionary lookup to determine the word(s) drawn.

Figure 3.3 shows a high-level overview of Yu et al.'s *WritingHacker* system. To recognise letters, the number of continuous strokes drawn is first detected in the *Input Detector*'s *Stroke Detection* component; the audio signal is then diverted to one of three classifiers

in the *Letter Detection* component, each of which is specifically trained to predict letters with that given number of strokes. As such, the system groups letters up in terms of the number of continuous strokes required to draw each letter.

$$\begin{aligned} C_1 &= \{C, G, L, M, O, S, U, V, W, Z\} \\ C_2 &= \{B, D, J, K, P, Q, R, T, X\} \\ C_3 &= \{A, E, F, H, I, N, Y\} \end{aligned} \tag{3.1}$$

The letter groups recognised by each of the three classifiers are shown in listing 3.1 and are as follows: Groups $C_1$ through to $C_3$ contain, respectively, letters that require one, two and three continuous strokes to complete.

The main feature descriptor used by each of the three classifiers—which are SVMs—is the MFCC feature descriptor. Two sensors are used to capture data, generic smartphone microphone and accelerometer, with the accelerometer being used to minimise near-field noise only. The system was evaluated under three different scenarios pertaining to whether or not training data from the victim was available, and whether or not the same writing location as used in training was used in testing. Figure 3.4 shows the detailed accuracy results for each of the three scenarios which are clearly labelled in the figure, across all 26 letters produced by the three classifiers combined. Referring to the figure, with training data of the victim available and the writing location known (labelled as maroon in the figure), it is seen to be much easier to recognise the letters than when

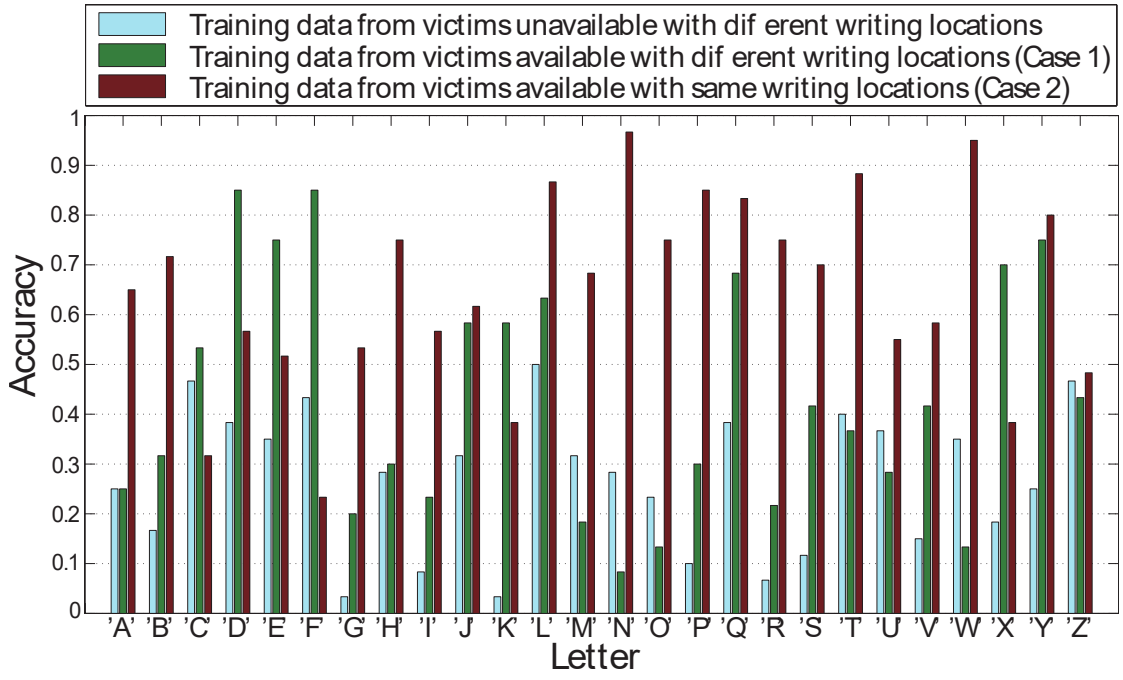


FIGURE 3.4: Yu et al.'s *WritingHacker* letter recognition accuracy results [69].

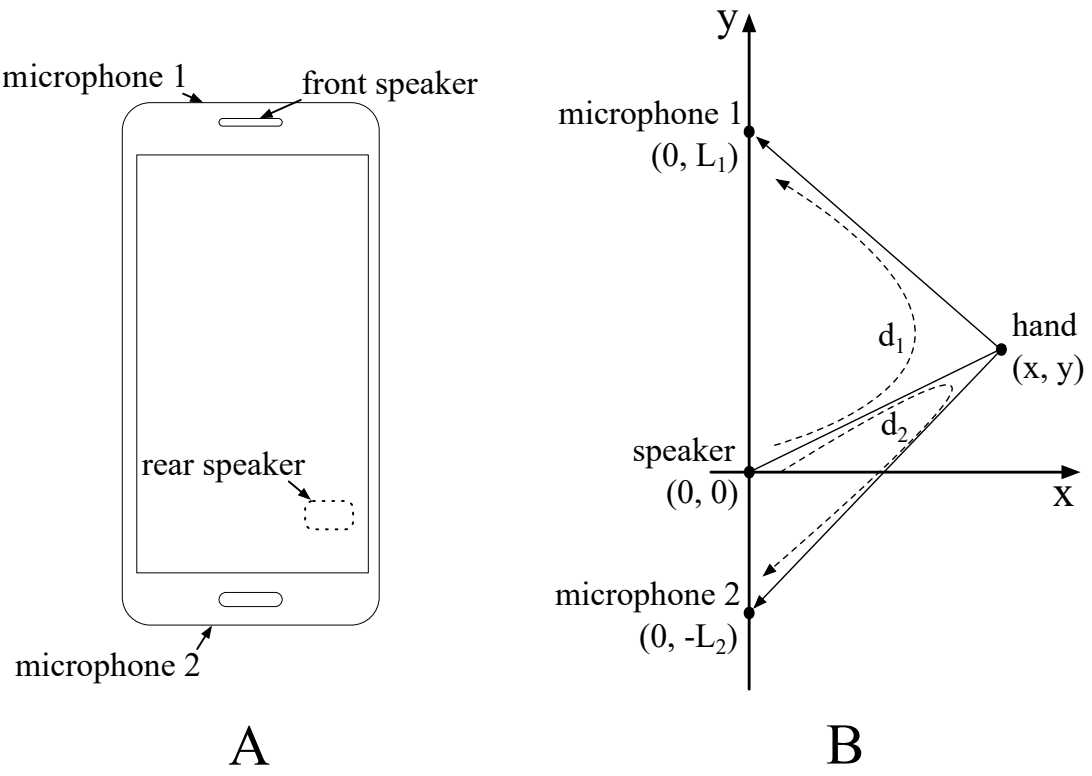


FIGURE 3.5: Wang et al.'s system taken from [62]: a) Typical hardware setup—the Samsung Galaxy S5; b) High-level overview of the tracking concept [62].

neither of those sources of information are available in training (labelled as light blue in the figure). With neither the victim's data available, nor the writing location known, the accuracies are seen to range between 5% and 50%.

Finally, a dictionary lookup is used with the recognised letters to determine the words written down, similar to auto-correct. The researchers state that the word recognition accuracy that they obtained was between 50–60%.

### 3.1.3 Gesture Tracking for Wearable Smart Devices

Wang et al. [62] constructed a framework capable of device-free gesture tracking, aimed at small wearable smart devices. The framework utilises the built-in speaker(s) and microphone(s) on a smart device. The speaker emits high frequencies that are beyond the human hearing range and are therefore not audible to humans. These are then reflected off of the user's hand and fingers and captured via the built-in microphone and processed. Audio processing techniques—in this case: Low-Latency Acoustic Phase (LLAP)—are used to accurately determine the location of the hand and fingers in time. By tracking the hand and fingers, it becomes possible to detect gestures that are traced in the airspace immediately above the device e.g. letters traced out with an extended

finger. The system is able to perform both one and two-dimensional gesture tracking using a pair of microphones and a speaker. This setup is used to recognise characters performed by the user.

Figure 3.5 illustrates a setup that utilises one speaker and a pair of microphones to track two-dimensional hand positions. Figure 3.5a is an example of a smart device that can be used which is, in this case, the Samsung Galaxy S5. Figure 3.5b shows the geometric abstraction of hand position tracking using only one of the speakers on the phone, which is sufficient to recognise gestures. Such gestures are performed mid-air within the range of the capturing and emission sensors i.e: microphone and speaker.

Referring to Figure 3.5b, $d_1$ and $d_2$ denote the reflection of signals towards microphones 1 and 2 respectively. Using $L_1$ and $L_2$ which denote the $y$ coordinates of microphones 1 and 2 respectively, and using the speaker as the origin of the coordinate system, the $(x, y)$ coordinates of the hand can be solved for using a series of equations provided in [62].

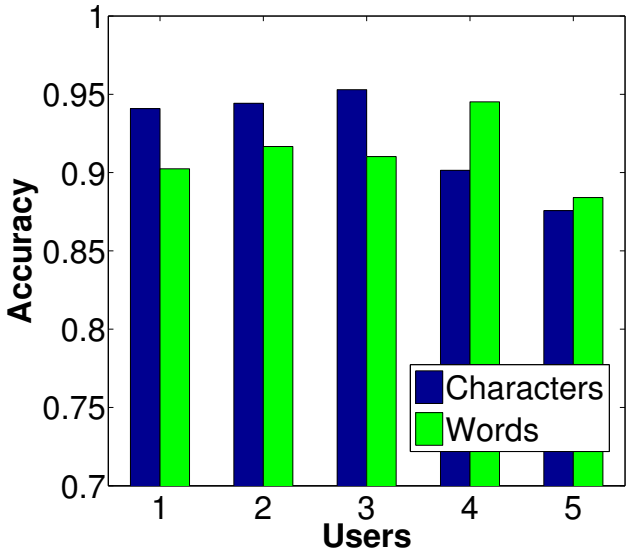


FIGURE 3.6: Wang et al.'s results of recognition accuracy on characters and words [62].

They conducted an experiment using template matching techniques to determine the gesture recognition accuracy of their system on letters and short words. A data set was collected, with five users performing each of the letters and words five times. The letters were all 26 letters of the Roman alphabet. Generally, lower case letters were used but lower case letters that can't be drawn in a single continuous stroke such as "i" were drawn in upper case i.e: "I". The words in the data set included a list of eleven short words; the authors only specifically mention the words "yes", "can" and "bye".

| Noise Reduction | Feature Set | Accuracy (%) |
|---|---|---|
| Endpoints | Mean amplitude | 76.1 |
| Endpoints | MFCC | 83.4 |
| Gaussian | Mean amplitude | 77.9 |
| Gaussian | MFCC | 81.2 |

TABLE 3.2: Li's accuracy results on experiment 1.

Overall, the system achieved 92.3% accuracy for two-dimensional letter recognition and 91.2% for short word recognition. Figure 3.6 provides a break down of their results per user for character and word recognition. In most cases, the recognition of characters is seen to be more accurate than that of words. This is expected since characters are shorter and less complex.

### 3.1.4 Letter Recognition Using a MacAir Microphone

Li [33] built a system that utilises the Dynamic Time Warping (DTW) approach to recognise the 26 uppercase letters—A–Z—from acoustic signals. The letters were drawn on a wooden desk using various blunt objects. The audio was collected from a MacAir microphone. Two feature descriptors were compared: the MFCC feature descriptor and the mean amplitude value form each time frame. Each time frame have a size of 45ms with a overlapping size of 30ms. Finally, The classifier consists of using DTW approach applied to a template matching technique. The authors clearly state that the system is user-dependent i.e. the system must be trained on the writing style of a user before it can be used.

A data set was collected using 6 users. Each user used each of three modes of drawing, namely, pen, fingernail and key, to draw each of the 26 letters five times i.e. each user had 26 letters $\times$ 3 modes $\times$ 5 samples = 390 samples.

Two experiments were conducted using the data set. In the first experiment, 1 of the 5 samples of each user were used as candidate templates for a user-specific classifier i.e. one classifier per user based on 26 letters $\times$ 3 modes $\times$ 1 sample = 78 template samples. Then, the remaining 4 samples per letter per mode for each user were used to test each user-specific classifier; this was done by comparing each of these samples to

| Feature Set | Accuracy (%) |
|---|---|
| Key | 85.9 |
| Pen | 86.8 |
| Fingernail | 78.3 |

TABLE 3.3: Li's accuracy results on experiment 2.

each of the candidate samples—78 samples in total—and finding the most likely match. 5-fold cross-validation was used to obtain a cross-validation accuracy per user. Finally, the cross-validation accuracy of all six subjects were averaged to get an overall accuracy for the technique. This whole process was replicated for combinations of the following: two noise reduction methods that were compared, namely, endpoints noise reduction and Gaussian noise reduction; and the MFCC and mean amplitude feature descriptors.

The results of the first experiment are summarised in Table 3.2. It can be noticed that MFCC features appear to yield a higher accuracy result than mean amplitude features. In terms of noise reduction, the two techniques appear to be more or less on-par with each other.

In the second experiment, only MFCC and endpoints noise reduction were considered. The three input modes mentioned previously were treated separately in this case to obtain an accuracy per mode i.e. separate classifiers were produced per subject and per mode. Therefore, a total of 3 classifiers were produced per subject. This was done by using 1 of the 5 samples of each user for each mode were used as candidate templates for a user-specific classifier i.e. one classifier per user per mode based on 26 letters $\times$ 1 sample = 26 template samples. The remaining 4 samples per letter per mode were used in testing and 5-fold cross-validation was again used to obtain an accuracy. Finally, the average accuracy across all subjects per mode was computed. The results obtained for this experiment are summarised in Table 3.3. According to the results, it appears that using a stylus or key would be recommended over using one's fingernail.

It should be noted, again, that this study made use of template matching which resulted in a very user-dependent system. The use of a machine learning classifier is expected to reduce or eliminate user-dependence.

### 3.1.5 Updated Letter Recognition Using a MacAir Microphone

A later work by Li, in collaboration with Hammond [32], expanded on his research described in Section 3.1.4. The MFCC feature descriptor was used, with several additional features to recognise the sketch of upper case alphabet letters. The additional features used were skewness, kurtosis, curviness and peak location. These global features were expected to further characterise the audio features and provide robustness. The below Equation 3.2 shows the calculation for skewness, which is the measurement of symmetry,

$$S(f) = \frac{\sum_{i=1}^{n}(f_i - \bar{f})^3}{(n-1)s^3}. \tag{3.2}$$

Where $s$ is the standard deviation of $f$, $f_i$ is the input frequency of the $i$th cepstral coefficient, $n$ is the number of coefficients and $\bar{f}$ is the mean of all frequencies $f_i$. Equation 3.3 defines kurtosis which describes the weight of the tails of a distribution,

$$K(f) = \frac{\sum_{i=1}^{n}(f_i - \bar{f})^4}{(n-1)s^4}. \tag{3.3}$$

Variables $s$,$f$,$n$ and $i$ are identical to equation 3.2. Equation 3.4 shows the calculation for curviness, which is the measurement of jerkiness of the audio features by summing all the local spikes.

$$C(f) = \sum_{i=2}^{n}(f_i \mid f_i > f_{i+1},\ \ f_i > f_{i-1}). \tag{3.4}$$

Equation 3.5 shows the calculation for peak location. The definition is the arguments of the maxima distributed among the number of coefficients, where all the symbols are the same as those that have been previously defined except $j$ which is within the set $n$,

$$P(f) = \frac{\arg\max_i\{i|f_i \geq f_j, \forall j \in n\}}{n}. \tag{3.5}$$

A variety of statistical matching techniques were used for classification and compared as follows:

- DTW with noise reduction
- Quadratic Discriminant Analysis (QDA) that uses a multivariate Gaussian distribution
- Copula Discriminant Analysis (CDA)
- QDA with DTW
- CDA with DTW
- Average distance approach using Euclidean distance
- Four global features along with distance measure i.e. a total of five features

Table 3.4 summarises the recognition accuracy of the classifiers compared. The best recognition accuracy achieved was 87.8% using CDA with DTW, which was closely followed by QDA with DTW and DTW. Similar to the study mentioned in the previous

| **Classifying type** | **Average accuracy (%)** |
|---|---|
| DTW | 87 |
| CDA | 76 |
| QDA with DTW | 87.6 |
| CDA with DTW | 87.8 |
| Average distance | 82.1 |
| Five features | 50.2 |

TABLE 3.4: Li's accuracy results on different classifiers.

section, the systems created in this section makes use of statistical matching techniques for classification and are therefore user-dependent.

The system further attempted to carry out word recognition, which was done with the concatenation of the alphabet segments. According to the study, the word recognition accuracy significantly improves when using a dictionary filter or auto-correct, which is similar to the method described in Section 3.1.2.

### 3.1.6 Cursive Writing Recognition

Seniuk and Blostein [53] created a recognition system to recognise cursive writing—as opposed to block letters—via acoustic emission from a pen or stylus. The system captures audio using a Labtec computer microphone 333 with the plastic housing removed. The microphone is taped to the middle of the writing stylus, with the recording end facing the writing tip. All writing was done on a single sheet of standard 60lb HP laser printer paper attached to a Masonite clipboard.

A template matching technique with a range of different features was used. The features were: the raw power signal; the discrete sequence of magnitudes obtained from the power signal's peaks; and an ordered tree in a scaled representation. Gaussian smoothing was used to reduce noise in the audio data. The feature descriptor is done between the template signal and testing signal: integrating the absolute difference between the signals, comparing peaks between the signals, and finally comparing the structures obtained from scale space representations which is done similarly to Gaussian smoothing that was used except with increased scale. The template matching technique served as the classifier.

A disadvantage of this grouping approach is the large amount of data required to carry out recognition; template data is required for every word contained in the dictionary. As a proof of concept, a pre-defined set of words were selected and templates of the words were acquired.

The top portion of Figure 3.7 shows a sample of words drawn in cursive writing, that the system caters for. The bottom of the figure shows three audio signatures of the word "*mango*" which would be used to recognise the word.

Two data sets were collected, both of a set of words written in cursive handwriting: the "ABC" data set consisting of uppercase letters; and the "FOOD" data set consisting of words which were names of various food items, as seen in the top of Figure 3.7. Using the matching techniques, the accuracy achieved was above 70% for letters while words achieved 90% accuracy. The authors attribute the higher accuracy achieved by words to the more feature-rich nature of words than single letters in cursive writing.

### 3.1.7 Combination of Pen's Tip Motion and Writing Sounds

Schrapel et al. [52] built a system called *Pentelligence* that combines the acoustic emission from writing on paper and a sensor device embedded within a pen to recognise handwritten digits. The pen houses an ATmega328p microcontroller which is Arduino

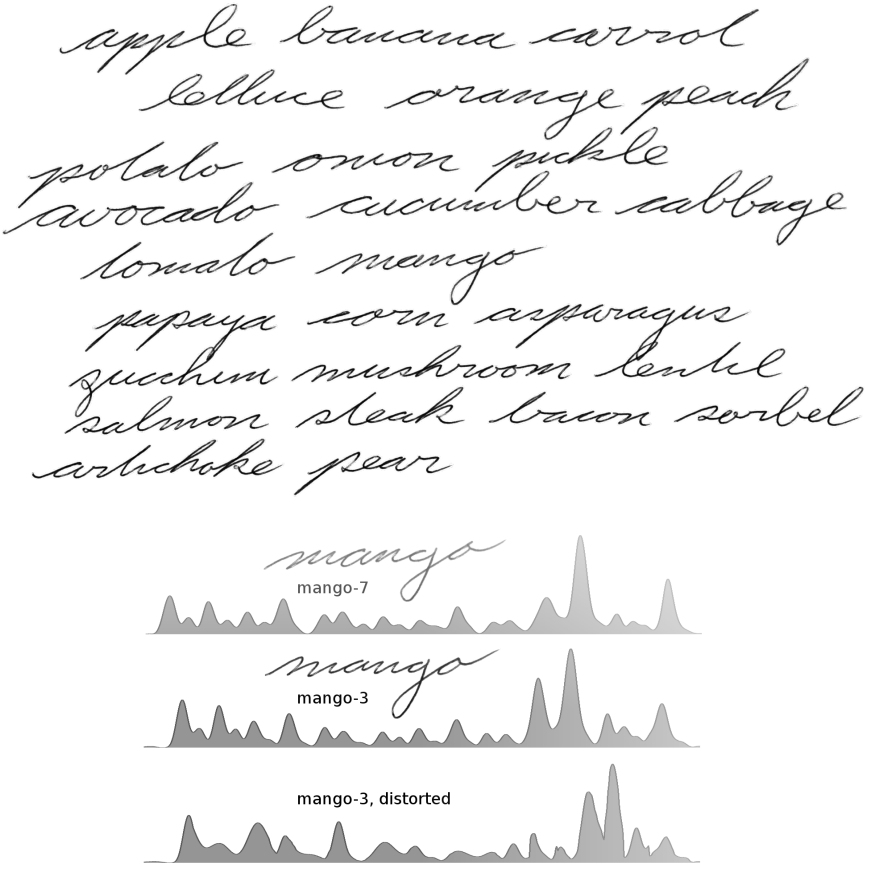


FIGURE 3.7: Sample of Seniuk and Blostein's data, from [53]: (top) example cursive words; (bottom) audio data in visual projection.

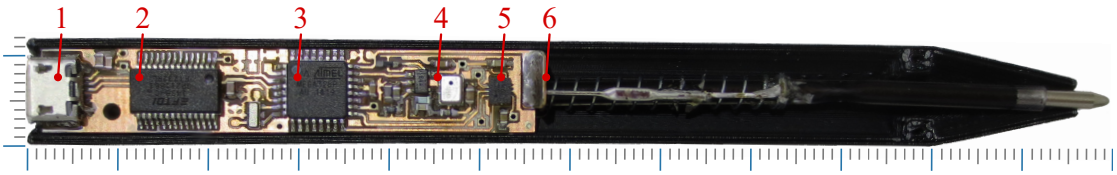


FIGURE 3.8: Schrapel et al.'s sensor embedded pen. 1: micro USB jack, 2: USB/UART converter, 3: microcontroller, 4: microphone with amplifier, 5: inertial measurement unit, 6: write or pressure sensor. [52].

bootloader-enabled for programming and is connected to a number of sensors. Figure 3.8 shows the close-up cross-section photograph of Schrapel et al.'s sensor pen. Referring to Figure 3.8, the sensors include the following:

- a pressure sensor mounted to the spring of the pen to detect contact with the paper or surface, labelled "6" in the figure.
- an analogue omni-directional microphone placed at the tip of the pen and connected to a circuit board, with a rail-to-rail operational amplifier used to then amplify and convert from analogue to digital the audio signal received at the microphone. This is labelled "4" in the figure.
- an inertial measurement unit which consists of an accelerometer and gyroscope to detect and measure motion, labelled "5" in the figure.

Audio processing is done by applying the Hilbert Transform that computes the Hilbert Envelope of the audio data. The transformed audio data together with the motion data from the inertial measurement unit forms the feature set for training and testing the classifier. A neural network is used as the classifier and the classes are digits from 0 to 9. The overall accuracy achieved was 78.38%.

Figure 3.9 shows the confusion matrix of their results per digit class recognised. The recognition accuracy across digit classes is observed to be relatively close, ranging between 69.8% and 88.5%. It is important to note that, even with a complex hardware setup employed and several sources of information, it is still possible for digits to be confused with each other.

### 3.1.8 Digit Recognition For Smart Watches

Chen et al. [10] created a system called *WritePad* to address the difficulties associated with device interaction on the small screens of smart watches. The system aimed to

recognise digits drawn on the back of a user's hand. Audio was collected using a microphone that was built in to the smart watch which, in this case, was the Huawei Watch I, running Android 4.3.

After a pre-processing which is done with wavelet transform algorithms, the system extracts features from the audio data via the Fast Fourier transform. The transform generates images which are used as input features to a classifier. Figure 3.10 shows examples of the wavelet transform images for digits 1,2 and 3, which are $(64 \times 64)$ pixels in size. These are used as features for classification.

A hybrid convolutional neural network with three layers of convolution each followed by a layer of max pooling was used for classification.

Two experiments were conducted to assess the accuracy of the proposed approach. The first experiment aimed to assess the accuracy of the system across 10 test subjects, with each subject performing each of the 10 digits 50 times. An overall accuracy of 92.75% was achieved. Figure 3.11 shows the confusion matrix for the first experiment. It is, however, unclear whether the testing subjects were included in the training set, in which case the system accuracy becomes dependent on a pre-training procedure.

A subsequent experiment was carried out to compare the effects of various noise levels emanating from different environments on the recognition accuracy. The three environments compared, and their corresponding final accuracies, were: a laboratory with

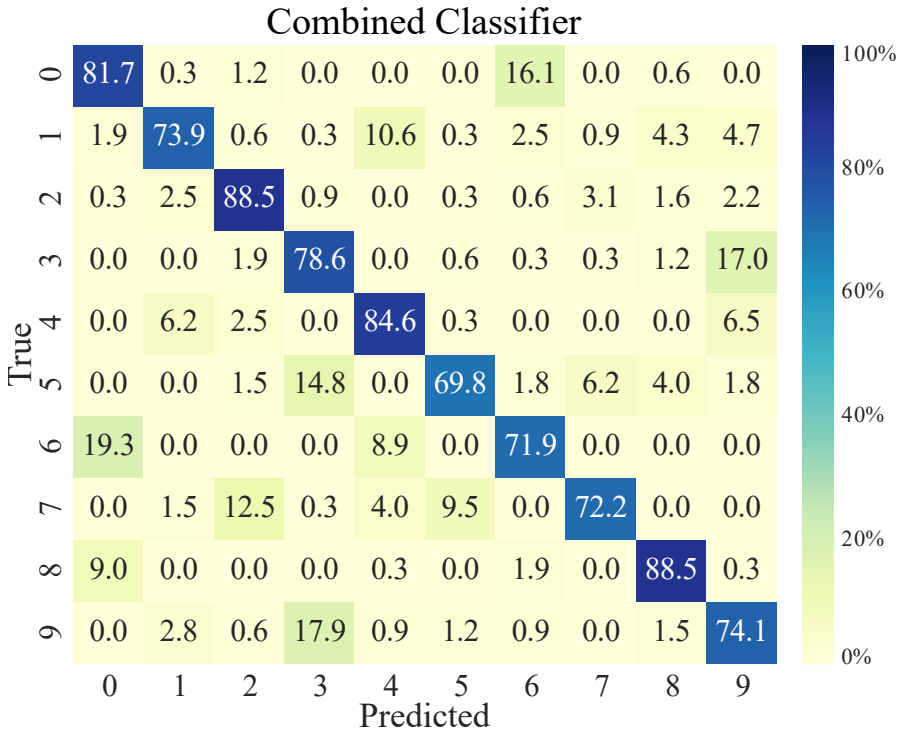


FIGURE 3.9: Confusion matrix of Schrapel et al.'s results [52].

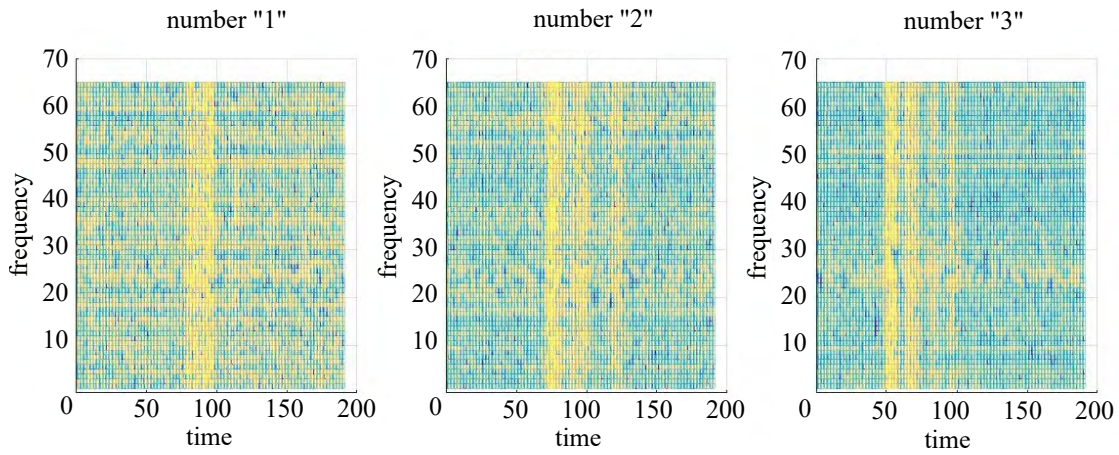


FIGURE 3.10: Chen et al.'s audio images as features for digits 1,2 and 3 [10].

94.46% accuracy, a dormitory with 90.55% accuracy and a canteen with 85.48% accuracy.

### 3.1.9 Letter Recognition for Smart Wearable Devices

Du et al. [16] built a system called *WordRecorder* that recognises written letters from acoustic signals generated by pens on a paper. The aim was to implement the system on small wearable devices like smart watches.

The audio input is captured from by a Huawei Watch I, running Android 4.3 and segmented into its constituent letters. The audio segment of each letter is sent further for processing. The system uses the Short Term Fourier Transform (STFT) as the main signal processing method. The features yielded from the STFT are transformed into a greyscale image which is then used as the source of input features to a classifier. Figure

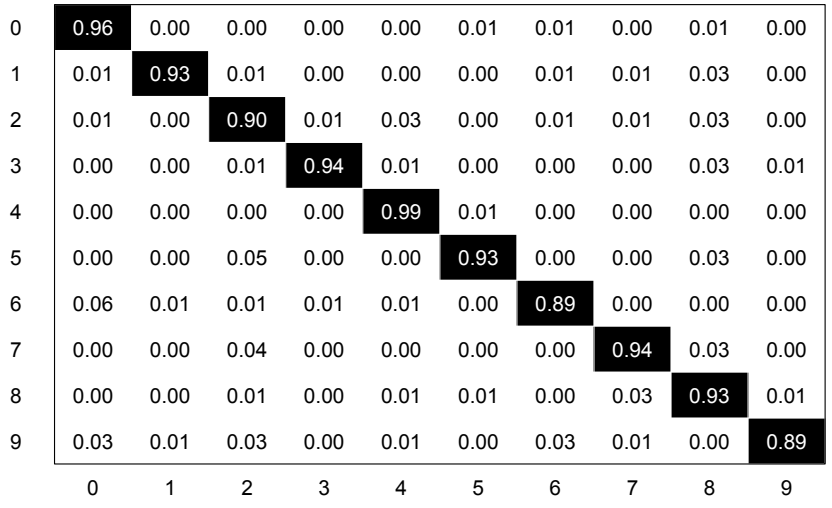


FIGURE 3.11: Chen et al.'s confusion matrix for the first experiment [10].

3.12 is an example of one such greyscale image produced from an audio signal, which is used to train the classifier to recognise uppercase letters of the alphabet.

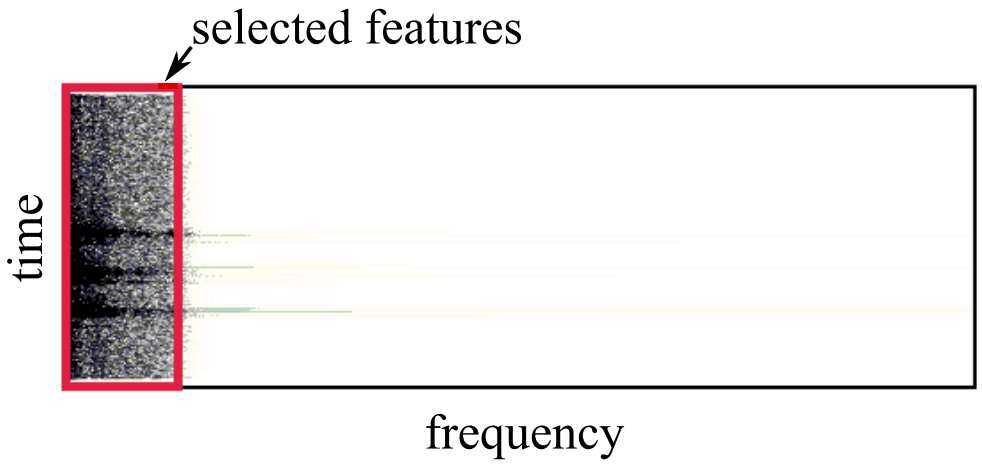


FIGURE 3.12: Du et al.'s greyscale image resulting from the application of the STFT to an audio signal.

The classifier used was a Convolutional Neural Network (CNN) which is a deep learning based approach. The aim was to take advantage of the analysis ability of CNNs on visual imagery. The structure of the CNN takes a hybrid approach which combines two types of CNNs, namely, LeNet-5 [29] and AlexNet [26] as follows: three convolutional layers with three pooling layers; two fully connected layers; finally, one output layer. The size of the kernels are 11 $\times$ 11, 5 $\times$ 5 and 3 $\times$ 3 for each layer respectively, and the pooling size is 3 $\times$ 3, with a stride of two. This intent behind using a hybrid approach is to utilise the structure of LeNet-5 and the convolutional layers of AlexNet as an advantage. A modified classifier of this type is also able to be implemented on smart devices.

On testing the system, an overall accuracy of 81% for pre-trained subjects and 75% overall accuracy was achieved for subjects that was not present during training. Figure 3.13 is a confusion matrix of the accuracies per letter. It is clear that the letters are generally recognised at a high accuracy.

Finally, a dictionary lookup is used with the predicted output, i.e. the chain of letters, to enhance the word recognition accuracy, similar to Yu et al. [69] mentioned previously.

