Air Force Institute of Technology
AFIT Scholar

Theses and Dissertations | Student Graduate Works

3-20-2001

Hidden Markov Models for Visual Speech Synthesis in Limited Data

Harold A. Arb

Follow this and additional works at: https://scholar.afit.edu/etd

Part of the Graphics and Human Computer Interfaces Commons

Recommended Citation

Arb, Harold A., "Hidden Markov Models for Visual Speech Synthesis in Limited Data" (2001). *Theses and Dissertations*. 4544.
https://scholar.afit.edu/etd/4544

This Dissertation is brought to you for free and open access by the Student Graduate Works at AFIT Scholar. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AFIT Scholar. For more information, please contact richard.mansfield@afit.edu.

HIDDEN MARKOV MODELS FOR VISUAL SPEECH SYNTHESIS IN LIMITED DATA ENVIRONMENTS

DISSERTATION

Harold Allan Arb, Captain, USAF

AFIT/DS/ENG/01-01

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

20010411 142

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the U.S. Air Force, Department of Defense or the United States Government.

AFIT/DS/ENG/01-01

HIDDEN MARKOV MODELS FOR VISUAL SPEECH SYNTHESIS IN LIMITED DATA ENVIRONMENTS

DISSERTATION

Presented to the Faculty of the Graduate School of Engineering and Management
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Harold Allan Arb, B.S.E.E.,M.S.E.E.
Captain, USAF

March 20, 2001

Approved for public release; distribution unlimited

AFIT/DS/ENG/01-01

HIDDEN MARKOV MODELS FOR VISUAL SPEECH SYNTHESIS IN LIMITED DATA ENVIRONMENTS

Harold Allan Arb, B.S.E.E.,M.S.E.E.

Captain, USAF

Approved:

Steven C. Gustafson (Chairman)	5 Mar 01 Date
Timothy R. Anderson (Member)	28 Feb 01 Date
Martin P. DeSimio (Member)	February 28 2001 Date
Gary B. Lamont (Member)	28 FEB 01 Date
Mark E. Oxley (Member)	28 Feb 01 Date
Mark N. Goltz (Dean's Representative)	5 Mar 01 Date

Robert A. Calico, Jr.
Dean

Acknowledgements

I first need to thank Dr. Marty DeSimio, Dr. Ray Slyh and Dr. Tim Anderson for believing in me enough to recommend me for the Ph.D. program at AFIT. Second, a special thanks goes out to Dr. Steve Gustafson for stepping in as my committee chair and taking the risk of a project he was unfamiliar with. Additionally, without the work of the remainder of the committee (Dr. Oxley, Dr. Lamont and the Dean's representative Dr. Goltz), this document would not be possible. A dissertation is not simply the outcome of a single person's work, but the result of influences from many people. This work is the result of help from my committee, sponsor, peers, friends, and family.

Finally, I wish to dedicate this dissertation to my late grandparents, Allan and Mildred Williston. It was their inspiration, guidance, and teaching throughout my life that helped me attain the abilities and personal drive to accomplish this work. I love you both. Thank you for all you have given me. Our dreams are realized.

Harold Allan Arb

Table of Contents

Page

Acknowledgements iii

List of Figures viii

List of Tables xiv

Abstract xvi

I. Introduction 1

1.1 Background 1

1.2 Problem Statement and Scope 3

1.2.1 Problem Statement 3

1.2.2 Scope 3

1.3 Research Contributions 5

1.4 Dissertation Composition 5

II. Literature Review 6

2.1 Why Visual Speech 6

2.2 Bimodal Speech Recognition 7

2.2.1 Audio/Visual Databases 8

2.3 Types of Visual Speech Synthesis 9

2.3.1 Character Animation or "Talking Heads" 9

2.3.2 Real Objects (Humans, Animals, etc.) 12

2.4 Summary 15

Page

III. Methodology 16

3.1 Introduction 16

3.2 Video Database Recording 16

3.2.1 Selection of Prompts 16

3.2.2 Video Capture 18

3.3 Processing of Video Files 18

3.3.1 Separating the audio from the video 18

3.3.2 Phonetic Labeling of Audio 19

3.3.3 Visemic Labeling 20

3.4 Control Point Estimation 20

3.5 Distance Features 29

3.5.1 Static Features 29

3.5.2 Dynamic Features 30

3.5.3 Upsampling 30

3.6 Hidden Markov Model Training 30

3.6.1 Initialize Viseme Models 31

3.6.2 Clone Viseme Models into Triseme Models 32

3.6.3 Decision Tree-Based State Tying 32

3.7 Synthesis of Control Points from HMMs 34

3.8 Summary 37

IV. Experimental Results 39

4.1 Introduction 39

4.2 Experimental Design 39

4.2.1 Data Organization 39

4.2.2 Process Variables 41

4.2.3 Structure of Experiment 42

4.3 Analysis of Individual Triseme Results 43

4.3.1 Individual Feature Analysis by Triseme 44

Page

4.3.2 Average Feature Analysis by Triseme 58
4.4 Analysis of results with blocking 59
4.4.1 Individual feature analysis 59
4.4.2 Average feature analysis with blocking 64
4.5 Mouthshapes 65
4.5.1 The Six "Analysis" Trisemes 67
4.5.2 Additional Examples 85
4.6 Chapter Summary 89

V. Conclusions and Recommendations 101
5.1 Conclusions 101
5.2 Recommendations for Future Research 102
5.3 Contributions 103
5.4 To the Future 103

Appendix A. Chosen Prompts from the TIMIT Speech Corpus 104

Appendix B. Decision Tree Questions 118

Appendix C. Normal Probability Plots, Main Effect Plots, and Interaction Plots for Individual Triseme Analysis 122
C.1 Triseme L-E+F 122
C.2 Triseme K-D+C 132
C.3 Triseme K-E+G 143
C.4 Triseme A-U+C 159
C.5 Triseme F-K+D 171
C.6 Triseme R-P+G 179

Appendix D. ANOVA Tables, Main Effect Plots and Interaction Plots for Individual Triseme Analysis with Blocking 189

Bibliography 211

Page

Vita 218

List of Figures

Figure		Page
1.	Video Rewrite Process	2
2.	Video Rewrite process with HMM synthesis	4
3.	The Institut de la Communication Parlee (ICP) facial model	10
4.	Talking Head Analysis	11
5.	Flowchart of all tasks accomplished in the research presented here.	17
6.	Example of the video captured for the database.	19
7.	A cropped image example	21
8.	Example of the location of the control points around the mouth area.	22
9.	Log plot of Singular Values in Eigenpoints computation	24
10.	Poor Eigenpoints Estimation Example 1	27
11.	Poor Eigenpoints Estimation Example 2	27
12.	Poor Eigenpoints Estimation Example 3	28
13.	Good Eigenpoints Estimation Example	28
14.	Distances derived from control points	29
15.	An example of decision tree-based state tying.	33
16.	HMM Type Examples	42
17.	An example of a normal probability plot of the data in Table 6	47
18.	Main effect and interaction plots for measurement data used to determine effects in Table 6.	49
19.	Normal probability plots for each triseme with the MSE totaled for each frame in the triseme.	60
20.	Main effect and interaction plots for each triseme using average MSE per feature per triseme.	61
21.	Blocking analysis plots in averaged feature analysis.	66
22.	L-E+F,F-K+D $T = 20, R = 10, \Theta = 1$, No Skip HMM	69
23.	K-E+G $T = 20, R = 10, \Theta = 1$, No Skip HMM.	69

Figure Page

24. K-D+C $T = 20, R = 10, \Theta = 1$, No Skip HMM. 70
25. A-U+C $T = 20, R = 10, \Theta = 1$, No Skip HMM. 71
26. R-P+G $T = 20, R = 10, \Theta = 1$, No Skip HMM. 72
27. L-E+F,F-K+D $T = 100, R = 10, \Theta = 1$, No Skip HMM 73
28. K-E+G $T = 100, R = 10, \Theta = 1$, No Skip HMM. 73
29. K-D+C $T = 100, R = 10, \Theta = 1$, No Skip HMM. 74
30. A-U+C $T = 100, R = 10, \Theta = 1$, No Skip HMM. 75
31. R-P+G $T = 100, R = 10, \Theta = 1$, No Skip HMM. 76
32. L-E+F,F-K+D $T = 20, R = 20, \Theta = 1$, No Skip HMM 77
33. K-E+G $T = 20, R = 20, \Theta = 1$, No Skip HMM. 77
34. K-D+C $T = 20, R = 20, \Theta = 1$, No Skip HMM. 78
35. A-U+C $T = 20, R = 20, \Theta = 1$, No Skip HMM. 79
36. R-P+G $T = 20, R = 20, \Theta = 1$, No Skip HMM. 80
37. L-E+F,F-K+D $T = 100, R = 20, \Theta = 1$, No Skip HMM 81
38. K-E+G $T = 100, R = 20, \Theta = 1$, No Skip HMM. 81
39. K-D+C $T = 100, R = 20, \Theta = 1$, No Skip HMM. 82
40. A-U+C $T = 100, R = 20, \Theta = 1$, No Skip HMM. 83
41. R-P+G $T = 100, R = 20, \Theta = 1$, No Skip HMM. 84
42. K-D+C Magnified and at Original Scale. 86
43. K-E+G Magnified and at Original Scale. 87
44. F-K+D Magnified and at Original Scale. 88
45. A-O+G Magnified and at Original Scale. 90
46. B-D+V Magnified. 91
47. B-D+V Original Scale. 92
48. B-W+A Magnified and at Original Scale. 93
49. C-D+X Magnified. 94
50. C-D+X at Original Scale. 95
51. E-W+B Magnified and at Original Scale. 96

Figure		Page
52.	I-S+G Magnified and at Original Scale.	97
53.	O-G+A Magnified.	98
54.	O-G+A at Original Scale.	99
55.	W-A+C Magnified and at Original Scale.	100
56.	Normal probability plots of effects for triseme L-E+F features 1-4. . .	122
57.	Normal probability plots of effects for triseme L-E+F features 5-10. . .	123
58.	Normal probability plots of effects for triseme L-E+F features 11-16. .	124
59.	Normal probability plots of effects for triseme L-E+F features 17-22. .	125
60.	Normal probability plots of effects for triseme L-E+F features 23-28. .	126
61.	Main effect plots for factor C in triseme L-E+F for first 9 features where the effect of C was significant.	127
62.	Main effect plots for factor C in triseme L-E+F for last 9 features where the effect of C was significant.	128
63.	Main effect plots for factor B in triseme L-E+F for features 3, 6, 11 and 23 where the effect of B was significant.	129
64.	Interaction plots for factors B and C in triseme L-E+F for features 3, 5, 7, 9, 23 and 28 where BC interaction was significant.	130
65.	Main effect and interaction plots for factor D, BD, CD and BCD for feature 28 in triseme L-E+F.	131
66.	Normal probability plots of effects for triseme K-D+C features 1-4. . .	132
67.	Normal probability plots of effects for triseme K-D+C features 5-10. .	133
68.	Normal probability plots of effects for triseme K-D+C features 11-16. .	134
69.	Normal probability plots of effects for triseme K-D+C features 17-22. .	135
70.	Normal probability plots of effects for triseme K-D+C features 23-28. .	136
71.	Main effect plots for factor C in triseme K-D+C for features 1 and 3-9.	137
72.	Main effect plots for factor C in triseme K-D+C for features 11, 13, 14 and 16-20.	138
73.	Main effect plots for factor C in triseme K-D+C for features 23-28. . .	139
74.	Main effects plots for factor A in features 13 and 14 of triseme K-D+C.	139
75.	Main effect plots for factor B in features 9 and 16 of triseme K-D+C. .	140

Figure		Page
76.	Main effect plot for factor D in feature 27 of triseme K-D+C.	140
77.	Plots of for the two factor interactions: BC (the 3 left plots), AB (top right), AC (middle right) and CD (bottom right) for triseme K-D+C.	141
78.	Three factor interaction plots for ABC (left) and BCD (right) interactions in features 12 and 5 respectively of triseme K-D+C.	142
79.	Normal probability plots of effects for triseme K-E+G features 1-4. . .	143
80.	Normal probability plots of effects for triseme K-E+G features 5-10. .	144
81.	Normal probability plots of effects for triseme K-E+G features 11-16. .	145
82.	Normal probability plots of effects for triseme K-E+G features 17-22. .	146
83.	Normal probability plots of effects for triseme K-E+G features 23-28. .	147
84.	Main effect plots for factor C in triseme K-E+G for features 1, 4-8 and 10 and 11.	148
85.	Main effect plots for factor C in triseme K-E+G for features 12-19. . .	149
86.	Main effect plots for factor C in triseme K-E+G for features 20-27. . .	150
87.	Main effect plot for factor C in triseme K-E+G for feature 28.	151
88.	Plots for BC interaction in triseme K-E+G for features 1, and 4-6. . .	151
89.	Plots for BC interaction in triseme K-E+G for features 7,8 and 10-15.	152
90.	Plots for BC interaction in triseme K-E+G for features 16-23.	153
91.	Plots for BC interaction in triseme K-E+G for features 24-28.	154
92.	Main effect plots for factor B in triseme K-E+G for features 1, 4-7, 13, 17 and 18.	155
93.	Main effect plots for factor B in triseme K-E+G for features 20, 21 and 23.	156
94.	Main effect plot for factor A in triseme K-E+G for feature 13.	156
95.	Main effect plots for factor D in triseme K-E+G for features 7 and 21.	157
96.	Plots of the AB interaction in triseme K-E+G for features 21 and 28. .	158
97.	Plots of the AC interaction in feature 13 (a) and ABC interaction in triseme K-E+G for feature 28 (b).	158
98.	Normal probability plots of effects for triseme A-U+C features 1-4. . .	159
99.	Normal probability plots of effects for triseme A-U+C features 5-10. .	160

