**INDEX**

abstract i

Table of Figures ii

Tables iii

Chapter 1 Introduction 1

1.1 Objectives: 2

1.2 Thesis Outline 3

Chapter 2 Literature Review 4

2.1 Speech Processing 4

2.1.1 Human Speech Production System 4

2.1.1.1 Phonetic Representation of Speech 6

2.1.1.2 Voiced vs. Voiceless 6

2.1.1.3 Nasal vs. oral 6

2.1.1.4 Manners of Speech 7

2.1.1.5 Places of Articulation 7

2.1.2 Modeling Speech 10

2.1.2.1 Speech Features 10

2.1.2.2 Modeling features 12

2.2 Visual Speech (Viseme) Fundamentals 13

2.2.1 Introduction 13

2.2.2 Lip Region Localization 15

2.3 Classification 17

2.3.1 Classifiers 19

2.3.1.1 Neural Network 20

2.3.1.2 GMM 22

2.3.1.3 VQ 23

2.3.1.4 HMM 23

Chapter 3 Feature Extraction 26

3.1.1 Feature Extraction 26

3.1.1.1 Transform the PCM digital audio 26

3.1.1.2 Context Free Grammar 31

3.1.1.3 Adaptation 35

3.2 Video Feature Extraction 36

3.2.1 Feature Extraction Techniques 37

3.2.1.1 Shape-based Feature Extraction 37

3.2.1.2 Intensity-based Feature Extraction 39

3.2.1.3 Image Based Techniques 40

3.2.1.4 Color Based Techniques. 40

3.2.1.5 Subspace Based Techniques. 40

3.2.1.6 Model Based Techniques 41

3.2.1.7 Hybrid Techniques 41

Chapter 4 Implementation 43

4.1 Overview 43

4.2 Phoneme Extraction 44

4.3 Viseme Extraction 46

4.3.1 Introduction 46

4.3.2 Proposed Method of Lip Extraction 46

4.4 Phoneme to Viseme Mapping 52

Chapter 5 Results and Discussion 58

5.1. Introduction 58

5.2 Results 58

5.3 Discussion 65

Chapter 6 Conclusion, Limitation and Future Enhancement 66

6.1 Conclusion 66

6.2 Limitations and Future Enhancement 67

References 69

Appendix A 70

**TABLE OF FIGURES**

Figure 1: Complete physiological mechanism of speech production [2] 5

[Figure 2: Side cut-out view of places of articulation.[5] 7](#_Toc314573952)

[Figure 3: Visemes for different phonemes [5] 9](#_Toc314573953)

[Figure 4: Spectrograph of word “Generation5” [6] 11](#_Toc314573954)

[Figure 5: Spectrograph of ‘ss’ [6] 12](#_Toc314573955)

[Figure 6: Markov Model for word "tomato"[6] 13](#_Toc314573956)

[Figure 7: Visual Feature Extraction Process 15](#_Toc314573957)

[Figure 8: A Nonlinear model of a Neuron[28] 20](#_Toc314573958)

[Figure 9: Flow of Operation 44](#_Toc314573959)

[Figure 10: Mouth Region height and Width for viseme 46](#_Toc314573960)

[Figure 11: Overview of the Proposed Algorithm for Lip Feature Extraction 47](#_Toc314573961)

[Figure 12: Steps of Algorithm 50](#_Toc314573962)

[Figure 13: Results for Viseme V 50](#_Toc314573963)

[Figure 14: Results of applying various image processing techniques for lip extraction on video recorded of speaker “HARSH” 51](#_Toc314573964)

[Figure 15:Viseme Classes for Speaker “KINJAL” 51](#_Toc314573965)

[Figure 16: Neural Network Architecture 53](#_Toc314573966)

[Figure 17: The two activation function tansig(x) and logsig(x)[28] 55](#_Toc314573967)

[Figure 18: Overview of the training process of Neural Network](#_Toc300327702) 56

**TABLES**

[Table 1:Phonemes and their relative Visual Mouth Shape Details](#_Toc300327705) 08

[Table 2: Mapping from phoneme to viseme. . 1](#_Toc300327706)4

[Table 3 : List of Phonemes & Visemes Used for Digits One to Ten](#_Toc300327707) 46

[Table 4 : Audio Property for each .wav file](#_Toc300327708) 58

[Table 5: List 0f Phonemes and Associated Visemes for phoneme-viseme mapping with unique index assigned for each viseme](#_Toc300327709) 59

[Table 6 : Output of Phoneme Extraction for one to ten Digits 6](#_Toc300327710)0

[Table 7: Success of Lip Feature Extraction Algorithm 6](#_Toc300327711)1

[Table 8 : Confusion Matrix for Phoneme based Digit Recognition 6](#_Toc300327712)2

[Table 9: Results of Viseme Extracted for Speaker1](#_Toc300327713) 63