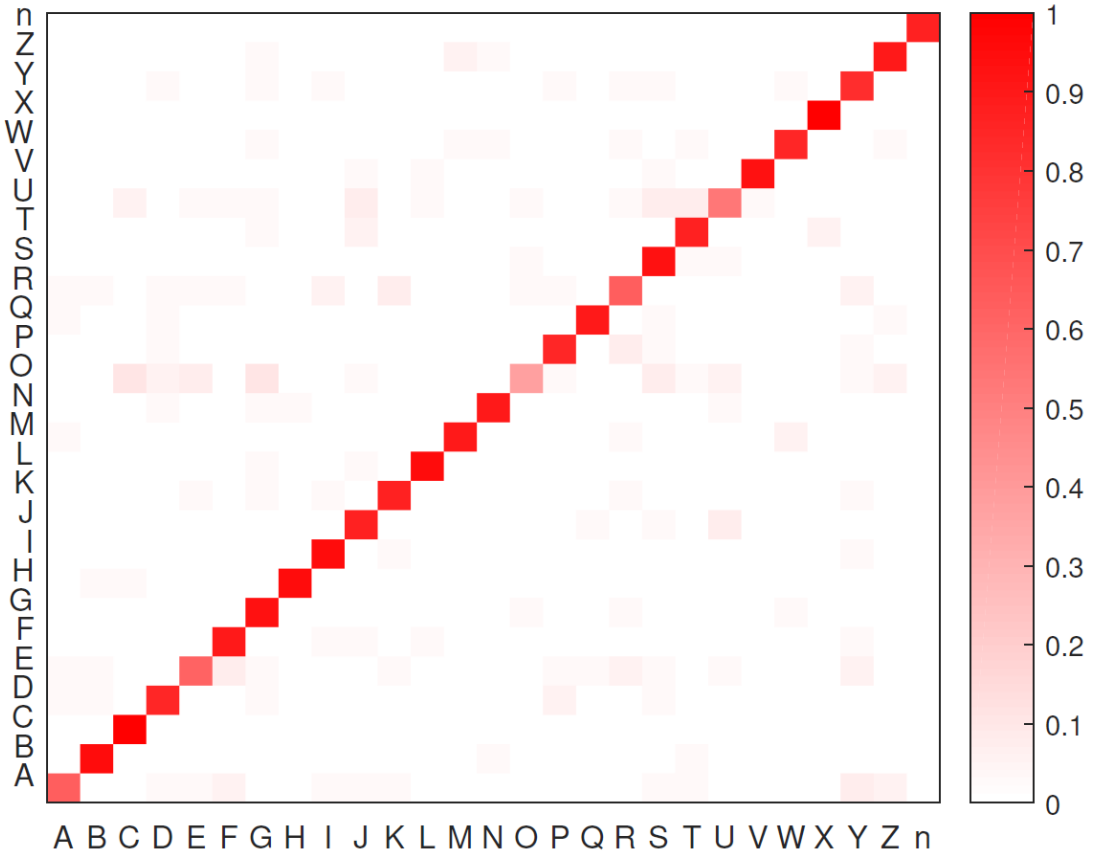


FIGURE 3.13: Du et al.'s confusion matrix of letter recognition accuracy [16].

## 3.2 Generic Audio Recognition Research

This section contains work that is less directly related as compared to Section 3.1. Instead, this section shows work using alternative feature descriptors etc., with speech recognition being the most prominent research field. While MFCC is quite popular in speech and non-speech, however, there are other descriptors and variations that are applied to certain problems.

The following subsections details these topics; pitch-range based feature set and gammatone frequency cepstral coefficients.

### 3.2.1 Environmental Sound Recognition Using Machine Learning Techniques

Uzkent et al. [57, 58] carried out a study that aimed to recognise various environmental sounds including: gunshots, glass breaking, screams, dogs barking, rain, engine sounds, and restaurant noise. They compared three feature extraction methods; the first two

methods were the MFCC and the range of pitch of the audio signal and the third method combined the first two methods. Furthermore, they compared several classifiers as follows: SVM with a linear kernel, SVM with a Gaussian kernel, a RBF neural network and a $k$ nearest neighbours classifier. The below Equation 3.6 shows the auto-correlation function,

$$\phi(t) = \sum_{n=-k}^{k} x(n)x(n+t). \tag{3.6}$$

$\phi(t)$ which is used to calculate the pitch range, where $x(n)$ is the input signal, $n$ is the moving window size, $t$ is the time offset and $k$ is the limiting bound of $n$. It measures the extent to which a signal correlates with a time-offset version of itself [57]. Equation 3.7 shows the appropriate definition of the auto-correlation function if the signal displays stationary random or periodic behaviour,

$$\phi(t) = \lim_{k\to\infty} \frac{1}{2k+1} \sum_{n=-k}^{k} x(n)x(n+t). \tag{3.7}$$

The pitch $P(i)$ is calculated when the auto-correlation is high,

$$P(i) = \frac{1}{T(i)}, \quad 1 < i < M. \tag{3.8}$$

Where $T(i)$ is the time delay between the first and second positive peak values of the auto-correlation function and $M$ is the total number of windows for any sound event. Equation 3.9 defines the two features $F_1$ and $F_2$ of the pitch range. $F_1$ is defined as the ratio of the maximum to the minimum of the pitch range. $F_2$ is defined as the ratio of the standard deviation $\sigma[P(i)]$ to the mean $\overline{P}$ of the pitch range which are defined in Equation 3.10:

$$F_1 = \frac{\max[P(i)]}{\min[P(i)]}, \quad F_2 = \frac{\sigma[P(i)]}{\overline{P}}, \quad \forall \;\; 1 \leq i \leq n. \tag{3.9}$$

$$\overline{P} = \frac{1}{k}\sum_{i=1}^{n} P(i), \quad \sigma[P(i)] = \sqrt{\frac{1}{k-1}\sum_{i=1}^{n}\left(P(i) - \overline{P}\right)^2}, \quad \forall \;\; 1 \leq i \leq n. \tag{3.10}$$

Table 3.5 summarises the results obtained by the feature descriptors and classifiers compared. According to the results, the MFCC outperformed pitch-range features in terms of accuracy. Furthermore, the combination of the MFCC with pitch-range features out-performed the individual feature descriptors across all classifiers. It is also observed

that the SVM with the Gaussian kernel out-performed all other classifiers regardless of the features used.

### 3.2.2 Environmental Sound Recognition Using Hidden Markov Models

Shao et al. [54] constructed a system that compared variants of the MFCC and the gammatone frequency cepstral coefficients (GFCC) feature descriptors in recognising speech under several types of environmental noise. The GFCC is an alternative to the MFCC that appears to provide greater robustness under high-to-severe environmental noise conditions, at a comparable computational efficiency. The GFCC was developed with more robustness and noise suppression in acoustic audio processing applications in mind. Equation 3.11 describes the gammatone filter $G_{f,t}$.

$$G_{f,t} = t^{a-1} \exp(-2\pi bt) \cos(2\pi f_c + \varphi), \qquad t \geq 0, \tag{3.11}$$

where $a$ is the peak value constant that controls gain, which is typically set to $a = 4$, $t^{a-1}$ is the time onset, where, $b$ is the rectangular bandwidth that is proportional to the characteristic centre frequency $f_c$ and $\varphi$ is the initial phase which is, however, typically set to zero i.e. $\varphi = 0$ [54]. Equation 3.12 shows the GFCC feature extraction after gammatone filtering $G_{f,t}$ is down sampled with a cubic root function resulting in $G_c(k)$, where $k$ is the frame index, $c$ is the channel index and $n$ is the total number of channels [54].

$$C_i(k) = \sqrt{\frac{2}{n}} \sum_{c=0}^{n-1} G_c(k) \cos\left(\frac{i\pi}{2n}(2c-1)\right), \qquad \forall \;\; 0 \leq i \leq n-1. \tag{3.12}$$

The following variants of the MFCC and GFCC were used and compared:

- a 36-dimensional MFCC feature labelled "MFCC_D_A".

| **Classifier** | **Accuracy (%)** | | |
|---|---|---|---|
| | **Pitch-range** | **MFCC** | **Pitch-range with MFCC** |
| SVM with linear kernel | 65.3 | 67.1 | 84.6 |
| SVM with Gaussian kernel | 72.4 | 82.6 | 87.9 |
| RBF neural network | 55.0 | 70.2 | 81.8 |
| $k$ Nearest Neighbours | 68.6 | 79.8 | 86.4 |

TABLE 3.5: Uzkent et al.'s accuracy results on different feature descriptors and classifiers.

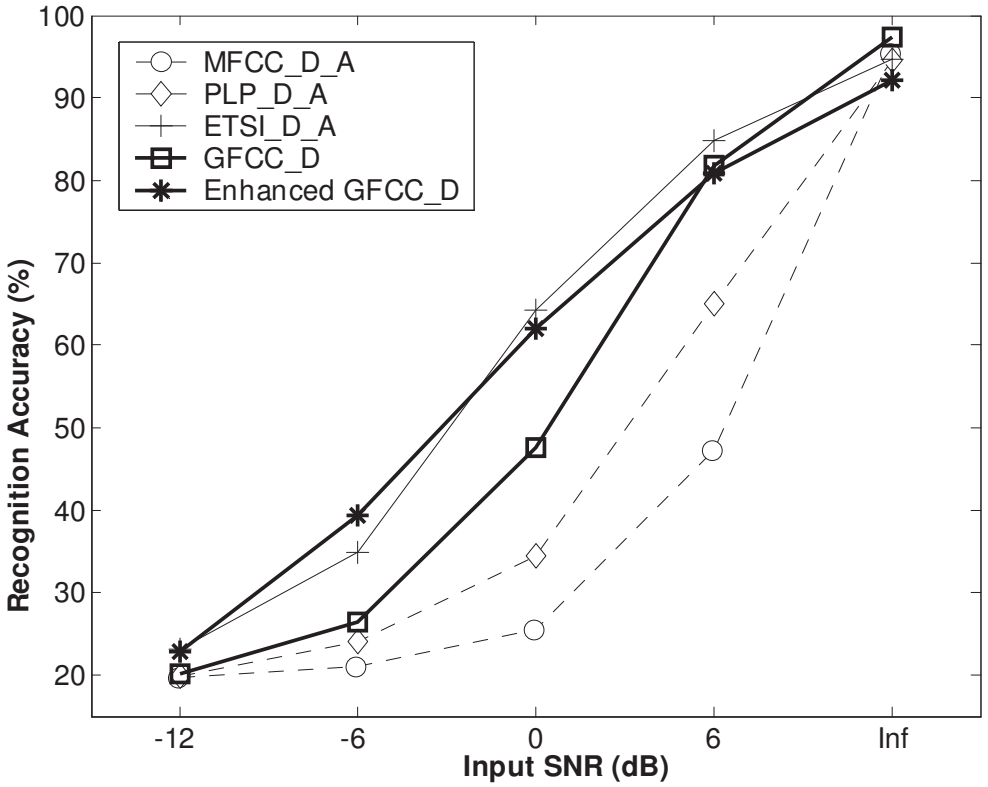


FIGURE 3.14: Shao et al.'s results on experiments with SSN.

- a 36-dimensional cepstral coefficients derived by perceptual linear predictive analysis labelled "PLP_D_A".
- an enhanced 36 dimensional MFCC feature derived by the ETSI-AFE speech recognition process labelled "ETSI_D_A".
- a 58-dimensional GFCC feature labelled "GFCC_D".
- An enhanced GFCC_D feature which uses feature reconstruction and uncertainty decoding which is a noise decoding process, labelled "Enhanced GFCC_D".

A hidden Markov Model (HMM) was used for classification. Two types of experiments were performed. In the first experiment, speech phrases from thirty-four speakers were mixed with speech-shaped-noise (SSN) at increasing signal-to-noise-ratio (SNR) values. Figure 3.14 shows the recognition accuracy of each of the feature descriptors at different SNR values. The results demonstrate the robustness of the GFCC in extremely noisy conditions.

The second experiment evaluated the accuracy of the various feature descriptors under alternative noisy conditions, namely, factory noise, speech babble, destroyer operation room and F-16 cockpit. The speech phrases from the thirty-four speakers was again mixed with each type of noise using noise generators, as in the previous experiment. The

| Feature set | Accuracy (%) | | | |
|---|---|---|---|---|
| | **Factory** | **Babble** | **Destroyer** | **F-16** |
| MFCC_D_A | 61.31 | 67.97 | 67.67 | 59.69 |
| PLP_D_A | 78.44 | 81.75 | 84.17 | 76.19 |
| ETSI_D_A | 88 | 90.39 | 90.44 | 90.03 |
| GFCC_D | 93.42 | 91.81 | 92.78 | 91.86 |
| Enhanced GFCC_D | 85.69 | 86.47 | 86.47 | 85.36 |

TABLE 3.6: Shao et al's accuracy results on different noise background.

results are summarised in Table 3.6. The table shows that the best overall performance is obtained by the GFCC feature set without enhancement.

The GFCC clearly out-performs the MFCC under very noisy conditions. However, the system proposed in this thesis is not currently expected to be used under very noisy conditions. The MFCC has also been demonstrated extensively to be robust in non-speech-based applications. Therefore, an investigation of the efficacy of the GFCC towards use in the proposed system is left to future work.

## 3.3 Summary

This chapter described related work in the field of audio recognition.

Section 3.1 discussed studies that are closely related to the work in this research i.e. related studies that are specifically focused on recognising shapes, digits or letters. Table 3.7 shows a summary of these related studies. The studies were of three types: those that recognise characters written on a surface by a stylus or another type of object; and those that recognise characters for smart wearable devices.

A variety of approaches were used to recognise characters, including recognising characters separately, decomposing characters into strokes and grouping characters based on their stroke counts. Furthermore, some studies grouped letters recognised and used dictionary lookups to recognise words. Various feature descriptors have been used, including various combined, modified and fine-tuned version of these descriptors. The descriptors investigated include: amplitude density function, mean amplitude, power signal, the MFCC and others. Classification techniques used included: $k$ nearest neighbours, template matching, regression and SVM.

By examining the hardware used for the various studies, it is observed that they unanimously used embedded condenser type, mostly those embedded into mobile or wearable smart devices. One of the goals of this research is to investigate whether an alternative low-cost hardware capture solution in the form of one or more piezo microphones can

be used to obtain accurate recognition accuracies. Hence, the main research question set forth in Chapter 1 involves investigating whether piezo microphones can provide sufficient sound definition and quality to enable audio character recognition.

Various feature extraction methods were compared, with varied and comparable levels of success, which depends on the classification technique used. The majority of the studies either made use of various forms of template matching or neural networks. Other studies either used SVMs or $k$NNs.

This research selects the MFCC as the feature extraction method of choice which is a suitable choice considering that it has shown promise in terms of effectiveness and robustness for various audio recognition problems [58], including the ones discussed in this chapter.

As regards classification, template matching was noted as being user-dependent, which is very undesirable. In contrast, machine learning techniques can be trained to be user-independent. Neural networks, especially deep neural networks, may be considered for this task but can be challenging to work with given their complexities in terms of hyperparameter tuning, overfitting, the possibility of not generalising and possibly diverging. On the other hand, SVMs always converge, perform well on even small-sized datasets, and can have their hyperparameters tuned with relative ease. Therefore, SVMs are selected for classification in this research. The application of neural networks may be implemented and compared to SVMs in future work.

Section 3.2 discussed more general audio recognition research studies that use feature descriptors and classifiers. The GFCC was shown to be an alternative feature descriptor and a derivative of the MFCC which is more robust under extreme noise conditions. The proposed system is expected to be used under conditions of low to moderate noise. The use of the GFCC will therefore be investigated in future.

This forms the basis for the next chapter which details the techniques that make up the two components for the proposed system's audio recognition process, namely, the MFCC feature descriptor and SVM classification.

| Study | Recognition classes | Audio Hardware | Feature Extraction | Classifier | User dependency | Overall Accuracy (%) |
|---|---|---|---|---|---|---|
| [70] | 7 strokes | Android mobile devices | Spectrum density function from FFT | $k$-nearest-neighbours | Heavy | 90.3 |
| [69] | 26 Letters | Generic mobile devices | MFCC | SVM | Moderate | 50-60 |
| [62] | 26 Letters and 11 Short Words | Mobile device with dual microphone and speaker | LLAP | Template matching | Heavy | 92.3 |
| [33] | 26 Letters | MacAir microphone | MFCC with mean amplitude | DTW | Heavy | 83.7 |
| [32] | 26 Letters | MacAir and Android mobile microphone | MFCC with additional features | CDA with DTW | Heavy | 87.8 |
| [53] | 26 cursive letters and 26 cursive words | Labtec computer microphone | Power signal derived features | Template matching | Heavy | 70 |
| [52] | Digits | ATmega328p microcontroller with sensors | Hilbert envelope with sensor data | Neural network | Minimal | 78.38 |
| [10] | Digits | Embedded microphone of a Huawei Watch I | Wavelet to Fourier transform then images | Convolutional neural network | Minimal | 92.75 |
| [16] | 26 letters | Embedded microphone of a Huawei Watch I | STFT | Convolutional neural network | Minimal | Seen: 81<br>Unseen: 75 |

Table 3.7: Summary of related works.

# Chapter 4

# Techniques of Audio Recognition

This chapter details the audio recognition techniques used by the system to extract features from audio data. The main focus of the proposed system, once audio data has been captured from a piezo microphone, is to utilise a feature extraction algorithm to extract audio features from the audio. For this task, the MFCC feature descriptor was selected as a robust and promising technique.

Given the feature vector produced by the feature descriptor, a machine learning technique is used to classify the data as being one of a pre-defined set of classes, in this case: the fundamental shapes described in Section 1.3, the digits 0 to 9 and the upper case letters from A to Z. The character prediction accuracy is certainly influenced by the choice of machine learning technique. This research uses Support Vector Machines (SVMs) which have been shown to be effective in a variety of classification problems [58].

This chapter is structured as follows: Section 4.1 provides the theories and processes of the MFCC feature descriptor. Section 4.2 provides information about the SVM used for classification in this research. The chapter is then summarised.

## 4.1 Mel-Frequency Cepstral Coefficients

The MFCC feature descriptor is designed via a methodology to simulate human auditory perception which resides in factors such as short-term changes in pitch and loudness which make up unique sounds [6], also referred to as phonemes. The objective is to extract a feature vector from the input audio signal which results in a representation of the dynamic nature of speech in the form of the sequence of phonemes that were produced.

Figure 4.1 is a flow diagram of the MFCC procedure depicting each step of the process applied to an audio signal received as input to produce the target feature vector. The audio signal goes through enhancements, then conversions, filtering and transformation. The following subsections follow the progression through this figure.

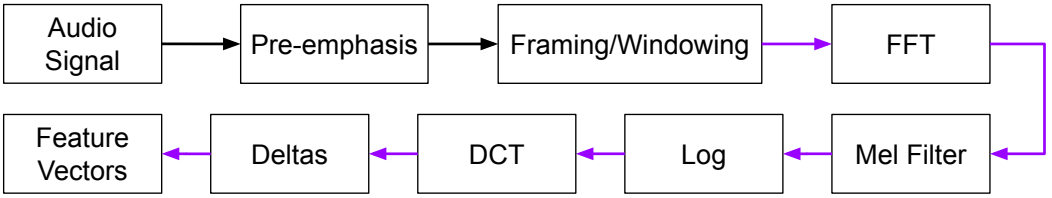


FIGURE 4.1: Overview of the MFCC feature descriptor.

### 4.1.1 Pre-emphasis

The pre-emphasis step is carried out to increase low energies of higher frequencies and reduce excessive energies of lower frequencies in the audio signal, this will suppress low frequencies that does not hold much information and can affect later steps in the process. This phenomenon tends to occur in natural audio signals. In this application, pre-emphasis takes the form of a high-pass filter which emphasises the energies at higher frequencies and the does the opposite to energies at lower frequencies [42]. This results in a better distribution of amplitudes at relative frequencies. Equation 4.1 defines the pre-emphasis process represented by $c^{(1)}$,

$$c_k^{(1)} = c_k^{(0)} - \alpha\, c_{k-1}^{(0)} \quad \forall \quad k \in \{1, 2, \ldots, N\}. \tag{4.1}$$

Where $c^{(0)}$ is the input audio signal, $k$ is a sampling point, $N$ is the total number of sampling points and $\alpha$ is a pre-emphasis factor that is usually set to between 0.9 and 1.0, and typically set to a default of 0.95 which is used here.

Figure 4.2b illustrates the effect of applying pre-emphasis to an audio signal which is shown in Figure 4.2a.

### 4.1.2 Framing and Windowing

The pre-emphasised signal $c^{(1)}$ is framed into $J$ overlapping frames with an overlap of $O$ samples between frames, and with each frame having a width of $W$ samples. This gives rise to frames $\{c_\tau^{(1)} | \tau \in \{1, 2, \ldots, J\}\}$, where $\tau$ is the frame number. Typically, the overlap between frames is set to $O = 10$ milliseconds (ms) and the frame width is set to

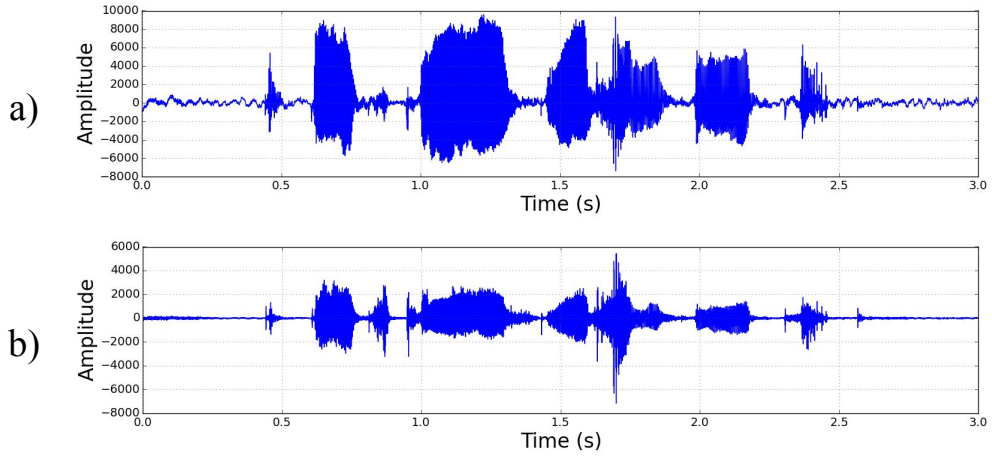


FIGURE 4.2: The pre-emphasis process: a) Original input audio signal; b) pre-emphasis applied to the audio signal.

$W = 25$ms, and these values were used in this implementation. All of the steps of the MFCC from this point onward apply to each frame separately. For illustrative purposes, Figure 4.3 provides an illustration of four overlapping frames, represented by the blue, yellow, green and red boxes super-imposed onto a portion of an input audio signal.

Given that frames have a width of $W$ samples, $k$ will from now be an index that specifies the sample number in a given frame $\tau$ and will take on values $k \in \{1, \ldots, W\}$, noting that every frame has the same size.

Windowing is the process in which a window function is applied to, i.e. convolved with, the samples of a signal. This helps reduce both spectral leakage, and breaks or static in

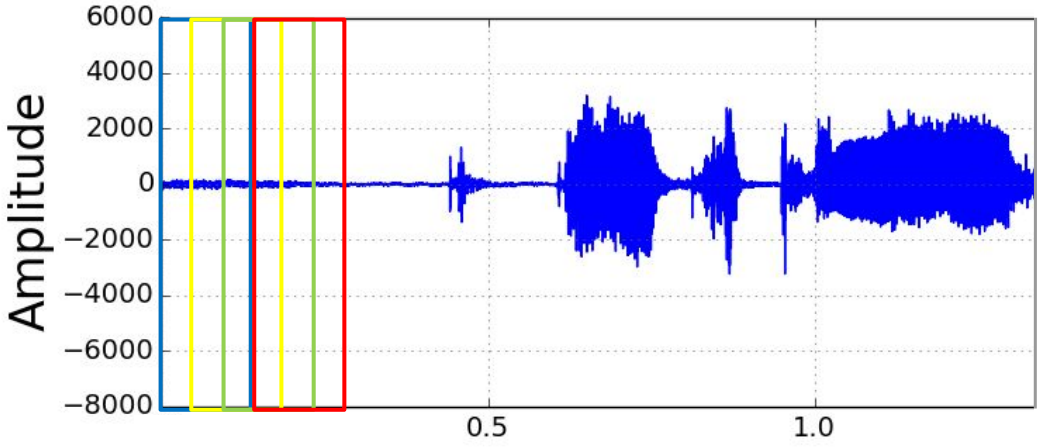


FIGURE 4.3: Part of the input audio signal with four example frames (blue, yellow, green and red boxes) super-imposed on it, for illustrative purposes.

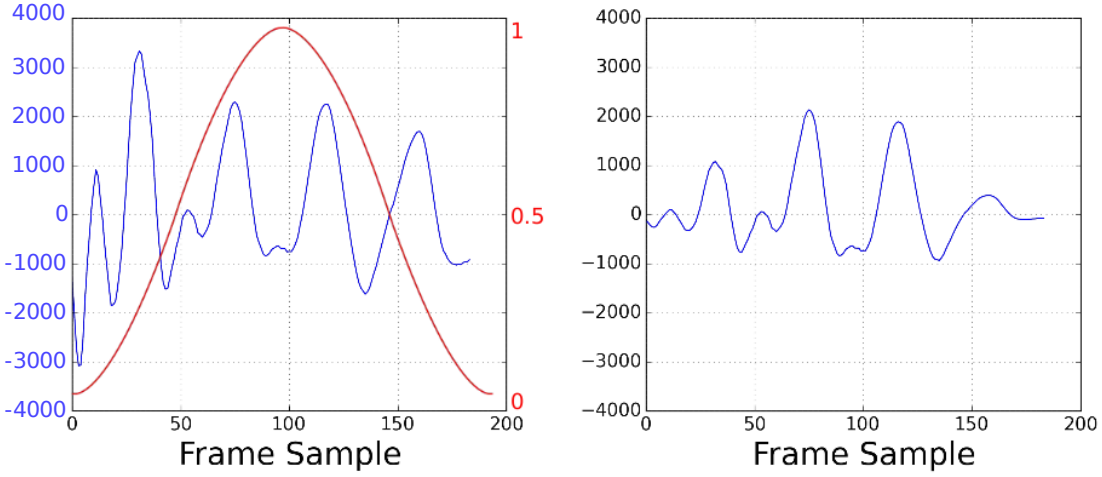


FIGURE 4.4: The windowing process: a) The audio signal of a single frame with a Hamming window super-imposed on it; b) The result of applying the Hamming window to the audio signal.

the signal [6]. In this case, the window function used was a Hamming window which was applied to each frame $\tau$. Figure 4.4a illustrates a Hamming window super-imposed on the audio signal of a frame, resulting in the signal shown in Figure 4.4b. Equation 4.2 describes a Hamming window of width $W$, and all other symbols have been previously defined,

$$H(k) = 0.54 - 0.46 \cos \frac{2\pi k}{W-1} \quad \forall \quad k \in \{0, ..., (W-1)\}. \tag{4.2}$$

This can be applied to each frame from the previous step $c_{\tau}^{(1)}$ to get $c_{\tau}^{(2)}$ as follows:

$$c_{\tau,k}^{(2)} = H(c_{\tau,k}^{(1)}) \quad \forall \quad k \in \{0, ..., (W-1)\}. \tag{4.3}$$

### 4.1.3 Fast Fourier Transform

The efficient form of the Discrete Fourier Transform is the Fast Fourier Transform (FFT). The FFT converts an audio signal from the time domain into its frequency domain. It is applied to the result of the previous step i.e. $c_{\tau}^{(2)}$ as follows:

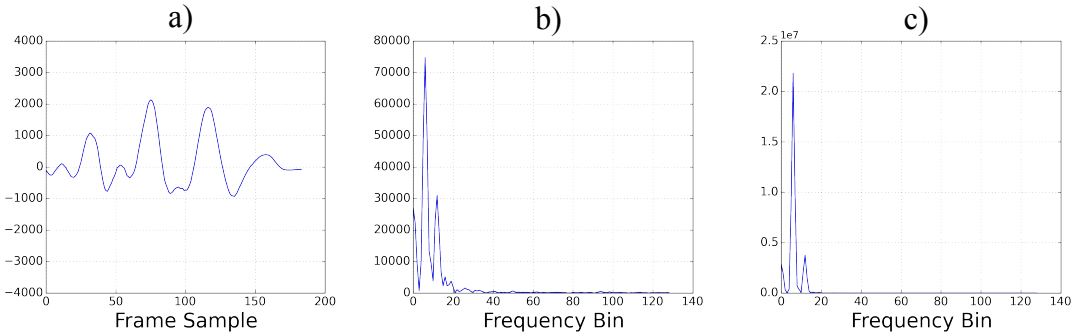


FIGURE 4.5: Application of the FFT: a) the windowed audio signal of the current frame; b) the frequency spectrum obtained by applying the FFT to the signal; c) estimate of the spectral power of each frequency in the audio signal.

$$c^{(3)}_{\tau,f} = \sum_{k=0}^{\frac{W}{2}-1} c^{(2)}_{\tau,\mathcal{EVEN}\{k\}} \exp\left(-i\ 2\pi \frac{kf}{2W}\right) + \exp\left(-i\ 2\pi \frac{kf}{W}\right) \sum_{k=0}^{\frac{W}{2}-1} c^{(2)}_{\tau,\mathcal{ODD}\{k\}} \exp\left(-i\ 2\pi \frac{kf}{2W}\right) \quad \forall \quad f \in \{0, 1, ..., F\}. \tag{4.4}$$

Where $f$ denotes each frequency bin, $F$ is number of frequency bins used which is typically $F = \left(\frac{W}{2} - 1\right)$, $c^{(3)}_{\tau,f}$ is the output frequency spectrum corresponding to frame $\tau$ which is clearly shown by the change in subscript from the time domain represented by $k$ to the frequency domain now represented by $f$, and all other symbols are identical to the same symbols in Equation 4.3.

Following the production of the frequency spectrum, estimates of the spectral power of each frequency in the frame $\tau$ is obtained as follows:

$$c^{(4)}_{\tau,f} = \frac{1}{W} \mid c^{(3)}_{\tau,f} \mid^2. \tag{4.5}$$

where all symbols have been previously defined. Figure 4.5 visually illustrates the absolute value of the frequency spectrum produced, and the estimate of the spectral power of each frequency.

### 4.1.4 Mel Scale Filtering

Mel-scale filtering is applied to the frequency power spectrum to obtain a mel-frequency spectrum which is modelled to mimic the human auditory perception in relation to pitch.

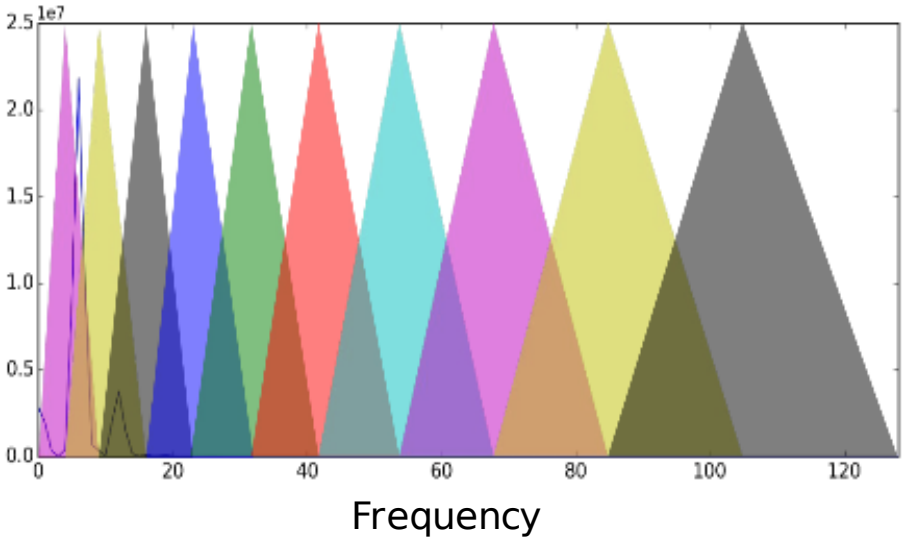


FIGURE 4.6: Visual illustration of Mel filters with $R = 10$, superimposed onto the frequency spectrum.

Mel-scale is a non-linear pyscho-acoustic scale that measures such pitch via auditory distance of humans [56].

Mel filtering involves assigning each frequency in the power spectrum into two consecutive sets of weighted bins on the Mel scale. These bins are also referred to as Mel filters. To determine the positions of the filters in the frequency spectrum, the smallest and largest frequencies in the spectrum are converted to the Mel scale and the interval between these frequencies is segmented into a desired number of equally spaced segments in the Mel scale, which define the bin values of the filters. Equation 4.6 describes the transformation of a frequency $f$ into its Mel-scale equivalent $m$,

$$m = 2595 \log \left(1 + \frac{f}{700}\right). \tag{4.6}$$

Converting the bin back to the frequency domain results in bin widths that increase proportional to the frequency, given that the Mel scale is logarithmic in nature. Equation 4.7 describes the transformation of a frequency in the Mel scale $f_m$ back into its frequency equivalent $f$,

$$f = 700\left(10^{\left(\frac{m}{2595}\right)} - 1\right). \tag{4.7}$$

Given that $R$ filters are required, a total of $R + 2$ bin edges, where the edges mark

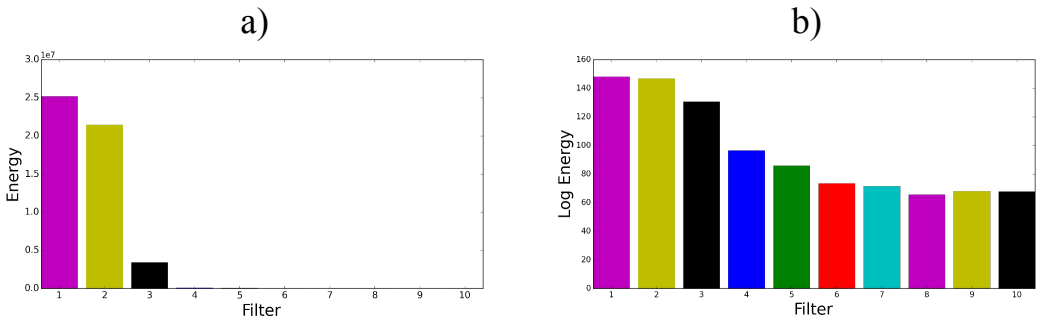


FIGURE 4.7: Illustration of Mel filtering for the case where $R = 10$: a) energies corresponding to each filter; b) log-energy corresponding to each filter.

the boundaries between bins, given by $\{t_0, t_1, \ldots t_R, t_{R+1}\}$ are obtained through this procedure. The filters are given by the function $T_r(f)$ below:

$$T_r(f) = \begin{cases} 0 & \text{if} \quad f < t_{r-1} \\ \dfrac{f - t_{r-1}}{t_m - t_{r-1}} & \text{if} \quad t_{r-1} \leq f < t_r \\ \dfrac{t_{r+1} - f}{t_{r+1} - t_m} & \text{if} \quad t_m < f \leq t_{r+1} \\ 0 & \text{if} \quad f > t_{r+1} \end{cases} \tag{4.8}$$

For illustrative purposes, Figure 4.6 visually depicts a case where $R = 10$ filters have been produced and have been super-imposed onto the frequency spectrum. It is observed that the width of the filters increases as the frequencies increase.