Figure		Page
100.	Normal probability plots of effects for triseme A-U+C features 11-16. .	161
101.	Normal probability plots of effects for triseme A-U+C features 17-22. .	162
102.	Normal probability plots of effects for triseme A-U+C features 23-28. .	163
103.	Main effect plots for factor C in triseme A-U+C for features 1-3, 5, 6, 8, 11 and 12.	164
104.	Main effect plots for factor C in triseme A-U+C for features 13-19 and 21.	165
105.	Main effect plots for factor C in triseme A-U+C for features 23, 24 and 27.	166
106.	Main effect plots for factor D in triseme A-U+C for features 1, 2, 7, 10 and 18.	167
107.	Main effect plots for factor B in triseme A-U+C for features 10, 14, 20 and 28.	168
108.	Plots of the BC interaction in triseme A-U+C for features 10, 14, 20 and 28.	168
109.	Plots of the BD and BCD interactions in triseme A-U+C for feature 10.	169
110.	Plots of the CD interaction in triseme A-U+C for features 14, 18 and 20.	170
111.	Normal probability plots of effects for triseme F-K+D features 1-4. . .	171
112.	Normal probability plots of effects for triseme F-K+D features 5-10. .	172
113.	Normal probability plots of effects for triseme F-K+D features 11-16. .	173
114.	Normal probability plots of effects for triseme F-K+D features 17-22. .	174
115.	Normal probability plots of effects for triseme F-K+D features 23-28. .	175
116.	Main effect plots for factor C in triseme F-K+D for features 2, 3, 9, 23, 24, 27 and 28.	176
117.	Main effect plots for factors A and D and interaction plot for the CD interaction in triseme F-K+D feature 27.	177
118.	Main effect plot for factor B in triseme F-K+D for feature 10.	178
119.	Normal probability plots of effects for triseme R-P+G features 1-4. . .	179
120.	Normal probability plots of effects for triseme R-P+G features 5-10. .	180

Figure		Page
121.	Normal probability plots of effects for triseme R-P+G features 11-16. .	181
122.	Normal probability plots of effects for triseme R-P+G features 17-22. .	182
123.	Normal probability plots of effects for triseme R-P+G features 23-28. .	183
124.	Main effect plots for factor C in triseme R-P+G for features 1, 5, 7-9, 18 and 19.	184
125.	Main effect plots for factor C in triseme R-P+G for features 20, 21 and 25-28.	185
126.	Main effects plots for factors A in features 1, 2 and 18 (a) and B in features 21, 25 and 26 (b) for triseme R-P+G.	186
127.	Interaction plots for triseme R-P+G for the BC interaction in feature 5 (top left), AC interactions in features 18 and 20 (left center and lower left), and AB interactions in features 14-16 (right top to bottom). . . .	187
128.	Interaction plots for triseme R-P+G for the ABC interaction in feature 25 (top), and the BD and CD interactions in feature 26 (middle/bottom).	188
129.	Main effect plots for factor B in features 6, 9, 10, 11, 13, 16-18 using block analysis	204
130.	Main effect plots for factor B in features 20, 21 and 23 using block analysis	205
131.	Main effect plots for factor C in features 1, 4-8, 10 and 11 using block analysis	206
132.	Main effect plots for factor C in features 12-16, 18, 21 and 22 using block analysis	207
133.	Main effect plots for factor C in features 23 and 25-28 using block analysis	208
134.	BC Interaction plots for features 9, 14, 18, 23-25, 27 and 28 using block analysis	209
135.	CD interaction plot for feature 28 using block analysis	210
136.	ACD and ABCD interaction plots for feature 2 using block analysis . .	210

List of Tables

Table		Page
1.	Phoneme to Viseme Groupings	18
2.	Trisemes occurring in the test sentence sx212	40
3.	Factors Evaluated in the ANOVA	42
4.	Experimental run labels for a 2^4 factorial experiment	47
5.	Contrast multiplicative constants from expansion of Equation 42	48
6.	Example of assigning normal probability for plotting to identify significant effects	48
7.	Summary of Results from Individual Triseme Analysis	57
8.	Analysis of Variance for a Four-Factor Factorial Design in a Randomized Complete Block	62
9.	ANOVA Summary for Pooled Feature Analysis.	65
10.	ANOVA Summary for Feature 1 Analysis	190
11.	ANOVA Summary for Feature 2 Analysis	190
12.	ANOVA Summary for Feature 3 Analysis	191
13.	ANOVA Summary for Feature 4 Analysis	191
14.	ANOVA Summary for Feature 5 Analysis	192
15.	ANOVA Summary for Feature 6 Analysis	192
16.	ANOVA Summary for Feature 7 Analysis	193
17.	ANOVA Summary for Feature 8 Analysis	193
18.	ANOVA Summary for Feature 9 Analysis	194
19.	ANOVA Summary for Feature 10 Analysis	194
20.	ANOVA Summary for Feature 11 Analysis	195
21.	ANOVA Summary for Feature 12 Analysis	195
22.	ANOVA Summary for Feature 13 Analysis	196
23.	ANOVA Summary for Feature 14 Analysis	196
24.	ANOVA Sumary for Feature 15 Analysis	197

Table		Page
25.	ANOVA Summary for Feature 16 Analysis	197
26.	ANOVA Summary for Feature 17 Analysis	198
27.	ANOVA Summary for Feature 18 Analysis	198
28.	ANOVA Summary for Feature 19 Analysis	199
29.	ANOVA Summary for Feature 20 Analysis	199
30.	ANOVA Summary for Feature 21 Analysis	200
31.	ANOVA Summary for Feature 22 Analysis	200
32.	ANOVA Summary for Feature 23 Analysis	201
33.	ANOVA Summary for Feature 24 Analysis	201
34.	ANOVA Summary for Feature 25 Analysis	202
35.	ANOVA Summary for Feature 26 Analysis	202
36.	ANOVA Summary for Feature 27 Analysis	203
37.	ANOVA Summary for Feature 28 Analysis	203

Abstract

This work presents a new approach for estimating control points (facial locations that control movement) to allow the artificial generation of video with apparent mouth movement (visual speech) time-synced with recorded audio. First, Hidden Markov Models (HMMs) are estimated for each visual speech category (viseme) present in stored video data, where a category is defined as the mouth movement corresponding to a given sound and where the visemes are further categorized as trisemes (a viseme in the context of previous and following visemes). Next, a decision tree is used to cluster and relate states in the HMMs that are similar in a contextual and statistical sense. The tree is also used to estimate HMMs that generate sequences of visual speech control points for trisemes not occurring in the stored data.

An experiment is described that evaluates the effect of several algorithm variables, and a statistical analysis is presented that establishes appropriate levels for each variable by minimizing the error between the desired and estimated control points. The analysis indicates that the error is lowest when the process is conducted with three-state left-to-right no skip HMMs trained using short-duration dynamic features, a high log-likelihood threshold, and a low outlier threshold. Also, comparisons of mouth shapes generated from the artificial control points and the true control points (estimated from video not used to train the HMMs) indicate that the process provides accurate estimates for most trisemes tested in this work. The research presented here thus establishes a useful method for synthesizing realistic audio-synchronized video facial features.

HIDDEN MARKOV MODELS FOR VISUAL SPEECH SYNTHESIS IN LIMITED DATA ENVIRONMENTS

I. Introduction

1.1 Background

It is well known that humans use more than auditory cues to perceive speech when additional cues are available. For example, deaf people can be taught to read lips (i.e., visual speech reading). Further, researchers have found that people with normal hearing use a visual mode of perception to enhance aural perception, especially in noisy environments.

Researchers have long been intrigued with the idea of generating a realistic video of a human or human-like character that speaks. Early attempts include the animated character Max Headroom and the television series based on it. A major problem in the synthesis of visual speech is ensuring that the mouth area moves in a realistic fashion according to the text of the spoken words. McGurk found in the 1970's that human perception of speech is significantly influenced by visual clues. The effect (known as the McGurk effect) manifests itself as a misperceived syllable induced by mouth motions that do not visually match the expected motion [44]. To ensure the absence of the McGurk effect in audio/visual speech, it is imperative that the movements of the mouth match the motion expected for the spoken sound.

Interval Research Corporation was one of several companies exploring the improvement of visual speech synthesis. Researchers at Interval developed a system to synthesize video of a human lip-syncing to a recorded audio track. Figure 1 illustrates their Video Rewrite process. The desired audio is analyzed to determine the appropriate visual categories required. Image sequences are selected from a database of recorded video that correspond to those categories or to a category similar to what is desired. Using those stored image sequences and image processing techniques, new video is generated that should be in sync with the desired audio.

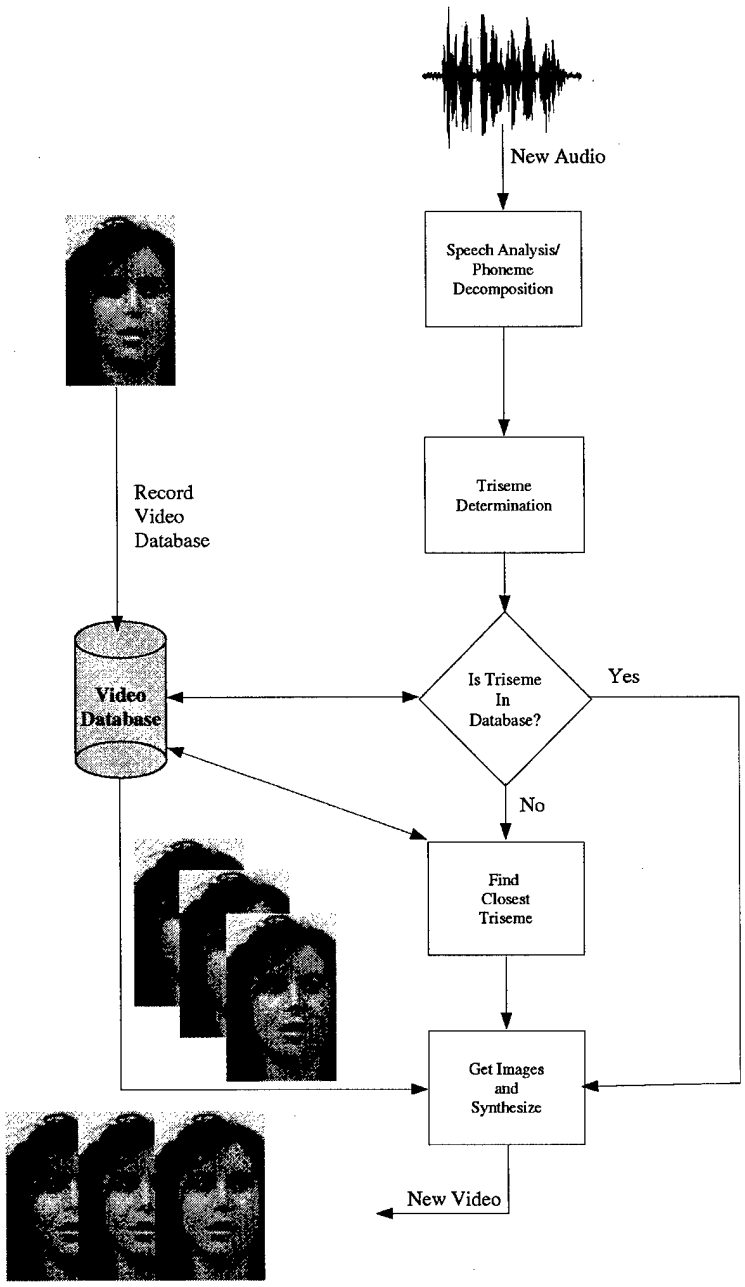


Figure 1. General process for audio-driven visual speech synthesis in Video Rewrite.

The Air Force is interested in research related to Video Rewrite for various applications, including low-bandwidth video teleconferencing. Using Video Rewrite to artificially generate the lip-synced video on reception would eliminate the need to transmit image data for every frame, thus lowering the amount bandwidth required. The problem with the Video Rewrite system is the selection of trisemes, a sequence of images corresponding to a visually distinct speech unit coupled with its context within the utterance, when the desired sequence is not available in a previously recorded and stored video database. The Air Force's application of Video Rewrite in teleconferencing will likely involve only having a short segment of recorded imagery, potentially less than three minutes. Bregler and his colleagues at Interval Research Corporation used eight minutes of training imagery [7]. The probability of encountering a desired triseme is much greater in eight minutes of training imagery than in three minutes.

Bregler's world wide web site (http://graphics.stanford.edu/ bregler/videorewrite/) shows some examples of the output from their Video Rewrite system. In some of these examples, there are instances where either the wrong triseme was selected or the blending process introduced artifacts. Incorrect triseme selection could be due to the particular distance metric chosen, or the closest triseme could be simply an unrealistic rendering of the desired mouth motion. This research addresses this problem and presents a solution. Figure 2 illustrates how the Video Rewrite system could be modified to incorporate this new solution. In lieu of using a distance metric to select a triseme "close" to the desired triseme when the triseme is not available in the video database, Hidden Markov Models trained for the desired triseme generate control point estimates for use in the processing of the new video.