### 4.1.5 Logarithmic function

A human's perception of loudness is on a logarithmic scale. Therefore, after computing the energy of the power spectrum $c_\tau^{(4)}$ obtained earlier corresponding to each of the filters $T_r(f)$ to get the log-power spectrum corresponding to each filter, the logarithm of the result is computed which results in a pseudo-human auditory perception of loudness of corresponding to the frequencies in each filter $r$ and frame $\tau$, as follows:

$$c_{\tau,r}^{(5)} = \log\left[\sum_{f=0}^{F} c_{\tau,f}^{(4)} \cdot T_r(f)\right] \quad \forall \quad r \in \{1, \ldots, R\} \tag{4.9}$$

Note that $c^{(5)}_{\tau,r}$ is the log-power spectrum corresponding to a specific filter of the current frame $\tau$ and is now therefore indexed by $r \in \{1, \ldots, R\}$, and $R$ is the number of desired filters.

For illustrative purposes, Figure 4.7a shows the energies of the audio signal in previous figures corresponding to each filter for a case where $R = 10$ filters have been used. Figure 4.7b visually illustrates the corresponding log-energies of the filter energies in Figure 4.7a.

### 4.1.6 Discrete Cosine Transform

The Discrete Cosine Transform (DCT) is commonly used for audio compression. The compression removes redundancies and duplicates from the audio. Applying this transform to the log-power spectrum results in a "spectrum of the spectrum" which is called a "cepstrum". The cepstrum $c^{(6)}_{\tau,p}$ of the current frame $\tau$ is a spectrum of the log-power spectrum with $R$ coefficients indexed by $p \in \{1, \ldots, R\}$ and given by:

$$c^{(6)}_{\tau,p} = \sum_{r=1}^{R} c^{(5)}_{\tau,r} \cos\left[\frac{p\,(2r+1)\pi}{2R}\right], \quad \forall \;\; p \in \{1, \ldots, R\} \tag{4.10}$$

where the set $\{c^{(6)}_{\tau,p} | p \in \{1, \ldots, R\}\}$ is a set of the desired cepstral coefficients (CCs) for the current frame, and all other symbols are the same as those in previous equations. In practice, $R = 26$ filters are used to obtain 26 final CCs, but only the first 13 are typically used [35], attributed to the fact that CCs at the higher end of the cepstrum represent extremely fast-paced changes in frequency that don't appear in practice and don't appear to assist in audio recognition. Therefore, the symbol $Y = 13$ will be used to refer to the number of CCs that were selected out of a total of $R$ CCs produced previously.

Figure 4.8b is an illustration of the CCs produced on the filter log-energies shown in Figure 4.8a.

### 4.1.7 Deltas Features

The CCs obtained in the previous step provide a good representation of the underlying phonemes in the input audio signal, but are confined to specific frames. The "deltas" or derivatives of the CCs can be computed and used to provide information about the

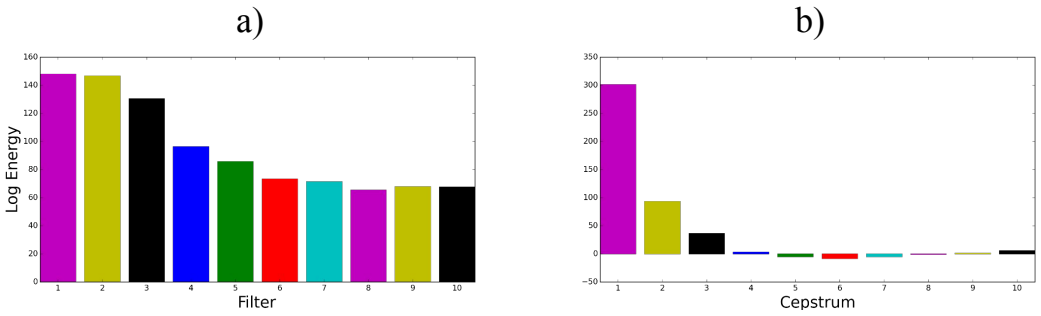


FIGURE 4.8: Production of cepstral coefficients for the case where $R = 10$: a) log-energy corresponding to each filter; b) final cepstral coefficients.

dynamics in the CCs i.e. how the CCs change across frames. They are therefore added to the final feature vector along with the CCs to represent the signature of the input audio. The first and second derivatives $\Delta c_{\tau,j}^{(7)}$ and $\Delta\Delta c_{\tau,j}^{(7)}$ are computed as follows:

$$\Delta c_{\tau,p}^{(7)} = c_{\tau+1,p}^{(6)} - c_{\tau-1,p}^{(6)} \tag{4.11}$$

$$\Delta\Delta c_{\tau,p}^{(7)} = \Delta c_{\tau+1,p}^{(7)} - \Delta c_{\tau-1,p}^{(7)} \tag{4.12}$$

### 4.1.8 Final Feature Vector

The CCs, first derivatives and second derivatives computed previously make up the final feature vector of frame $\tau$ given by $V_\tau$ shown below:

$$\mathbf{V}_\tau = \left\langle c_{\tau,p}^{(6)} \, , \; \Delta c_{\tau,p}^{(7)} \, , \; \Delta\Delta c_{\tau,p}^{(7)} \mid p \in \{1, \ldots, Y\} \right\rangle \tag{4.13}$$

Finally, the feature vectors of all frames $\tau \in \{1, 2, \ldots, J\}$ are concatenated to make up the final feature vector $\mathbf{V}$ given by:

$$\mathbf{V} = \left\langle \mathbf{V}_{\tau=1} \, , \; \mathbf{V}_{\tau=2} \, , \; \ldots \, , \; \mathbf{V}_{\tau=J} \right\rangle \tag{4.14}$$

## 4.2 Support Vector Machines

The Support Vector Machine (SVM) classification technique is a supervised machine learning technique which is a composite of different statistical learning models. It was initially created for binary classification [11] but has been adapted for multi-class classification problems as well. The SVM classification technique has many advantages [48] including the following:

- It is effective in high-dimensional spaces.
- It maintains its effectiveness when the number of dimensions exceeds the number of samples, it utilises only a subset of training points—called "support vectors"—in the decision function, which makes it memory efficient.
- It always leads to a converged model.
- Its hyper-parameters can be tuned with relative ease.
- It is versatile due to the possibility of using a number of different Kernel functions that can be specified for the decision function.
- Other than the common kernels provided, it is also possible to specify custom kernels.

SVMs are therefore a favoured technique for data classification. Although SVM is considered simpler to use than Neural Networks, and can be used to get superior results in many contexts if used correctly [22].

The following subsections will discuss the theory behind support vector classification used in SVMs, kernels, feature scaling and grid-search using cross-validation.

### 4.2.1 Support Vector Classification

Support vector classification aims to find a solution in the form of an optimal discriminant—a decision boundary—function or a hyperplane with maximum margin within a high-dimensional feature space [12]. Figure 4.9 shows a basic illustration of a binary linear support vector classification problem. Within the Cartesian plane of the $(X_1, X_2)$ feature space are data points of two classes represented by red crosses and blue circles. An optimal hyperplane is drawn as the discriminant function which separates and classifies the two classes, with the margin maximised until the first subset of data points in each

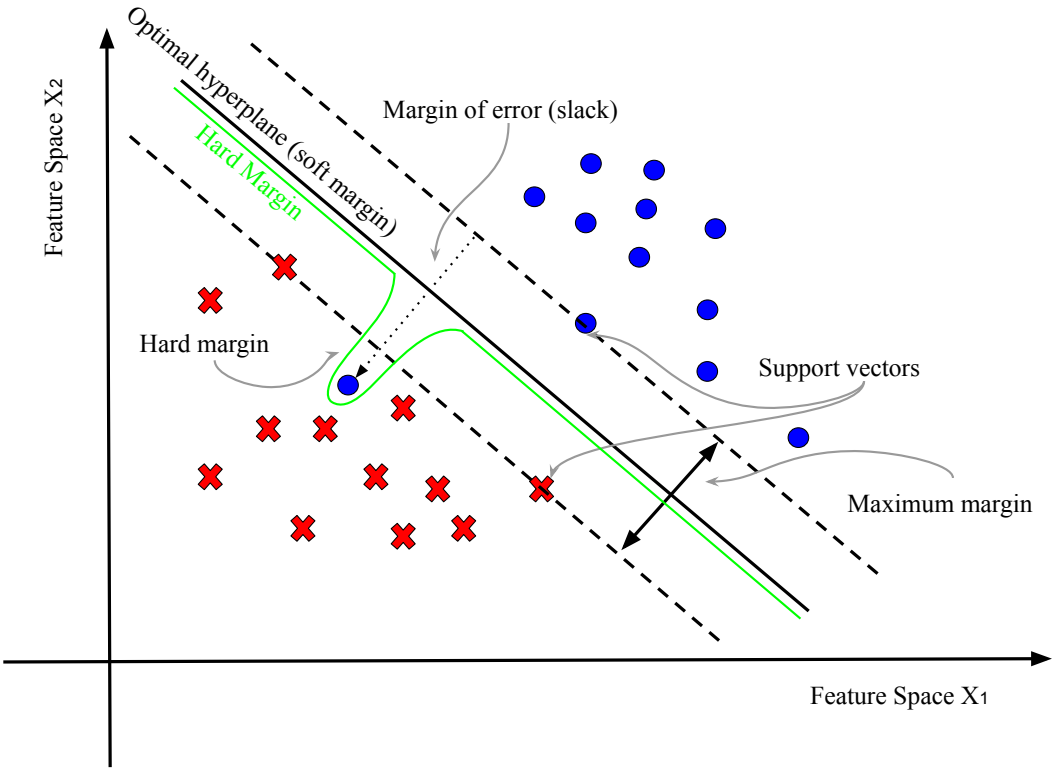


FIGURE 4.9: Basic illustration of a support vector classification. [50]

class is reached. This subset of data points in each class support the position of the discriminant function and are hence given the name "support vector".

When classifying realistic data, noise is definitely expected to occur in the form of data points from either class overlapping with data points of the opposing class. In terms of the illustration in Figure 4.9, it is seen that a blue data point is located in, and overlaps with, the region mostly occupied by red data points, and is therefore considered to be noise. Under hard margin optimisation—represented by the light green line in the figure—the margins are determined absolutely with no data points within the margin while maintaining class separation. This poses no issue with data sets that are perfectly separable. However, data with noise will result in the decision boundary to have undesired extreme contortions resulting in a heterogeneous class space.

Soft margin optimisation allows data points—presumably noise—to be placed outside their own class space. When a data point is located outside its class space, it is given an error margin or slack. A penalty meta-parameter—usually denoted $C$—that tunes the scale of error margin allowed, contributes to the overall discriminant function to compensate for noise. This will allow the decision boundary to be smoother, resulting in a more homogeneous class space instead [50]. This is illustrated in Figure 4.9 by the

optimal hyperplane (soft margin).

### 4.2.2 Kernels

In Section 4.2.1, Figure 4.9 used a linearly separable data set as an example. However, with realistic problems, the data sets are not likely to be linearly separable, which therefore requires existing data to be mapped to a higher-dimensional feature space in which the data can be separated linearly. SVM kernels become of great use in these situations. Kernels create an alternative solution for obtaining a more complex space. Mapping data onto higher dimensions helps make data linearly separable [50].

A kernel's validity is determined by its conformance to Mercer's theorem [37] such that the continuous symmetric kernel $K \in \mathbb{R}$ exhibits positive definite behaviour where convergence is absolute, thus a valid kernel qualifies as Mercer's kernel. This takes place when it is guaranteed that there indeed exists a mapping for such a feature space when a kernel function $K$ is applied [50]. The following four basic Kernels are a common choice:

- Linear kernel: $K(X_i, X_j) = X_i^T X_j$.
- Polynomial kernel: $K(X_i, X_j) = (\gamma X_i^T X_j + b)^d, \gamma > 0$.
- Radial basis function (RBF) kernel: $K(X_i, X_j) = \exp(-\gamma ||X_i - X_j||^2), \gamma > 0$.
- Sigmoid kernel: $K(X_i, X_j) = \tanh(\gamma X_i^T X_j + b)$.

Where $\gamma$ is a kernel parameter, $b$ is the bias or offshoot and $d$ is the polynomial kernel's degree parameter and $X$ is the set of $N$ data points in feature space where index $i, j \in \{1, \ldots, N\}$ [22].

From a general perspective, the RBF kernel is a reasonable choice for many problems. In comparison, the polynomial kernel has more hyper-parameters than the RBF kernel and is more difficult to optimise, and the RBF kernel is not divergent, in other words, it doesn't yield an infinite value if the hyper-parameter value used is too large. As regards the sigmoid kernel, this kernel can yield invalid values depending on the hyper-parameter values used [22]. Therefore, the RBF kernel is selected for this research. This kernel satisfies Mercer's theorem and is therefore a valid kernel.

### 4.2.3 Scaling

Data scaling provides an advantage of avoiding situations where data values of features with a much larger numeric range is dominating those with smaller numeric ranges.

Another advantage of scaling is that it helps alleviate the computational difficulties associated with working with large numerical values. Since kernel values such as values resulting from the linear and polynomial kernels explained in Section 4.2.2, generally focus on the inner products of feature vectors, large data values for such kernels might cause computational stress [22]. This is useful if switching between different kernels is required.

There are different strategies to scale data. A common choice is to centre the data of each feature around its mean, as well as reducing its range to as close to $[-1, +1]$ as possible. Equation 4.15 shows the function $N$ which converts $X_{ij}$ which is the $i$th data point of feature $j$ to its scaled equivalent given by $X'_{ij}$:

$$X'_{ij} = N(X_{ij}) = \frac{X_{ij} - \mu_j}{S_j} \tag{4.15}$$

where $\mu_j$ is the mean of feature $j$ and $S_j$ is the or range of feature $j$. An alternative strategy also uses Equation 4.15 where $S_j$ is the standard deviation of the feature.

### 4.2.4 Grid-Search With Cross-Validation

Section 4.2.2 specified that the RBF kernel is to be used for this research. Thus two hyper-parameters are required to be optimised to ensure optimal classification accuracy; $C$ and $\gamma$. Parameter $C$ adjusts the magnitude of the margin for the decision boundary. Given larger $C$ values, a smaller margin will be yielded if the discriminant function is better at correctly classifying all training points, potentially avoiding under-fitting. A smaller $C$ value will yield a larger margin, therefore a more general discriminant function, at the sacrifice of training accuracy, at the advantage of avoiding over-fitting. Parameter $\gamma$ determines the radius of influence that a single data point—training sample—has on classification. The higher the $\gamma$ value, the smaller the radius of influence and vice versa [48].

Since optimal values for both $C$ and $\gamma$ are not known beforehand, a hyper-parameter search is required to obtain suitable values for the model in order to accurately predict unseen data at a later stage. It is recommended to perform a grid-search using cross-validation to determine optimal values for these parameters. The process involves systemically trying out combinations of $C$ and $\gamma$ and selecting the pair with the best cross-validation accuracy.

A conventional way of setting the range and sequence for $C$ and $\gamma$ values to try out is to exponentially increasing values, usually in powers of 2. A common range and sequence

for $C$ and $\gamma$ is: $C = (2^{-5}, 2^{-3}, ..., 2^{15})$ and $\gamma = (2^{-15}, 2^{-13}, ..., 2^{3})$. Since these ranges are indicative and not exhaustive, a contour plot of the grid-search results proves to be of great help in identifying if there are areas on the edges of or outside of the scope of the given range that could possibly yield a higher accuracy by means of an extended grid-search.

A grid-search, however, requires computational resources, as a range of parameter combinations are evaluated. This issue may be addressed by using parallelised processing techniques coupled with a graphics processing unit (GPU). A GPU-enabled SVM implementation proved to be of immense help in speeding up this process in this implementation [2].

## 4.3 Summary

This chapter discussed the audio recognition techniques used by the system to extract features from audio data, as well as classify them into one of the required audio signature categories.

Section 4.1 provided details on the MFCC audio feature descriptor. The feature descriptor simulates the human auditory perception and provides the unique features of audio data in the form of phonemes. The section described the algorithm which takes place in the following stages: a pre-emphasis stage helps amplify and modify the audio signal; the signal is framed in order to extract short-term phonemes; windowing is applied to reduce spectral leakage; the data is then converted to the frequency domain via the FFT; the frequencies obtained are placed into the Mel scale and placed into a series of Mel-scale filters; the DCT is applied to obtain the CCs; finally, the first and second derivatives of the CCs are obtained; the CCs and derviatives are taken as the final feature vector for training and testing.

Section 4.2 justified the use of SVMs and discussed the theory behind SVM classification. The basic concept behind support vector classification was discussed. The use of kernels for data that is not linearly separable was detailed. A discussion around feature scaling was then undertaken; it was described that feature scaling is carried out to reduce or eliminate the domination of features with large ranges over those with small ranges, and to reduce the overall computational overhead involved with classifier training and classification. Finally, a discussion around the use of the grid-search technique coupled with cross-validation to find optimal parameters the SVM was provided.

Based on this theoretical foundation, the next chapter describes the design and implementation of the system proposed in this research for the recognition of shapes, digits and letters from acoustic emissions made by writing these characters on a surface.

# Chapter 5

# Design and Implementation

This chapter focuses on how the audio recognition system is designed and implemented ahead of experimentation described in the next chapter.

Users produce audio data by writing letters with a stylus on a fixed surface. This data is used to build and train a classifier that can recognise a required set of written shapes.

The audio data is captured first with one microphone, then two and finally with three microphones. Three sets of audio data are collected: a) the audio data of one microphone; b) the audio data of two of the microphones; and c) the audio data of all three microphones. Each of the sets of audio data are processed with the MFCC feature descriptor algorithm which yields feature vectors corresponding to each set. These vectors are used to train a SVM which yields prediction models. The prediction models perform classification with unseen data to test for the accuracy of such a model. Figure 5.1 shows the high-level overview of the processing pipeline.

This chapter consists of the following sections: Section 5.1 details the set up for audio processing, from the microphone to the feature vectors; Section 5.2 details the design of the classification method on the vector data; the chapter is then summarised.

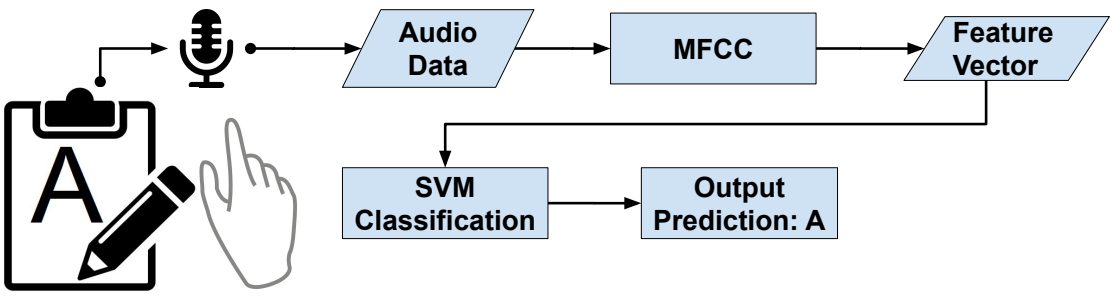


FIGURE 5.1: High-level overview of recognition process.

## 5.1 Feature Extraction

The feature extraction of the audio data within this experimental set up is a capture and compute process. Audio capture involves capturing microphone data created by drawing sounds on the drawing surface. The recording is done with an open platform software which then writes the audio data to file in `.wav` format as attributes of the recording process. The data is pre-processed to standardise it in length, and is then fed to the MFCC feature descriptor described in Section 4.1.

The following subsections give details of audio capture, audio processing and feature extraction.

### 5.1.1 Audio Capture

A piezo microphone contains a small piezoelectric crystal element. It is a more convenient capture tool for the experiment than conventional microphones because they detect sound waves transmitted directly through solid material, rather than through air. Figure 5.2 shows a photograph of a piezo microphone which is used in the experiments in this research, which can be attached to the surface on which letters are drawn.

FIGURE 5.2: A piezo disk microphone.

Combinations of one, two and three microphones are used. The exact placement of the microphones for audio capture differs in each combination used. Figure 5.3 shows the three types of set up: Figures 5.3a-5.3c visually illustrate the set up for one, two and

three microphones, respectively. The cardioid shape shown in each of the figures is a visual representation of each microphone's area of influence, with the writing surface kept as a constraint for the sake of the illustration. The purpose of using more than one microphone is to investigate if an increase in the number of microphones used has a positive influence on the accuracy of classification.

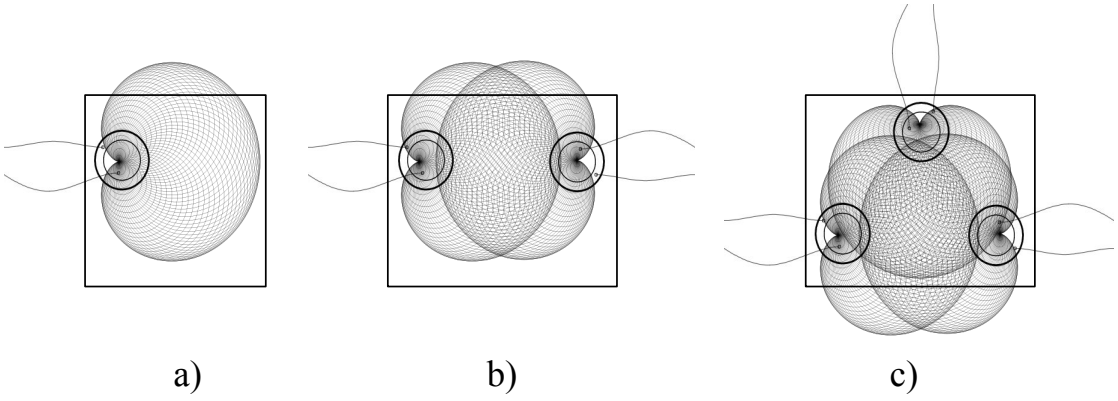


FIGURE 5.3: Microphone configuration with (a) one, (b) two and (c) three microphones.

The platform used for the recording was `Audacity`, an open source and cross-platform audio tool. All audio recordings were `Int16` signed `.wav` files with a sampling rate of 44100Hz. The single-, double- and triple-microphone configuration used, respectively, a mono, dual and triple channel input. At this stage, Research Objective 1 set out in Section 1.3 has been achieved.

### 5.1.2 Audio Processing

Audio data needs to be pre-processed before applying the MFCC to it. This pre-process focuses on standardising the data. The audio files that the user creates are potentially of different lengths i.e. durations. In order to standardise the length of audio files, a uniform length of two seconds is chosen, that is, a total of 88200 data samples. Any audio files that are shorter or longer than this length will, respectively, be stretched or compressed by means of interpolation into the standard length.

After pre-processing, the MFCC feature descriptor described in detail in Section 4.1 is applied to each audio file. For the FFT, the parameter $F$ described in Section 4.1 was set to $F = 2^{12} = 4096$. Setting this value to a large value in relation to the input signal length results a higher and finer resolution spectrogram. This is useful in capturing small frequency differences [46]. Figure 5.4 shows the application of the MFCC to an audio signal: a) is the original isolated audio signal of a letter "A" written on a surface and b) is the signal's feature vector in high-dimension after the application of MFCC. At this stage, Research Objective 2 set out in Section 1.3 has been achieved.

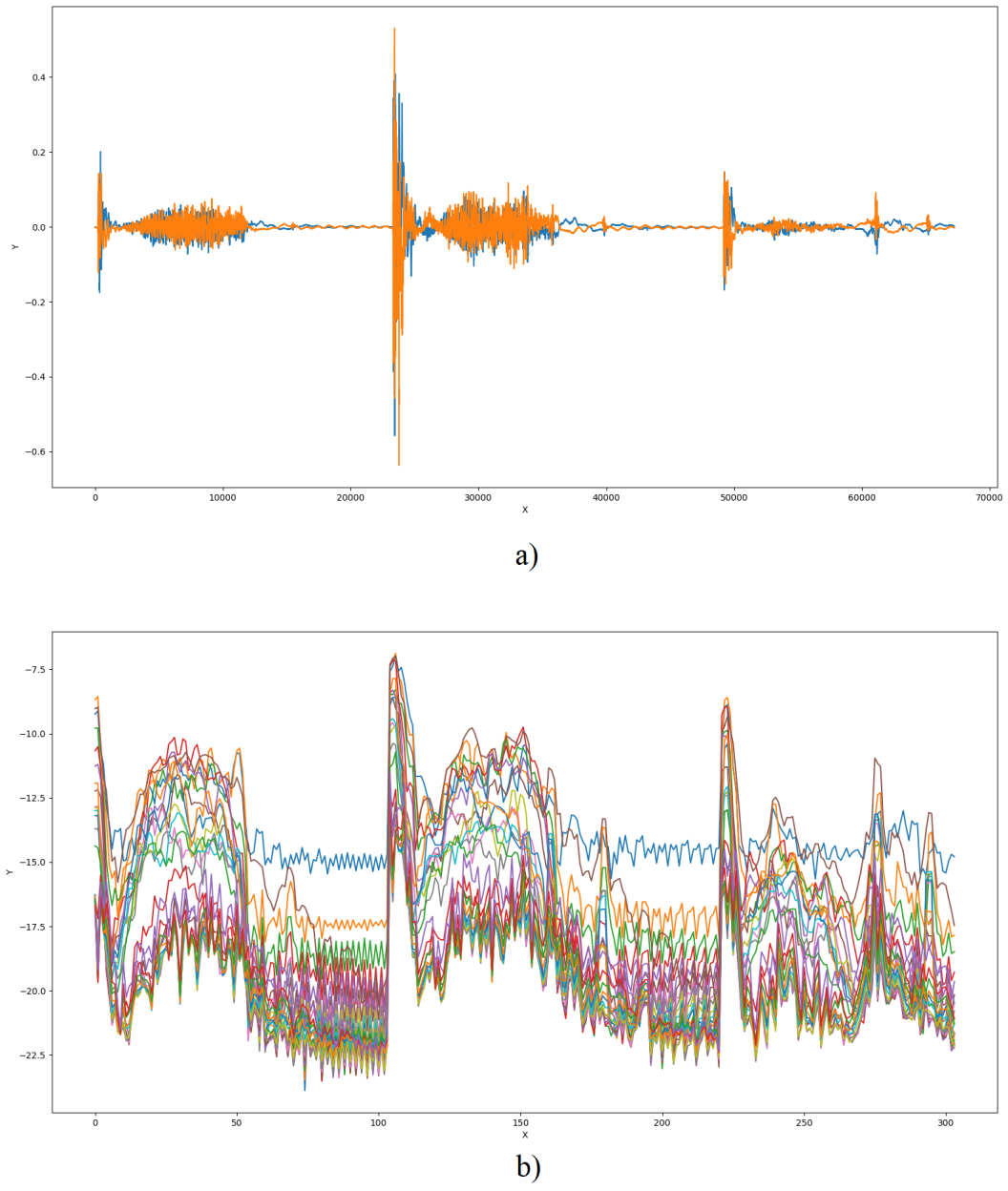


FIGURE 5.4: a) Audio signal of letter "A"; b) visual representation of the resulting feature vector after applying the MFCC feature decsriptor to the audio signal.

## 5.2 Classification

After audio processing has been carried out, all the data is in feature vector format. The next step is to train and test the SVM classifier on the data. The objective here is to build a prediction model based on existing data which involves optimisation to yield the best performing parameters; all of this collectively makes up the training phase.

The trained prediction model is then tested by verifying its accuracy by testing it on a different batch of unseen data. The results can then be used to conclude whether or not

the model is able to generalise and perform well. This leads to the successful completion of Research Objective 3 set out in Section 1.3.

The following subsections detail the following topics: classes and data; training and testing data sets; and optimisation.

### 5.2.1 Classes and Data

The experiment aims to separately and progressively recognise seven fundamental generic shapes, the digits 0–9, and the uppercase Roman alphabetic letters, as explained in Chapter 1. This results in 7 classes for the first phase of experimentation, 10 classes for the second phase of experimentation, and 26 classes for the final phase of experimentation.

Given that the data originates from users, the individuality of users' styles of writing will likely result in a wide diversity in audio signatures for each class [52]. Thus, similar to Li and Hammond [32], users were shown a pre-determined guideline for each character to be drawn, which they had to approximately—but not precisely—follow. The guidelines for the shapes, digits and letters are shown in Figure 5.5. Although users were given some freedom, with the requirement to approximately follow the guidelines, the possibility and viability of providing complete freedom to users can be an interesting area of exploration in future work. At this stage, Research Objectives 4–6 set forth in Section 1.3 have been successfully achieved.

A single sample in any one of the data sets used in this system consists of the feature vector corresponding to the audio recording of that sample, as well as its respective ground-truth label. The ground-truth labels pertaining to each data set are described below:

- For the fundamental shapes data set, the ground-truth labels of the shape classes were set to numeric values ranging from 1–7.
- For the digit shapes data set, the ground-truth labels of each digit class were set to the same digit as the class i.e. 0–9 as relevant.
- For the letter shapes data set. the ground-truth label of samples was based on the `ASCII` code of the relevant alphabet character, starting at `65` for "A" until `90` for "Z". This makes it easier to directly display the actual letter once a prediction is made.

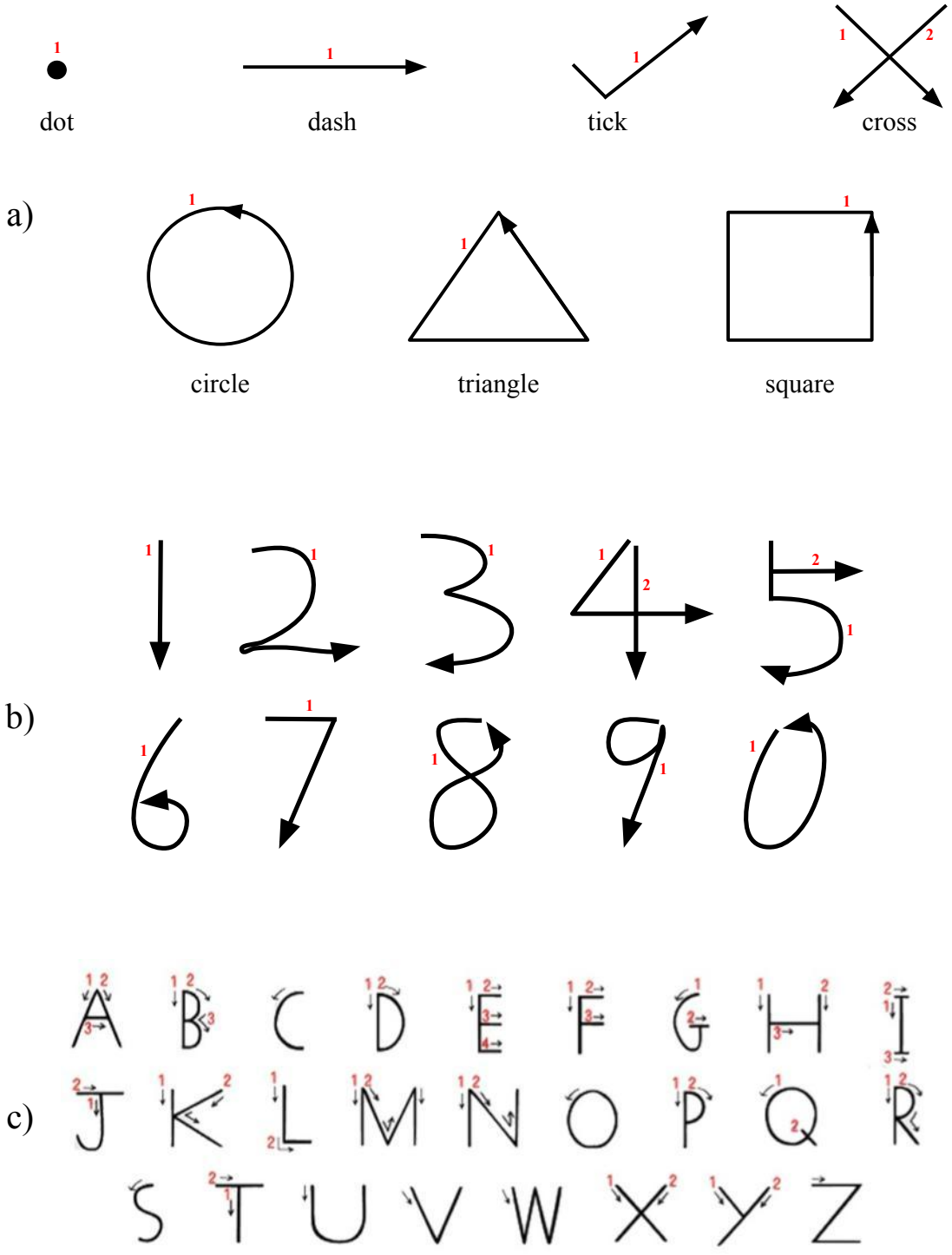


FIGURE 5.5: All fundamental shapes, digits and letters, showing the standardised guideline of writing.

Figure 5.6 is an example illustration of the composition of a feature vector and label pair for the letter "A". Such an instance is used by the SVM in training and testing. Figure 5.7 shows the feature vector of letter "A" in graphical format.

```
[ -7.8844702   -8.3063794   -8.47080319 -10.04409227 -11.48690624
 -11.67157054 -12.89097862 -14.62854074 -12.96423474 -12.736822
 -12.2508538  -11.02031779 -13.08336203 -14.89321009 -15.10718035
 -14.98639846 -14.96239137 -14.852714   -14.73796459 -14.61079301
 -14.4916636  -14.36716893 -14.24049646 -14.11063764 -13.9856151
 -13.85397963  -7.98451955  -8.87024064  -9.78734837 -11.34147792
 -14.18709461 -13.45498867 -13.86133538 -14.12637082 -13.73235798
 -14.05950341 -14.49603125 -13.93715011 -14.12981856 -14.73013489
 -15.1339163  -15.39873767 -15.53836718 -15.72826787 -15.88297332
 -16.03137688 -16.16119414 -16.25924702 -16.34111224 -16.39167495 ...][65]
```

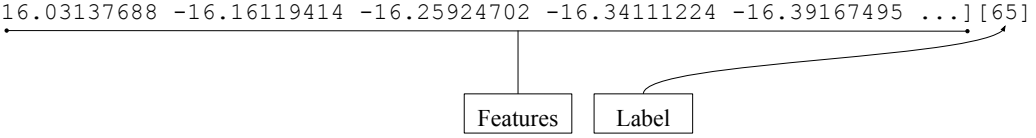


FIGURE 5.6: Illustration of a data sample composition, in this case, of the letter "A".

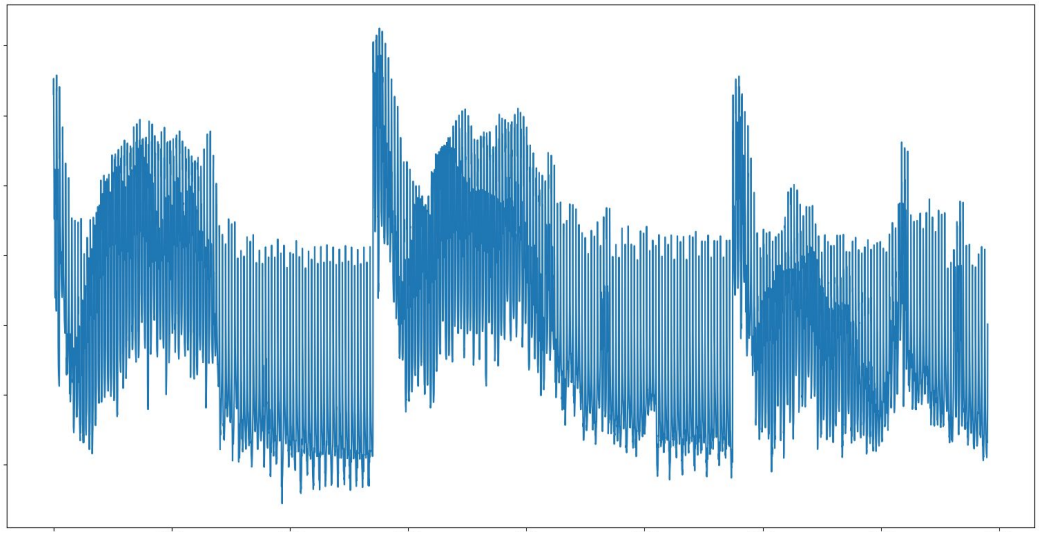

FIGURE 5.7: Feature vector of letter "A".

### 5.2.2 Training and Testing Data Sets

In order to build and verify a prediction model, an SVM model needs to be trained on the data, and later tested with unseen testing data. Thus the entire data set will be partitioned into two parts, one for training and the rest for testing.

The percentage of the partitioning between training and testing varies, and depends on the context of application. A general convention is to use a training set size that is sufficiently large to allow the model to generalise, while ensuring that the quantity of testing data left over is sufficiently large that allows for the model to be thoroughly tested.

The following subsections describe the collection of each of the data sets, namely, the

fundamental shapes (FS) data set, the digit shapes (DS) data set and the letter shapes (LS) data set.

#### 5.2.2.1 Fundamental Shapes Data Set

For this data set, a total of 10 subjects were used. Five subjects were asked to draw each of the seven fundamental shapes a total of six times each, resulting in (5 subjects $\times$ 7 shapes $\times$ 6 samples) = 210 audio recordings.

Of the 6 samples collected per subject per shape, 4 samples were used for training i.e. (5 subjects $\times$ 7 shapes $\times$ 4 samples) = 140 recordings for training. This is henceforth referred to as the "FS training set".

The remaining 2 samples per subject per shape were set aside for testing, resulting in (5 subjects $\times$ 7 shapes $\times$ 2 samples) = 70 audio recordings. Although these specific samples were not used or seen during training at all, other samples of the same subjects and shapes were used in training. Therefore, the classifier may be "familiar" with the writing style of these subjects, and we therefore refer to this set as the "FS semi-seen testing set".

Testing with this kind of data provides a good indication of the system's accuracy provided that a user performs a pre-training procedure prior to use. Doing so may be justified if the pre-training procedure is relatively convenient and once-off, the pay-off—accurate recognition—is high, and if there is no other option to providing high-accuracy recognition.

A completely different set of five subjects were used to construct an "FS unseen testing set"; the data in this set was collected in the same way as with the training and semi-seen data sets. The result was a data set with (5 new subjects $\times$ 7 shapes $\times$ 10 samples) = 350 recordings for unseen testing. Testing with unseen data in this way helps evaluate the ability of the classifier to generalise to completely new subjects without the need for a pre-training procedure of any kind, which is the preferred option.

#### 5.2.2.2 Digit Shapes Data Set

In this case, a total of eight test subjects were used. Five of these subjects were asked to write each of the ten digits a total of six times each. This resulted in a total of (5 subjects $\times$ 10 digits $\times$ 6 samples) = 300 audio recordings. Of the 6 samples per subject per digit, 4 samples were used for training i.e. (5 subjects $\times$ 10 digits $\times$ 4 samples) =

200 audio recordings for training. This data set is henceforth referred to as the "DS training set".

The remaining 2 samples per subject per digit were set aside for semi-seen testing i.e. (5 subjects $\times$ 10 digits $\times$ 2 samples) = 100 audio recordings for semi-seen testing. This data set is henceforth referred to as the "DS semi-seen testing set".

Once again, a completely different set of three subjects were used to construct an "unseen" testing set, with the same justification as explained previously. The data was collected using the same method as with the training and semi-seen data sets. The result was (3 new subjects $\times$ 10 digits $\times$ 6 samples) = 180 audio recordings for unseen testing. This data set is henceforth referred to as the "DS unseen testing set".

#### 5.2.2.3 Letter Shapes Data Set

This data set consists of 10 subjects. Five of the subjects were each asked to write down each of the 26 upper-case letters in the alphabet on the writing surface a total of seven times each i.e. (5 subjects $\times$ 26 letters $\times$ 7 samples) = 910 audio recordings. Of the 7 samples of each letter per subject, 5 samples were used for training i.e. (5 subjects $\times$ 26 letters $\times$ 5 samples) = 650 audio recordings for training. This data set is henceforth referred to as the "LS training set".

The remaining 2 samples per letter per subject were set aside for the "LS semi-seen testing set" i.e. a total of (5 subjects $\times$ 26 letters $\times$ 2 samples) = 260 audio recordings in this data set.

Finally, five new unseen subjects were used to construct a "LS unseen testing set" consisting of (5 subjects $\times$ 26 letters $\times$ 7 samples) = 910 unseen audio recordings.

Figure 5.8 provides an example visual illustration of the way in which the data is partitioned according to training, semi-seen testing, and unseen testing sets. Each box in the figure represents a specific sample number of a specific subject for all the classes recognised in the data set e.g. the top-left box of the figure, as applied to the LS data set, represents collectively the first sample of all 26 letters as drawn by the first subject in the data set. Note that, in general, the figure applies equally to all three data sets, with the exception of the specific number of samples per class per subject which differs between data sets.

Table 5.1 summarises the number of subjects, classes and samples in each partition of each data set, as described in detail above.

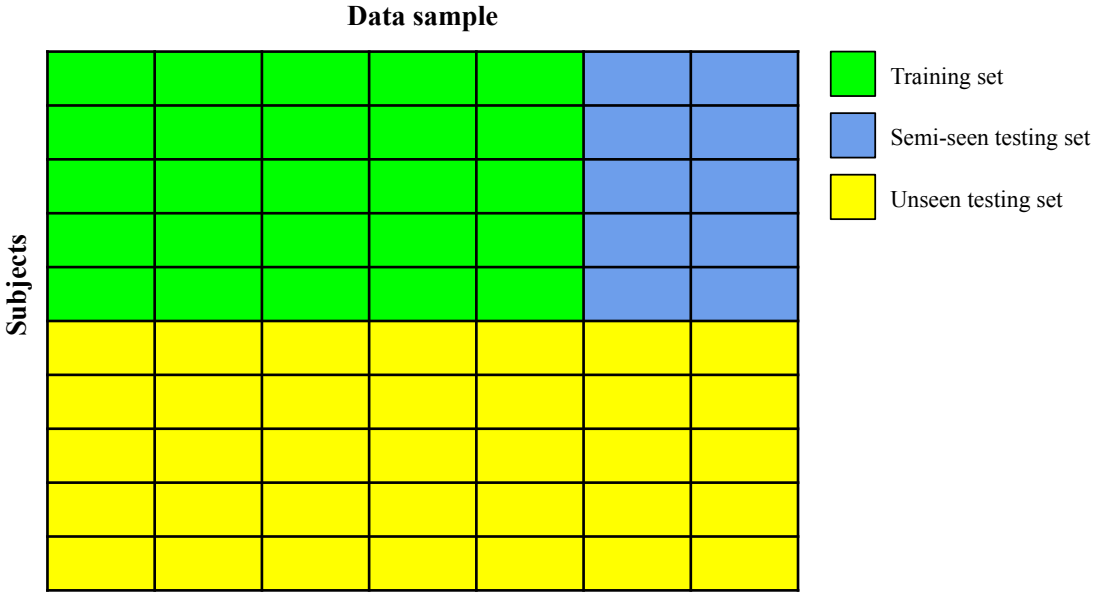


FIGURE 5.8: Example visual illustration of the partitioning of the data sets on training, semi-seen testing and unseen testing sets. Each box in the figure represents a specific sample number of a specific subject for all the classes recognised in the data set. This specific partitioning is applied to the letter shapes data set, however the same approach is used on other data sets as well.

| Data Set | Set | Subjects | Classes | Samples | Total |
|---|---|---|---|---|---|
| Fundamental Shapes | Training | 5 | 7 | 4 | 140 |
| | Semi-Seen Testing | 5 | 7 | 2 | 70 |
| | Unseen Testing | 5 | 7 | 10 | 350 |
| Digit Shapes | Training | 5 | 10 | 4 | 200 |
| | Semi-Seen Testing | 5 | 10 | 2 | 100 |
| | Unseen Testing | 3 | 10 | 6 | 180 |
| Letter Shapes | Training | 5 | 26 | 5 | 650 |
| | Semi-Seen Testing | 5 | 26 | 2 | 260 |
| | Unseen Testing | 5 | 26 | 7 | 910 |

TABLE 5.1: Summary of partitioning of the three data sets for training, semi-seen testing and unseen testing.

### 5.2.3 Optimisation

The optimisation was done by means of 5-fold cross validation via grid-search on a GPU, specifically the NVIDIA® GeForce® GTX 1060 GPU with 6GB VRAM at 1506MHz. This is the application of the technique described in Section 4.2.4. The objective is to obtain the optimal $C$ and $\gamma$ parameter values for the RBF kernel of each SVM trained, each of which leads to a prediction model. The search process is applied exclusively on

the training sets of each respective data set, while cross validating results with combination of $C$ and $\gamma$ values in each case. The best performing $C$ and $\gamma$ parameter pair is selected when the given $C$ and $\gamma$ range is exhausted.

As mentioned in Chapter 1, recognition of the fundamental shapes and digits was a precursor to the eventual goal of recognising letters. Therefore, for the scope of this research and to limit time and computational resources to a feasible amount, it was decided to only apply SVM optimisation to the LS training set, and to rather use default SVM parameters for the FS and DS training data sets, since these can be considered preliminary experiments. Below, a description of the optimisation procedure applied to the LS data set is provided. The default parameters used for both the FS and DS data sets were $(C = 1, \gamma = \frac{1}{n})$, where $n = 7761$ is the number of features per audio recording within the final vector from Equation 4.1.8 in Section 4.1.8.

Three classifiers were optimised and trained, one that made use of only one audio channel data of the LS training set data, another that used two audio channels of the same data set, and the final one that used the data from all three microphone channels. Each classifier was optimised separately, as described below. These classifiers will be referred to as the "one-microphone classifier", "two-microphone classifier" and "three-microphone classifier" for ease of reference below.

When training the one-microphone classifier, to start off, the conventional ranges of $C = (2^{-5}, 2^{-3}, \ldots, 2^{15})$ and $\gamma = (2^{-15}, 2^{-13}, \ldots, 2^{3})$ were used. As discussed in Section 4.2.4, a contour plot of the cross-validation accuracies pertaining to the parameter pairs helps assist in the optimisation of the parameters $C$ and $\gamma$. Figure 5.9a is a contour plot of the cross-validation accuracies, which are colour-coded, for the conventional range of $C$ and $\gamma$ values. By observing the contour plot, it was determined that an adjustment to the parameter search range could help explore better-performing parameter values. Therefore, the ranges were adjusted to $C = (2^{-4}, 2^{-2}, \ldots, 2^{15})$ and $\gamma = (2^{-8}, 2^{-6}, \ldots, 2^{8})$ . Figure 5.9b shows the same contour plot with the adjusted range, with the best performing parameters determined to be $(C = 2^{8}, \gamma = 2^{6})$ with a cross-validation accuracy of 83.38%. The interested reader is referred to Table B.1 in Appendix B which provides a comprehensive log file output of all $C$ and $\gamma$ pairs, along with the cross-validation accuracy attained for each pair.

Based on the experience of the one-microphone classifier optimisation procedure, the adapted ranges of $C = (2^{-2}, 2^{-4}, \ldots, 2^{15})$ and $\gamma = (2^{-8}, 2^{-6}, \ldots, 2^{8})$ were used to optimise the two-microphone and three-microphone classifiers.

For the two-microphone classifier, the best parameters yielded were $(C = 2^{2}, \gamma = 2^{6})$ with a cross-validation accuracy of 86.62%. Figure C.1 in Appendix C shows the contour

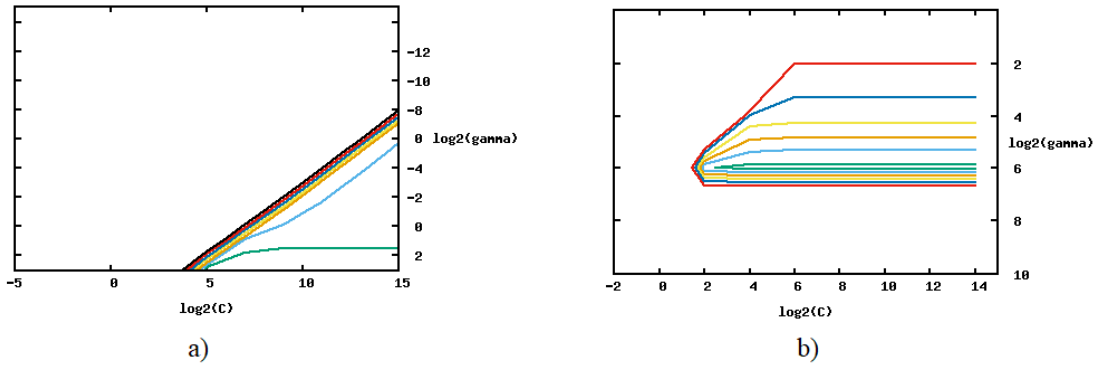


FIGURE 5.9: Contour plot of the grid-search results for the LS data set: a) using the conventional range of $C$ and $\gamma$; b) using the adjusted range of $C$ and $\gamma$.

plot of the grid-search optimisation results for this procedure. Table C.1 in appendix C shows the optimisation output log of the grid-search algorithm for $C$ and $\gamma$ values along with their respective cross-validation accuracies.

For the three-microphone classifier, the optimisation procedure yielded optimal parameters ($C = 2^4, \gamma = 2^6$), with a cross-validation accuracy of 84.92%. Figure D.1 in Appendix D shows the contour plot of the grid-search optimisation results for this procedure. Table D.1 in appendix D shows the comprehensive optimisation output log of the grid-search procedure.

## 5.3 Summary

This chapter detailed how the audio recognition system proposed in this research is designed and implemented ahead of experimentation described in the next chapter.

Section 5.1 discussed the set up used to capture data, the method of capturing data via one, two or three microphones, the pre-processing of the data, and the application of feature extraction via the MFCC feature descriptor.

Section 5.2 discussed classification. The section detailed: the target classification classes that the system aims to recognise; the three data sets that were collected and how they were partitioned into training, semi-seen testing, and unseen testing sets; and the process used to optimise the SVM of the LS data set via grid-search with cross-validation.

At this stage, it can be concluded that Research Objectives 1–6 have been fully achieved, but Research Objectives 7–9 have been *partially* achieved, since the implementation described in this chapter resulted in trained models for the FS, DS and LS data sets, but did not evaluate them.

The next chapter will complete these objectives by evaluating the models on the semi-seen testing sets and unseen testing sets, thereby providing answers to the main research question and sub-questions posed in Chapter 1.

# Chapter 6

# Experimental Results and Analysis

This chapter details the experiments carried out, as well as the results and analysis of each experiment, towards providing answers Research Sub-Questions 1–4, the answers to which will help provide an answer the main research question posed in Chapter 1.

All experiments were performed in a `Python` environment. The hardware specifications were: Intel® Core™i7-7700 at 3.60GHz CPU with 16GB DDR4 at 2133MHz RAM and an NVIDIA® GeForce® GTX 1060 with 6GB VRAM at 1506MHz GPU, although the GPU was only used during optimisation explained in the previous chapter.

The classification success of each of the classifiers in each of the respective experiments detailed below will be analysed by means of the analyses below:

1. Overall accuracy, precision, recall and $f_1$ score.
2. Accuracy per class recognised: This analysis will reveal the extent to which the proposed classification approach is consistent across classes. In the ideal case, classes should be recognised at approximately the same accuracy—tightly centred around the mean accuracy—which would indicate class-independence of the classifier.
3. Accuracy per subject: This analysis helps reveal the extent to which the proposed classification approach is robust to variations in test subjects. As with classes, the ideal case is one in which all subjects have approximately the same recognition accuracy—tightly centred around the mean. On a semi-seen testing set, this would imply subject-independence of the classifier provided that a pre-training procedure is carried out. On an unseen testing set, this would indicate that the classification approach is subject independent without the need for any pre-training at all.

4. Error analysis of fringe cases and outliers: Various analyses may be carried out to provide insight into the causes of classification errors, depending on whether or not outliers are observed. However, as noted in [20], error analyses on classifier predictions are indicative at best, and it is difficult to determine the exact cause of the predictions of the classifier observed.

It should be noted that each experiment will evaluate the classifier success progressively by first evaluating success on the respective semi-seen testing sets, which will be followed by evaluations using the unseen testing sets.

Furthermore, as detailed in Chapter 1, each experiment will start by using audio captured by a single-microphone configuration, and will not resort to the use of double- or triple-microphone configurations if the results of the experiment prove to be sufficiently high.

A "sufficiently high" accuracy may be taken in relation to the number of classes to be recognised as follows. For a $k$-class recognition problem, the probability of success of a naive classifier is $\frac{1}{k}$. Statistically speaking, a classifier with an accuracy greater than the naive classifier accuracy is considered to be effective. The degree of effectiveness of the classifier can therefore be evaluated according to the amount by which the classifer's accuracy exceeds the naive classifier accuracy. For the FS, DS and LS data sets, the naive classifier accuracies are approximately 14%, 10% and 4% respectively. However, for the purposes of this research, a "sufficiently high accuracy" will be taken as significantly higher than the naive classifier accuracy: it will taken as being at least 60% in each case, which approximately represents an accuracy that is 4, 6 and 16 times greater than the naive classifier accuracy, respectively for the FS, DS and LS data sets. For ease of reference, this accuracy will be referred to as the "paradigm accuracy" in the rest of this chapter.

This chapter consists of the following sections: Sections 6.1, 6.2 and 6.3 detail the respective experiments carried out to recognise the fundamental shapes, digits and letters, in each case starting with the data from a single piezo microphone input, in order to answer Research Sub-Questions 1, 2 and 3, respectively, and progressively develop an understanding of the effect of the number of microphone inputs on the recognition accuracy towards finally answering Research Sub-Question 4. In so doing, all remaining Research Objectives 7–9 will be successfully met. Section 6.4 then details a final experiment carried out to further investigate the extent to which the use of additional microphone inputs affect the recognition accuracy of letters on an unseen data set, as compared to the use of data from only one microphone input, towards providing a more comprehensive response to Research Sub-Question 4. Finally, Section 6.5 compares the

| Set | Subjects | Classes | Samples | Total |
|---|---|---|---|---|
| Training | 5 | 7 | 4 | 140 |
| Semi-Seen Testing | 5 | 7 | 2 | 70 |
| Unseen Testing | 5 | 7 | 10 | 350 |

TABLE 6.1: Summary of partitioning of the FS data set for training, semi-seen testing and unseen testing.

proposed system to related systems detailed in Chapter 3 to contextualise the results obtained. The chapter is then summarised and concluded.

## 6.1 Fundamental Shape Recognition Experiment, Results and Analysis Using One Microphone

An experiment to recognise the seven fundamental shapes was carried out first [63] in order to answer Research Sub-Question 1, and provide an indication about Research Sub-Question 4.

For reference purposes in the analysis below, and for the reader's convenience: the fundamental shapes of Figure 5.5 are repeated here in Figure 6.1; and the partitions of the FS data set provided previously in Table 5.1 is repeated here in Table 6.1. Note that only the data from one microphone was used, as an initial trial.

The subsections below describe, respectively, the results and their analysis obtained using the FS semi-seen and the FS unseen testing sets described in the previous chapter in Section 5.2.2.3 using one microphone.

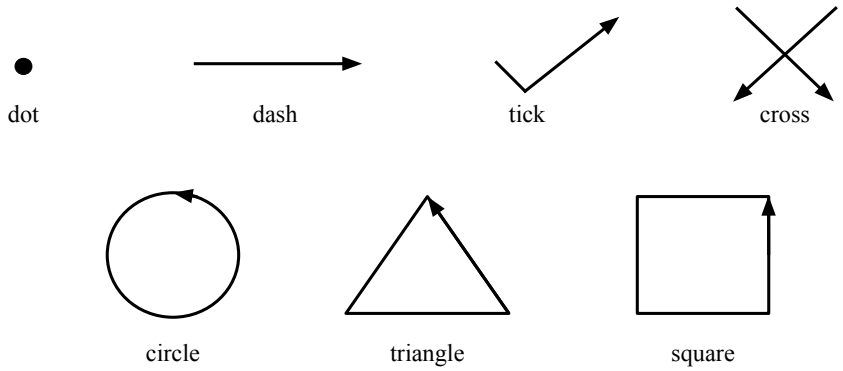


FIGURE 6.1: Fundamental shape classes recognised.

| **Shape** | **Correctly Predicted (10)** | **Accuracy (%)** |
|---|---|---|
| Dot | 10 | 100 |
| Dash | 10 | 100 |
| Tick | 10 | 100 |
| Cross | 10 | 100 |
| Circle | 10 | 100 |
| Triangle | 7 | 70 |
| Square | 9 | 90 |
| **Overall** | — | **94** |

TABLE 6.2: Accuracy results per recognised shape on the semi-seen testing set.

### 6.1.1 Semi-Seen Testing Results and Analysis

Overall, a near-perfect accuracy of 94% was achieved on the semi-seen testing set, with 66 of the 70 semi-seen test samples correctly classified. A precision score of 95% and recall score of 94% were achieved, which implies an $f_1$ score is at 94%. This accuracy exceeds the paradigm accuracy by 34% which is extremely pleasing.

Table 6.2 summarises the accuracy of each shape class. An almost faultless ability to recognise the shapes is demonstrated. Five of the seven shapes achieved a perfect recognition accuracy of 100% i.e. 10 out of 10 samples correctly recognised. The two remaining shapes only had a total of 4 incorrectly classified samples between them. This demonstrates that the system is robust to variations in shapes provided that a pre-training procedure is carried out.

The two shapes *Triangle* and *Square* had a very small number of incorrectly classified samples. Figure 6.2 is a confusion matrix of the semi-seen results in the form of a heat map. The actual table corresponding to the heat map is provided in Table A.1 in Appendix A for the interested reader. Observing the figure reveals that *Triangle* is mistaken with *Tick* consistently (a total of 3 cases) and *Square* is mistaken with *Circle* (a total of 1 case). All things equal i.e. stylus, writing surface, environment etc., misclassified cases can be attributed to the manner in which specific samples of shapes were drawn that may have caused them to sound the same.

For example, if *Triangle* is performed by drawing two edges rapidly, followed by the remaining edge, it may sound like *Tick*. In essence, one or two edges are "phased out" as illustrated in Figure 6.3.

Similarly, if *Square* is performed continuously without pausing at the edges, it may be indistinguishable from *Circle*. To demonstrate this further, Figure 6.4 shows audio

signal plots of *Square* and *Circle*, as well as a plot of *Square* that has been drawn with curved corners. The square in Figure 6.4a is drawn correctly with pauses at the corners; these are visible in the plot as very brief moments of little to no vibration in the audio signal, clearly indicated in the figure by the dotted red lines. On the other hand, the square in Figure 6.4c has been drawn in one continuous movement with curved edges. It is observed that the audio signal of this drawing lacks the pauses observed in Figure 6.4a, and rather appears as one continuous vibration which bears similarity to the audio signal of *Circle* which is shown in Figure 6.4b. This therefore supports the belief and assertion that misclassified cases are attributed to the manner in which specific users drew specific shapes in a very small number of samples.

At this stage, it can be concluded that Research Objective 7 has been partially achieved, pending the experiment using the FS unseen testing set.

Furthermore, pending the experiment using the FS unseen testing set, preliminary and conditional answers to the following research sub-questions can be provided as follows:

- In partial response to Research Sub-Question 1, it can be concluded that, provided that a pre-training procedure is carried out by a user, it is possible to recognise the seven fundamental shapes with an almost perfect accuracy.

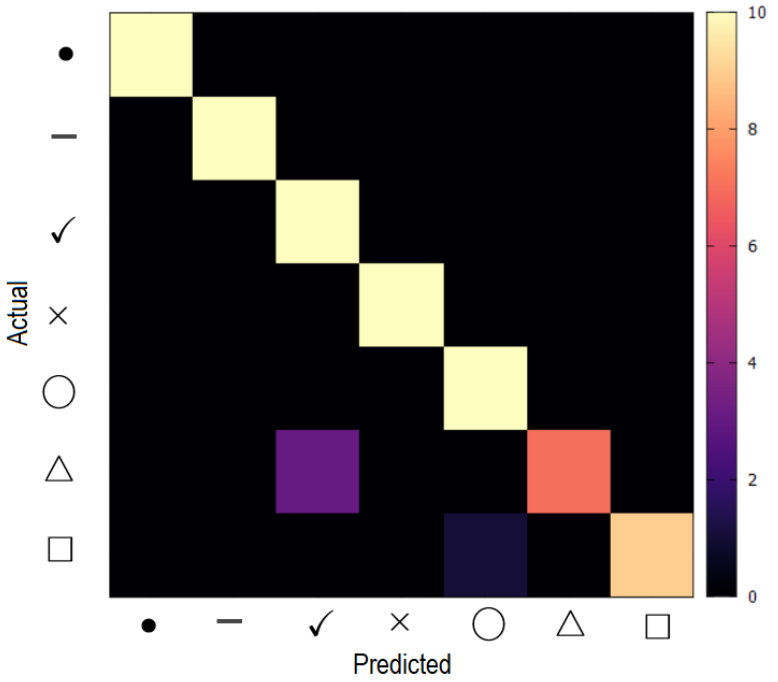


FIGURE 6.2: Confusion matrix in the form of a heat map for the FS semi-seen testing set results.

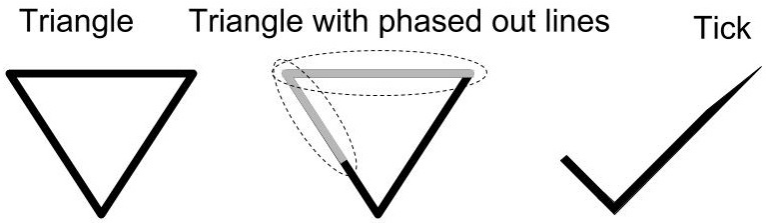


FIGURE 6.3: Similarity between *Triangle* and *Tick*.

- In partial response to Research Sub-Question 4, provided that a pre-training procedure is carried out by a user, it can be stated that high-accuracy recognition of the fundamental shapes can be achieved using only a single microphone.

The next subsection completes the experiment on the FS data set.

### 6.1.2 Unseen Testing Results and Analysis

On the FS unseen testing set, a very high overall accuracy of 83% was achieved, with 292 of the 350 test samples correctly classified. Overall, a precision of 91% and recall of 83% were achieved, resulting in an $f_1$ score of 85%. Despite the introduction to completely new test subjects, the system still maintains a very high level of accuracy. This result is extremely encouraging and exceeds the paradigm accuracy by a substantial amount of just over 23%. This result demonstrates that the approach generalises splendidly.

Table A.2 in Appendix A summarises the accuracy performance by every subject for every shape, as well as aggregates across shapes and subjects. Figure 6.5 graphically illustrates the average accuracy of each shape class across all tests subjects, taken from the right-most column of the table. The figure demonstrates that the accuracies of all seven shapes are at a high level: three of the shapes are at or above 90% accuracy; two shapes are between 80% and 90% accuracy; and two shapes are between 70% and 80% accuracy. None of the accuracies can be considered as outliers. As such, it is clear that the proposed strategy is consistent across shapes, although some shapes are easier to recognise than others.

The differences in accuracy are expected as a normal part of class variations in classification; some classes are easier to recognise than others, depending on a number of factors. As with the experiment with the FS semi-seen testing set, *Triangle* and *Square* are comparatively more difficult to recognise than other classes, although the accuracies of both of these classes are still well above the paradigm accuracy.

Table A.3 in Appendix A is the confusion matrix for the FS unseen data, and the matrix is graphically provided as a heat map in Figure 6.6. As with the semi-seen results, it is

observed that *Triangle* was confused with *Tick*, and *Square* was confused with *Circle* in almost all erroneous cases. Furthermore, in this confusion matrix as well, it is observed that *Circle* appears to be confused with *Square*. The potential cause of these similarities was explained in the previous experiment.

An analysis of the robustness of the proposed approach to variations in test subjects was carried out. Figure 6.7 is a graphical representation of the average accuracies of each test subject across all shape classes. The data in the graph is based on the bottom

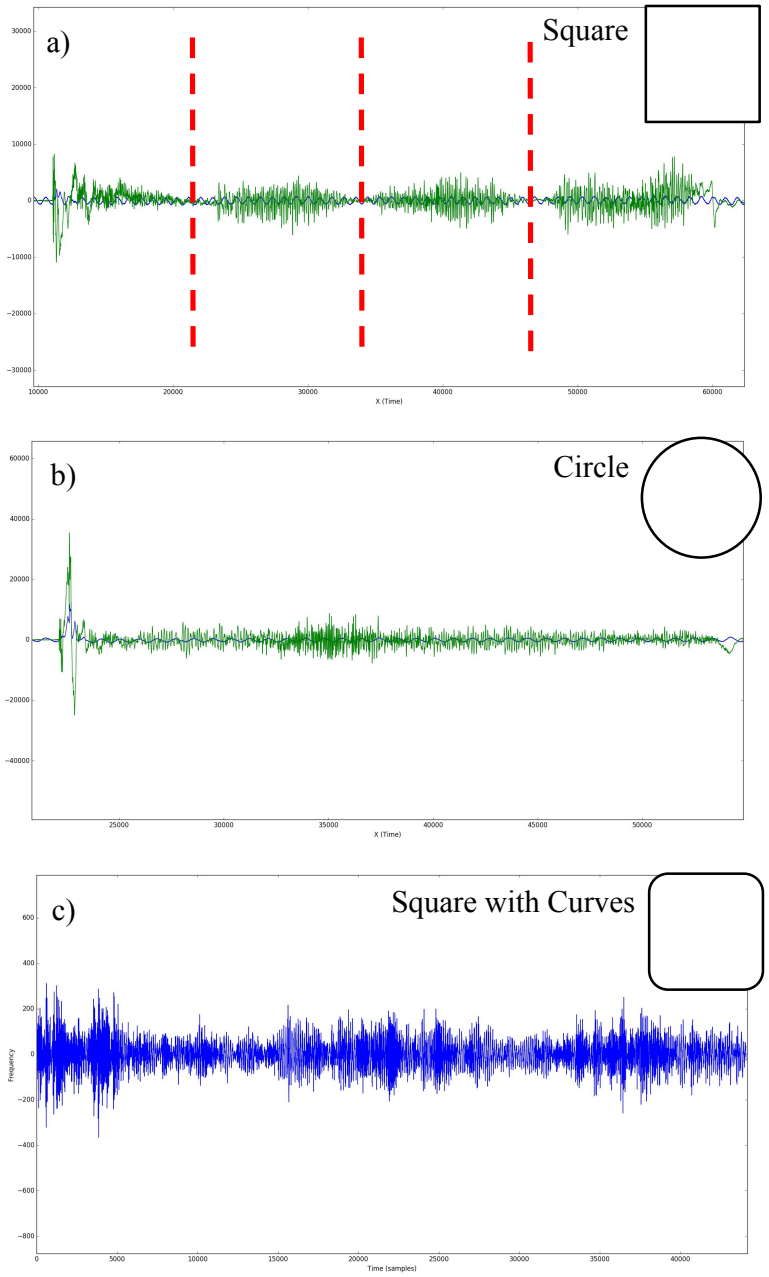


FIGURE 6.4: Audio signal plot of: a) *Square*; b) *Circle*; and c) *Circle* drawn with curved edges.