1.2 Problem Statement and Scope

1.2.1 Problem Statement. This dissertation presents and evaluates a method of synchronizing video with new audio tracks using Hidden Markov Models (HMMs) as viseme synthesis parameter generators for sequences not contained in the available recorded imagery.

1.2.2 Scope. This research is limited in scope to the synthetic estimation of control points for use in generating new image sequences not available in training imagery, where

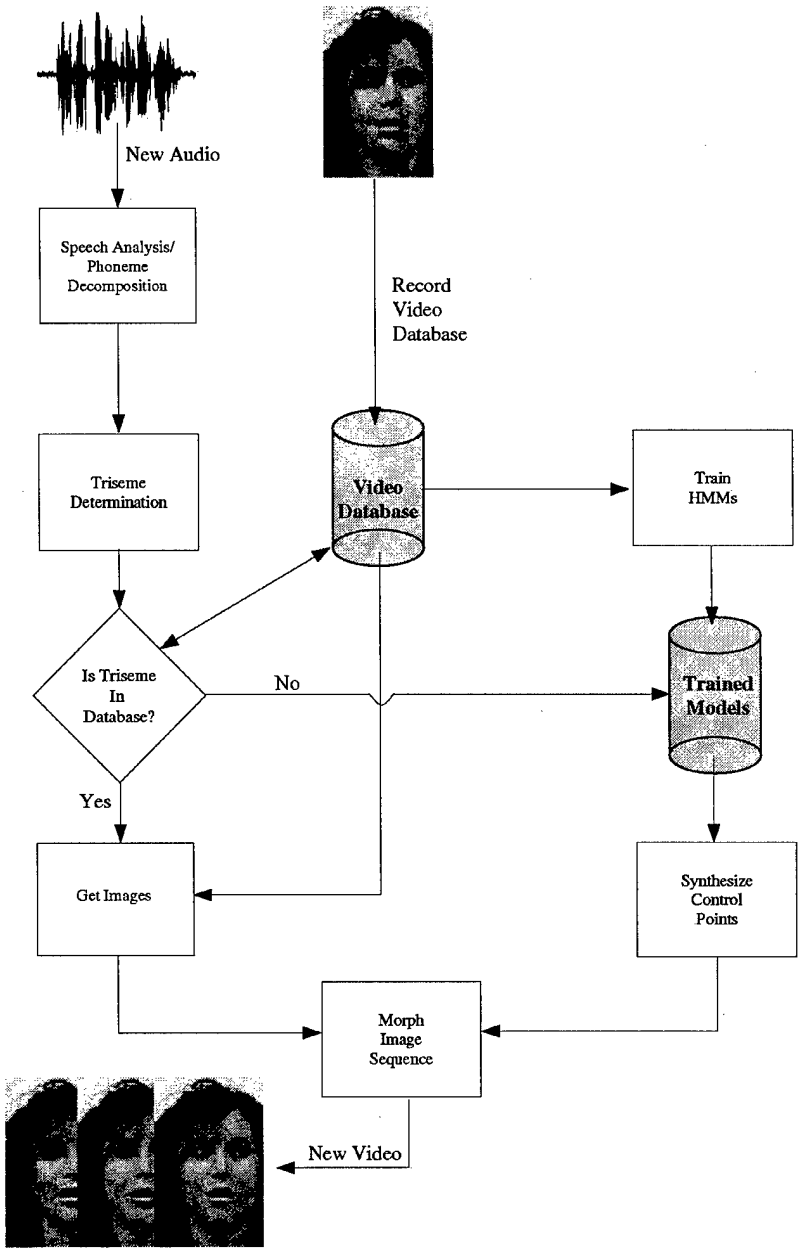


Figure 2. General process for audio-driven visual speech synthesis in Video Rewrite with HMM synthesis of missing triseme control points.

mouth shape sequences which lip-sync to the original audio are synthesized. Additionally, this research focuses only on synthesizing parameters for the mouth area and not for any other areas such as the chin, nose, eyes, etc.

1.3 Research Contributions

HMMs have not previously been applied to generate the synthesis parameters for English trisemes missing from training data. They have, however, been used to generate triphone (a phoneme (the basic unit of speech) in context with the previous and following phoneme) synthesis parameters in text-to-speech synthesis (acoustic only) and to generate sequences of synthetic mouth shapes corresponding to Japanese syllables (all of which were included in data used to train the HMMs). The use of decision tree-based HMMs in estimating visual speech synthesis control points is a new application. It is used in various acoustic speech recognition and synthesis applications, but has never been used for visual speech synthesis.

Additionally, no statistical analysis exists on the effects of parameters used to estimate new sequences using HMMs. This research presents an analysis of the effects of variable parameters in the estimation of new visual speech synthesis features using HMMs.

1.4 Dissertation Composition

This dissertation is organized into five chapters. Chapter II gives historical background on visual speech analysis and synthesis. Chapter III presents a description of video analysis/decomposition, HMM training algorithms, and parameter synthesis using trained HMMs. Chapter IV focuses on analysis of the experimental results. Finally, Chapter V summarizes the research, reviews new contributions, and provides recommendations for further research.

II. Literature Review

2.1 Why Visual Speech

From the talking head/character "Max Headroom" to the present day "Taco Bell Chihuahua", applications requiring the addition of realistic mouth movements to the audio in motion pictures have become too numerous to count. Also, visual speech synthesis can be used for dubbing of movies into other languages, teleconferencing, and special effects in movies [7]. Finally, artificial and real faces lip syncing to either real or synthetic speech can be used to enhance human-computer interactions [4, 10–12, 16, 22, 23, 33, 34, 57, 90].

In the mid-1970's, McGurk and MacDonald from the Department of Psychology at the University of Surrey in the UK discovered that the perception of speech is significantly influenced by visual cues [44]. For example, they found that people presented with a film of a talking head making lip movements of the syllable [ga], with audio of the syllable [ba] dubbed in, perceived the syllable [da]. Their research showed that the perception of speech can be drastically impacted by the presence of the visual mode of speech. This effect has been termed the "McGurk Effect". While McGurk and MacDonald's research focused mainly on English listeners and the English language, Sekiyama and Tohkura [74] demonstrated that Japanese subjects hearing Japanese syllables do not experience this effect to a significant degree. Therefore, at least for English visual speech, it is very important that the correct mouth movements are used.

Research by Cohen, Massaro, and others since the early 1980's [12, 15, 16, 38, 39, 44, 46, 55, 56, 67, 74, 77] has shown that humans, especially those with normal hearing, have difficulty recognizing spoken language in face-to-face conversations when the visual speech does not match the acoustic speech [15, 16, 38, 39]. This effect further illustrates the need for accurate temporal and spatial synchronization between visual and acoustic speech representations.

One recent example illustrating problems in asynchrony between mouth movements and audio is the re-release of the Italian film "La Vida È Bella (Life is Beautiful)" dubbed in English by Miramax Films [46]. This movie in its original version (in Italian with subtitles) won seven Academy Awards and was extremely popular in the United States and North

America. However, the release of the English dubbed version in August, 1999 was less successful and was criticized by reviewers. Michael San Filippo, a movie critic for "About.com", stated in the following about the premier in New York City [72] :

> Even excepting for the tracking problems inherent in outdoor audio projection, the always busy New York traffic, and the sound bouncing around the skyscrapers, watching and listening to *Life Is Beautiful* in English was a disorienting, disappointing experience. The superbly hilarious moment in which Guido intentionally screws up in translating orders from a German soldier was an incoherent, linguistic muddle. Many of the audience were baffled by an Italian speaker translating German into English while the lips were out of sync as in a cartoon. [Roberto] Benigni's wonderful play on words while posing as a waiter were a classic in the original language but a dud in the dubbed version.

The following sections describe research on the idea that the two modes of speech contribute to accurate recognition of language. This idea has applications in bimodal speech recognition and visual speech synthesis.

2.2 Bimodal Speech Recognition

Given that people can be taught to read lips in order to understand language [26, 58], it is reasonable to make use of technology to train a computer to do the same task.

In 1984, Petajan recognized the potential offered by automatic lipreading for enhancing automatic speech recognition (ASR), especially in the presence of noise [55, 56]. His experiments illustrated a 13% improvement in word recognition accuracy ($65\% - 78\%$), a 2% improvement in letter recognition ($64\% - 66\%$), and a 5% improvement in digit recognition ($95\% - 100\%$) by bimodal ASR [55, 56].

Improvements have been made over the years since Petajan's initial research. Petajan continues to look for various ways to improve Audio/Visual ASR [8, 10, 11, 27, 56, 57, 90]. In 1993, Silsbee used newer computer technology to improve Petajan's original lipreading system, generating LEAPS (Lipreading to Enhance Automatic Perception of Speech) [77]. His work involved a more detailed testing of the system with more than a single speaker (as Petajan had used). He also incorporated more image processing steps in LEAPS that

further enhanced the robustness of the system by allowing for varied lighting conditions and multiple speakers.

Other researchers have focused on varied processing in the bimodal ASR systems. Prasad, Stork, and Wolff [62] focused on preprocessing the video images and using neural networks for the lipreading classification. Yuhas, Goldstein, and Sejnowski [89] also investigated neural networks in ASR. However, they used the networks to assist in estimating the vocal tract transfer function in an attempt to "denoise" the spectrum prior to classification. Reveret and Benoit [66] considered a viseme-based approach to the video portion of the bimodal ASR system. As another approach to automatic lipreading, Bregler and Konig [5] investigated "Eigenlips" as the basis for their system. More recently, Dupont and Leuttin [21] used features similar to those considered in this research for synthesis. Their features were fused with acoustic features using multi-stream Hidden Markov Models.

2.2.1 Audio/Visual Databases. Few extensive audio-visual databases exist for visual speech synthesis testing or for bimodal ASR. The AT&T corporation has an Audio-Visual (A/V) database of 4 parts. Part 1 contains a small vocabulary of highly confusable mostly monosyllabic consonant-vowel-consonant words; Part 2 has connected letters; Part 3 phonetically balanced North American Business sentences; and Part 4 spontaneously spoken sentences [61]. Parts 3 and 4 of the AT&T database would work well for this research, but are not yet publicly available.

Matthews, *et al.* [43], mention recording an A/V database of four talkers (two male and two female) repeating each of the letters A to Z three times. However, this database is not yet publicly available.

Finally, a small A/V database (Tulips) consisting of 12 subjects' mouth areas uttering the first four digits (1-4) is described by Movellan [49]. The Tulips database contains only a small number of trisemes and generally does not contain enough data. Due to the fact that the databases most appropriate for this research are not publicly available, an A/V database with complete sentences was collected for this research.

2.3 Types of Visual Speech Synthesis

There are essentially two broad classes of visual speech synthesis: character animation and synthesis of real objects (humans, animals, etc) lip-syncing to audio.

2.3.1 Character Animation or "Talking Heads". One type of synthetic visual speech is the animation of computer generated images. These images can take the form of human faces (also referred to as "talking heads"), cartoon animals, etc.

Since this research deals with the second type of visual speech synthesis (real human faces), only a brief discussion is given here of "talking head" or character animation. There is a plethora of reading material in computer graphics and computer vision literature on various aspects of character animation (for a good summary, see [54]). Many of the articles referenced in this section, however, are useful in understanding the dynamics of facial analysis and applying them to realistic synthesis of human visual speech.

While there are various methods for actually generating the image sequences of talking heads, most begin with the same fundamental principle – the use of an active shape model similar to a wire frame which can be manipulated using control parameters. Figure 3 from [33] illustrates this concept. The location of each vertex of the model can be controlled to change the shape of the image. Computer graphics techniques are used to generate the "skin" over the frame to synthesize the new image. The original model of the whole face was first developed by Parke [52, 53] in 1974. Cohen and Massaro [15, 16] later made improvements to the model for speech production.

Using these wire-frame models, various approaches have been taken to control the motion of the frame. In Le Goff and Benoit's [33] text-to-audiovisual speech synthesizer for French, eight parameters are used to control the motion. These parameters are selected from twenty or so target mouth shapes that best describe the structure of the jaw and lips in French. The synthesizer then calculates trajectories between the various mouth shapes using mathematical rules. Le Goff and Benoit [34] used a five parameter model to perform real-time analysis of incoming video and synthesis of talking head visual speech.

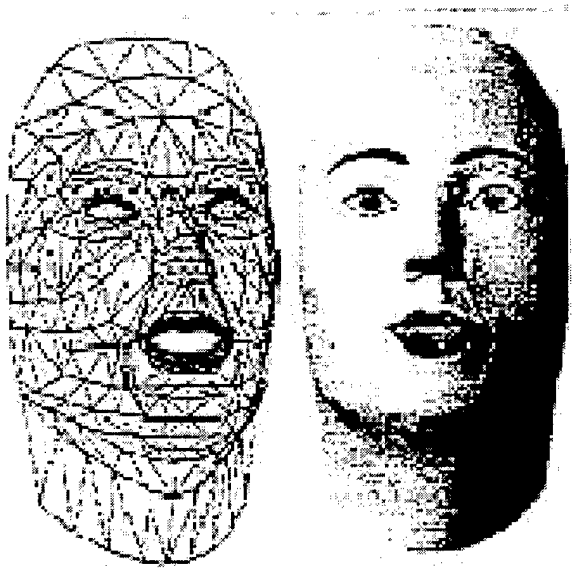


Figure 3. The Institut de la Communication Parlee (ICP) facial model. Left: The wire-framed structure underneath the face. Right: A shaded view after graphics processing [33].