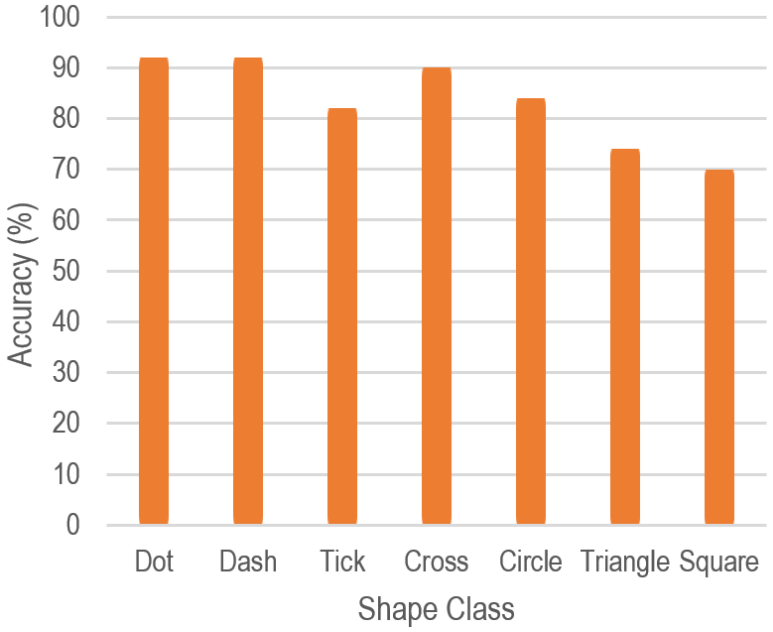


FIGURE 6.5: Accuracy (%) per shape class for the FS unseen testing set.

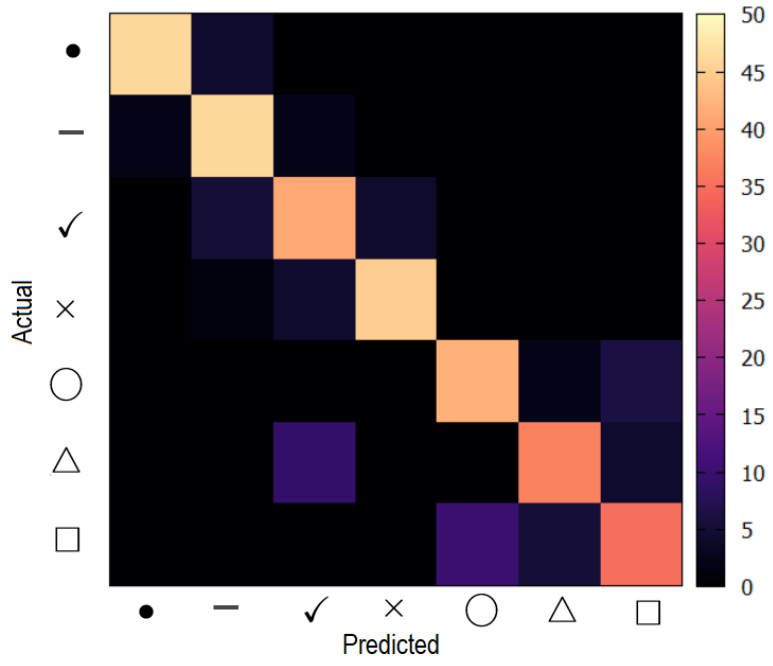


FIGURE 6.6: Confusion matrix in the form of a heat map for the FS unseen testing set results.

row of Table A.2 in Appendix A. The figure demonstrates that the implementation is generally very robust to variations in test subjects, with every subject achieving 76% accuracy or above. This clearly demonstrates that the proposed approach successfully generalises to unseen subjects. The style in which subjects draw shapes varies, and this does appear to affect individual accuracies, however it only influences the accuracy for specific shapes.

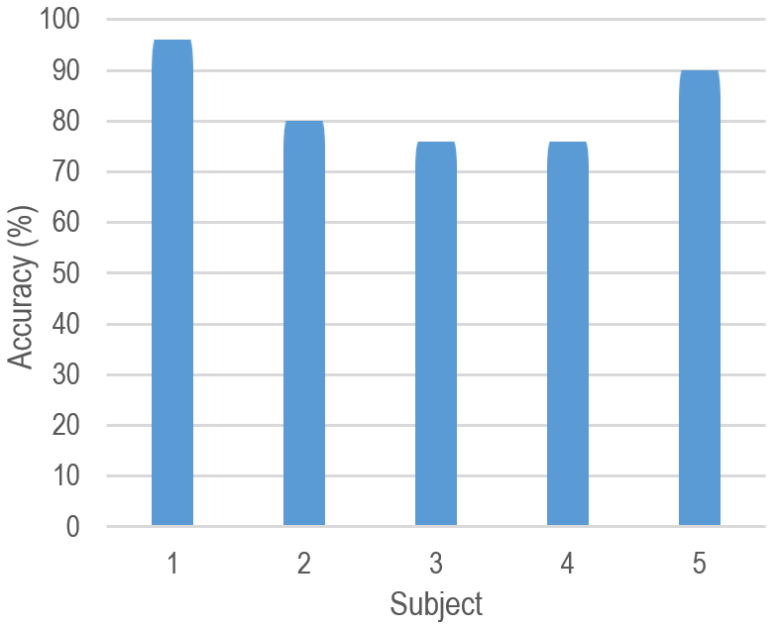


FIGURE 6.7: Accuracy (%) per test subject for the FS unseen testing set.

| **Classification metric** | **FS Testing Set** | |
|---|---|---|
| | **Semi-seen** | **Unseen** |
| Accuracy (%) | 94 | 83 |
| Precision (%) | 95 | 91 |
| Recall (%) | 94 | 83 |
| $f_1$ score (%) | 94 | 85 |

TABLE 6.3: Summary of overall classification metric scores on the FS data set for semi-seen and unseen testing data.

Specifically, the shapes *Triangle* and *Square* as performed by Subjects 3 and 4 had relatively lower accuracies—60%—than other individual accuracies in the results. It should, however, be noted that Subjects 3 and 4 achieved accuracies in excess of 80% and 90% for all other shapes. Therefore, it becomes clear that the contrast in accuracies across subjects can mostly be attributed to the incorrect manner in which *Triangle* and *Square* were performed by these subjects, which caused these shapes to be a little bit more challenging to recognise. Nevertheless, an accuracy of 60% is not considered low by any means, and is at least as good as the paradigm accuracy.

At this stage, it can be concluded that Research Objective 7 has been fully achieved. Given the high-accuracy recognition obtained on both the semi-seen and unseen testing sets, the use of the two- and three-microphone configurations will not be undertaken in this research. It may, however, be an interesting area of investigation for future work.

Answers to the following research sub-questions can be provided as follows:

| Set | Subjects | Classes | Samples | Total |
|---|---|---|---|---|
| Training | 5 | 10 | 4 | 200 |
| Semi-Seen Testing | 5 | 10 | 2 | 100 |
| Unseen Testing | 3 | 10 | 6 | 180 |

TABLE 6.4: Summary of partitioning of the DS data set for training, semi-seen testing and unseen testing.

- As a final response to Research Sub-Question 1, it can be concluded that it is possible to recognise the seven fundamental shapes with an almost perfect accuracy, without the need for any pre-training procedure.
- In partial response to Research Sub-Question 4, pending the experiments on the DS and LS data sets, it can be stated that high-accuracy recognition of the fundamental shapes can be achieved using only a single microphone.

Table 6.3 summarises the overall results of the experiments on the FS semi-seen and unseen testing sets.

## 6.2 Digit Recognition Experiment, Results and Analysis Using One Microphone

An experiment to recognise the digits was carried out [64] in order to answer Research Sub-Question 2, and provide an indication about Research Sub-Question 4. For reference purposes in the analysis below, and for the reader's convenience: the digits of Figure 5.5 are repeated here in Figure 6.8; and the partitions of the DS data set provided previously in Table 5.1 are repeated here in Table 6.4. Note that only the data from one microphone was used, as an initial trial.

The subsections below describe, respectively, the results and their analysis obtained using the DS semi-seen and unseen testing sets using one microphone.

### 6.2.1 Semi-Seen Testing Results and Analysis

For the semi-seen experiment, a very high overall accuracy of 91% was achieved, with 91 of the 100 semi-seen samples correctly classified. The precision, recall and $f_1$ score were 92%, 90% and 90%, respectively. Despite the increase in the number of classes recognised, the accuracy achieved is still substantially above the paradigm accuracy, by 31%. This result is very encouraging.

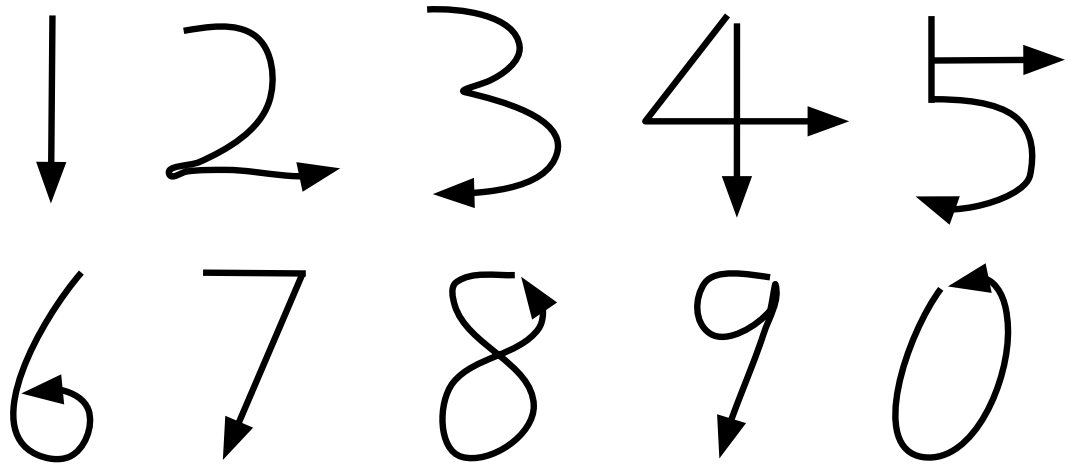

FIGURE 6.8: Digit shapes classes recognised.

| **Shape** | **Correctly Predicted (10)** | **Accuracy (%)** |
|---|---|---|
| 0 | 10 | 100 |
| 1 | 10 | 100 |
| 2 | 8 | 80 |
| 3 | 10 | 100 |
| 4 | 10 | 100 |
| 5 | 9 | 90 |
| 6 | 6 | 60 |
| 7 | 9 | 90 |
| 8 | 10 | 100 |
| 9 | 9 | 90 |
| **Overall** | — | **91** |

TABLE 6.5: Accuracy results per recognised shape on the semi-seen testing set.

Table 6.5 summarises the average accuracy of each digit class recognised across all subjects. Observing the table, it is seen that five (half) of the digits achieve a perfect accuracy of 10 out of 10 samples correctly recognised, with four more digits achieving 8 or 9 correct predictions out of 10. Only digits *2* and *6* have more than one incorrect prediction.

A confusion matrix in the form of a heat map is provided in Figure 6.9, which is based on the confusion matrix in Table A.4 in Appendix A. Observation of the matrix shows that digit *6* is consistently confused with digits *2* and *3*. Purely from an audio perspective i.e. in terms of the number of strokes and pauses in the audio signal when these digits are drawn, they can sound the same. All three digits are drawn in two distinct strokes, with one pause. Although they are distinct, as evidenced by a large number of correct predictions on each of these digits, it is possible to draw them in a manner that makes

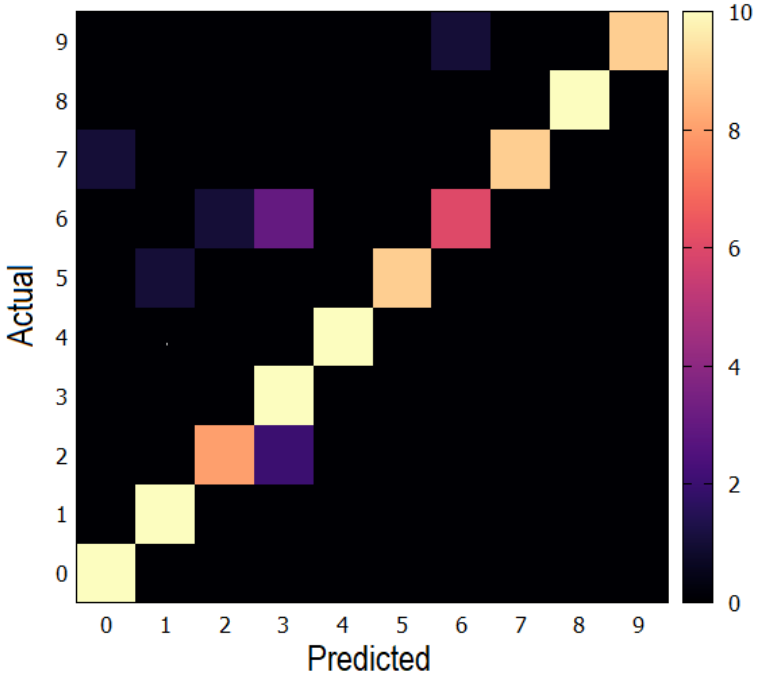


FIGURE 6.9: Confusion matrix in the form of a heat map for the DS semi-seen testing set results.

them sound co-similar. This can come about when, for example, the speed of writing is such that the length of pauses is either too short or too long.

As a demonstration of this, Figure 6.10 shows an audio signal plot of actual samples of digits *2* and *3* that were confused with each other. The plots have similar features in terms of the number of pauses and the number of strokes, as indicated by the dotted red lines in the figure. This supports the belief that the manner of writing has caused these samples to have similar audio signatures, resulting in an incorrect prediction.

Therefore, these errors are likely caused by the similarity in the digits from an audio perspective. It is promising to note that the incorrect predicted samples are insignificant in proportion to the number of correctly predicted samples. Noting that these specific samples were not seen during training means that it is possible to accomplish a high-accuracy performance provided that the user of the system performs a once-off initial pre-training procedure.

At this stage, it can be concluded that Research Objective 8 has been partially achieved, pending the experiment using the DS unseen testing set, which is explained in the next subsection.

In addition, prior to the experiment using the DS unseen testing set, the following

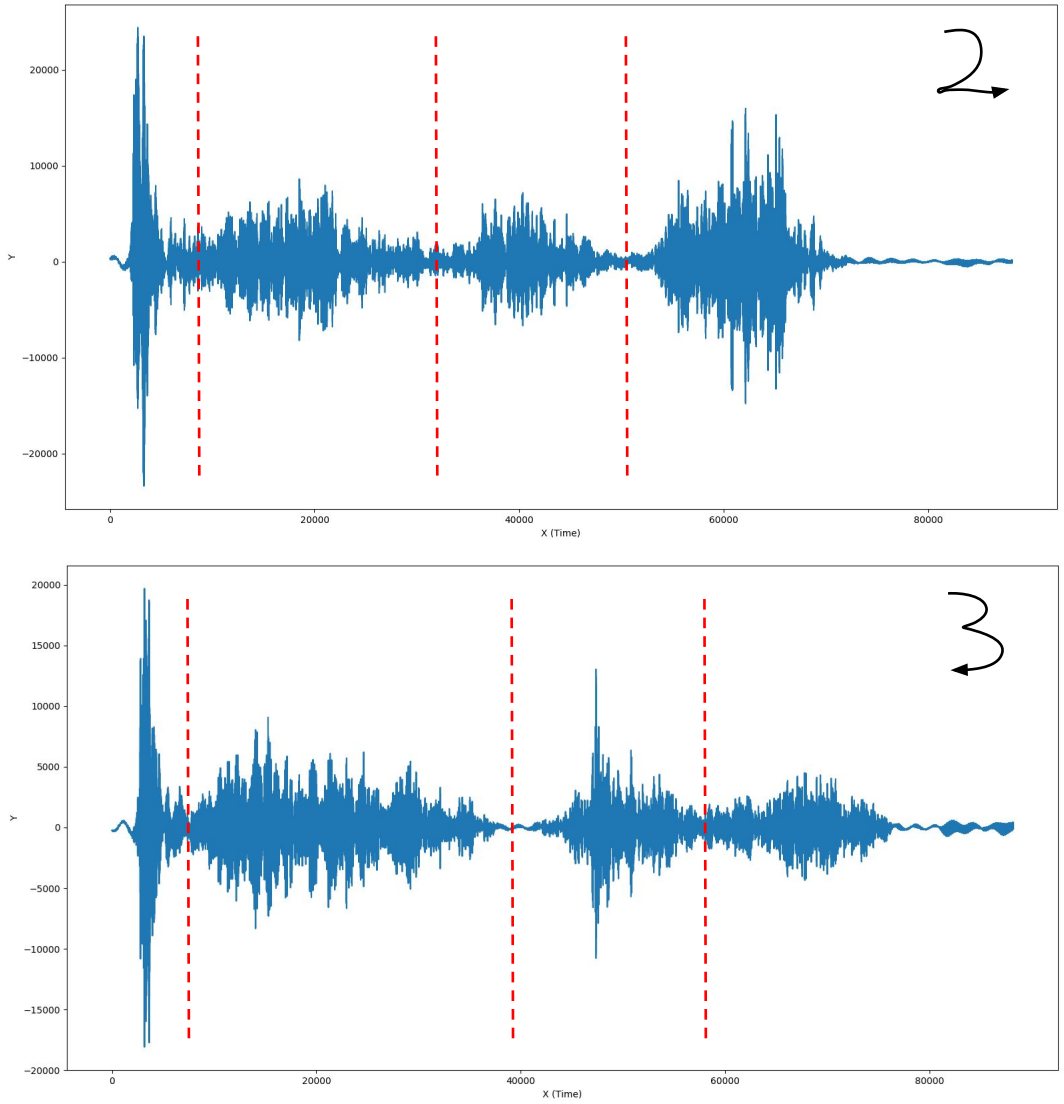


FIGURE 6.10: Audio signal plot of digits *2* and *3* showing similarity in the pauses and strokes.

preliminary and conditional answers to the following research sub-questions can be provided:

- As a partial response to Research Sub-Question 2, it can be concluded that, on condition that a pre-training procedure is carried out by a user prior to using the system, it is possible to recognise the digits with a very high accuracy.
- As a preliminary response to Research Sub-Question 4, it can be stated that, on condition that a pre-training procedure is carried out by a user prior to using the system, high-accuracy recognition of the the digit shapes can be achieved using only a single microphone.

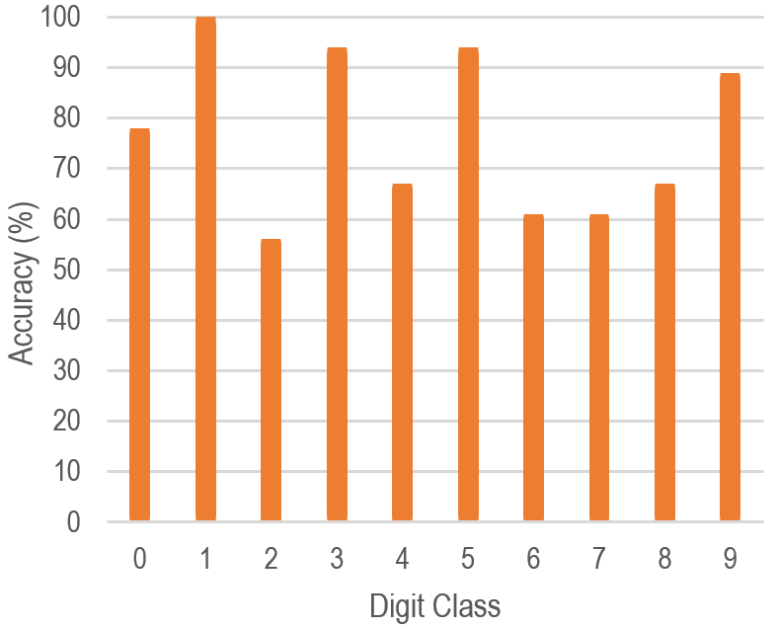


FIGURE 6.11: Accuracy (%) per digit class for the DS unseen testing set.

The experimentation on the DS data set is completed in the next subsection.

### 6.2.2 Unseen Testing Results and Analysis

The DS unseen testing set experiment's overall accuracy is 77%, with 138 of the 180 unseen samples correctly classified. This accuracy is 17% above the paradigm accuracy and is a very promising outcome. Despite the increase in the number of classes and the use of completely new test subjects, the approach still generalises effectively. The precision, recall and $f_1$ score were 78%, 76% and 76% respectively.

| **Digit** | **Accuracy (%)** | | | **Overall (%)** |
|---|---|---|---|---|
| | **Subj. 1** | **Subj. 2** | **Subj. 3** | |
| 0 | 100 | 100 | 33 | **78** |
| 1 | 100 | 100 | 100 | **100** |
| 2 | 50 | 67 | 50 | **56** |
| 3 | 100 | 100 | 83 | **94** |
| 4 | 83 | 50 | 67 | **67** |
| 5 | 100 | 83 | 100 | **94** |
| 6 | 50 | 83 | 50 | **61** |
| 7 | 50 | 83 | 50 | **61** |
| 8 | 33 | 67 | 100 | **67** |
| 9 | 100 | 100 | 67 | **89** |
| **Overall (%)** | **77** | **83** | **70** | **77** |

TABLE 6.6: Accuracy and subject results of the digits recognition for unseen data.

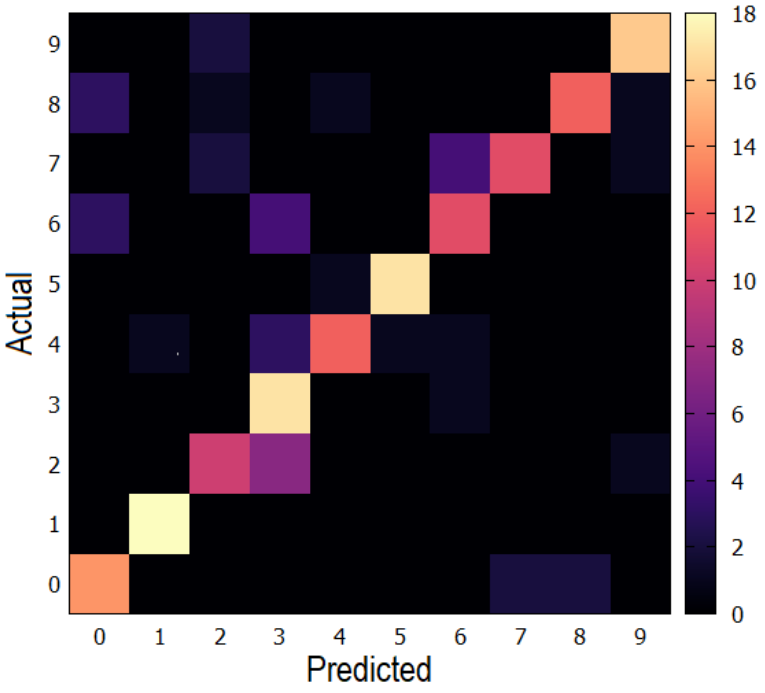


FIGURE 6.12: Confusion matrix in the form of a heat map for the DS unseen testing set results.

Figure 6.11 graphically illustrates the average accuracy of each test subject across all digits. The data in the graph is based on the right-most column of the comprehensive table of results in Table 6.6 which summarises the accuracy of every subject for every digit, and the aggregates across shapes and subjects. Referring to Figure 6.11, it is noticed that specific digits are easier to recognise than others. Digits *0*, *1*, *3*, *5* and *9* achieved very high or even perfect accuracies ranging between 77% and 100%, while other digits achieved accuracies that can be considered as high, ranging from 61% to 66%, noting that these are all above the paradigm accuracy. With the exception of digit *2*, no digit's accuracy reading falls below the paradigm accuracy. It is important to note that digit *2* achieves an accuracy of 56% which is only 4% below the paradigm accuracy, and is nevertheless still 5.6 times larger than the accuracy of a naive classifier for a 10-class problem. It therefore is not by any means considered to be a low accuracy or a poor result.

It is also reassuring to note, when observing the individual accuracies per subject-digit combination in the table, that 70% of these accuracies—21 of the 30 accuracies in the table—have two-thirds or more of the predictions correct. Only 2 of the 30 accuracies fall below 50% recognition. This is very encouraging considering the setup. It should also be noted that the classifier in this case was not optimised but used default parameters, for the reason explained in Section 5.2.3. Optimisation will most likely provide a significant boost in accuracy, which can be investigated in future.

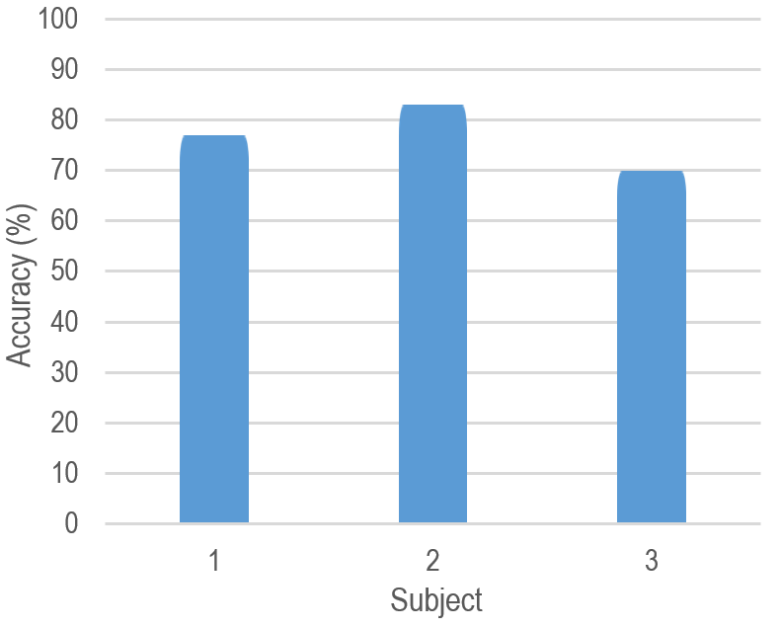


FIGURE 6.13: Accuracy (%) per test subject for the DS unseen testing set.

To analyse the predictions further, a confusion matrix in the form of a heat map is provided in Figure 6.12, which is based on the tabulated confusion matrix in Table A.5 in Appendix A. In general, it is observed that the majority of predictions for every digit lie on the diagonal i.e. they are correct predictions. Trends that were observed with the DS semi-seen testing are also observed in this confusion matrix i.e. the inter-confusion between a few cases of *2*, *3* and *6*. This further supports the previous assertion and demonstration that these cases, and likely those of others, are attributed to the manner in which these shapes are drawn by specific test subjects.

To analyse the robustness of the approach to new and unseen test subjects, Figure 6.13 provides a bar graph of the average accuracy of each test subject across all shapes, based on the bottom row of Table 6.6. Observing the bar graph, it is seen that the accuracies of all three subjects are well above the paradigm accuracy by between 10–23%. Furthermore, all three accuracies are relatively consistent, all being close to the overall average of 77%, with no outliers observed. This is a very pleasing result and demonstrates that, even without any pre-training procedure, the proposed approach is robust to variations in subjects, and generalises very well in this regard.

It is of value to repeat here that the classifier used in this experiment made use of default parameters. It is very likely that even better results can be obtained if the classifier is optimised. This can be investigated in future.

At this stage, it can be concluded that Research Objective 8 has been fully achieved. Given the high-accuracy recognition obtained on both the semi-seen and unseen testing

| **Classification metric** | **DS Testing Set** | |
|---|---|---|
| | **Semi-seen** | **Unseen** |
| Accuracy (%) | 91 | 77 |
| Precision (%) | 92 | 78 |
| Recall (%) | 90 | 76 |
| $f_1$ score (%) | 90 | 76 |

TABLE 6.7: Summary of overall classification metric scores on the DS data set for semi-seen and unseen testing data.

sets, the use of the two- and three-microphone configurations will not be undertaken in this research. It may, however, be an interesting area of investigation for future work.

The following answers to the research sub-questions can be provided as follows:

- As a final response to Research Sub-Question 2, it is concluded that it is possible to recognise the digits with a high accuracy, without the need for any pre-training procedure.
- As a partial response to Research Sub-Question 4, pending the experiment on the LS data set, it is stated that high-accuracy recognition of the digits can be achieved using only a single microphone.

Table 6.7 summarises the overall results of the experiments on the FS semi-seen and unseen testing sets.

## 6.3 Letter Recognition Experiment, Results and Analysis Using One Microphone

This section discusses the experiment carried out to recognise letters with a single piezo microphone in order to answer Research Sub-Question 3, and potentially provide a final answer on Research Sub-Question 4, depending on the results. This experiment is the most complex classification task out of the three considered, with more than double the number of classes of the experiment with digits, and almost triple those of the experiment with fundamental shapes. It is the culmination of the work in this research, and if high-accuracy results are obtained, all research questions can be successfully answered.

For ease of reference in the analysis below, and for the reader's convenience: the letters of Figure 5.5 are repeated here in Figure 6.14; and the partitions of the LS data set provided previously in Table 5.1 are repeated here in Table 6.8.

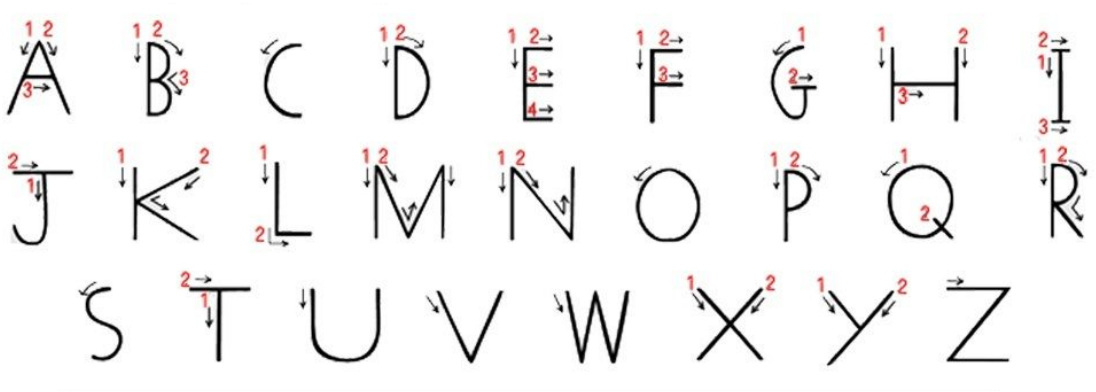

FIGURE 6.14: Letter shapes classes recognised.

The subsections below describe, respectively, the results and their analysis obtained using the DS semi-seen and unseen testing sets using the data of one microphone.

### 6.3.1 Semi-Seen Testing Results and Analysis

Using only a single microphone, an excellent overall accuracy of 86.2% was obtained, with 224 of the 260 semi-seen samples correctly classified. Despite the difficulty of the classification problem, the accuracy obtained exceeded the paradigm accuracy by a very substantial amount of just over 26%. The precision, recall and $f_1$ score were 87.4%, 86.2% and 86.1% respectively.

Figure 6.15 visually presents the average accuracy per letter on this semi-seen testing set. The figure is based on Table B.2 in Appendix B. Despite the difficulty of the classification problem, the graph demonstrates that every letter is recognised with at least 70% accuracy, which is 10% higher than the paradigm accuracy, but as high as 100% for some letters. No outliers are observed, but it does appear that the increase in classes has made it increasingly challenging for the system to recognise some letters, while some letters are almost perfectly recognised. Nevertheless, the system clearly performs well.

| Set | Subjects | Classes | Samples | Total |
|---|---|---|---|---|
| Training | 5 | 26 | 5 | 650 |
| Semi-Seen Testing | 5 | 26 | 2 | 260 |
| Unseen Testing | 5 | 26 | 7 | 910 |

TABLE 6.8: Summary of partitioning of the LS data set for training, semi-seen testing and unseen testing.


FIGURE 6.15: Accuracy results per letter class on the semi-seen testing set using one microphone.

To analyse clashes between letter classes, Figure 6.16 provides a confusion matrix in the form of a heat map for these results, which is based on Table B.4 in Appendix B. Scanning through the heat map row-by-row, it can be observed that letter classes with incorrectly classified samples are mostly confused with a small number of other letters. This points to the fact that, as with the FS and DS data sets, incorrectly classified samples are most likely attributed to the audio similarity, i.e. the number of strokes and pauses, of letters with a small number of other letters, but that the classifier is generally performing well. This will be analysed further with the LS unseen testing set in the next sub-section. At this stage, there are no significantly bright spots or outliers in the heat map that warrant any further analysis.

At this point, it can be concluded that Research Objective 9 has been partially achieved, prior to performing the experiment using the LS unseen testing set, which is explained in the next sub-section.

Furthermore, the following preliminary and conditional answers to the following research sub-questions can be provided, pending the experiment with the LS unseen testing set:

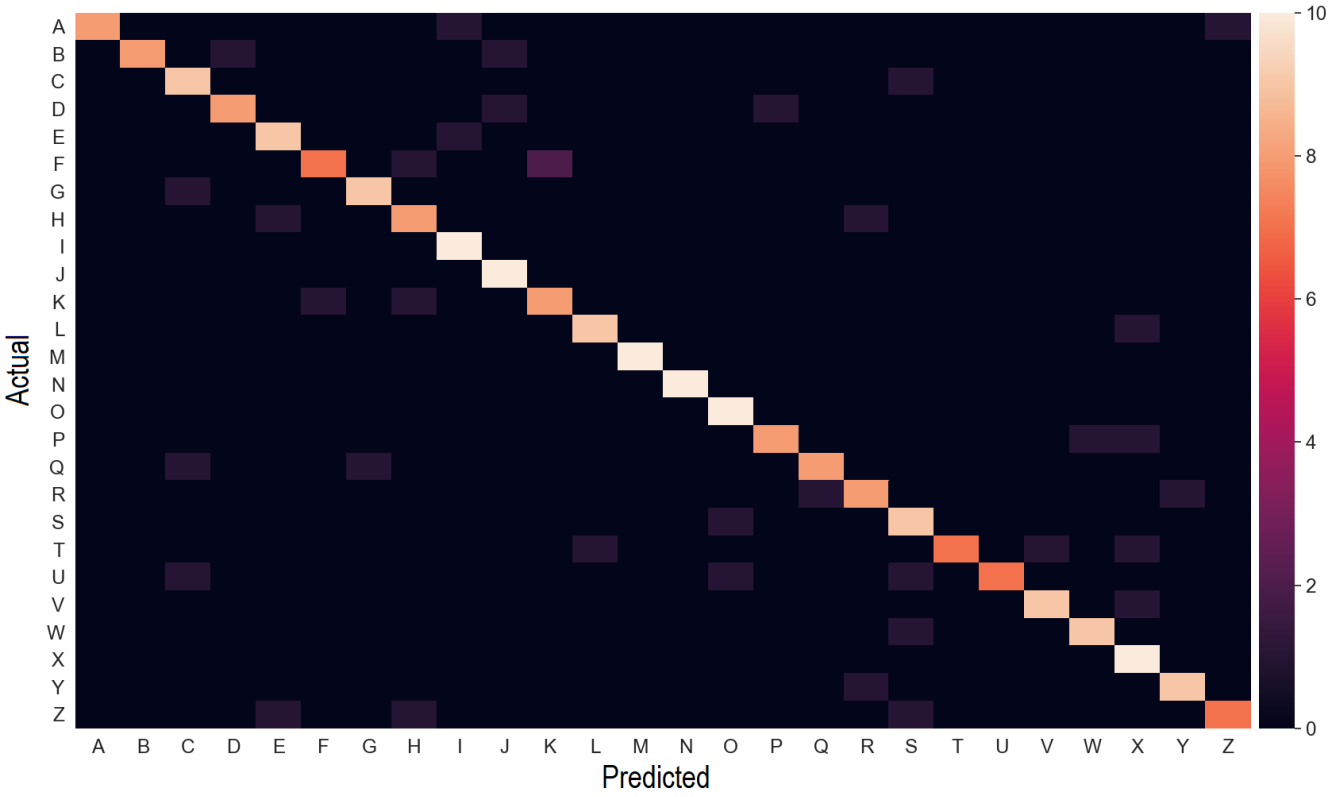


FIGURE 6.16: Confusion matrix in the form of a heat map for the LS semi-seen testing set results using one microphone.

- As a partial response to Research Sub-Question 3, a partial conclusion can be reached as follows: provided that a pre-training procedure is carried out by a user prior to using the system, it is possible to recognise the letters with a high accuracy.
- As a preliminary response to Research Sub-Question 4, it can be stated that, provided that a pre-training procedure is carried out by a user prior to using the system, high-accuracy recognition of the letter shapes can be achieved using only a single microphone.

The next sub-section details the experiment using the LS unseen testing set. If the results of the next sub-section maintain a high accuracy, all research questions can be successfully answered and objectives, met.

### 6.3.2 Unseen Testing Results and Analysis

Using only a single microphone, and using data from completely unseen test subjects, it was extremely pleasing to note that an excellent overall accuracy of 79.0% was obtained, with 719 of the 910 unseen samples correctly classified. This represents an accuracy that

is 19.0% better than the paradigm accuracy, despite the large number of classes recognised, and despite the fact that the test subjects were completely new. It is encouraging to note that changing from the semi-seen to the unseen testing set only resulted in a reduction of 7.0% in overall accuracy. This will be analysed further, but demonstrates that the proposed system can most likely function at a high accuracy without the need for any pre-training procedure. The precision, recall and $f_1$ score were 80.5%, 79.0% and 78.9% respectively.

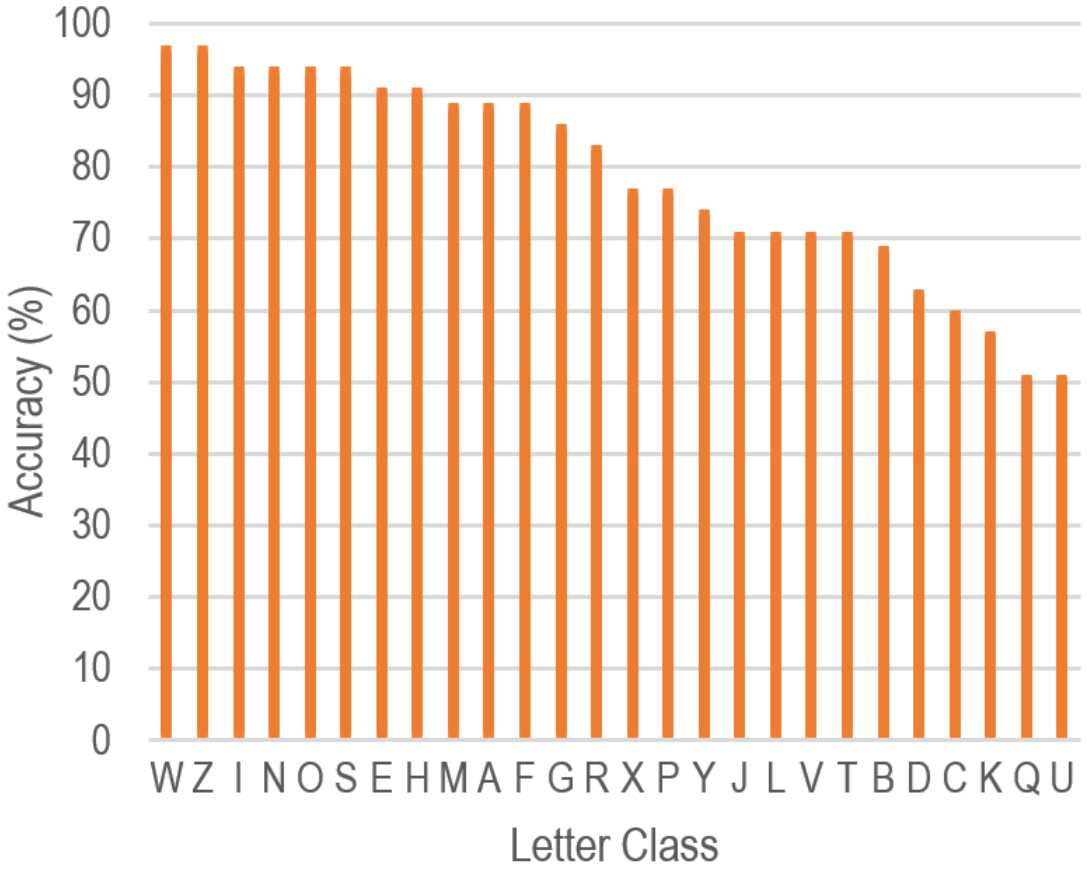


FIGURE 6.17: Accuracy results per letter class on the unseen testing set using one microphone, sorted in descending order of accuracy.

Figure 6.17 visually illustrates the average accuracy per letter on the unseen testing set, with the letters sorted in descending order of accuracy. The data in the graph is based on Table B.5 in Appendix B. The graph has been sorted in order of accuracy in order to be able to more easily compare the accuracy of letters to the paradigm accuracy. Referring to the graph, it is observed that 12 letters—almost half of the letters—have accuracies above 85%, with a further 7 letters achieving 70% accuracy or above. Collectively, it can be said that 19 of the 26 letters—almost three quarters of the letters—are above the 70% accuracy line.

Only 3 of the 26 letters fall below the paradigm accuracy, but even then only by a very small to moderate amount of 3% for $K$ and 9% for $Q$ and $U$. Its important to mention that the accuracies of these letters—57% for $K$ and 51% for $Q$ and $U$—still represent accuracies that are still many times higher than the accuracy of a naive classifier on a complex 26-class problem, which is approximately 4%. Therefore, the proposed system can still be considered to be exceptionally effective under the circumstances. All of this serves to demonstrate that with only one microphone and even with completely unseen test subjects, the proposed system is capable of high-accuracy recognition of letters, which is a remarkable result.

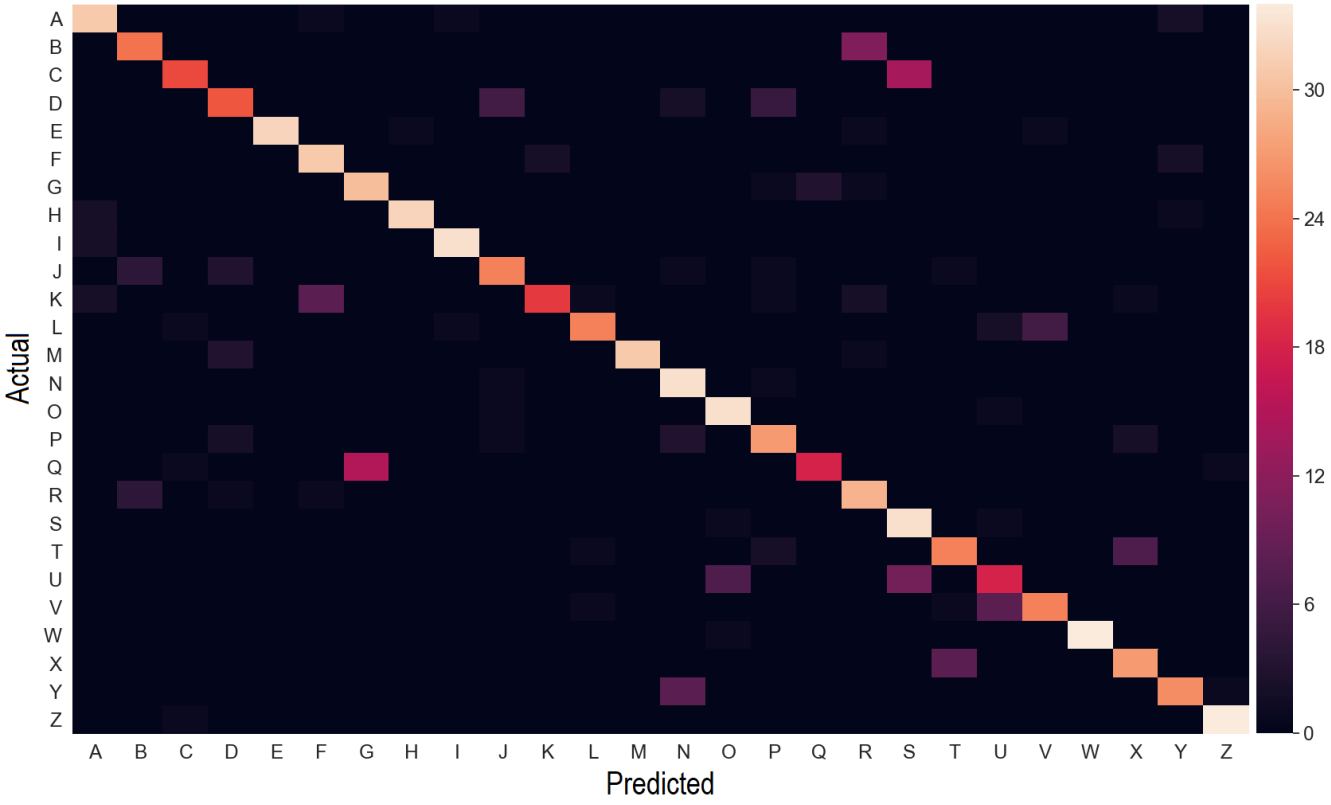


FIGURE 6.18: Confusion matrix in the form of a heat map for the LS unseen testing set results using one microphone.

To analyse the possible causes of incorrect predictions, mainly for letters $K$, $Q$ and $U$, Figure 6.18 provides a confusion matrix in the form of a heat map for these results, and Table B.7 in Appendix B provides the original confusion matrix. It can be observed from the heat map that:

- $K$ is almost consistently confused with $F$. Both of these letters are drawn with three strokes, with the first stroke longer than the two last strokes, and two pauses between the strokes.

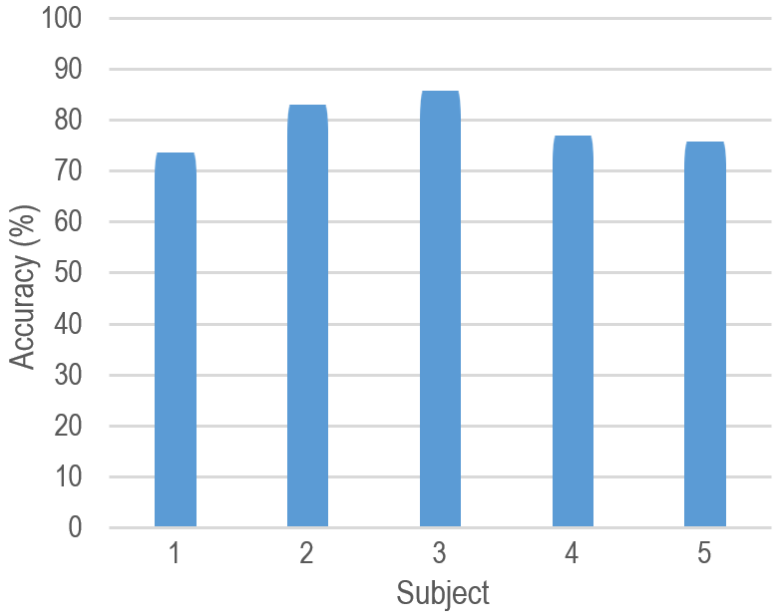


FIGURE 6.19: Accuracy (%) per test subject for the LS unseen testing set using one microphone.

- $Q$ is almost exclusively confused with $G$. These letters are both round and are completed with a single stroke approximately in the middle of the letter.
- $U$ is consistently confused with $O$ and $S$. These shapes are all curved and drawn in one continuous stroke, without any pauses.

The analysis above, along with the analyses of the FS and DS testing sets, strongly supports the idea that the proposed system has learnt to very accurately model the phonemes in the various shape classes, as well as associate given sets of phonemes with specific shape classes. As has been demonstrated repeatedly in this chapter, incorrectly classified samples appear to be directly tied to a similarity in the audio forms of classes, rather than to confusion with random unrelated classes or other unexplainable errors. For the most part, regardless of the correctness of predictions, the classifier is consistent; it either makes a correct classification or, in a much smaller number of cases, when it makes a mistake, it does so in a predictable and consistent way. This inspires confidence in its predictions, and it makes it possible to devise a strategy to mitigate its mistakes in future, such as a dictionary lookup as used in [32, 69] or by grouping up letters with similar features into smaller groups to help the classifier more effectively learn their differences as done in [69]. This is a desirable outcome.

An analysis of the robustness of the approach to new and unseen test subjects is carried out. Figure 6.19 is a bar graph of the average accuracy of each unseen test subject across all letter classes. The graph is based on Table B.8 in Appendix B. The graph shows

that the accuracies across all 5 unseen test subjects are high, with 3 subjects achieving above 70% accuracy, and 2 subjects exceeding 80% accuracy. There are no outliers in the graph, whatsoever. The graph indicates a relatively small amount of variation in average accuracy across test subjects, with a very small standard deviation of 5% across all subjects. This demonstrates that the proposed system is very robust to variations in test subjects, has absolutely no need for a pre-training procedure, and exhibits strong subject-independence, which is a very desirable quality for a system of this type.

At this stage, it can be concluded that Research Objective 9, as the final remaining objective, has been successfully completed.

The following final answers to the research sub-questions can be provided as follows:

- As a final and complete response to Research Sub-Question 3, it is concluded that it is possible to recognise the letters with a high accuracy, without the need for any pre-training procedure.
- As a final response to Research Sub-Question 4, it is stated that high-accuracy recognition of the letters, digits and fundamental shapes can be achieved using only a single microphone, without the need for additional microphone inputs. This is a remarkable achievement as it represents the most minimal, non-complex and low-cost set up that doesn't require any hardware modifications whatsoever.

Given the high-accuracy recognition obtained on both the semi-seen and unseen testing sets in this experiment, the use of the two- or three-microphone configurations are not deemed to be necessary for high-accuracy recognition. However, for the sake of interest, to investigate the extent to which additional microphone inputs can help increase the recognition accuracy, a final comparative experiment on the LS unseen testing set was carried out to compare the use of one, two and three microphone inputs towards recognition accuracy. This is explained in the next section. Note, however, that results of using two and three microphone inputs on the LS semi-seen testing set have been provided in Appendices B, C and D for the interested reader, but are not analysed.

## 6.4 Experiment to Compare One, Two and Three Microphone Inputs Towards Letter Recognition on Unseen Data, Results and Analysis

For ease of reference in this section, and henceforth, the one-microphone, two-microphone and three-microphone letter classifiers will be referred to as "1Cls", "2Cls" and "3Cls",

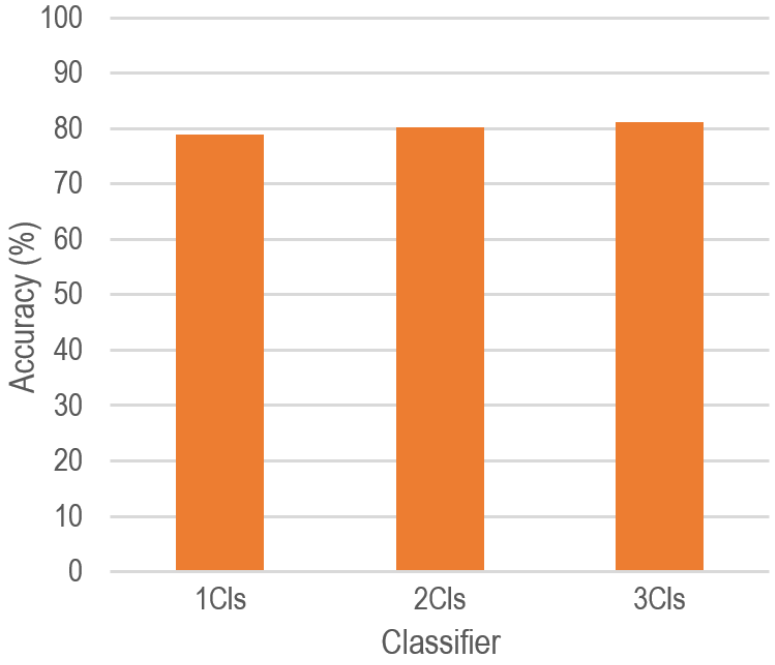


FIGURE 6.20: Comparison of the overall accuracy (%) of the three classifiers across all test subjects and letters.

respectively.

The comprehensive set of results for 2Cls and 3Cls on both the semi-seen and unseen testing sets are provided in Appendices C and D, respectively. Relevant excerpts of the results in these appendices will be presented below.

Figure 6.20 graphically illustrates the overall accuracies of the three classifiers across all test subjects and letters. On average, it is observed that the accuracy achieved when using 1, 2 and 3 microphone inputs appears to be approximately comparable, although a larger number of microphone input does appear to provide a very small increase in overall accuracy—from all test subjects and all letters—by about 1%. This result was unexpected, since the initial expectation was that more inputs would help provide significantly higher-accuracy recognition overall. The investigation for this small margin accuracy will be assigned to future work.

However, when considering that all three classifiers have been optimised, the similarity in accuracy is most likely attributed to the fact that, with the available data set, the most optimally recognisable features have been learned by 1Cls, while 2Cls and 3Cls have not been able to fully converge on the available data. Therefore, these classifiers have not been able to significantly improve on the accuracy. It is very likely that making use of a significantly larger data set can help 2Cls and 3Cls converge further, thereby providing even higher accuracies. This will require the collection of a much larger number of samples to be added to the respective data sets, which can be taken on in future.

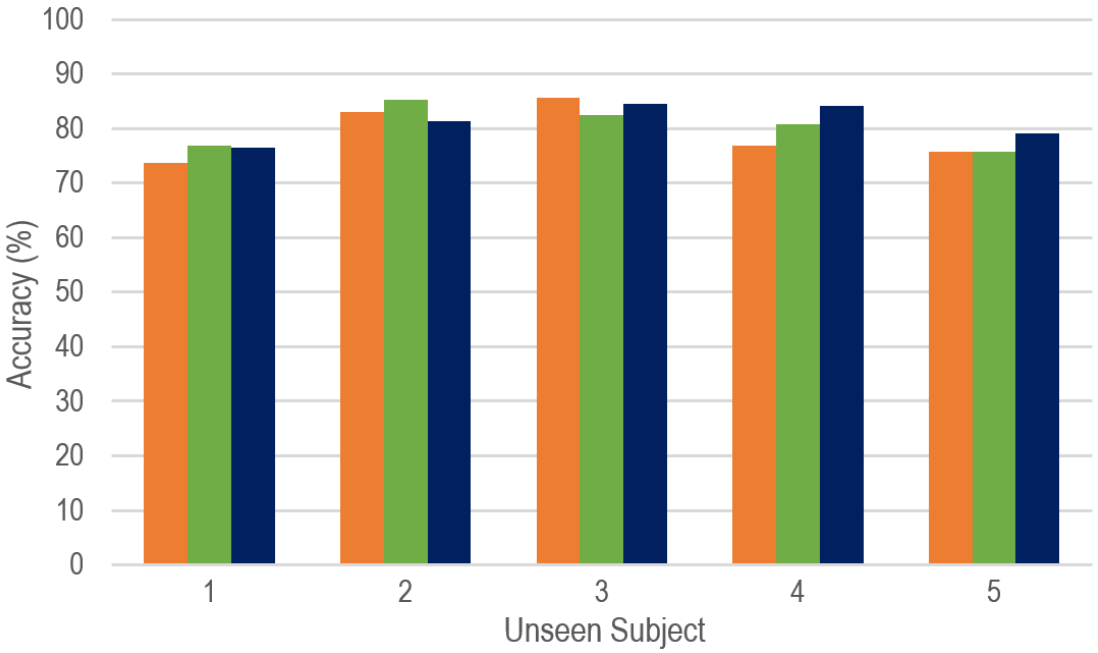


FIGURE 6.21: Comparison of the average accuracy (%) of the three classifiers per subject across all letters.

Figure 6.21 is based on Tables B.8, C.8 and D.8 in the respective appendices, and is a graphical comparison in the average accuracy per test subject across all letters for each of the three classifiers. At the highest level, the graph demonstrates that the use of 1, 2 or 3 microphone inputs appear to be very comparable in this regard as well, in all three cases yielding classifiers that are highly robust to variations in test subjects and exhibiting strong user-independence. In general, the accuracies achieved across 1, 2 or 3 inputs appear to be approximately at same level for each subject. No outliers or extreme values are observed in this regard.

Observing the accuracy of each of the subjects versus the number of microphones, it is observed that:

- in most cases, using 2 inputs instead of 1 appears to provide a small benefit to accuracy, although Subject 5 does not register this benefit, and Subject 3 actually registers a small reduction in accuracy using 2 inputs.
- in majority cases, using 3 inputs instead of 1 or 2 inputs appears to also provide a small benefit to accuracy, although this trend is again not completely consistent, with some subjects registering either no benefit or even a small reduction in accuracy.

Therefore, as regards test subjects, it can be concluded that the use of any of the number of inputs considered provides strong user-independence, but a larger number of inputs

does appear to provide a general but small accuracy benefit. Following on the previous discussion, it is expected that making use of a larger training set can help increase the accuracies per subject, but the accuracies across test subjects would likely remain at an approximately consistent level as they rise i.e. subject independence is expected to be maintained for a given classifier. Again, this may be investigated for future work.

The final comparison in this analysis concerns the accuracies per letter of the three classifiers. Given the relatively large number of letters recognised, it was decided to conduct the analyses with 1Cls treated as a baseline to compare 2Cls and 3Cls to. Figure 6.22 is a bar graph of the average accuracy of 1Cls per letter across all test subjects, sorted in descending order of accuracy, which is based on Table B.5 in Appendix B. The graph has been segmented into three regions corresponding to the top, middle and bottom 1/3 of accuracies achieved by the letters. The comparison of 1Cls to 2Cls and 3Cls in this regard will aim to determine how letters in each of the three accuracy segments are affected by additional inputs, 2 or 3. The three groups i.e. the top, middle and bottom 1/3 accuracy groups will henceforth be referred to as "top group", "middle group" and "lower group", respectively.


FIGURE 6.22: Average accuracy (%) of 1Cls per letter across all test subjects sorted in descending order of accuracy, with demarcations indicating the top, middle and lower 1/3 groupings of accuracies.

It is observed that exactly half of the letters fall in the top group which is in the range between about 81% to 97%. Figure 6.23 is a bar graph that presents the comparison between 1Cls, 2Cls and 3Cls for letters in the top group, based on Tables B.5, C.5 and D.5 in the respective appendices. A careful observation of each of the letters reveals the following:

- For 8 letters (*I*, *N*, *O*, *E*, *H*, *M*, *G* and *R*), using 3 inputs—as opposed to 1—provides a moderate benefit to accuracy ranging from 3–6%, and in most of these cases, the use of 2 inputs provides the same or a slightly higher benefit to accuracy. Either way, more inputs appears to help accuracy in this case.
- For only 2 letters (*W* and *A*), using 3 inputs offers no benefit to accuracy, while using 2 inputs does appear to offer a benefit to accuracy.
- For the remaining 3 letters (*Z*, *S* and *F*), using 3 inputs appears to be detrimental to accuracy, with the use of 2 inputs either bettering the baseline 1Cls accuracy, or offering no benefit, or causing a detriment to accuracy, but never any worse than with 3 inputs.


FIGURE 6.23: Comparison of the average accuracy (%) of letters in the top group for 1Cls (orange), 2Cls (green) and 3Cls (navy-blue) across all test subjects.

All in all, the above points suggest that for letters in the top group i.e. letters that are very easily recognisable with 1 input, the use of more inputs generally offers a small to moderate benefit to accuracy, with a few exceptions, and the use of 2 inputs appears to be more beneficial than the use of 3 inputs.

Moving on to the letters in the middle group, which includes 8 letters, Figure 6.24 is a bar graph that visualises the comparison between 1Cls, 2Cls and 3Cls for letters in this group. Carefully observing each of the letters in the figure uncovers the following:

- For half of these letters ($J$, $L$, $V$ and $B$), the use of 3 inputs provides a benefit to accuracy, which ranges from moderate increases of 6% to one extremely large increase of 20%, and in these cases 2 inputs has very varied behaviour ranging from offering a small or moderate benefit (2 cases), leaving the accuracy unchanged (1 case), or causing a significant drop in accuracy of about 12% (1 case).
- For 2 letters ($P$ and $Y$), the use of 3 inputs leaves the accuracy unchanged, while the use of 2 inputs either offers a small benefit to accuracy (1 case) or harms the accuracy (1 case).

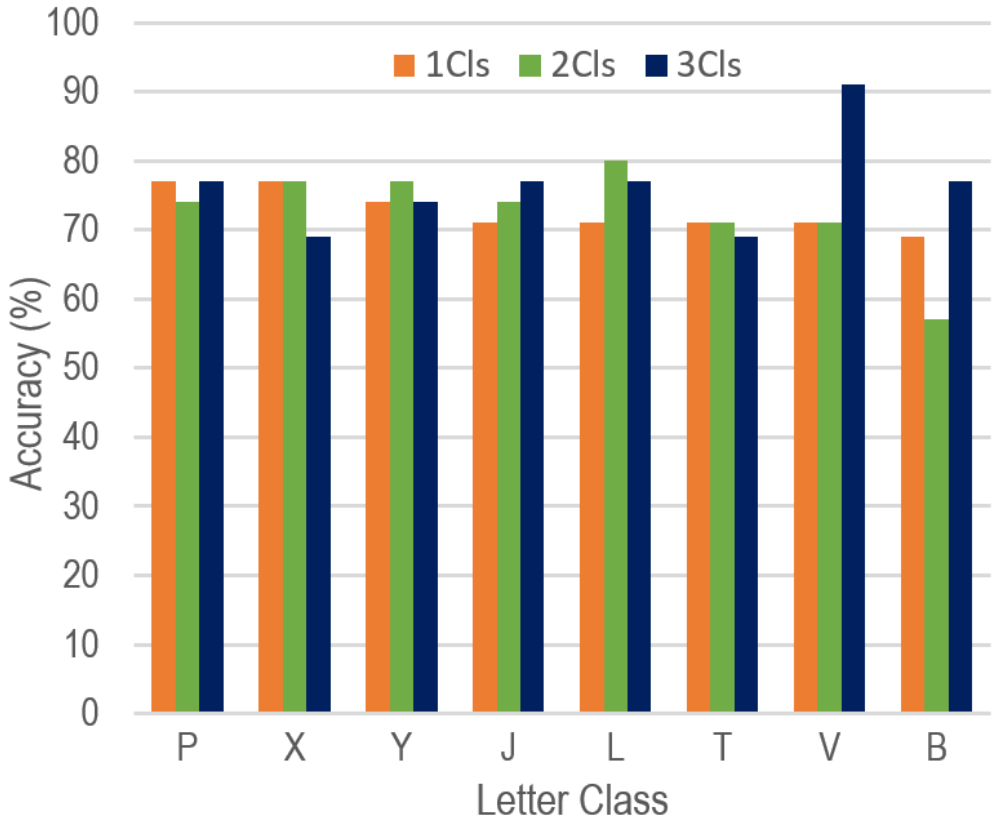


FIGURE 6.24: Comparison of the average accuracy (%) of letters in the middle group for 1Cls (orange), 2Cls (green) and 3Cls (navy-blue) across all test subjects.

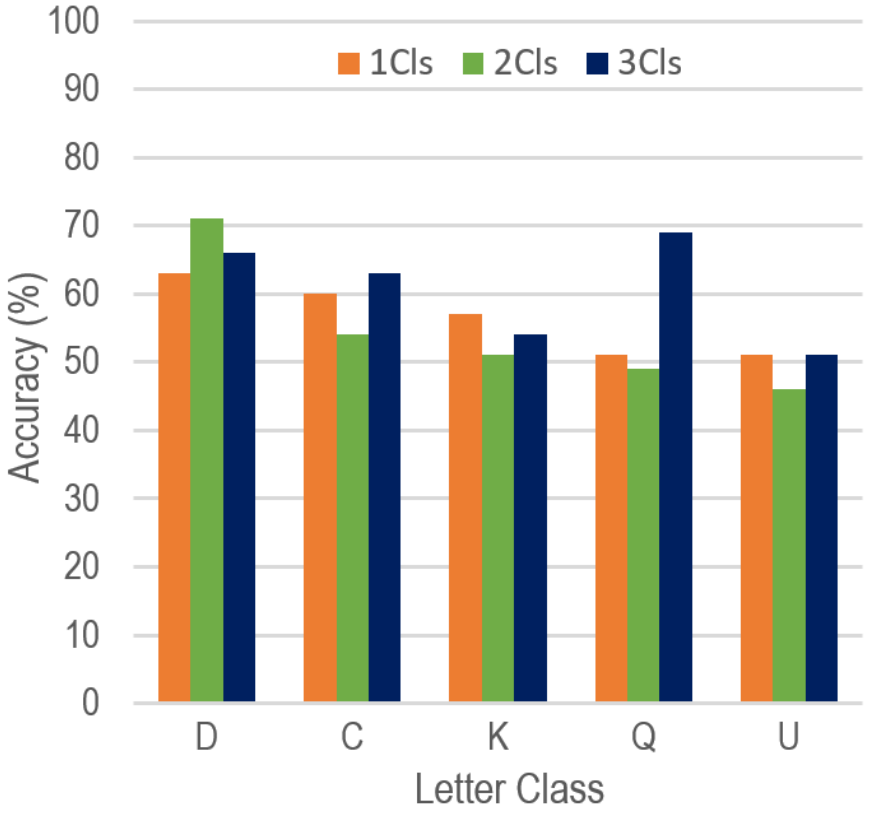


FIGURE 6.25: Comparison of the average accuracy (%) of letters in the lower group for 1Cls (orange), 2Cls (green) and 3Cls (navy-blue) across all test subjects.

- Finally, for 2 letters ($X$ and $T$), the use of 3 inputs is detrimental to accuracy, ranging from a small (2%) to moderate reduction (8%) while the use of 2 inputs leaves the accuracy unchanged.

All in all, the above points suggest that, once again, the use of more inputs is generally beneficial to accuracy, with some exceptions, but for letters in this group i.e. letters that are highly recognizable, the use of 3 inputs appears to be more beneficial to accuracy than the use of 2 inputs, which is opposite to the trend observed in the top group.

Finally, analysing the letters in the lower group, which only consists of 5 of the 26 letters, Figure 6.25 is a bar graph that visualises the comparison between 1Cls, 2Cls and 3Cls for letters in this group. Observing this graphs reveals that:

- As in previous groups, for most of the letters in this group ($D$, $C$, $Q$), the use of 3 inputs provides a benefit to accuracy, but in this case the use of 2 inputs is either of detriment to the accuracy (2 cases) or offers a slightly higher benefit than the use of 3 inputs (1 case).

- For 1 letter ($U$), which is the letter at the bottom of the group and has the minimum accuracy out of all the letters, the use of 3 inputs is the same as using 1 input, and the use of 2 inputs slightly reduces the accuracy.
- For the final remaining letter ($K$), the use of 3 inputs causes a small decrease in accuracy, while the use of 2 inputs causes an even bigger decrease in accuracy than with 3 inputs.

So overall, the points above yet again suggest that the use of 3 inputs is mostly beneficial to accuracy, but in this case i.e. for letters that are less easily recognisable that the previous two groups, it appears that the use of 2 inputs proves to be of a small detriment to the accuracy.

The overarching theme in the analysis above is that additional inputs generally prove to be beneficial to accuracy and are on average comparable. However, the exact number of inputs that provides the best benefit is letter-specific, with most letters in this configuration taking more benefit from 3 inputs than from 2. The previous note made about the possible need for 2Cls and 3Cls to be trained further on a larger number of samples is important to repeat here. It is expected that, with a sufficient number of training samples, the use of 3 inputs will most likely provide at least the same benefit to accuracy across all or most of the letters, or a better accuracy than with 2 inputs. This will be an interesting area of future work.