While using the models developed by Parke and enhanced by Cohen and Massaro, Petajan and Graf [57] use a different method for generating the parameters for the talking head control. They identify features on the face in the incoming video such as the nostril location, eye location, and mouth details. These features are then used to change the talking head's wire-frame shape and thus the image of the artificial human head. Figure 4 illustrates this concept. While it is difficult to see in the image, the pupil and nostril locations are in the left image and are used for the talking head's rotation parameters. An outline of the inner mouth contour is used to generate the mouth parameters.

Chen, Graf, and others have used this technique of driving animated characters with human head motion-tracking to reduce the amount of data needed for transmission of images in video telephony. In [9–12], analysis of the speech information includes interpolating the talking head synthesis parameters at the decoder and generating new talking head frames between those transmitted. This analysis effectively increases the frame rate and helps to keep the lips in sync with the audio.

Figure 4. The image on the left is analyzed to generate the parameters driving the talking head on the right (from [57]).

More recently, Cosatto and Graf [17] describe using image samples as a basis for mouth motion in text-to-speech (TTS) talking-head animation. In their application, a TTS system provides the audio track and bitmaps of facial parts from recorded video of real people are projected and blended onto an image of the whole head using pose information.

While HMM's are used in [9–12] to generate the talking head synthesis parameters, Lavagetto *et al.* [32] use an Active Shape Model for the articulatory parameters. These parameters are found using principal component analysis similar to that described in the next chapter (i.e., "Eigenpoints"). Once the visual parameters in the training data are extracted, Time Delay Neural Networks (TDNNs) are trained for each parameter using the normalized cepstral coefficients obtained from the acoustic signal.

Recently, Pixar animation studios released its "RenderMan" software kit [59]. RenderMan can be used to generate a multitude of computer animations. The package was used in feature films such as: Disney's "A Bug's Life", "Toy Story 2" and "Inspector Gadget"; LucasFilm's "Star Wars Episode 1: The Phantom Menace"; and many others. This software, while obviously very capable at generating fictitious images and effects, is not suited for the syncing of mouth movements in real human video to audio.

The importance of coupling analysis of the audio signal with the video for talking head animation is apparent. The next section discusses the use of acoustic characteristics to generate synthetic images of real human heads with lip and face movements synchronized to the audio signal.

2.3.2 Real Objects (Humans, Animals, etc.) While animated characters can be useful tools for human-computer interaction, there are several applications (movies, television, video teleconferencing, etc.) where real objects need to lip sync to audio. Real objects in this context refer to humans and animals. Recent television commercials (Taco Bell, etc.) and movies ("Babe") illustrate animals lip syncing to English speech. Movies such as "Forrest Gump" and others where dubbing is required show the need for realistic lip sync to audio tracks other than those originally recorded (*e.g.*, the scenes of President Nixon talking to Forrest in "Forrest Gump"). Video teleconferencing is becoming more popular as technology improves. This section discusses two techniques capable of achieving the A/V synthesis required by these applications.

The first technique is Example Based Image Analysis and Synthesis (EBIAS) [4,22,23] developed at M.I.T.'s Media Lab. In 1993, Beymer, Shashua and Poggio [4] presented their idea of using images representative of various extreme conditions in the picture. For instance, an image of a face viewed from the front might be one example/reference image while another might be the same head rotated to the right. These example images form what might be considered the extrema on axes in some "example space".

M.I.T.'s work on EBIAS considers each pixel in the image as a feature. Correspondences, or mappings, are calculated from each pixel in a source image to new locations in a target image. These correspondences take the form of optical flow vectors. Once the pixel level correspondences from each example image to every other example image are formed, the EBIAS algorithms can generate any image within the example space using morphing algorithms [1].

Ezzat [22,23] has continued the work of Beymer, *et al.*, and applied it to generating realistic talking faces (real faces, not animations). He has recently presented M.I.T.'s new

text-to-visual speech synthesizer, MikeTalk [24,25]. This system essentially uses each viseme in the video corpus as a separate example and builds the example space as described above for EBIAS. It then determines correspondences between each of the visemes. MikeTalk determines which viseme it needs and the transitions to use based on synthetic speech generated with a text-to-speech synthesizer. The phoneme durations determine the speed at which the morphing occurs. The result is a realistic video of a human face lip syncing to synthetic speech.

Another approach to visual speech synthesis of human faces (or other animals) is to use a set of control points which determine the movement from one image to another. This set of control points is smaller in number than the features in EBIAS (i.e., in EBIAS every pixel is a control point). Scott, *et al.* [73], at the Jet Propulsion Laboratory have developed a visual speech synthesis system called Actors^{TM}. This system uses a recorded audio signal as the output speech. It uses this signal to generate the visemes and timing information to be used. The Actors system stores mouth shapes for each viseme only at the point in the viseme where the shape is at its extreme (i.e., fully open, fully closed, etc.). The system then synthesizes visemes using morphing algorithms that use control points around key areas of the head (e.g., the eyes, mouth, outline of the head, etc). Movement of the control points within a viseme or from one viseme to another is interpolated linearly. These control points however, must be hand marked on each image. The result is a video of a human head lip syncing to real speech. The primary drawback to Actors is the hand-labeling required for the control points.

One major problem with the previous two systems (MikeTalk and Actors), is failure to handle the effects of coarticulation (the motion of the mouth and production of sound in anticipation of the next sound to be produced) in synthetic visual speech. Cohen and Massaro [15] have studied the effects of coarticulation in visual speech in detail and point out that poor modeling of this effect substantially degrades the "realism" of the synthetic visual speech. Ezzat [25] acknowledges that the current MikeTalk system does not account for coarticulation effects and recognizes that future research should look into handling coarticulation. The next system described does attempt to compensate for coarticulation effects.

Bregler, Covell, and Slaney [7] of Interval Research Corporation have developed a system very similar to Actors: Video Rewrite. Unlike Actors, Video Rewrite only requires manual control point labeling of a select few images. Control points on the remaining images are estimated automatically using Covell's [18] algorithm, Eigenpoints, a principal components analysis algorithm. Additionally, it differs from Actors and MikeTalk in that instead of synthesizing on a viseme basis, it segments the training video into trisemes (context dependent visemes) and synthesizes at the triseme level. Video Rewrite overlaps and blends each viseme in successive trisemes. The use of trisemes as a synthesis unit helps account for coarticulation effects. Trisemes are selected to match the phonetic decomposition of the source audio. If the desired triseme is not found in the reference video database, the system attempts to find a sequence that approximates the desired transitions and mouth shapes. To accomplish this search, Video Rewrite computes an distance metric to each triseme in the data base:

$$d = \alpha D_p + (1 - \alpha) D_s, \tag{1}$$

where D_p is a phoneme-context distance based on categorical distances between phoneme categories and viseme classes, D_s measures how closely the mouth contours match in overlapping segments of the adjacent trisemes, and α $(0 \leq \alpha \leq 1)$ is a proportionality constant that can be adjusted to give more weight to either of the distance parameters.

Once the trisemes have been selected, time aligned, and blended, the resulting mouth sequences are "stitched" onto the mouth area of the face in the background video. The result is a fairly smooth (assuming the needed trisemes were available in the stored database) video of the person in the background video lip syncing to a new audio track. That audio track (as in Actors) need not be of the same person in the video. Examples of Video Rewrite output can be found at http://graphics.stanford.edu/ bregler/videorewrite/.

The use of HMMs to synthesize acoustic speech or acoustic speech synthesis parameters is considered in [20,40,41,80]. Donovan and Woodland's system [20] uses contextual decision trees to generate models for acoustic triphones not included in the HMM training data. HMMs have also been used for text-to-visual speech synthesis by Masuko and associates [42]

in a Japanese syllable synthesis system. However, their work did not include synthesis of visual speech segments unseen in the training data.

This research combines the two ideas of HMM visual speech synthesis parameter estimation and contextual decision trees for estimating HMMs for segments not included in the training data.

2.4 Summary

The need for accurate synchronization between mouth motion and acoustic speech is apparent from the literature. While researchers continue to work on potential solutions, there is much room for improvement. Many existing solutions require a vast amount of data to ensure video realism. The next chapter presents the use of HMMs to generate visual speech synthesis parameters when the amount of data are limited.

III. Methodology

3.1 Introduction

This chapter details the methods used in this research. Figure 5 serves as a broad outline of the chapter. First, Sections 3.2 and 3.3 describe the recording of an audio/visual (A/V) database and the processing of the recorded A/V files into separate images, audio, and acoustic transcription files. Section 3.4 covers the estimation of control points around the mouth area in each image. After control point estimation, the points are converted to corresponding features used in training HMMs. Section 3.5 details the conversion process. Section 3.6 describes the HMM training which begins after the feature calculation. Finally, Section 3.7 presents the process for synthesizing features from the trained HMMs for use in morphing synthetic sequences for trisemes that were not present in the recorded database.

3.2 Video Database Recording

3.2.1 Selection of Prompts. Prompts (*i.e.*, the sentences/text spoken) for the video database were chosen from the prompts in the DARPA TIMIT speech corpus [19]. The selection of prompts for use in the video database began with a search through the TIMIT list to acquire a subset containing a minimal amount of trisemes occurring at a chosen threshold count. The target amount of trisemes was a minimum of 100 occurring a minimum of 25 times in the data.

There are 2,342 prompts in the full TIMIT corpus. Each prompt was first decomposed into its phonetic transcription. Table 1 illustrates the phoneme-to-viseme mapping for converting the phonetic transcription to a visemic transcription. In this table, "phoneme(s)" refers to the basic units of English speech and "viseme" is the group label of the mouth movement corresponding to the appropriate phoneme(s). Next, the order of the prompts was randomized and a search through the visemic transcription continued until 100 distinct trisemes were found a minimum of 25 times. This process was repeated 10 times and the subset with the smallest number of sentences was chosen. The chosen subset contained 405 prompts. Appendix A lists the text for the 405 chosen prompts.

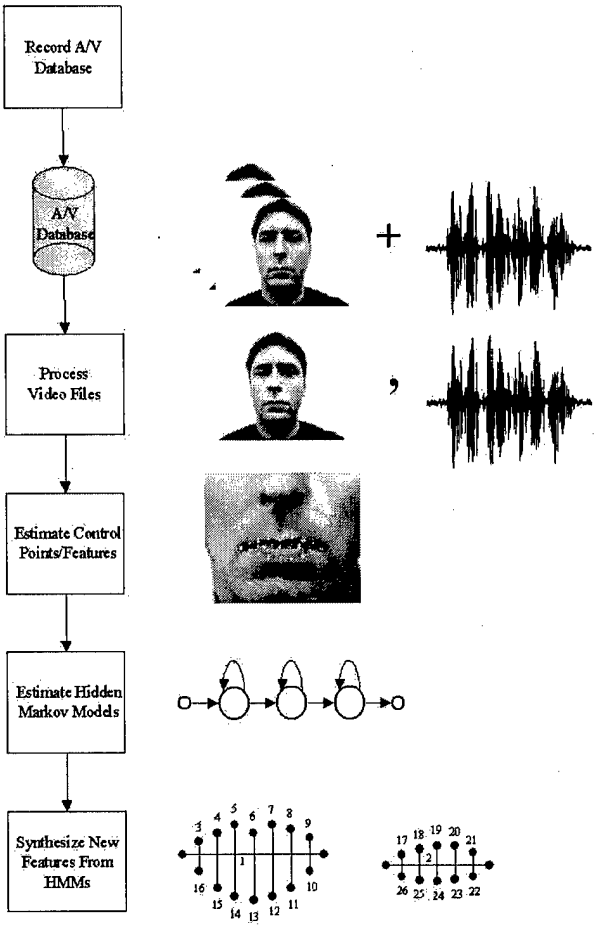


Figure 5. Flowchart of all tasks accomplished in the research presented here.

Table 1. Phoneme to Viseme Groupings

Viseme	Phoneme(s)		Viseme	Phoneme(s)
A	/ch/, /jh/, /sh/, /zh/		B	/k/, /g/, /n/, /l/
C	/t/, /d/, /s/, /z/		D	/p/, /b/, /m/
E	/f/, /v/		F	/th/, /dh/
G	/w/, /r/		H	/hh/
I	/y/		J	/ng/
K	/eh/		L	/ey/
M	/er/		N	/uh/
O	/aa/		P	/ao/
Q	/aw/		R	/ay/
S	/uw/		T	/ow/
U	/oy/		V	/iy/
W	/ih/		X	/ae/
Y	/ah/		Z	/sil/

3.2.2 Video Capture. Recording was accomplished in three sessions in January and February 1999. All sessions were done in the same location with the same setup. A 3Com PCI video camera captured the 240x320 pixel 24 bit color video at 30 frames per second with monaural audio sampled at 22 kHz with 16 bit quantization. The subject (the researcher) was seated approximately 19 inches from the camera and the entire head and shoulders were captured. Figure 6 illustrates the view captured by the camera. Captured video was stored in uncompressed Microsoft AVI format to allow future processing.

As recording and processing of the video progressed, it became evident that the process was extremely slow and time was a factor. Therefore only the first 97 prompts listed in Appendix A were recorded.

3.3 Processing of Video Files

3.3.1 Separating the audio from the video. To separate the audio from the video, two freeware applications were used. The first, "Videdit", a Microsoft product, is a GUI based tool for editing video. Using it, the audio portion of each AVI file was copied to the Windows clipboard and pasted into the Windows "Sound Recorder" allowing the audio to

Figure 6. Example of the video captured for the database.

be saved as a 22 kHz Microsoft WAV file. Videdit also allowed the deinterleaving of audio from video frames in preparation for the next step of the process.

The second application, "AVIRIP", is part of the freeware package "convmpg3" which is widely available on the internet (ftp://x2ftp.oulu.fi/pub/msdos/programming/convert for example). AVIRIP converts the input full frame, uncompressed AVI file into a sequence of 24 bit color Tagged Image Format (TIF) images. These TIF images are then converted to JPEG format for file size considerations and to allow easy input into MATLAB .

3.3.2 Phonetic Labeling of Audio. There were two steps in phonetically labeling the audio. First, the audio was downsampled to 16kHz using the application "CoolEdit" manufactured by Syntrillium and available at http://www.syntrillium.com. The phonetic labeling program used in this research employed speech models based on 16 kHz audio which required downsampling from 22 kHz.

The program "Aligner", from Entropic Research Labs produced a phonetic transcript of the 16 kHz audio files. The transcripts contained a list of the phonemes present in the audio and their start and stop times.

3.3.3 Visemic Labeling. Once Aligner produced the phonetic transcripts, the phoneme labels were converted to the proper viseme class as listed in Table 1.

3.4 Control Point Estimation

To maintain continuity with the Video Rewrite process, this research used the Eigenpoints algorithm described by Covell [18] to initially estimate the mouth control point locations on the images. This section describes the algorithm in more detail, explains what specific steps of the process this research accomplished, and presents examples of the algorithm performance in this research and remedies taken for errors.

The eigenpoints algorithm estimates multiple locations simultaneously by "exploiting" the inter-dependence between these locations. This algorithm first locates a "feature" of interest, say, the lips. The location of this feature defines the subimage and image-plane origin for all subsequent processing. Next, control points are placed around the feature by, for example, marking the locations on the image that signify the outer part of the lips. The point locations are estimated using an affine manifold which represents the relationship between the grayscale values within the feature and the control point locations associated with the feature. The assumptions made by this model are that there is a single k-dimensional vector x which drives both the vector of grayscale values and control point locations and that the transformation functions are affine.

Once the feature of interest is located, each image used to estimate the model is reshaped into a vector (*i.e.*, each the matrix of pixel intensity values is reshaped columnwise to form a single vector). Each of these vectors is then augmented with the manually estimated control point coordinates. A singular value decomposition of the matrix whose columns are these augmented vectors of pixel intensities and control point coordinates provides the model for estimating control points in additional images.

Since this research focuses only on the mouth area, it is the only "feature" of interest. Therefore, each image was cropped to a 60x90 pixel area with the mouth at the center. With the assumption that the face remains in the same image plane throughout a sentence, the location of the center of the mouth was hand annotated in the initial image in each sentence

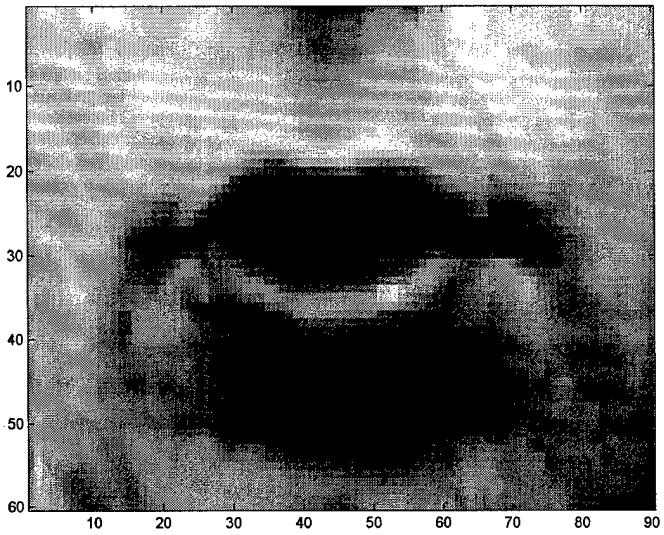


Figure 7. An example cropped image from the audio/visual database. The image is frame 33 of utterance si458.

and was used as the center of the crop box for the appropriate sentence. Figure 7 shows an example of the cropped area.

The images were converted to 256 level grayscale and processed so that the histogram of the grayscale intensities matched a target "normalized" histogram. The target histogram was roughly Gaussian in shape with a mean at 128 and a standard deviation of 32.

In Covell [18], the images used in the training set are hand-labeled with 235 control points: 56 outlining the head, face, and ears, 29 around each of the eyes, irises and eyebrows; 31 around the nose and nostrils; and 90 around the boundaries of the lips, teeth, gums, and "smile lines". This research uses 30 control points around the mouth area as illustrated in Figure 8, since the scope of the research is the mouth area only.

The labeled control points consist of one each at the outer and inner mouth corners, seven each on the upper and lower outer lip contours, five each on the upper and lower inner mouth contours, and one each for the lower edge of the upper teeth and the upper edge of the lower teeth, respectively, for a total of 30 control points. Including additional (more

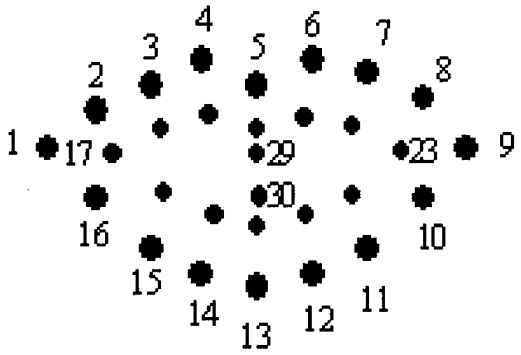


Figure 8. Example of the location of the control points around the mouth area.

than the 30) control points around either the inner or outer contour results in very crowded, difficult to estimate and/or locate control points.

This research estimates the eigenpoints model using 200 images taken from five utterances (si462, si480, sx8, sx312, sx423) so that examples from each recording session are included. The set is also designed so all 26 visemes are covered. The original images are 240×320 pixels. Since only the control points on the mouth are needed, the k-means clustering in Covell [18] is not required. The mouth area is cropped to 60×90 pixels as described previously. The control points are hand labeled using a mouse to click on the corresponding locations. The training set of 200 was expanded to 400 using the mirrors of each of the 200 original images, where a mirror is a horizontal inversion about the median.

The model for the grayscale values and control-point locations is defined in [18] as

$$\begin{aligned} \mathbf{f} &= \mathbf{M}_f\mathbf{x} + \overline{\mathbf{f}} + \mathbf{n}_f \\ \mathbf{p} &= \mathbf{M}_p\mathbf{x} + \overline{\mathbf{p}} + \mathbf{n}_p \end{aligned} \tag{2}$$

where $\mathbf{f}$ is the vector of the grayscale values in the subimage, $\mathbf{p}$ is the vector of the control-point coordinates relative to the subimage origin, $\overline{\mathbf{f}}$ and $\overline{\mathbf{p}}$ are the mean grayscale values and control-point locations, and $\mathbf{x}$ is a vector of zero-mean, independent identically distributed (iid) random variables that drive both appearance and shape. Also, $\mathbf{n}_f$ and $\mathbf{n}_p$ are iid random

variables that represent additive noise, and $\mathbf{M}_f$ and $\mathbf{M}_p$ are manifolds. The two equations are jointly analyzed to give a coupled manifold model.

First, the grayscale values are projected onto the coupled manifold to give an estimate for $\mathbf{x}$. This estimated $\mathbf{x}$ is then re-projected into the control-point subspace to give estimates of the control-point locations.

Once the control points in the training set have been hand-labeled, the coupled manifold model must be trained, that is, the parameters describing the model must be estimated.

First, each image, i, is reshaped into a $N_x \cdot N_y$-length vector, $\mathbf{f}_i$ of grayscale image values, and $\overline{\mathbf{f}}$ is estimated as the mean of these vectors. Then a matrix $\mathbf{F}$ is formed by making each vector, $\mathbf{f}_i$ a column of $\mathbf{F}$. Similarly, the (x, y) control point coordinate vectors for each image are analyzed to find $\overline{\mathbf{p}}$, and the vectors form the matrix $\mathbf{P}$.

Covell's algorithm forms the augmented matrix $\begin{bmatrix} \mathbf{F} \\ \mathbf{P} \end{bmatrix}$ from the two matrices $\mathbf{F}$ and $\mathbf{P}$. Singular Value Decomposition (SVD) is performed on this augmented matrix,

$$\begin{bmatrix} \mathbf{F} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{U_F} \\ \mathbf{U_p} \end{bmatrix} \mid \mathbf{U}_\perp \begin{bmatrix} \mathbf{\Sigma}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_\perp \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [\mathbf{V} \mid \mathbf{V}_\perp]^T \tag{3}$$

where the first K components of the SVD (*i.e.*, the components that correspond to the K largest singular values) are considered significant, and the rest are considered noise dimensions. The matrix Σ_K contains the K most significant singular values on its main diagonal and zeros off the main diagonal. Figure 9 shows a logarithmic plot of the singular values computed by the SVD of the $\begin{bmatrix} \mathbf{F} \\ \mathbf{P} \end{bmatrix}$ formed from the hand-labeled training images. A value for K of 50 was chosen since that is approximately at the knee of the curve. Thus, 50 singular values and their singular vectors are retained for the model and used in estimating control points for new images as described next.

Assume that the noise is uncorrelated with $\mathbf{x}$ and that the covariance of the noise term in the coupled manifold model is $\sigma_{cn}^2\mathbf{I}$, then

$$\mathbf{M}_f = \mathbf{U_F}(\mathbf{\Sigma}_K^2 - \sigma_{cn}^2\mathbf{I})^{\frac{1}{2}} \tag{4}$$

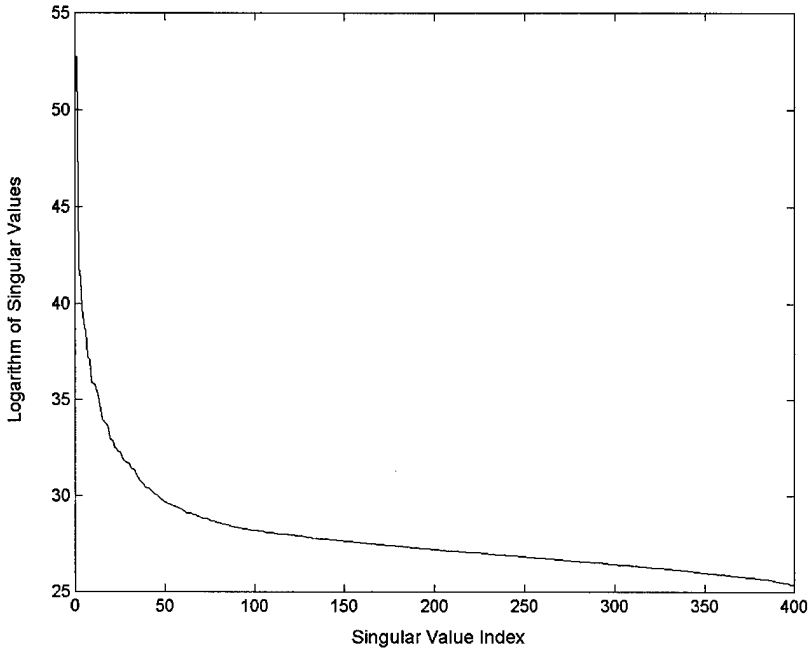


Figure 9. Logarithmic plot of the singular values calculated by the SVD process.

and

$$\mathbf{M}_p = \mathbf{U_P}(\mathbf{\Sigma}_K^2 - \sigma_{cn}^2\mathbf{I})^{\frac{1}{2}} \tag{5}$$

to within a unitary transform.

If it is assumed that the training data conforms to the affine manifold model, that they are noiseless ($\sigma_{cn}^2 = 0$), and that the image on which control points are to be located are noise free, then using Equation (2) with Equations (4) and (5) yields

$$\hat{\mathbf{x}} = \mathbf{\Sigma_K^{-1}}\mathbf{U_F^{-1}}(\mathbf{f} - \overline{\mathbf{f}}) \tag{6}$$

and the estimator for the control point locations is

$$\hat{\mathbf{p}} = \mathbf{U_P}\mathbf{\Sigma_K}\hat{\mathbf{x}} + \overline{\mathbf{p}} = \mathbf{U_P}\mathbf{U_F^{-1}}(\mathbf{f} - \overline{\mathbf{f}}) + \overline{\mathbf{p}}. \tag{7}$$

The notation above is the same as in Covell [18], however, the matrices $\mathbf{U_F}$ and $\mathbf{U_P}$ are not necessarily square, so this research used the pseudo-inverse.