It can be concluded that, more inputs for the most part appear to help contribute towards a higher accuracy per letter, and the extent of the benefit provided by additional inputs is letter-specific, but some letters are still easier to recognise than others, regardless of the number of inputs used.

Therefore, as an addition to the answer provided above for Research Sub-Question 4, it is stated that high-accuracy recognition of the letters, digits and fundamental shapes can be achieved using only a single microphone, without the need for additional microphone inputs, **but making use of additional inputs—upto 3 considered here—can provide a benefit to the overall accuracy of the system, and specifically help to enhance the accuracy of many, but not all, of the individual letters.**

## 6.5 Comparison of the Proposed Approach to Related Studies

Table 6.9 is a reworked version of Table 3.7 that was provided in Chapter 3, that summarised the related studies that were reviewed in that chapter. The studies in the re-worked table provided here have been organised according to, respectively: (1) the classes recognised; (2) the robustness to test subjects; and (3) the overall accuracy achieved. In each class category, the table also includes the proposed system along with the results obtained on the respective **unseen** testing sets for each category, for comparison.

Comparing the proposed system to [70] at the top of the table, it is observed that both recognise a set of seven classes, but the proposed system is capable of providing strong robustness to subject variations at small reduction in accuracy as compared to [70] which is heavily user-dependent. It is important to note that the fundamental shapes classifier was not optimised and doing so will likely improve on the accuracy even further. Therefore, the proposed strategy is more effective and beneficial.

In the digits class category, the proposed system is compared to [52] and [10], both of which are very robust to test subjects, as is the proposed system, and [52] obtains a very slightly higher accuracy than the proposed system, while [10] surpasses the accuracy of the proposed system by a significant amount. In this case, it appears as though the two related systems out-perform the proposed system, but the proposed system still performs at a very good level. Once again, it should be noted that the digit classifier was also not optimised and doing so would most likely provide a much higher accuracy, although this comparison would need to be repeated in future to determine this definitively.

Finally, in the letters class category, the proposed system is compared to several related systems [16, 33, 53, 62, 69]. Three of these systems [33, 53, 62] are heavily user-dependent, and of these, two have higher accuracies than the proposed system. The robustness to users of the proposed system coupled with its high accuracy can be considered to be more advantageous and effective than all three of these systems. The remaining two systems [16, 69] have lower or minimal user-dependence, but both of these systems achieve lower accuracies than the proposed system. Therefore, the proposed approach is more effective and beneficial than all of the related systems in this category.

This comparison indicates that the use of piezo microphones, coupled with the MFCC and SVM, not only provide a rich source of audio information for recognising audio classes at a high accuracy, but also that this setup can strongly compete with, and

even surpass, more sophisticated systems with specialised and costly hardware, although optimisation of the classifiers used in the proposed system is very important to doing so.

## 6.6 Summary

This chapter detailed the experiments carried out, as well as the results and analysis of each experiment, towards providing answers to the Research Sub-Questions 1–4, and finally providing an answer to the main research question posed in Chapter 1.

Experiments were carried out to recognise the fundamental shapes, the digits and letters, first on the respective semi-seen testing sets, followed by the same experiments on the respective unseen testing sets, both using data from only one microphone input. In all of these experiments, it was shown definitely that, without the need for any pre-training procedure, high-accuracy recognition of the recognised classes is possible, although as expected, pre-training provides higher-accuracy recognition than when completely unseen data is used.

For all three experiments, it was shown that the proposed technique exhibits strong user-independence, and in general performs very well across all test subjects, with a small amount of variation. This is a very desirable quality for a system of this type.

In terms of the classes recognised, in all three experiments, it was apparent that, while in most cases the proposed approach showed a strong ability to recognise most classes at a high accuracy, the proposed approach is class-specific, with some classes being easier to recognise than others.

Further analyses of the accuracies of classes across all three experiments revealed a recurring observation that incorrectly classified samples mostly appeared to be directly tied to a similarity in the audio forms of classes, rather than to confusion with random unrelated classes or other unexplainable errors. It was noted repeatedly that this strongly supports the idea that the proposed system learnt in every case to very accurately model the phonemes in the respective recognised classes, as well as then associate sets of these phonemes with specific classes. This points to the fact that, for the most part, the proposed approach is shown to be consistent; in most cases a correct classification is made and in the smaller number of cases when an incorrect classification is made, this happens in a predictable and consistent way. This inspires confidence in its predictions, and it makes it possible to devise a strategy to mitigate classification errors in future, such as a dictionary lookup as used in [32, 69] or by grouping up letters with similar

features into smaller groups to help the approach more effectively learn their differences as done in [69].

As a result of the observations above, it was noted that Research Objectives 7–8 were successfully achieved, and final answers to the four research sub-questions were progressively provided as follows:

- As a final response to Research Sub-Question 1, it was concluded that it is possible to recognise the seven fundamental shapes with an almost perfect accuracy, without the need for any pre-training procedure. It was also noted that default parameters were used in this classifier and that an optimisation of the classifier in future can provide even better results.
- As a final response to Research Sub-Question 2, it was concluded that it is possible to recognise the digits with a high accuracy, without the need for any pre-training procedure. In this case, also, it was noted that an optimisation procedure can further enhance these results.
- As a final response to Research Sub-Question 3, it was concluded that it is possible to recognise the letters with a high accuracy, without the need for any pre-training procedure.
- As a response to Research Sub-Question 4, it was stated that high-accuracy recognition of the letters, digits and fundamental shapes can be achieved using only a single microphone, without the need for additional microphone inputs. It was noted that this is a remarkable achievement as it represents the most minimal, non-complex and low-cost set up that doesn't require any hardware modifications whatsoever.

Given the high-accuracy recognition obtained for each of the experiments, the use of the two- or three-microphone configurations was not deemed to be necessary for high-accuracy recognition. However, for the sake of interest, and to investigate the extent to which additional microphone inputs can help increase the recognition accuracy, a final comparative experiment was carried out to compare the use of one, two and three microphone inputs towards the recognition of unseen letters.

When comparing different numbers of inputs in terms of overall accuracy, it was found that, in general, the use of more microphone inputs—out of the options compared—provides a benefit to the overall accuracy, but, contrary to expectation, the benefit is relatively small. It was noted that one possible reason for this is that, with the available data set, the most optimally recognisable features were learned by 1Cls, while 2Cls and

3Cls were most likely not able to fully converge on the available data, leading to a small improvement in accuracy by 2Cls and 3Cls. It was noted that it is very likely that making use of a significantly larger data set can help 2Cls and 3Cls converge further, thereby providing even higher accuracies, which can be investigated in future.

Comparing the effect of increasing the number of inputs on the accuracy of individual test subjects showed that the use of any of the number of inputs considered provides strong user-independence, but a larger number of inputs appears to provide a small accuracy benefit to all subjects approximately uniformly.

Finally, comparing the effect of increasing the number of inputs on the accuracy of individual letters revealed that a larger number of inputs mostly and generally proves to be beneficial to the accuracy of individual letters, but the optimal number of inputs used appears to be letter-specific, with some letters doing better with 2 inputs, and others doing better with 3 inputs. This was noted as, once again, most likely being associated with the fact that 2Cls and 3Cls may require more data to generalise further, and doing so will likely cause 3Cls to perform better than 2Cls.

Therefore, a more complete answer to Research Sub-Question 4 was provided as: high-accuracy recognition of the letters, digits and fundamental shapes can be achieved using only a single microphone, without the need for additional microphone inputs, but making use of additional inputs—upto 3 considered here—can provide a benefit to the overall accuracy of the system, and specifically help to enhance the accuracy of many, but not all, of the individual letters.

With all results obtained and analysed in detail, a comparison of results with related systems discussed in Chapter 3 was carried out in order to contextualise these results. The comparison indicated that the use of piezo microphones, coupled with the MFCC and SVM, not only provide a rich source of audio information for recognising audio classes at a high accuracy, but also strongly compete with more sophisticated systems with specialised and costly hardware, and in most cases surpass these systems, although optimisation of the classifiers used is important to doing so.

Based on these analyses and conclusions, the next chapter concludes the thesis with an answer to the main research question.

| Study | Recognition classes | Audio Hardware | Feature Extraction | Classifier | User dependency | Overall Accuracy (%) |
|---|---|---|---|---|---|---|
| Proposed | 7 fundamental shapes | piezo microphone | MFCC | SVM | Minimal | 83 |
| [70] | 7 strokes | Android mobile devices | Spectrum density function from FFT | $k$-nearest-neighbours | Heavy | 90.3 |
| Proposed | Digits | piezo microphone | MFCC | SVM | Minimal | 77 |
| [52] | Digits | ATmega328p microcontroller with sensors | Hilbert envelope with sensor data | Neural network | Minimal | 78.38 |
| [10] | Digits | Embedded microphone of a Huawei Watch I | Wavelet to Fourier transform then images | Convolutional neural network | Minimal | 92.75 |
| Proposed | 26 Letters | piezo microphone | MFCC | SVM | Minimal | 81.1 (3 mics) |
| [16] | 26 letters | Embedded microphone of a Huawei Watch I | STFT | Convolutional neural network | Minimal | 75 (Unseen) |
| [69] | 26 Letters | Generic mobile devices | MFCC | SVM | Moderate | 50-60 |
| [62] | 26 Letters and 11 Short Words | Mobile device with dual microphone and speaker | LLAP | Template matching | Heavy | 92.3 |
| [33] | 26 Letters | MacAir microphone | MFCC with mean amplitude | DTW | Heavy | 83.7 |
| [53] | 26 cursive letters and 26 cursive words | Labtec computer microphone | Power signal derived features | Template matching | Heavy | 70 |

TABLE 6.9: Summary of related studies reviewed, and the results obtained by the proposed system.

# Chapter 7

# Conclusion

This research investigated the creation of an audio-shape recognition system that allows the user to draw audio shapes—fundamental shapes, digits and/or letters—on a given surface such as a table-top using a generic stylus such as the back of a pen. The system then aimed to make use of one, two or three piezo microphones to capture the sound of the audio gestures, and a combination of the MFCC feature descriptor and SVMs to recognise audio shapes. The framework was initially applied to a set of seven fundamental shapes, followed by the 10 digits ranging from 0–9, and finally to uppercase alphabet characters. The novelty of the system is in the use of piezo microphones which are low cost, light-weight and portable, and the main investigation was around determining whether these microphones are able to provide sufficiently rich information to recognise the audio shapes mentioned.

In response to the main research question "How accurately can the proposed configuration, i.e. a combination of one or more piezo microphones, the MFCC feature descriptor and SVMs, recognise audio shapes drawn by a generic stylus on a given surface?", it can be stated that the proposed configuration makes it possible to recognise audio shapes drawn in this manner with a very high accuracy, as summarised in Table 7.1. Futhermore, proposed set up was shown to also be able to out-perform almost all related systems that made use of costly specialised hardware, which is remarkable.

The proposed system can be used as a low-cost and highly portable novel input device that can be used without the need for any hardware modifications, and without the need for pre-training. Furthermore, Chapter 3 mentioned two studies [10, 16] that aimed to create novel interfaces for smart watches that offer very limited screen space. The limited screen space of smart watches is a significant challenge from an input and interface perspective. Were a miniature piezo microphone to be embedded into smart watches, this research has shown that it may be possible to adapt the system proposed

| Class | Accuracy (%) | |
|---|---|---|
| | **Semi-Seen** | **Unseen** |
| Fundamental Shapes | 94 | 83 |
| Digits | 91 | 77 |
| Letters (1 mic) | 86.2 | 79.0 |
| Letters (2 mics) | 88.5 | 80.2 |
| Letters (3 mics) | 91.5 | 81.1 |

TABLE 7.1: Summary of the average accuracies obtained.

in this research to that context whereby, for example, gestures can be drawn on the back of the user's hand or arm and captured in, and recognised on, the smart watch, similar to [10]. This can significantly expand on the capabilities of smart watches, from an interface perspective, with relatively low cost hardware, and this research has made a strong case for manufacturers to begin investigating the feasibility of doing so.

## 7.1 Future Work

The following provides several directions for future work.

**Writing-style variations:** The writing-style standard for subjects may be lifted, and the experiments repeated on a more comprehensive data set, to investigate the provision of subjects with complete freedom of writing style.

**Use of a more comprehensive data set:** As mentioned in the previous chapter, it is possible that a more comprehensive data set can help the 2-microphone and 3-microphone letter classifiers further generalise, thereby providing higher-accuracy recognition. This can be investigated.

**Optimisation of the fundamental shapes and digit classifiers:** These classifiers were used with their default parameters in this research. When optimised, it is expected that these classifiers will perform significantly better.

**Accuracy enhancement methods:** The use of a dictionary lookup as used in [32, 69] or the use of a letter-grouping method as done in [69], can both be investigated, towards higher-accuracy recognition of words.

**Use an alternative classifier:** An alternative classifier like an Artificial Neural Network or Deep Learning approaches may be applied and compared for performance.

**Use an alternative feature descriptor:** An alternative feature descriptor like the GFCC may be applied and compared for performance, especially under higher-noise

environments.

## 7.2 Concluding Comments

This has given a wonderful academic experience and journey for the researcher. May the basis of the content from this research provide any form of assistance to other fellow researchers within the field of acoustic audio, HCI and machine learning.

# Appendix A

# Additional Results for the Fundamental Shape and Digit Recognition Experiments

## A.1 Shapes

| Actual Shape | Predicted Shape | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Dot** | **Dash** | **Tick** | **Cross** | **Circle** | **Triangle** | **Square** |
| **Dot** | **10** | 0 | 0 | 0 | 0 | 0 | 0 |
| **Dash** | 0 | **10** | 0 | 0 | 0 | 0 | 0 |
| **Tick** | 0 | 0 | **10** | 0 | 0 | 0 | 0 |
| **Cross** | 0 | 0 | 0 | **10** | 0 | 0 | 0 |
| **Circle** | 0 | 0 | 0 | 0 | **10** | 0 | 0 |
| **Triangle** | 0 | 0 | 3 | 0 | 0 | **7** | 0 |
| **Square** | 0 | 0 | 0 | 0 | 1 | 0 | **9** |

TABLE A.1: Confusion matrix of shape recognition results for semi-seen data.

| Shape | Accuracy (%) | | | | | Overall (%) |
|---|---|---|---|---|---|---|
| | **Subj. 1** | **Subj. 2** | **Subj. 3** | **Subj. 4** | **Subj. 5** | |
| Dot | 100 | 90 | 90 | 80 | 100 | 92 |
| Dash | 100 | 90 | 90 | 90 | 90 | 92 |
| Tick | 90 | 70 | 80 | 70 | 100 | 82 |
| Cross | 100 | 80 | 70 | 100 | 100 | 90 |
| Circle | 100 | 90 | 80 | 70 | 80 | 84 |
| Triangle | 90 | 70 | 60 | 60 | 90 | 74 |
| Square | 90 | 70 | 60 | 60 | 70 | 70 |
| **Overall (%)** | **96** | **80** | **76** | **76** | **90** | **83.4** |

TABLE A.2: Accuracy and subject results of the shapes recognition for unseen data.

| **Actual Shape** | **Predicted Shape** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Dot** | **Dash** | **Tick** | **Cross** | **Circle** | **Triangle** | **Square** |
| **Dot** | **46** | 4 | 0 | 0 | 0 | 0 | 0 |
| **Dash** | 2 | **46** | 2 | 0 | 0 | 0 | 0 |
| **Tick** | 0 | 5 | **41** | 4 | 0 | 0 | 0 |
| **Cross** | 0 | 1 | 4 | **45** | 0 | 0 | 0 |
| **Circle** | 0 | 0 | 0 | 0 | **42** | 2 | 6 |
| **Triangle** | 0 | 0 | 9 | 0 | 0 | **37** | 4 |
| **Square** | 0 | 0 | 0 | 0 | 10 | 5 | **35** |

TABLE A.3: Confusion matrix of shape recognition results for unseen data.

## A.2 Digits

| **Actual Digit** | **Predicted Digit** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | **8** | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 |
| **5** | 0 | 1 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 |
| **6** | 0 | 0 | 1 | 3 | 0 | 0 | **6** | 0 | 0 | 0 |
| **7** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | **9** |

TABLE A.4: Confusion matrix of digits recognition results for semi-seen data.

| **Actual Digit** | **Predicted Digit** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | **14** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 |
| **1** | 0 | **18** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | **10** | 7 | 0 | 0 | 0 | 0 | 0 | 1 |
| **3** | 0 | 0 | 0 | **17** | 0 | 0 | 1 | 0 | 0 | 0 |
| **4** | 0 | 1 | 0 | 3 | **12** | 1 | 1 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 1 | **17** | 0 | 0 | 0 | 0 |
| **6** | 3 | 0 | 0 | 4 | 0 | 0 | **11** | 0 | 0 | 0 |
| **7** | 0 | 0 | 2 | 0 | 0 | 0 | 4 | **11** | 0 | 1 |
| **8** | 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | **12** | 1 |
| **9** | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **16** |

TABLE A.5: Confusion matrix of digits recognition results for unseen data.

# Appendix B

# Additional Results for the Letter Recognition Experiments Using One Microphone

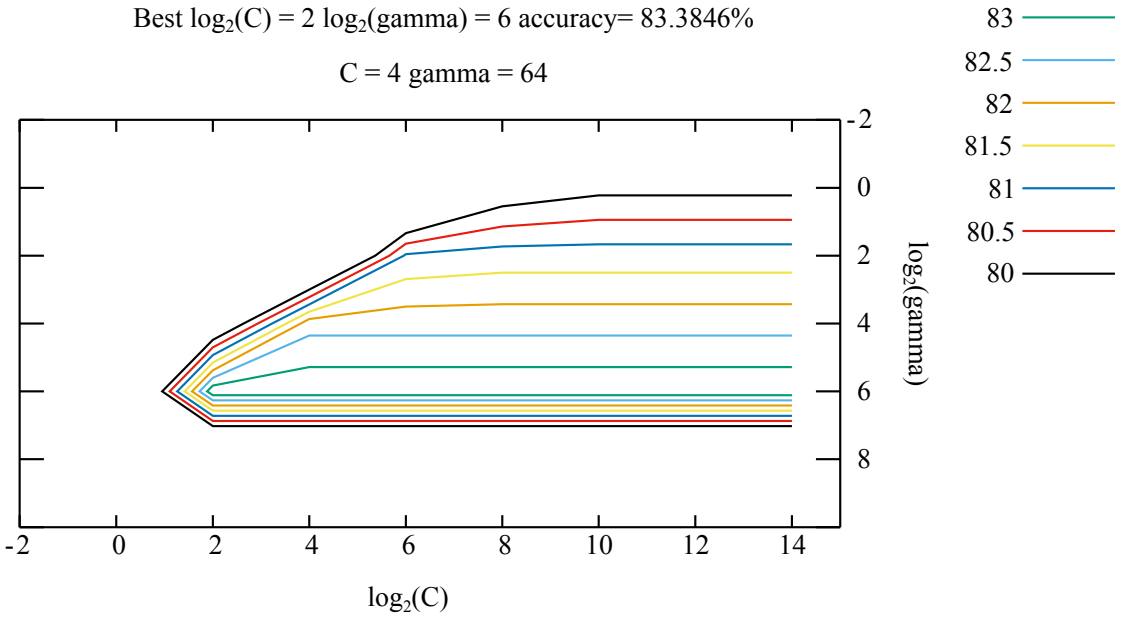


FIGURE B.1: Contour plot of grid-search results for the LS data set using one microphones.

```
log2c=6 log2g=4 rate=82.3077
log2c=2 log2g=4 rate=78.9231
log2c=6 log2g=8 rate=76.7692
log2c=2 log2g=8 rate=76.7692
log2c=12 log2g=4 rate=82.3077
log2c=12 log2g=8 rate=76.7692
log2c=6 log2g=0 rate=77.8462
log2c=2 log2g=0 rate=56.4615
log2c=12 log2g=0 rate=79.8462
log2c=0 log2g=4 rate=66.9231
log2c=0 log2g=8 rate=75.3846
log2c=0 log2g=0 rate=56.6154
log2c=6 log2g=10 rate=42.1538
log2c=2 log2g=10 rate=42.1538
log2c=12 log2g=10 rate=42.1538
log2c=0 log2g=10 rate=39.3846
log2c=10 log2g=4 rate=82.3077
log2c=10 log2g=8 rate=76.7692
log2c=10 log2g=0 rate=79.8462
log2c=10 log2g=10 rate=42.1538
log2c=4 log2g=4 rate=82.3077
log2c=4 log2g=8 rate=76.7692
log2c=4 log2g=0 rate=67.5385
log2c=4 log2g=10 rate=42.1538
log2c=6 log2g=2 rate=81.0769
log2c=2 log2g=2 rate=67.3846
log2c=12 log2g=2 rate=81.2308
log2c=0 log2g=2 rate=57.8462
log2c=10 log2g=2 rate=81.2308
log2c=4 log2g=2 rate=77.6923
log2c=14 log2g=4 rate=82.3077
log2c=14 log2g=8 rate=76.7692
log2c=14 log2g=0 rate=79.8462
log2c=14 log2g=10 rate=42.1538
log2c=14 log2g=2 rate=81.2308
log2c=6 log2g=6 rate=83.3846
log2c=2 log2g=6 rate=83.3846
log2c=12 log2g=6 rate=83.3846
log2c=0 log2g=6 rate=76.9231
log2c=10 log2g=6 rate=83.3846
log2c=4 log2g=6 rate=83.3846
log2c=14 log2g=6 rate=83.3846
log2c=-2 log2g=4 rate=59.8462
log2c=-2 log2g=8 rate=60.3077
log2c=-2 log2g=0 rate=56.6154
log2c=-2 log2g=10 rate=22.6154
log2c=-2 log2g=2 rate=57.6923
log2c=-2 log2g=6 rate=67.3846
log2c=6 log2g=-2 rate=67.5385
log2c=2 log2g=-2 rate=56.4615
log2c=12 log2g=-2 rate=79.5385
log2c=0 log2g=-2 rate=56.4615
log2c=10 log2g=-2 rate=79.3846
log2c=4 log2g=-2 rate=56.4615
log2c=14 log2g=-2 rate=79.5385
log2c=-2 log2g=-2 rate=56.4615
log2c=8 log2g=4 rate=82.3077
log2c=8 log2g=8 rate=76.7692
log2c=8 log2g=0 rate=79.5385
log2c=8 log2g=10 rate=42.1538
log2c=8 log2g=2 rate=81.2308
log2c=8 log2g=6 rate=83.3846
log2c=8 log2g=-2 rate=77.6923
```

TABLE B.1: Grid-search optimisation log file output for the one-microphone letter classifier, showing $C$ and $\gamma$ parameter values and the percentage cross-validation accuracy ("rate") for each pair.

| Letter | Correctly Predicted (10) | Accuracy (%) |
|---|---|---|
| **A** | 8 | 80 |
| **B** | 8 | 80 |
| **C** | 9 | 90 |
| **D** | 8 | 80 |
| **E** | 9 | 90 |
| **F** | 7 | 70 |
| **G** | 9 | 90 |
| **H** | 8 | 80 |
| **I** | 10 | 100 |
| **J** | 10 | 100 |
| **K** | 8 | 80 |
| **L** | 9 | 90 |
| **M** | 10 | 100 |
| (cont. right) | | |

(cont. from left)

| Letter | Correctly Predicted (10) | Accuracy (%) |
|---|---|---|
| **N** | 10 | 100 |
| **O** | 10 | 100 |
| **P** | 8 | 80 |
| **Q** | 8 | 80 |
| **R** | 8 | 80 |
| **S** | 9 | 90 |
| **T** | 7 | 70 |
| **U** | 7 | 70 |
| **V** | 9 | 90 |
| **W** | 9 | 90 |
| **X** | 10 | 100 |
| **Y** | 9 | 90 |
| **Z** | 7 | 70 |
| **Average** | — | **86.2** |

TABLE B.2: Average accuracy per letter for the LS semi-seen testing set for one microphone.

| Letter | Precision | Recall | $f_1$ Score |
|---|---|---|---|
| **A** | 100 | 80 | 89 |
| **B** | 100 | 80 | 89 |
| **C** | 75 | 90 | 82 |
| **D** | 89 | 80 | 84 |
| **E** | 82 | 90 | 86 |
| **F** | 88 | 70 | 78 |
| **G** | 90 | 90 | 90 |
| **H** | 73 | 80 | 76 |
| **I** | 83 | 100 | 91 |
| **J** | 83 | 100 | 91 |
| **K** | 80 | 80 | 80 |
| **L** | 90 | 90 | 90 |
| **M** | 100 | 100 | 100 |
| (cont. right) | | | |

(cont. from left)

| Letter | Precision | Recall | $f_1$ score |
|---|---|---|---|
| **N** | 100 | 100 | 100 |
| **O** | 83 | 100 | 91 |
| **P** | 89 | 80 | 84 |
| **Q** | 89 | 80 | 84 |
| **R** | 80 | 80 | 80 |
| **S** | 69 | 90 | 78 |
| **T** | 100 | 70 | 82 |
| **U** | 100 | 70 | 82 |
| **V** | 90 | 90 | 90 |
| **W** | 90 | 90 | 90 |
| **X** | 71 | 100 | 83 |
| **Y** | 90 | 90 | 90 |
| **Z** | 88 | 70 | 78 |
| **Average** | **87** | **86** | **86** |

TABLE B.3: Percentage (%) performance metrics per letter for the LS semi-seen testing set for a single microphone.

| Actual Letter | Predicted Letter | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
| **A** | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |
| **B** | 0 | **8** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **F** | 0 | 0 | 0 | 0 | 0 | **7** | 0 | **1** | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **H** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **I** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **J** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **K** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | **1** | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **L** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| **M** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **N** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **1** | 0 | 0 |
| **Q** | 0 | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **R** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **8** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| **S** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | **1** | 0 | **1** | 0 | 0 |
| **U** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **1** | 0 | **7** | 0 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | **1** | 0 | 0 |
| **W** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 |
| **X** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 |
| **Z** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **7** |

TABLE B.4: Confusion matrix of letter recognition results for semi-seen data for the LS unseen testing set for one microphone.

| Letter | Correctly Predicted (35) | Accuracy (%) |
|---|---|---|
| **A** | 31 | 89 |
| **B** | 24 | 69 |
| **C** | 21 | 60 |
| **D** | 22 | 63 |
| **E** | 32 | 91 |
| **F** | 31 | 89 |
| **G** | 30 | 86 |
| **H** | 32 | 91 |
| **I** | 33 | 94 |
| **J** | 25 | 71 |
| **K** | 20 | 57 |
| **L** | 25 | 71 |
| **M** | 31 | 89 |
| (cont. right) | | |

(cont. from left)

| Letter | Correctly Predicted (35) | Accuracy (%) |
|---|---|---|
| **N** | 33 | 94 |
| **O** | 33 | 94 |
| **P** | 27 | 77 |
| **Q** | 18 | 51 |
| **R** | 29 | 83 |
| **S** | 33 | 94 |
| **T** | 25 | 71 |
| **U** | 18 | 51 |
| **V** | 25 | 71 |
| **W** | 34 | 97 |
| **X** | 27 | 77 |
| **Y** | 26 | 74 |
| **Z** | 34 | 97 |
| **Average** | — | **79.0** |

TABLE B.5: Average accuracy per letter for the LS unseen testing set for one microphone.

| Letter | Precision | Recall | $f_1$ Score |
|---|---|---|---|
| **A** | 84 | 89 | 86 |
| **B** | 75 | 69 | 72 |
| **C** | 88 | 60 | 71 |
| **D** | 71 | 63 | 67 |
| **E** | 100 | 91 | 96 |
| **F** | 76 | 89 | 82 |
| **G** | 67 | 86 | 75 |
| **H** | 97 | 91 | 94 |
| **I** | 94 | 94 | 94 |
| **J** | 74 | 71 | 72 |
| **K** | 91 | 57 | 70 |
| **L** | 89 | 71 | 79 |
| **M** | 100 | 89 | 94 |
| (cont. right) | | | |

(cont. from left)

| Letter | Precision | Recall | $f_1$ score |
|---|---|---|---|
| **N** | 70 | 94 | 80 |
| **O** | 79 | 94 | 86 |
| **P** | 71 | 77 | 74 |
| **Q** | 86 | 51 | 64 |
| **R** | 64 | 83 | 73 |
| **S** | 58 | 94 | 72 |
| **T** | 71 | 71 | 71 |
| **U** | 60 | 51 | 55 |
| **V** | 78 | 71 | 75 |
| **W** | 100 | 97 | 99 |
| **X** | 73 | 77 | 75 |
| **Y** | 84 | 74 | 79 |
| **Z** | 94 | 97 | 96 |
| **Average** | **81** | **79** | **79** |

TABLE B.6: Percentage (%) performance metrics per letter for the LS unseen testing set for a single microphone.

| Actual Letter | Predicted Letter | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
| **A** | **31** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| **B** | 0 | **24** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **11** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | **21** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **14** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0 | 0 | 0 | **22** | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | **2** | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 0 | 0 | 0 | 0 | **32** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| **F** | 0 | 0 | 0 | 0 | 0 | **31** | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| **G** | 0 | 0 | 0 | 0 | 0 | 0 | **30** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **3** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **H** | **2** | 0 | 0 | 0 | 0 | 0 | 0 | **32** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| **I** | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **33** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **J** | 0 | **4** | 0 | **3** | 0 | 0 | 0 | 0 | 0 | **25** | 0 | 0 | 0 | **1** | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 |
| **K** | **2** | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | **20** | **1** | 0 | 0 | 0 | **1** | 0 | **2** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| **L** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **25** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | **6** | 0 | 0 | 0 | 0 |
| **M** | 0 | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **31** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **N** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **33** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | **33** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **3** | 0 | **27** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 |
| **Q** | 0 | 0 | **1** | 0 | 0 | 0 | **15** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **18** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |
| **R** | 0 | **4** | 0 | **1** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **29** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **S** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **33** | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **2** | 0 | 0 | 0 | **25** | 0 | 0 | 0 | **7** | 0 | 0 |
| **U** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | **10** | 0 | **18** | 0 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **8** | **25** | 0 | 0 | 0 | 0 |
| **W** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **34** | 0 | 0 | 0 |
| **X** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | **27** | 0 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **26** | **1** |
| **Z** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **34** |

TABLE B.7: Confusion matrix of letter recognition results for unseen data for the LS unseen testing set for one microphone.

| **Subject** | **Correctly Predicted (182)** | **Average Accuracy (%)** |
|---|---|---|
| **1** | 134 | 73.6 |
| **2** | 151 | 83.0 |
| **3** | 156 | 85.7 |
| **4** | 140 | 76.9 |
| **5** | 138 | 75.8 |
| **Average** | — | **79.0** |
| **Std. Dev.** | — | **5.1** |

TABLE B.8: Average accuracy (%) per unseen test subject for the LS unseen testing set using one microphone.

# Appendix C

# Additional Results for the Letter Recognition Experiments Using Two Microphones

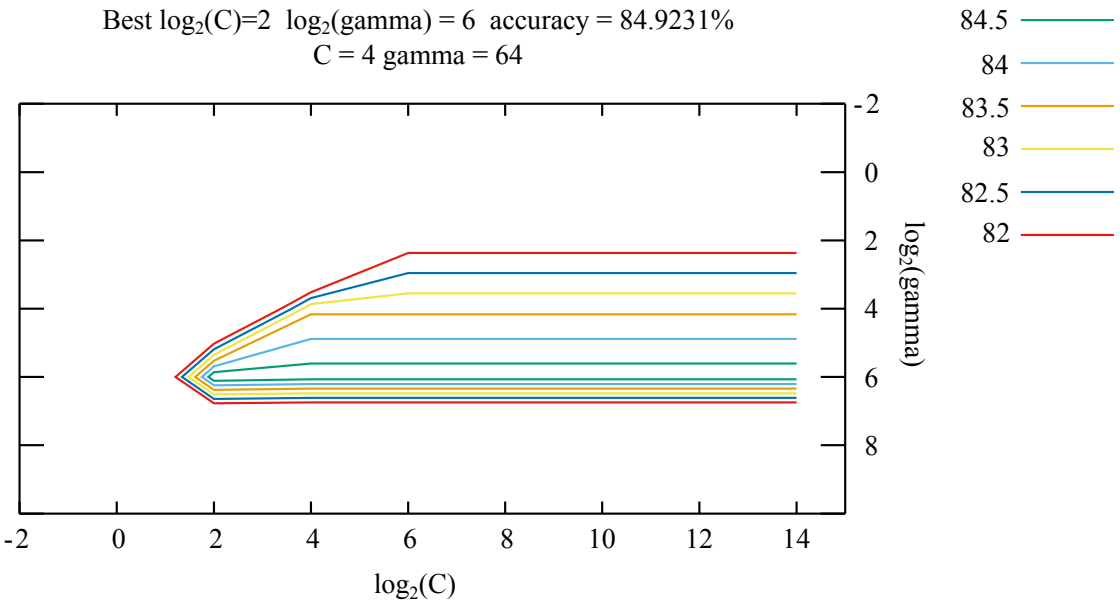


FIGURE C.1: Contour plot of grid-search results for the LS data set using two microphones.

```
 log2c=6 log2g=4 rate=83.3846     log2c=4 log2g=8 rate=77.3846     log2c=-2 log2g=4 rate=59.5385
log2c=2 log2g=4 rate=78.9231      log2c=4 log2g=0 rate=66.7692     log2c=-2 log2g=8 rate=59.3846
log2c=6 log2g=8 rate=77.3846      log2c=4 log2g=10 rate=40.7692    log2c=-2 log2g=0 rate=56.7692
log2c=2 log2g=8 rate=77.3846      log2c=6 log2g=2 rate=81.6923     log2c=-2 log2g=10 rate=22.6154
log2c=12 log2g=4 rate=83.3846     log2c=2 log2g=2 rate=67.5385     log2c=-2 log2g=2 rate=57.5385
log2c=12 log2g=8 rate=77.3846     log2c=12 log2g=2 rate=81.6923    log2c=-2 log2g=6 rate=66.7692
log2c=6 log2g=0 rate=77.6923      log2c=0 log2g=2 rate=57.8462     log2c=6 log2g=-2 rate=67.0769
log2c=2 log2g=0 rate=56.9231      log2c=10 log2g=2 rate=81.6923    log2c=2 log2g=-2 rate=56.7692
log2c=12 log2g=0 rate=81.0769     log2c=4 log2g=2 rate=77.6923     log2c=12 log2g=-2 rate=80.9231
log2c=0 log2g=4 rate=66.1538      log2c=14 log2g=4 rate=83.3846    log2c=0 log2g=-2 rate=56.7692
log2c=0 log2g=8 rate=75.6923      log2c=14 log2g=8 rate=77.3846    log2c=10 log2g=-2 rate=80.9231
log2c=0 log2g=0 rate=56.7692      log2c=14 log2g=0 rate=81.0769    log2c=4 log2g=-2 rate=56.9231
log2c=6 log2g=10 rate=40.7692     log2c=14 log2g=10 rate=40.7692   log2c=14 log2g=-2 rate=80.9231
log2c=2 log2g=10 rate=40.7692     log2c=14 log2g=2 rate=81.6923    log2c=-2 log2g=-2 rate=56.7692
log2c=12 log2g=10 rate=40.7692    log2c=6 log2g=6 rate=84.7692     log2c=8 log2g=4 rate=83.3846
log2c=0 log2g=10 rate=37.8462     log2c=2 log2g=6 rate=84.9231     log2c=8 log2g=8 rate=77.3846
log2c=10 log2g=4 rate=83.3846     log2c=12 log2g=6 rate=84.7692    log2c=8 log2g=0 rate=81.0769
log2c=10 log2g=8 rate=77.3846     log2c=0 log2g=6 rate=77.5385     log2c=8 log2g=10 rate=40.7692
log2c=10 log2g=0 rate=81.0769     log2c=10 log2g=6 rate=84.7692    log2c=8 log2g=2 rate=81.6923
log2c=10 log2g=10 rate=40.7692    log2c=4 log2g=6 rate=84.7692     log2c=8 log2g=6 rate=84.7692
log2c=4 log2g=4 rate=83.3846      log2c=14 log2g=6 rate=84.7692    log2c=8 log2g=-2 rate=77.3846
```

TABLE C.1: Grid-search optimisation log file output for the two-microphone letter classifier, showing $C$ and $\gamma$ parameter values and the percentage cross-validation accuracy ("rate") for each pair.