Covell [18] mentions that under these assumptions, the above estimators suffer from computational noise. To alleviate this problem, Covell makes use of the property

$$\mathbf{U_F^T U_F} + \mathbf{U_P^T Y_P} = \mathbf{I} \tag{8}$$

and a C-S decomposition [28] of the two matrices

$$\begin{aligned} \mathbf{U_F} &= [\mathbf{Q_F} \mid \mathbf{Q_{F\perp}}] \begin{bmatrix} \mathbf{\Sigma_F} \\ \mathbf{0} \end{bmatrix} \mathbf{V_{FP}^T} \\ \mathbf{U_P} &= [\mathbf{Q_P} \mid \mathbf{Q_{P\perp}}] \begin{bmatrix} \mathbf{\Sigma_P} \\ \mathbf{0} \end{bmatrix} \mathbf{V_{FP}^T} \end{aligned} \tag{9}$$

Making the appropriate substitutions, Equation (7) becomes

$$\hat{\mathbf{p}} = \mathbf{Q_P}(\mathbf{\Sigma_P \Sigma_F^{-1}})\mathbf{Q_F^T}(\mathbf{f} - \overline{\mathbf{f}}) + \overline{\mathbf{p}}. \tag{10}$$

If the noise covariance is not a scaled identity matrix as was the case above, then a bound to the variance of the additive noise in the estimate of x is $\sigma_{cn}^2(\mathbf{\Sigma_K}^2 - \sigma_{cn}^2\mathbf{I})^{-1}$. Using this constraint, the minimum mean-squared-estimate (MMSE) of x is

$$\hat{\mathbf{x}} = \mathbf{\Sigma_K}^{-2}(\mathbf{\Sigma_K}^2 - \sigma_{cn}^2\mathbf{I})\mathbf{U_F}^{-1}(\mathbf{f} - \overline{\mathbf{f}}). \tag{11}$$

Resubstituting into the control point location estimate equation yields

$$\hat{\mathbf{p}} = \mathbf{Q_P}(\mathbf{\Sigma_P V_{FP}^T \Sigma_K^{-2}}(\sigma_{\mathbf{K}}^2 - \sigma_{cn}^2\mathbf{I})\mathbf{V_{FP}\Sigma_F^{-1}})\mathbf{Q_F^T}(\mathbf{f} - \overline{\mathbf{f}}) + \overline{\mathbf{p}}. \tag{12}$$

Finally, to account for noise in the new image data (the image presented to the Eigenpoints system), the system adjusts the estimate for x by the signal-to-signal plus noise ratio. The signal variance in the feature image is represented by the training data in the matrix

$\mathbf{FF^T} - \sigma_{cn}^2\mathbf{I}$. The noise variance in the new image data is referred to in the following equations as the covariance matrix $\mathbf{R_{fn}}$.

If $\mathbf{N} = (\mathbf{Q_F}(\mathbf{FF^T} - \sigma_{cn}^2\mathbf{I})\mathbf{Q_F})^{-1}(\mathbf{Q_F^T}\mathbf{R_{fn}}\mathbf{Q_F})$, then the feature-image signal-to-signal-plus-noise ratio is

$$(\mathbf{Q_F^T} - \sigma_{cn}^2\mathbf{I})(\mathbf{Q_F} + \mathbf{Q_F^T}\mathbf{R_{fn}}\mathbf{Q_F})^{-1}\mathbf{Q_F^T}(\mathbf{FF^T} - \sigma_{cn}^2\mathbf{I})\mathbf{Q_F} \tag{13}$$

$$= (\mathbf{I} + \mathbf{N})^{-1},$$

which changes the control point location estimate to

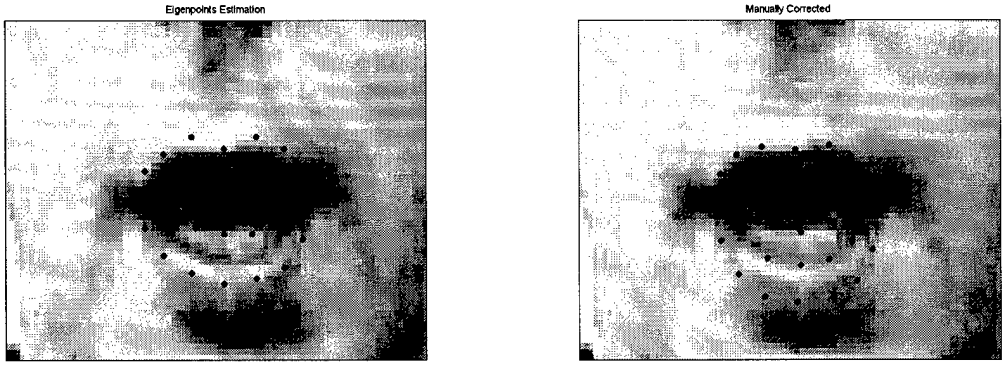
$$\hat{\mathbf{p}} = \mathbf{Q_P}(\mathbf{\Sigma_P}\mathbf{V_{FP}^T}\mathbf{\Sigma_K^{-2}}(\sigma_{\mathbf{K}}^2 - \sigma_{cn}^2\mathbf{I})\mathbf{V_{FP}}\mathbf{\Sigma_F^{\dagger}})\mathbf{Q_F^T}(\mathbf{f} - \overline{\mathbf{f}}) + \overline{\mathbf{p}}. \tag{14}$$

The estimate modifications described by Equations (9)-(14) were implemented, but the was no apparent difference in empirical trials in the estimates compared with the simple estimate of equation (7) in approximately 15 test images representing an array of mouth shapes ranging from a closed mouth to a wide open mouth. Therefore, the estimator of Equation (7) was used for this research.

In evaluating the performance of the eigenpoints algorithm on the image database used here, it was discovered that the process did not perform well on a large majority of the images. The process occasionally estimated lip control points along teeth, as Figure 10(a) illustrates, or estimated control points along shadows below lips as shown in Figure 11(a). Figure 12(a) illustrates a third type of error—namely, estimating a closed mouth when the mouth is open. Figures 10(b), 11(b), and 12(b) show the correct points determined by manual correction. Figure 13 shows an example where Eigenpoints adequately estimated the control points.

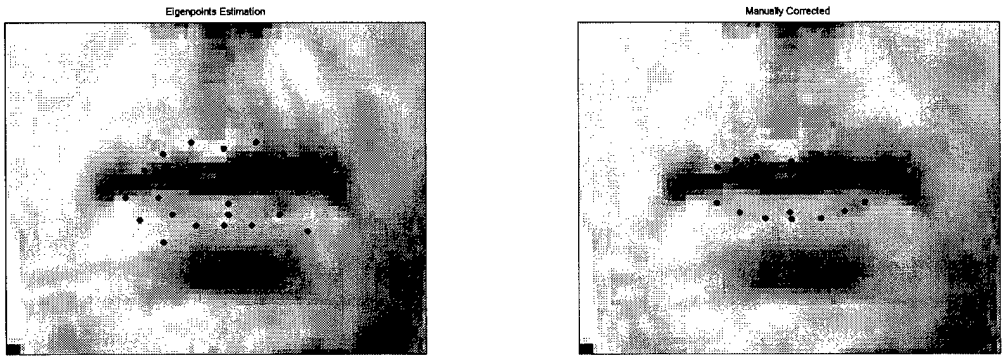
The discovery of poor performance led to a review of the control points estimated for images in the database and manual correction when required. Since a large majority of the control point sets required manual correction (a long and tedious process), only 50 sentences were reviewed and corrected. The bold faced prompts in Appendix A are the 50 used for the remainder of this research. This set was chosen simply because they are the first 50 recorded

in alphanumerical order. The set of 50 sentences contained 5031 images, 3644 (72%) of which required manual correction.



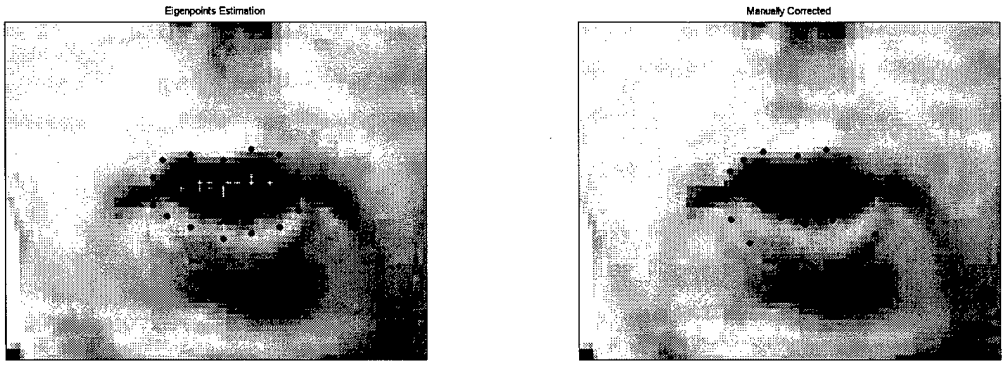
(a) (b)

Figure 10. Example of (a) eigenpoints algorithm control point estimation along teeth and (b) manual correction.



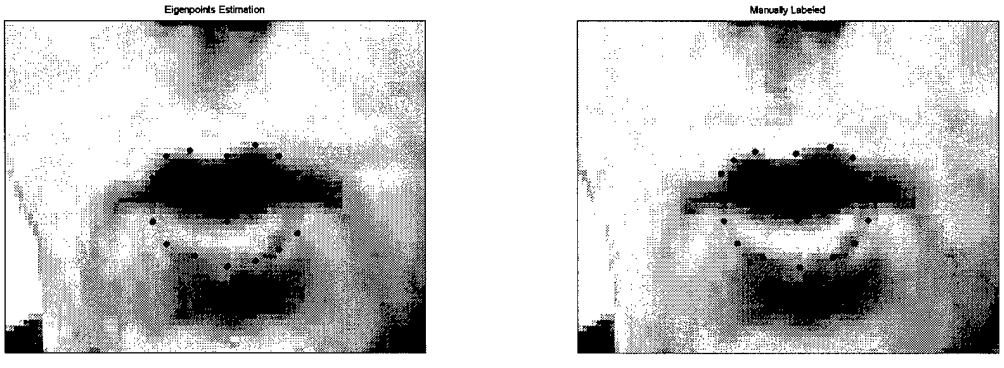
(a) (b)

Figure 11. Example of (a) eigenpoints algorithm control point estimation along shadows below the lips and (b) manual correction.



(a) (b)

Figure 12. Example of eigenpoints algorithm control point estimation of (a) a closed mouth when actually open and (b) the manual correction.



(a) (b)

Figure 13. Example of (a) good eigenpoints algorithm control point estimation and (b) the manually labeled points.

3.5 Distance Features

In lieu of using the control point coordinates directly, the work presented here uses distance features derived from these coordinates. Additionally, features capturing the dynamic characteristics of the static features for each frame are calculated. Finally, since there are many occasions where there are not enough examples (frames) in a given sentence to completely model each viseme in the sentence with three-state HMMs, the 30 frames/second data is upsampled to 90 frames/second. This upsampling proved to be crucial for this work.

3.5.1 Static Features. Figure 14 illustrates the 28 static (single frame) distance features derived from the control point coordinates.

The horizontal distances 1 and 2 are simply the outer and inner mouth widths. The vertical distances are the orthogonal distances from the control point coordinate to the line connecting the mouth corners (line 1). All distances are in pixels and are Euclidean distances.

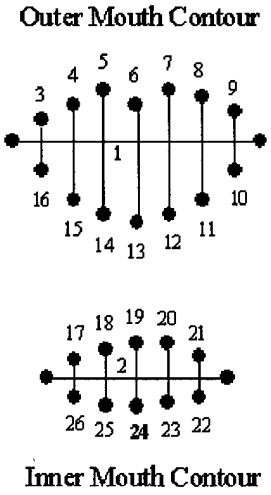


Figure 14. Distances derived from control points

3.5.2 Dynamic Features. Using features which characterize the temporal dynamics of the static features at some time t tends to improve performance in acoustic speech recognition. Also, dynamic features were found to be important in the synthesis of the Japanese syllables in [80]. Therefore, the algorithms presented here use these dynamic features to help capture temporal dynamics within and across visemes.

Let $\mathbf{c}_t$ be the static distance feature vector at time t. Using The HTK Book's [88] definitions of its functions, the dynamic features at time t are

$$\Delta \mathbf{c}_t = \frac{\Sigma_{\theta=1}^{\Theta} \theta(\mathbf{c}_{t+\theta} - \mathbf{c}_{t-\theta})}{2\Sigma_{\theta=1}^{\Theta} \theta^2}, \tag{15}$$

where Θ is the number of past and future features included in the computation. This research investigates values of Θ of 1 and 3.

3.5.3 Upsampling. A frame rate of 30 frames/second (fps) does not provide enough samples in most utterances to allow for three-state no-skip models for each viseme. Upsampling to 90 fps ensures enough data per utterance for embedded training of three-state left-to-right models for each viseme.

Upsampling to 90 fps required the use of the MATLAB "resample" function. This function resamples using a polyphase implementation. It applies an anti-aliasing FIR filter during the resampling process and compensates for filter delay. The function uses a weighted sum of $2*N*\max\{1, Q/P\}$ samples of the input vector to compute each sample of the output vector. For this research, $Q = 30$, $P = 90$, and $N = 2$. Experimentation with various values of N led to the choice of $N = 2$. This value resulted in the closest upsampled representation of the 30 fps data. Higher values introduced additional peaks and valleys in the feature sequences that were not evident in the 30 fps data.

3.6 Hidden Markov Model Training

Training of the HMMs requires several steps:

- Initialize viseme models.

- Clone visemes to form context-dependent triseme models for trisemes occurring in the training set, then re-estimate the triseme models.
- Use contextual decision trees to cluster model states.
- Use the trees and clustering statistics to develop models for trisemes that do not occur in the training set.

This section covers each step in detail.

3.6.1 Initialize Viseme Models. The training algorithm used in this research begins by estimating a global three-state HMM from all the training data. It then uses the feature segments of each utterance corresponding to each particular viseme to perform 20 iterations of basic Baum-Welch re-estimation of the viseme HMM parameters.

Let

$$\mathbf{O} = \{\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_T}\} \tag{16}$$

be a sequence of observations or feature vectors, where

$$\mathbf{o}_t = [\mathbf{c}_t \ \mathbf{\Delta c}_t]^T, \tag{17}$$

(where ^{T} denotes the transpose operation) and let λ be an N state HMM defined by the Gaussian state densities b_j, $1 \leq j \leq N$, b_j follows the multivariate normal distribution $\mathcal{N}(\left[\mu_j \ \mathbf{\Delta}\mu_j\right], [\mathbf{\Sigma}_j \ \mathbf{\Delta\Sigma}_j])$ and state transition probabilities

$$\mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}, \tag{18}$$

where a_{ij} is the probability of transitioning from state i to state j. Baum-Welch re-estimation determines the model parameters and sequence through the states $\mathbf{q} = \{q_1, q_2, \ldots, q_T\}$ that maximizes

$$P(\mathbf{q}, \lambda|\mathbf{O}) \tag{19}$$

The output of this initial stage are three-state HMMs for each of the 26 visemes.

3.6.2 Clone Viseme Models into Triseme Models. The end results of HMM training are context-dependent viseme models. These models are called triseme models as they are formed from a core viseme and are contextually distinct based on the visemes before and after the core viseme. Thus, the set of three visemes defines the triseme.

The cloning stage of the training algorithm creates models for each triseme contained in the training set from the 26 core visemes. For example, if the triseme A-B+C (the visemes A and C are to the left and right of B) exists in the training set, one estimates an HMM for A-B+C using the HMM for the viseme B as the initialization. Likewise, if A-B+D exists, one estimates a separate HMM, again initialized from the core viseme B model.

The algorithm also ties the state transition matrix (the matrix of transition probabilities) of each model to that of the core viseme. In the above example, the state transition matrices are identical for A-B+C and A-B+D.

Once context dependent models exist, an embedded version of the Baum-Welch algorithm is used which trains the HMMs using the context information by moving through the sentences sequentially forward and backward, not by simply grouping the data by distinct triseme class. For each training utterance, a sentence HMM is formed by concatenating each triseme model together given a labeled transcription of the utterance. A single pass through each training utterance simultaneously updates each of the triseme model parameters for the trisemes in the particular utterance. This process results in distinct trained three-state HMMs for each triseme in the training set.

3.6.3 Decision Tree-Based State Tying. In cases where one has only a limited amount of training data (as assumed for this research), one can use common states across models to reduce the number of overall model parameters requiring estimation. That is, states from multiple models can be clustered together into a common state, essentially increasing the amount of data available for estimating the state parameters. While there exist several methods for determining how to cluster the states, this research uses a decision tree to identify the states that should be clustered and tied together.

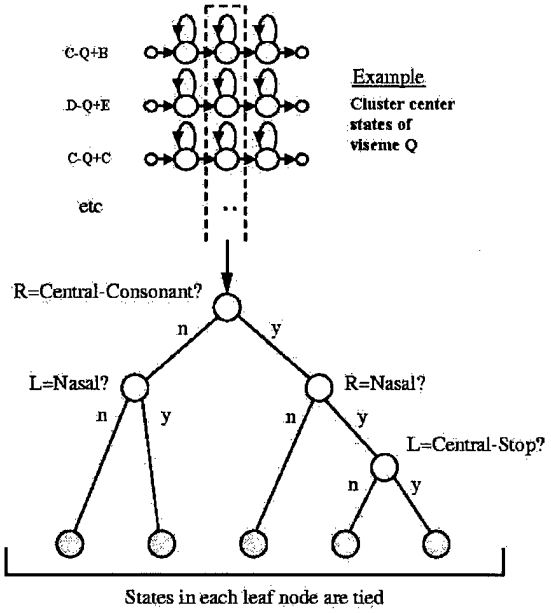


Figure 15. An example of decision tree-based state tying.

A visemic decision tree is a binary tree where a yes or no question is attached to each node. Initially, all states in a specific viseme state position are placed at the root node. Depending on the answer to each question, the pool of states is successively split until the states reduce to leaf-nodes. All states in the same leaf-node are tied together. Figure 15 adapted from [88] illustrates the concept for questions related to the type of phonemes the visemes represent.

Trees are built using a top-down sequential optimization process. Initially, all models are placed in a single cluster at the root of the tree. The question is found which gives the best (maximum) increase in log likelihood given the pool of states/models, the training data, and the state occupancy statistics of the training data. The pool is then split in two based on the answer to the question (yes or no). Once the increase in log likelihood falls below a threshold, splitting stops and states remaining at each new node merge into a single state.

Appendix B lists the decision tree questions used here. The questions come from an example provided with HTK as representative of a decision tree commonly used in continuous

acoustic speech recognition. Mapping phonemes to visemes and eliminating questions with duplicate "yes" lists resulted in the visemic questions listed in the appendix.

The advantage to decision tree-based state tying is that it allows for development of models for trisemes not present in the training set [51, 87]. State models are chosen for the new trisemes given the decision tree statistics and contextual questions.

A final embedded Baum-Welch pass through the training data refines the newly clustered and tied models. This final pass provides a more accurate estimate of the model parameters, since more training vectors are available for many of the models after clustering. This last pass produces the final HMMs for use in synthesis.

3.7 Synthesis of Control Points from HMMs

The primary goal of this research is synthesis of sequences of control points for trisemes not occurring in the training data. Using the definitions for $\mathbf{O}$ and $\mathbf{o}$ (functions of $\mathbf{c}$ and $\mathbf{\Delta c}$) in Equations 16 and 17, the problem reduces to finding the static feature sequence $\mathbf{c} = [\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_T]^T$ (assuming that $\mathbf{c}$ and $\mathbf{\Delta c}$ are statistically independent) which maximizes

$$P[\mathbf{O}|\lambda] = \sum_{\text{all } \mathbf{q}} P[\mathbf{q}, \mathbf{O}|\lambda], \tag{20}$$

where

$$P[\mathbf{q}, \mathbf{O}|\lambda] = P[\mathbf{q}|\lambda] \cdot P[\mathbf{O}|\mathbf{q}, \lambda]. \tag{21}$$

Since $P[\mathbf{q}|\lambda]$ is independent of $\mathbf{c}$, maximizing $P[\mathbf{q}, \mathbf{O}|\lambda]$ given a state sequence $\mathbf{q}$ with respect to $\mathbf{c}$ is equivalent to maximizing $P[\mathbf{O}|\mathbf{q}, \lambda]$ with respect to $\mathbf{c}$. Finding the optimal $\mathbf{c}$ and corresponding $\mathbf{q}$ that maximizes $P[\mathbf{q}, \mathbf{O}|\lambda]$ essentially gives the $\mathbf{c}$ which maximizes $P[\mathbf{O}|\lambda]$ in Equation (20).

The probability $P[\mathbf{O}|\mathbf{q}, \lambda]$ is

$$P[\mathbf{O}|\mathbf{q}, \lambda] = b_{q_1}(\mathbf{o}_1) b_{q_2}(\mathbf{o}_2) \cdots b_{q_T}(\mathbf{o}_T) \tag{22}$$

where the output probabilities at state j and time t are

$$b_j(\mathbf{o}_t) \sim \mathcal{N}(\mathbf{c}_t; \mu_j, \mathbf{\Sigma}_j) \cdot \mathcal{N}(\mathbf{\Delta c}_t; \mathbf{\Delta}\mu_j, \mathbf{\Delta\Sigma}_j). \tag{23}$$

Substituting Equation (23) into Equation (22) and taking the logarithm of both sides results in

$$\begin{aligned} \log P\left[\mathbf{O}|\mathbf{q}, \lambda\right] &= -\tfrac{1}{2}(\mathbf{c}-\mu)^T\mathbf{U}^{-1}(\mathbf{c}-\mu) \\ &\quad -\tfrac{1}{2}(\mathbf{\Delta c}-\mathbf{\Delta}\mu)^T\mathbf{\Delta U}^{-1}(\mathbf{\Delta c}-\mathbf{\Delta}\mu) \\ &\quad -\tfrac{1}{2}\sum_{t=1}^{T}\log|\mathbf{\Sigma}_{q_t}| - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{\Delta\Sigma}_{q_t}| - TM\log 2\pi \end{aligned} \tag{24}$$

where $M = 28$ (the number of static features per frame),

$$\mu = \left[\mu_{q_1}^T\ \mu_{q_2}^T\ \cdots\ \mu_{q_T}^T\right]^T \tag{25}$$

$$\mathbf{U} = \mathrm{diag}\left[\mathbf{\Sigma}_{q_1}\ \mathbf{\Sigma}_{q_2}\ \cdots\ \mathbf{\Sigma}_{q_T}\right] \tag{26}$$

$$\mathbf{\Delta}\mu = \left[\mathbf{\Delta}\mu_{q_1}^T\ \mathbf{\Delta}\mu_{q_2}^T\ \cdots\ \mathbf{\Delta}\mu_{q_T}^T\right]^T \tag{27}$$

$$\mathbf{\Delta U} = \mathrm{diag}\left[\mathbf{\Delta\Sigma}_{q_1}\ \mathbf{\Delta\Sigma}_{q_2}\ \cdots\ \mathbf{\Delta\Sigma}_{q_T}\right]. \tag{28}$$

Rewriting Equation (15) as

$$\mathbf{\Delta c}_t = \sum_{i=-\Theta}^{\Theta} w(i)\mathbf{c}_{t+i} \tag{29}$$

where

$$w(i) = \frac{i}{2\sum_{\theta=1}^{\Theta}\theta^2} \tag{30}$$

results in

$$\mathbf{\Delta c} = \mathbf{Wc} - \mathbf{b}, \tag{31}$$

where

$$W = \begin{bmatrix} w(0)\mathbf{I}_M & \cdots & w(\Theta)\mathbf{I}_M & & \mathbf{0} \\ \vdots & w(0)\mathbf{I}_M & & \ddots & \\ w(-\Theta)\mathbf{I}_M & & \ddots & & w(\Theta)\mathbf{I}_M \\ \mathbf{0} & \vdots & & \vdots & \\ \mathbf{0} & & w(-\Theta)\mathbf{I}_M & \cdots & w(0)\mathbf{I}_M \end{bmatrix} \tag{32}$$

with $\mathbf{I}_M$ as the $M \times M$ identity matrix and

$$\mathbf{b} = \begin{bmatrix} -w(-\Theta)\mathbf{c}_{-\Theta}^T - w(-\Theta+1)\mathbf{c}_{-\Theta+1}^T - \cdots - w(-1)\mathbf{c}_{-1}^T \\ -w(-\Theta)\mathbf{c}_{-\Theta+1}^T - w(-\Theta+1)\mathbf{c}_{-\Theta+2}^T - \cdots - w(-2)\mathbf{c}_{-1}^T \\ \vdots \\ -w(-\Theta)\mathbf{c}_{-1}^T \\ \mathbf{0}_{1\times M(T-2\Theta)} \\ -w(\Theta)\mathbf{c}_{T+1}^T \\ \vdots \\ -w(\Theta)\mathbf{c}_{T+\Theta-1}^T - w(\Theta-1)\mathbf{c}_{T+\Theta-2}^T - \cdots - w(2)\mathbf{c}_{T+1}^T \\ -w(\Theta)\mathbf{c}_{T+\Theta}^T - w(\Theta-1)\mathbf{c}_{T+\Theta-1}^T - \cdots - w(1)\mathbf{c}_{T+1}^T \end{bmatrix}, \tag{33}$$

where $\mathbf{c}_{-i}$ are the distance features derived from the control points in the Θ frames prior to the current triseme, and $\mathbf{c}_i$ are from the control points in the Θ frames after the current triseme and $\mathbf{0}$ is the zero vector.