(cont. from left)

| Letter | Correctly Predicted (10) | Accuracy (%) |
|---|---|---|
| **A** | 7 | 70 |
| **B** | 9 | 90 |
| **C** | 9 | 90 |
| **D** | 9 | 90 |
| **E** | 10 | 100 |
| **F** | 8 | 80 |
| **G** | 9 | 90 |
| **H** | 8 | 80 |
| **I** | 10 | 100 |
| **J** | 10 | 100 |
| **K** | 8 | 80 |
| **L** | 10 | 100 |
| **M** | 10 | 100 |

(cont. right)

| Letter | Correctly Predicted (10) | Accuracy (%) |
|---|---|---|
| **N** | 10 | 100 |
| **O** | 10 | 100 |
| **P** | 8 | 80 |
| **Q** | 8 | 80 |
| **R** | 8 | 80 |
| **S** | 9 | 90 |
| **T** | 8 | 80 |
| **U** | 8 | 80 |
| **V** | 10 | 100 |
| **W** | 9 | 90 |
| **X** | 9 | 90 |
| **Y** | 9 | 90 |
| **Z** | 7 | 70 |
| **Average** | — | **88.5** |

TABLE C.2: Average accuracy per letter for the LS semi-seen testing set for two microphones.

| Letter | Precision | Recall | $f_1$ Score |
|---|---|---|---|
| **A** | 100 | 70 | 82 |
| **B** | 100 | 90 | 95 |
| **C** | 90 | 90 | 90 |
| **D** | 90 | 90 | 90 |
| **E** | 91 | 100 | 95 |
| **F** | 100 | 80 | 89 |
| **G** | 90 | 90 | 90 |
| **H** | 62 | 80 | 70 |
| **I** | 77 | 100 | 87 |
| **J** | 91 | 100 | 95 |
| **K** | 100 | 80 | 89 |
| **L** | 91 | 100 | 95 |
| **M** | 100 | 100 | 100 |
| (cont. right) | | | |

| (cont. from left) | | | |
|---|---|---|---|
| **Letter** | **Precision** | **Recall** | $f_1$ **score** |
| **N** | 100 | 100 | 100 |
| **O** | 77 | 100 | 87 |
| **P** | 89 | 80 | 84 |
| **Q** | 89 | 80 | 84 |
| **R** | 89 | 80 | 84 |
| **S** | 75 | 90 | 82 |
| **T** | 100 | 80 | 89 |
| **U** | 89 | 80 | 84 |
| **V** | 91 | 100 | 95 |
| **W** | 90 | 90 | 90 |
| **X** | 82 | 90 | 86 |
| **Y** | 90 | 90 | 90 |
| **Z** | 88 | 70 | 78 |
| **Average** | **90** | **88** | **88** |

TABLE C.3: Percentage (%) performance metrics per letter for the LS semi-seen testing set for a two microphones.

| Actual Letter | Predicted Letter | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
| **A** | **7** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |
| **B** | 0 | **9** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **F** | 0 | 0 | 0 | 0 | 0 | **8** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| **G** | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| **H** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **I** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **J** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **K** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **L** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **M** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **N** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **1** | 0 | 0 |
| **Q** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **R** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **8** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| **S** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | **1** | 0 | 0 |
| **U** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **1** | 0 | **8** | 0 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 |
| **W** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 |
| **X** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 |
| **Z** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **7** |

TABLE C.4: Confusion matrix of letter recognition results for unseen data for the LS semi-seen testing set for two microphones.

| Letter | Correctly Predicted (35) | Accuracy (%) |
|---|---|---|
| **A** | 33 | 94 |
| **B** | 20 | 57 |
| **C** | 19 | 54 |
| **D** | 25 | 71 |
| **E** | 33 | 94 |
| **F** | 31 | 89 |
| **G** | 32 | 91 |
| **H** | 34 | 97 |
| **I** | 34 | 97 |
| **J** | 26 | 74 |
| **K** | 18 | 51 |
| **L** | 28 | 80 |
| **M** | 33 | 94 |

(cont. right)

(cont. from left)

| Letter | Correctly Predicted (35) | Accuracy (%) |
|---|---|---|
| **N** | 35 | 100 |
| **O** | 33 | 94 |
| **P** | 26 | 74 |
| **Q** | 17 | 49 |
| **R** | 31 | 89 |
| **S** | 34 | 97 |
| **T** | 25 | 71 |
| **U** | 16 | 46 |
| **V** | 25 | 71 |
| **W** | 35 | 100 |
| **X** | 27 | 77 |
| **Y** | 27 | 77 |
| **Z** | 33 | 94 |
| **Average** | — | **80.2** |

TABLE C.5: Average accuracy per letter for the LS unseen testing set for two microphones.

| Letter | Precision | Recall | $f_1$ Score |
|---|---|---|---|
| **A** | 87 | 94 | 90 |
| **B** | 74 | 57 | 65 |
| **C** | 86 | 54 | 67 |
| **D** | 81 | 71 | 76 |
| **E** | 100 | 94 | 97 |
| **F** | 72 | 89 | 79 |
| **G** | 67 | 91 | 77 |
| **H** | 97 | 97 | 97 |
| **I** | 94 | 97 | 96 |
| **J** | 81 | 74 | 78 |
| **K** | 82 | 51 | 63 |
| **L** | 97 | 80 | 88 |
| **M** | 94 | 94 | 94 |

(cont. right)

(cont. from left)

| Letter | Precision | Recall | $f_1$ score |
|---|---|---|---|
| **N** | 71 | 100 | 83 |
| **O** | 75 | 94 | 84 |
| **P** | 76 | 74 | 75 |
| **Q** | 89 | 49 | 63 |
| **R** | 66 | 89 | 76 |
| **S** | 57 | 97 | 72 |
| **T** | 71 | 71 | 71 |
| **U** | 57 | 46 | 51 |
| **V** | 86 | 71 | 78 |
| **W** | 100 | 100 | 100 |
| **X** | 77 | 77 | 77 |
| **Y** | 90 | 77 | 83 |
| **Z** | 97 | 94 | 96 |
| **Average** | **82** | **80** | **80** |

TABLE C.6: Percentage (%) performance metrics per letter for the LS unseen testing set for two microphones.

| Actual Letter | Predicted Letter | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
| **A** | **33** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| **B** | 0 | **20** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | **13** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | **19** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **15** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0 | 0 | 0 | **25** | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | **3** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 0 | 0 | 0 | 0 | **33** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| **F** | 0 | 0 | 0 | 0 | 0 | **31** | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| **G** | 0 | 0 | 0 | 0 | 0 | 0 | **32** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **H** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **34** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **I** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **34** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **J** | 0 | **4** | 0 | **2** | 0 | 0 | 0 | 0 | 0 | **26** | 0 | 0 | 0 | **1** | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 |
| **K** | **2** | 0 | 0 | 0 | 0 | **11** | 0 | 0 | 0 | 0 | **18** | 0 | 0 | 0 | 0 | **2** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| **L** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **28** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | **3** | 0 | 0 | 0 | 0 |
| **M** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **33** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **N** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **35** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **33** | 0 | 0 | 0 | **1** | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **3** | 0 | **26** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 |
| **Q** | 0 | 0 | **1** | 0 | 0 | 0 | **16** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **17** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |
| **R** | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **31** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **S** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **34** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **4** | 0 | 0 | 0 | **25** | 0 | 0 | 0 | **5** | 0 | 0 |
| **U** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | **10** | 0 | **16** | 0 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **9** | **25** | 0 | 0 | 0 | 0 |
| **W** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **35** | 0 | 0 | 0 |
| **X** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | **27** | 0 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **27** | 0 |
| **Z** | **1** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **33** |

TABLE C.7: Confusion matrix of letter recognition results for unseen data for the LS unseen testing set for two microphones.

| **Subjects** | **Correctly Predicted (182)** | **Average Accuracy (%)** |
|---|---|---|
| **1** | 140 | 76.9 |
| **2** | 155 | 85.2 |
| **3** | 150 | 82.4 |
| **4** | 147 | 80.8 |
| **5** | 138 | 75.8 |
| **Average** | — | **80.2** |
| **Std. Dev.** | — | **3.9** |

TABLE C.8: Average accuracy (%) per unseen test subject for the LS unseen testing set using two microphones.

# Appendix D

# Additional Results for the Letter Recognition Experiments Using Three Microphones

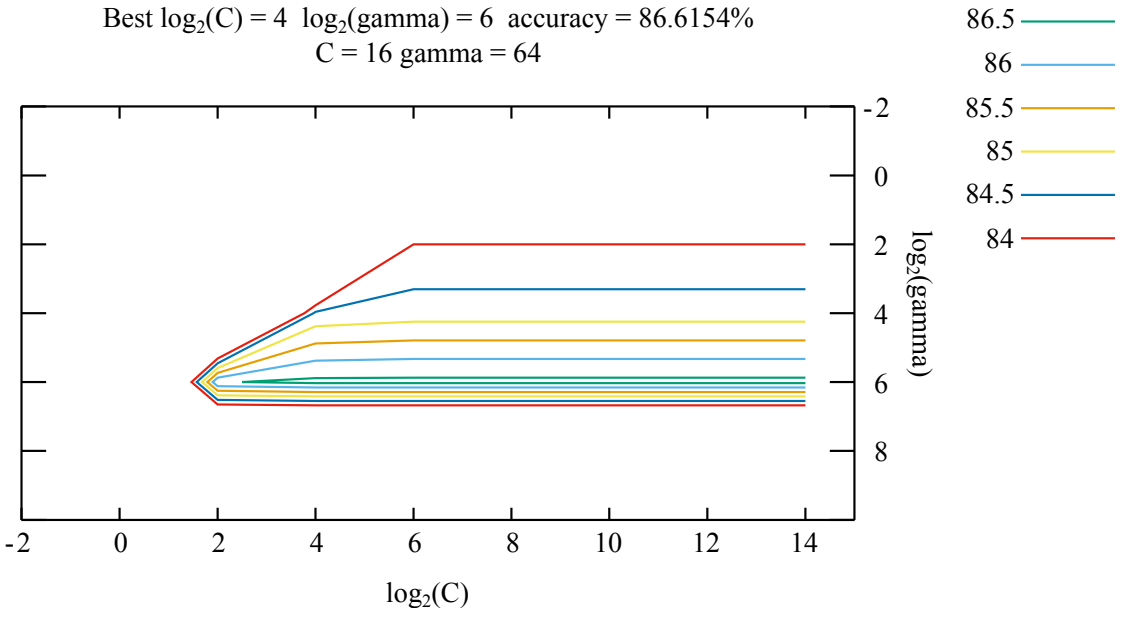


FIGURE D.1: Contour plot of grid-search results for the LS data set using three microphones.

```
 log2c=6 log2g=4 rate=84.7692     log2c=4 log2g=8 rate=78.9231     log2c=-2 log2g=4 rate=51.0769
log2c=2 log2g=4 rate=79.2308      log2c=4 log2g=0 rate=60.6154     log2c=-2 log2g=8 rate=52.7692
log2c=6 log2g=8 rate=78.9231      log2c=4 log2g=10 rate=39.2308    log2c=-2 log2g=0 rate=47.5385
log2c=2 log2g=8 rate=78.9231      log2c=6 log2g=2 rate=84.0        log2c=-2 log2g=10 rate=26.6154
log2c=12 log2g=4 rate=84.7692     log2c=2 log2g=2 rate=60.4615     log2c=-2 log2g=2 rate=48.0
log2c=12 log2g=8 rate=78.9231     log2c=12 log2g=2 rate=84.0       log2c=-2 log2g=6 rate=60.3077
log2c=6 log2g=0 rate=78.7692      log2c=0 log2g=2 rate=48.0        log2c=6 log2g=-2 rate=60.7692
log2c=2 log2g=0 rate=47.5385      log2c=10 log2g=2 rate=84.0       log2c=2 log2g=-2 rate=47.3846
log2c=12 log2g=0 rate=83.6923     log2c=4 log2g=2 rate=79.2308     log2c=12 log2g=-2 rate=83.6923
log2c=0 log2g=4 rate=59.5385      log2c=14 log2g=4 rate=84.7692    log2c=0 log2g=-2 rate=47.3846
log2c=0 log2g=8 rate=76.6154      log2c=14 log2g=8 rate=78.9231    log2c=10 log2g=-2 rate=83.2308
log2c=0 log2g=0 rate=47.5385      log2c=14 log2g=0 rate=83.6923    log2c=4 log2g=-2 rate=47.3846
log2c=6 log2g=10 rate=39.2308     log2c=14 log2g=10 rate=39.2308   log2c=14 log2g=-2 rate=83.6923
log2c=2 log2g=10 rate=39.2308     log2c=14 log2g=2 rate=84.0       log2c=-2 log2g=-2 rate=47.3846
log2c=12 log2g=10 rate=39.2308    log2c=6 log2g=6 rate=86.6154     log2c=8 log2g=4 rate=84.7692
log2c=0 log2g=10 rate=36.9231     log2c=2 log2g=6 rate=86.4615     log2c=8 log2g=8 rate=78.9231
log2c=10 log2g=4 rate=84.7692     log2c=12 log2g=6 rate=86.6154    log2c=8 log2g=0 rate=83.3846
log2c=10 log2g=8 rate=78.9231     log2c=0 log2g=6 rate=77.2308     log2c=8 log2g=10 rate=39.2308
log2c=10 log2g=0 rate=83.6923     log2c=10 log2g=6 rate=86.6154    log2c=8 log2g=2 rate=84.0
log2c=10 log2g=10 rate=39.2308    log2c=4 log2g=6 rate=86.6154     log2c=8 log2g=6 rate=86.6154
log2c=4 log2g=4 rate=84.6154      log2c=14 log2g=6 rate=86.6154    log2c=8 log2g=-2 rate=78.7692
```

TABLE D.1: Grid-search optimisation log file output for the three-microphone letter classifier, showing $C$ and $\gamma$ parameter values and the percentage cross-validation accuracy ("rate") for each pair.

| Letter | Correctly Predicted (10) | Accuracy (%) |
|---|---|---|
| **A** | 7 | 70 |
| **B** | 10 | 100 |
| **C** | 9 | 90 |
| **D** | 9 | 90 |
| **E** | 10 | 100 |
| **F** | 9 | 90 |
| **G** | 9 | 90 |
| **H** | 9 | 90 |
| **I** | 10 | 100 |
| **J** | 10 | 100 |
| **K** | 9 | 90 |
| **L** | 10 | 100 |
| **M** | 10 | 100 |
| (cont. right) | | |

(cont. from left)

| Letter | Correctly Predicted (10) | Accuracy (%) |
|---|---|---|
| **N** | 10 | 100 |
| **O** | 10 | 100 |
| **P** | 8 | 80 |
| **Q** | 10 | 100 |
| **R** | 8 | 80 |
| **S** | 9 | 90 |
| **T** | 9 | 90 |
| **U** | 9 | 90 |
| **V** | 9 | 90 |
| **W** | 9 | 90 |
| **X** | 10 | 100 |
| **Y** | 9 | 90 |
| **Z** | 7 | 70 |
| **Average** | — | **91.5** |

TABLE D.2: Average accuracy per letter for the LS semi-seen testing set for three microphones.

| Letter | Precision | Recall | $f_1$ Score |
|---|---|---|---|
| **A** | 100 | 70 | 82 |
| **B** | 100 | 100 | 100 |
| **C** | 100 | 90 | 95 |
| **D** | 100 | 90 | 95 |
| **E** | 83 | 100 | 91 |
| **F** | 100 | 90 | 95 |
| **G** | 90 | 90 | 90 |
| **H** | 69 | 90 | 78 |
| **I** | 77 | 100 | 87 |
| **J** | 100 | 100 | 100 |
| **K** | 100 | 90 | 95 |
| **L** | 91 | 100 | 95 |
| **M** | 100 | 100 | 100 |
| (cont. right) | | | |

(cont. from left)

| Letter | Precision | Recall | $f_1$ score |
|---|---|---|---|
| **N** | 100 | 100 | 100 |
| **O** | 71 | 100 | 83 |
| **P** | 89 | 80 | 84 |
| **Q** | 91 | 100 | 95 |
| **R** | 100 | 80 | 89 |
| **S** | 82 | 90 | 86 |
| **T** | 100 | 90 | 95 |
| **U** | 100 | 90 | 95 |
| **V** | 100 | 90 | 95 |
| **W** | 100 | 90 | 95 |
| **X** | 83 | 100 | 91 |
| **Y** | 90 | 90 | 90 |
| **Z** | 100 | 70 | 82 |
| **Average** | **93** | **92** | **92** |

TABLE D.3: Percentage (%) performance metrics per letter for the LS semi-seen testing set for three microphones.

| Actual Letter | Predicted Letter | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
| **A** | **7** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **B** | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **F** | 0 | 0 | 0 | 0 | 0 | **9** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **H** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **I** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **J** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **K** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **L** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **M** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **N** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| **Q** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **R** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **8** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| **S** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 | 0 |
| **U** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **9** | 0 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 | **1** | 0 | 0 |
| **W** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **9** | 0 | 0 | 0 |
| **X** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **9** | 0 |
| **Z** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **7** |

TABLE D.4: Confusion matrix of letter recognition results for unseen data for the LS semi-seen testing set for three microphones.

| Letter | Correctly Predicted (35) | Accuracy (%) |
|---|---|---|
| **A** | 31 | 89 |
| **B** | 27 | 77 |
| **C** | 22 | 63 |
| **D** | 23 | 66 |
| **E** | 33 | 94 |
| **F** | 25 | 71 |
| **G** | 31 | 89 |
| **H** | 33 | 94 |
| **I** | 34 | 97 |
| **J** | 27 | 77 |
| **K** | 19 | 54 |
| **L** | 27 | 77 |
| **M** | 33 | 94 |

(cont. right)

(cont. from left)

| Letter | Correctly Predicted (35) | Accuracy (%) |
|---|---|---|
| **N** | 35 | 100 |
| **O** | 34 | 97 |
| **P** | 27 | 77 |
| **Q** | 24 | 69 |
| **R** | 31 | 89 |
| **S** | 32 | 91 |
| **T** | 24 | 69 |
| **U** | 18 | 51 |
| **V** | 32 | 91 |
| **W** | 34 | 97 |
| **X** | 24 | 69 |
| **Y** | 26 | 74 |
| **Z** | 33 | 94 |
| **Average** | — | **81.1** |

TABLE D.5: Average accuracy per letter for the LS unseen testing set for three microphones.

| Letter | Precision | Recall | $f_1$ Score |
|---|---|---|---|
| **A** | 91 | 89 | 90 |
| **B** | 82 | 77 | 79 |
| **C** | 96 | 63 | 76 |
| **D** | 85 | 66 | 74 |
| **E** | 100 | 94 | 97 |
| **F** | 61 | 71 | 66 |
| **G** | 76 | 89 | 82 |
| **H** | 94 | 94 | 94 |
| **I** | 94 | 97 | 96 |
| **J** | 90 | 77 | 83 |
| **K** | 73 | 54 | 62 |
| **L** | 96 | 77 | 86 |
| **M** | 94 | 94 | 94 |

(cont. right)

(cont. from left)

| Letter | Precision | Recall | $f_1$ score |
|---|---|---|---|
| **N** | 69 | 100 | 81 |
| **O** | 71 | 97 | 82 |
| **P** | 71 | 77 | 74 |
| **Q** | 92 | 69 | 79 |
| **R** | 79 | 89 | 84 |
| **S** | 60 | 91 | 73 |
| **T** | 67 | 69 | 68 |
| **U** | 82 | 51 | 63 |
| **V** | 82 | 91 | 86 |
| **W** | 100 | 97 | 99 |
| **X** | 65 | 69 | 67 |
| **Y** | 81 | 74 | 78 |
| **Z** | 100 | 94 | 97 |
| **Average** | **83** | **81** | **81** |

TABLE D.6: Percentage (%) performance metrics per letter for the LS unseen testing set for three microphones.

| Actual Letter | Predicted Letter | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
| **A** | **31** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **1** | 0 |
| **B** | 0 | **27** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | **22** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0 | 0 | 0 | **23** | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 | **3** | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **E** | 0 | 0 | 0 | 0 | **33** | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **F** | **1** | 0 | 0 | 0 | 0 | **25** | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 |
| **G** | 0 | 0 | 0 | 0 | 0 | 0 | **31** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 |
| **H** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **33** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| **I** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **34** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **J** | 0 | **3** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **27** | 0 | 0 | 0 | **2** | 0 | **1** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 |
| **K** | 0 | 0 | 0 | 0 | 0 | **14** | 0 | 0 | 0 | 0 | **19** | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **L** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **27** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **5** | 0 | 0 | 0 | 0 |
| **M** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **33** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **N** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **35** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **34** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **3** | 0 | **27** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 |
| **Q** | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **24** | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| **R** | 0 | **3** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **31** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **S** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | **32** | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | **24** | 0 | 0 | 0 | **9** | 0 | 0 |
| **U** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | **7** | 0 | **18** | 0 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | **32** | 0 | 0 | 0 | 0 |
| **W** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **34** | 0 | 0 | 0 |
| **X** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **11** | 0 | 0 | 0 | **24** | 0 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **26** | 0 |
| **Z** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **1** | **33** |

TABLE D.7: Confusion matrix of letter recognition results for unseen data for the LS unseen testing set for three microphones.

| **Subjects** | **Correctly Predicted (182)** | **Average Accuracy (%)** |
|---|---|---|
| **1** | 139 | 76.4 |
| **2** | 148 | 81.3 |
| **3** | 154 | 84.6 |
| **4** | 153 | 84.1 |
| **5** | 144 | 79.1 |
| **Average** | — | **81.1** |
| **Std. Dev.** | — | **3.6** |

TABLE D.8: Average accuracy (%) per unseen test subject for the LS unseen testing set using three microphones.

# Bibliography


[1] A. Ahmad, C. Viard-Gaudin, M. Khalid, and E. Poisson, "Comparison of support vector machine and neural network in character level discriminant training for online word recognition," *UNITEN Students Conference on Research and Development, Malaysia*, 2004.

[2] A. Athanasopoulos, A. Dimou, V. Mezaris, and I. Kompatsiaris, "Gpu acceleration for support vector machines," in *Procs. 12th Inter. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011), Delft, Netherlands*, 2011, pp. 17–55.

[3] D. Bhalke, C. R. Rao, and D. S. Bormane, "Automatic musical instrument classification using fractional Fourier transform based-MFCC features and counter propagation neural network," *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 425–446, 2016.

[4] V. Bountourakis, L. Vrysis, and G. Papanikolaou, "Machine learning algorithms for environmental sound recognition: Towards soundscape semantics," in *Proceedings of the Audio Mostly 2015 on Interaction With Sound*. ACM, 2015, p. 5.

[5] A. C. Braun, U. Weidner, and S. Hinz, "Classification in high-dimensional feature spacesassessment using SVM, IVM and RVM with focus on simulated EnMAP data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 436–443, 2012.

[6] W. Burgos, "Gammatone and MFCC features in speaker recognition," Ph.D. dissertation, Florida Institute of Technology, 2014.

[7] G. Burrell and G. Morgan, *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*. Routledge, 2017.

[8] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*. IEEE, 2007, pp. 293–298.

[9] S. Chandra, G. Sharma, S. Malhotra, D. Jha, and A. P. Mittal, "Eye tracking based human computer interaction: Applications and their uses," in *2015 International Conference on Man and Machine Interfacing (MAMI)*. IEEE, 2015, pp. 1–5.

[10] M. Chen, P. Yang, S. Cao, M. Zhang, and P. Li, "Writepad: Consecutive number writing on your hand with smart acoustic sensing," *IEEE Access*, vol. 6, pp. 77 240–77 249, 2018.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[13] B. Curtis, "Introduction to empirical research on the design process in MCC's software technology program," in *Empirical Studies of the Design Process: Papers for the Second Workshop on Empirical Studies of Programmers*, 1987, pp. 1–4.

[14] P. J. Denning, D. E. Comer, D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, and P. R. Young, "Computing as a discipline," *Computer*, vol. 22, no. 2, pp. 63–70, 1989.

[15] M. Dhuliawala, J. Lee, J. Shimizu, A. Bulling, K. Kunze, T. Starner, and W. Woo, "Smooth eye movement interaction using eog glasses," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 307–311.

[16] H. Du, P. Li, H. Zhou, W. Gong, G. Luo, and P. Yang, "Wordrecorder: Accurate acoustic-based handwriting recognition using deep learning," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1448–1456.

[17] A. Giddens, *New rules of sociological method: A positive critique of interpretative sociologies*. John Wiley & Sons, 2013.

[18] J. Gill and P. Johnson, *Research methods for managers*. Sage, 2002.

[19] C. Harrison and S. E. Hudson, "Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces," in *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*. ACM, 2008, pp. 205–208.

[20] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The Elements of Statistical Learning*. Springer, 2009, pp. 485–585.

[21] M. T. Holden and P. Lynch, "Choosing the appropriate methodology: Understanding research philosophy," *The Marketing Review*, vol. 4, no. 4, pp. 397–409, 2004.

[22] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Computer Science Department, National Taiwan University, Tech. Rep., 2003.

[23] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1–2, pp. 116–134, 2007.

[25] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms," *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1311–1325, 2018.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] S. E. Kruger, M. Schaffoner, M. Katz, E. Andelic, and A. Wendemuth, "Mixture of support vector machines for hmm based speech recognition," in *18th International Conference on Pattern Recognition*. IEEE, 2006, pp. 326–329.

[28] B.-H. Kwon, T.-W. Kim, and J.-H. Youm, "A novel SVM-based hysteresis current controller," *IEEE Transactions on Power Electronics*, vol. 13, no. 2, pp. 297–307, 1998.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[30] W. S. Lee and S. S. Lee, "Piezoelectric microphone built on circular diaphragm," *Sensors and Actuators A: Physical*, vol. 144, no. 2, pp. 367–373, 2008.

[31] M.-J. D. Levers, "Philosophical paradigms, grounded theory, and perspectives on emergence," *Sage Open*, vol. 3, no. 4, p. 2158244013517243, 2013.

[32] W. Li, "Acoustic based sketch recognition," Ph.D. dissertation, Texas A & M University, 2012.

[33] W. Li and T. A. Hammond, "Recognizing text through sound alone." in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI, 2011, pp. 1481–1486.

[34] M. Lojka, M. Pleva, E. Kiktová, J. Juhár, and A. Čižmár, "Efficient acoustic detector of gunshots and glass breaking," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10 441–10 469, 2016.

[35] K. Mannepalli, P. N. Sastry, and M. Suman, "Mfcc-gmm based accent recognition system for telugu speech signals," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 87–93, 2016.

[36] X. Mao, L. Chen, and B. Zhang, "Mandarin speech emotion recognition based on a hybrid of HMM/ANN," *International Journal of Computers*, vol. 1, no. 4, pp. 321–324, 2007.

[37] J. Mercer, "Functions ofpositive and negativetypeand theircommection with the theory ofintegral equations," *Philosophicol Trinsdictions ofthe Rogyal Society*, pp. 4–415, 1909.

[38] A. Mohammed, A. Seeam, X. Bellekens, K. Nieradzinska, and V. Ramsurrun, "Gesture based iot light control for smart clothing," in *2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech)*. IEEE, 2016, pp. 139–142.

[39] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 73–76.

[40] G. Morgan and L. Smircich, "The case for qualitative research," *Academy of Management Review*, vol. 5, no. 4, pp. 491–500, 1980.

[41] M. S. S. Morton, *State of the art of research in Management Support Systems*. Center for Information Systems Research, Sloan School of Management, 1983, no. 107.

[42] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[43] G. R. Naik, D. K. Kumar, V. P. Singh, and M. Palaniswami, "Hand gestures for hci using ica of emg," in *Proceedings of the HCSNet Workshop on Use of Vision in Human-Computer Interaction*, vol. 56. Australian Computer Society, 2006, pp. 67–72.

[44] F. V. Nelwamondo, T. Marwala, and U. Mahola, "Early classifications of bearing faults using hidden Markov models, Gaussian mixture models, mel-frequency

cepstral coefficients and fractals," *International Journal of Innovative Computing, Information and Control*, vol. 2, no. 6, pp. 1281–1299, 2006.

[45] J. F. Nunamaker Jr, M. Chen, and T. D. Purdin, "Systems development in information systems research," *Journal of Management Information Systems*, vol. 7, no. 3, pp. 89–106, 1990.

[46] E. Nykaza, S. Bunkley, and M. G. Blevins, "Objectively choosing spectrogram parameters to classify environmental noises," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 253, no. 1. Institute of Noise Control Engineering, 2016, pp. 7336–7343.

[47] Y. Pan, S. S. Ge, F. R. Tang, and A. Al Mamun, "Detection of epileptic spike-wave discharges using SVM," in *IEEE International Conference on Control Applications, 2007. CCA 2007.* IEEE, 2007, pp. 467–472.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[49] L. Peikoff, *Objectivism: the philosophy of Ayn Rand.* Penguin, 1993.

[50] A. D. Pietersma, "Feature space learning in support vector machines through dual objective optimization," *Order*, vol. 501, p. 3295, 2010.

[51] C. Pittman, P. Wisniewski, C. Brooks, and J. J. LaViola Jr, "Multiwave: Doppler effect based gesture recognition in multiple dimensions," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, 2016, pp. 1729–1736.

[52] M. Schrapel, M.-L. Stadler, and M. Rohs, "Pentelligence: Combining pen tip motion and writing sounds for handwritten digit recognition," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 2018, p. 131.

[53] A. Seniuk and D. Blostein, "Pen acoustic emissions for text and gesture recognition," in *10th International Conference on Document Analysis and Recognition, 2009. ICDAR'09.* IEEE, 2009, pp. 872–876.

[54] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 4625–4628.

[55] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music." in *ISMIR*, 2006, pp. 286–289.

[56] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[57] B. Uzkent and B. D. Barkana, "Pitch-range based feature extraction for audio surveillance systems," in *Information Technology: New Generations (ITNG), 2011 Eighth International Conference on*. IEEE, 2011, pp. 476–480.

[58] B. Uzkent, B. D. Barkana, and H. Cevikalp, "Non-speech environmental sound classification using SVMs with a new set of features," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 5, pp. 3511–3524, 2012.

[59] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[60] R. H. Venkatnarayan and M. Shahzad, "Gesture recognition using ambient light," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 40, 2018.

[61] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with SOLI: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 851–860.

[62] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 2016, pp. 82–94.

[63] Q. Wu, M. Ghaziasgar, R. Dodds, and J. Connan, "Audio recognition of contact surface gestures," in *Southern Africa Telecommunication Networks and Applications Conference 2017*, 2017, pp. 260–265.

[64] Q. Wu, M. Ghaziasgar, R. M. Dodds, and J. Connan, "Robust audio-based digit recognition," in *Southern Africa Telecommunication Networks and Applications Conference 2018*, 2018, pp. 250–255.

[65] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, and Z. Zhou, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[66] Y. Xing, J. Wang, and Z. Zhao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," in *International Conference on Convergence Information Technology, 2007.* IEEE, 2007, pp. 868–872.

[67] S. Yang and Y.-p. Guan, "Audio–visual perception-based multimodal HCI," *The Journal of Engineering*, vol. 2018, no. 4, pp. 190–198, 2018.

[68] W. Yongmin, "Chinese wubi input method," 1983.

[69] T. Yu, H. Jin, and K. Nahrstedt, "Writinghacker: audio based eavesdropping of handwriting via mobile devices," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 2016, pp. 463–473.

[70] M. Zhang, P. Yang, C. Tian, L. Shi, S. Tang, and F. Xiao, "Soundwrite: Text input on surfaces through mobile acoustic sensing," in *Proceedings of the 1st International Workshop on Experiences with the Design and Implementation of Smart Objects.* ACM, 2015, pp. 13–17.

[71] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on.* IEEE, 2011, pp. 559–564.