Substituting Equation (31) into Equation (24) results in

$$\begin{aligned} \log P\left[\mathbf{O}|\mathbf{q},\lambda\right] \;=\; & -\tfrac{1}{2}(\mathbf{c}-\mu)^T\mathbf{U}^{-1}(\mathbf{c}-\mu) \\ & -\tfrac{1}{2}(\mathbf{W}\mathbf{c}-\mathbf{b}-\mathbf{\Delta}\mu)^T\mathbf{\Delta}\mathbf{U}^{-1}(\mathbf{W}\mathbf{c}-\mathbf{b}-\mathbf{\Delta}\mu) \\ & -\tfrac{1}{2}\sum_{t=1}^{T}\log|\mathbf{\Sigma}_{q_t}| - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{\Delta}\mathbf{\Sigma}_{q_t}| - TM\log 2\pi \end{aligned} \tag{34}$$

Expansion of the two product terms that are functions of $\mathbf{c}$ yields

$$\begin{aligned}\log P[\mathbf{O}|\mathbf{q},\lambda] &= -\tfrac{1}{2}\left[\mathbf{c^T U^{-1} c} - \mathbf{c^T U^{-1}}\mu - \mu^{\mathbf{T}}\mathbf{U^{-1} c} + \mu^{\mathbf{T}}\mathbf{U^{-1}}\mu\right] \\ &\quad -\tfrac{1}{2}[\mathbf{c^T W^T \Delta U^{-1} W c} - \mathbf{c^T W^T \Delta U^{-1} b} - \mathbf{c^T W^T \Delta U^{-1}}\Delta\mu \\ &\quad -\mathbf{b^T \Delta U^{-1} W c} + \mathbf{b^T \Delta U^{-1} b} + \mathbf{b^T \Delta U^{-1}}\Delta\mu \\ &\quad -\Delta\mu^{\mathbf{T}}\mathbf{\Delta U^{-1} W c} + \Delta\mu^{\mathbf{T}}\mathbf{\Delta U^{-1} b} + \Delta\mu^{\mathbf{T}}\mathbf{\Delta U^{-1}}\Delta\mu] \\ &\quad -\tfrac{1}{2}\sum_{t=1}^{T}\log|\mathbf{\Sigma}_{q_t}| - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{\Delta\Sigma}_{q_t}| - TM\log 2\pi\end{aligned} \quad (35)$$

To determine the necessary optimality condition for maximizing this expression we assume there exists a $\mathbf{c}^*$ such that the partial derivative of the expression evaluated at $\mathbf{c}^*$ yields $\mathbf{0}_{TM\times 1}$.

$$\frac{\partial \log P[\mathbf{O}|\mathbf{q},\lambda]}{\partial \mathbf{c}} = -\mathbf{U^{-1} c} - \mathbf{W^T \Delta U^{-1} W c} + \mathbf{U^{-1}}\mu + \mathbf{W^T \Delta U^{-1}}\Delta\mu + \mathbf{W^T \Delta U^{-1} b} \quad (36)$$

Solving for $\mathbf{c}^*$ gives a candidate for the optimal $\mathbf{c}$ that maximizes $\log P[\mathbf{O}|\mathbf{q},\lambda]$

$$\mathbf{c}^* = (\mathbf{U^{-1}} + \mathbf{W^T \Delta U^{-1} W})^{-1}(\mathbf{U^{-1}}\mu + \mathbf{W^T \Delta U^{-1}}(\Delta\mu + \mathbf{b})). \quad (37)$$

Synthesis of a triseme requires solving Equation (37) for every possible state sequence $\mathbf{q}$. Additionally, solution of Equation (21) requires the computation of $P[\mathbf{q}|\lambda]$ for every possible $\mathbf{q}$:

$$P[\mathbf{q}|\lambda] = \prod_{t=1}^{T-1} a_{q_t\, q_{t+1}} \quad (38)$$

Therefore, the $\mathbf{c}^*$ and $\mathbf{q}$ combination that maximizes $P[\mathbf{q},\mathbf{O}|\lambda]$ in Equation (21) yields the new distance features $\mathbf{c}$. These features can then be converted to control points for generation of new video sequences.

3.8 Summary

This chapter details the methods used in the research presented here. The recording of a small audio/visual speech corpus of 97 sentences formed the initial data set. The eigenpoints algorithm automatically estimated control points around the mouth area for

each image in the data set. All image/control point estimations in 50 of the 97 sentences were manually reviewed and corrections made where necessary forming the data set used in the analysis presented in the next chapter. These control points were converted to static and dynamic distance features. Trained HMMs for each of the 26 visemes resulted from Baum-Welch estimation using the corresponding segments from the training data set. The use of decision tree-based clustering allowed creation of HMMs for trisemes (visemes in context) that did not occur in the data set used for HMM training. Finally, the models were then used to estimate/synthesize new features through solution of a set of linear equations for every possible transition sequence through the HMM.

The next chapter presents experimental results of the synthesis algorithm applied to the 50 sentence A/V corpus subset.

IV. Experimental Results

4.1 Introduction

Since the ultimate goal of this research is to generate an image sequence lip synced to audio for viewing by humans, subjective viewing tests would have been preferable but were beyond the scope of this research for cost reasons, time constraints and subject unavailability. Additionally, the levels of four variable factors in the synthesis process can affect the resultant quality. To evaluate the performance of the algorithms presented in Chapter III, mouth shape features were synthesized while setting each of the four variable factors to one of two values. Statistical analysis of the mean-squared error between the features synthesized using the algorithm here and the true values of the features lends insight into the appropriate values for each factor. Finally, mouth outlines are presented using the features generated from the experimental runs with the variable factors set to the most appropriate levels to investigate visually the "closeness" of the synthetic results to the original mouth outlines.

Section 4.2 describes the design of the statistical analysis experiment and how the 50 sentences were segmented into a training set and a test set. Sections 4.3-4.4 present the analysis of the experimental results. Finally, Section 4.5 presents comparisons between mouth outlines generated from the synthetic features to those generated from the original features illustrating the performance of this technique visually.

4.2 Experimental Design

This section presents the breakdown of the data into a training set and a test set and the variable parameters in the synthesis process used as factors in the statistical analyses.

4.2.1 Data Organization. The data used in the experiment consist of the static and distance features defined in Section 3.5 for the 50 sentences manually reviewed for control point estimation errors. The static features were upsampled to 90 fps and dynamic features calculated based on the upsampled static features. The data set was divided into a test

Table 2. Trisemes occurring in the test sentence sx212
("I gave them several choices and let them set the priorities")

Index	Triseme	Triphone	In Training Set?	Index	Triseme	Triphone	In Training Set?
1	Z	silence	Y	23	B-B+K	n-l-eh	N
2	Z-R+B	sil-ay-g	N	24	B-K+C	l-eh+t	Y
3	R-B+L	ay-g+ey	N	25	K-C+F	eh-t+th	N
4	B-L+E	g-ey+v	N	26	C-F+K	t-th+eh	N
5	**L-E+F**	ey-v+th	N	27	**F-K+D**	th-e+m	N
6	E-F+K	v-th+eh	N	28	K-D+C	e-m+s	N
7	F-K+D	th-eh+m	N	29	D-C+K	m-s+eh	Y
8	**K-D+C**	eh-m+s	N	30	C-K+C	s-eh+t	Y
9	D-C+K	m-s+eh	Y	31	K-C+F	eh-t+th	N
10	C-K+E	s-eh+v	Y	32	C-F+Y	t-th+ax	Y
11	**K-E+G**	eh-v+r	N	33	F-Y+D	th-ax+p	Y
12	E-G+Y	v-r+ax	Y	34	Y-D+G	ax-p+r	Y
13	G-Y+B	r-ax+l	Y	35	D-G+R	p-r+ay	N
14	Y-B+Z	ax-l+sil	Y	36	G-R+P	r-ay+ao	N
15	B-Z+A	l-sil+ch	N	37	**R-P+G**	ay-ao+r	N
16	Z-A+U	sil-ch+oy	N	38	P-G+Y	ao-r+ax	Y
17	**A-U+C**	ch-oy+s	N	39	G-Y+C	r-ax+t	Y
18	U-C+Y	oy-s+ax	N	40	Y-C+V	ax-t+iy	Y
19	C-Y+C	s-ax+s	Y	41	C-V+C	t-iy+s	Y
20	Y-C+K	ax-s+eh	Y	42	V-C+Z	iy-s+sil	Y
21	C-K+B	s-eh+n	Y	43	Z	silence	Y
22	K-B+B	eh-n+l	N				

subset of features from sentence sx212. The remaining 49 sentences' features comprise the subset used to train the HMMs.

The 49 sentence training set contained 164.93 seconds (2.75 minutes) of audio producing 14,844 (at 90 fps) training vectors for the HMMs. Distinct trisemes numbered 975 (maximum theoretical amount for 26 visemes = $26^3 = 17,576$), and the silence viseme also existed among the 49 training sentences.

The test sentence, sx212, was 3.55 seconds in length and contained 37 distinct trisemes and the silence viseme. Four trisemes were duplicated in the sentence. Table 2 lists the trisemes in the sentence, whether or not the triseme was included in the training set, and

the acoustic triphone corresponding to the visual triseme. The six trisemes in bold are the six trisemes not occurring in the training data that were synthesized and used in the analysis presented here.

HMMs were trained using the process described in Section 3.6 using the feature vectors from the 49 training sentences for each triseme occurring in the 50 sentences. The visemic transcription file for the test sentence sx212 generated by "Aligner" contains the duration (in seconds) of each of the six trisemes tested here. This duration determines the number of feature vectors required for each triseme (frame rate $\times$ duration). Using the process described in Section 3.7, synthetic feature vectors for each of the six test trisemes were generated at a 90 fps frame rate, then downsampled to 30 fps again using MATLAB's resample function. The values used in the vector $\mathbf{b}$ defined in Equation (31) from Chapter III come from the true feature values for the adjacent frames for that particular triseme in sx212, since these adjacent frames were present in the stored data and would be used in the normal synthesis process.

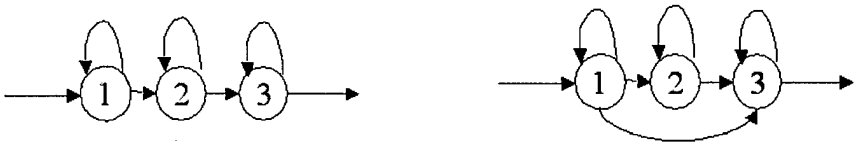
4.2.2 Process Variables. The synthesis process has essentially four parameters which can be varied to produce different results: log likelihood increase threshold (T), outlier threshold (R), dynamic feature window length (Θ), and HMM structure (three-state left-to-right with a skip from state 1 to state 3 either allowed or not allowed).

The log-likelihood increase threshold (T) sets the level below which an increase in log likelihood of the training data causes termination of the given path through the decision tree [88].

The outlier threshold (R) determines the minimum occupancy of any cluster in the decision-tree state clustering and prevents a single outlier state forming a cluster with one occupant because it is different from the other states [88].

The dynamic feature window length Θ is described in Chapter III.

Two HMM structures are evaluated in this research. The first is a three-state, left-to-right model with no state skips allowed. The second is the same three-state left-to-right



(a) (b)

Figure 16. Examples of a three-state left-to-right model with no skips allowed (a), and a skip from state 1 to state 3 allowed (b).

model, but a skip of state 2 is allowed. Figures 16(a) and (b) show these two structures respectively.

4.2.3 Structure of Experiment. The experiment consisted of a four factor, two level (i.e. 2^4) factorial design. Two types of analysis were conducted on the results. The first is an unreplicated 2^4 experiment for each of the six trisemes. The second uses the six trisemes as blocking factors in the factorial analysis. The blocking reduces the experimental error caused by a known and controllable nuisance source (in this case differing trisemes/models) [48]. For each analysis, results are presented using the average MSE per feature in each triseme and again averaging the MSE for all features for the given triseme.

Each of the factors were tested at both a high and a low level to evaluate the extreme effects of each factor. Table 3 lists the levels chosen for each of the factors. The levels for T and R were chosen as representative extremes after experimentation with several values.

Table 3. Factors Evaluated in the ANOVA

Factor	Low Level (-)	High Level (+)
T	20	100
R	10	20
Θ	1	3
HMM Type	No Skip	Skip Allowed

The levels of 1 and 3 for Θ represent a range that ensures that no synthetic data is used in the new feature estimation. A value of 0 eliminates the dynamic features and only static features are present. A value larger than 3 implies the potential use of static features in the prior and following triseme relative to the triseme under analysis (assuming an upsampled rate of 90 fps) which could realistically not be in the stored video database and thus be synthetic (see Equation (31)). The two HMM types reflect a commonly used type (the no skip model), and a type that attempts to allow for short trisemes or trisemes varying little from frame to frame and not requiring three states.

All six test trisemes were synthesized using each combination of the four factors at each level (16 runs).

4.3 Analysis of Individual Triseme Results

This section details the analysis performed on the static distance features for sx212 synthesized using HMMs trained from the 49 sentence data set.

The variable under analysis is the mean squared error (MSE) between the original and synthetic features. The squared error is averaged across all N frames of a given triseme. The research presents two types of analyses. The first type of analysis considers the effects each of the four factors and their higher order interactions have on the MSE for each individual feature in each triseme using MSE defined by

$$\mathbf{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{c}_o(i) - \mathbf{c}_s(i))^2, \tag{39}$$

where $\mathbf{c}_o(i)$ is the vector of true feature values for the i^{th} frame of the triseme, $\mathbf{c}_s(i)$ is the vector of synthesized features for the i^{th} frame, and the squaring operation is performed on the individual elements of the resultant vectors.

The second type of analysis investigates the effects of the 4 factors on the MSE for the average MSE per feature per frame per triseme (*i.e.*, the errors for each feature in a frame are essentially pooled). That is, the squared error is averaged over all features in each frame

as described by

$$\mathbf{MSE} = \frac{1}{28N}\sum_{i=1}^{N}\mathbf{1}^{\mathbf{T}}(\mathbf{c}_o(i) - \mathbf{c}_s(i))^2, \tag{40}$$

where $\mathbf{1}$ is a 28 element column vector of ones.

4.3.1 Individual Feature Analysis by Triseme. This section details the analysis of the MSE for each of the 28 features one triseme at a time. The experiment for each of the features is a 2^4 factorial design with only a single replication. In analyzing this type of experiment, there is no estimate of experimental error since there is only one replication [48]. Examination of a normal probability plot of the estimates of each factor's effects identifies those factors with significant effect on the response variable (MSE). Plots of the average MSE for the significant factors/interactions at both the low and high levels reveals which level produces the lower MSE on average. The next section gives an example of the technique and plots used in the analysis.

4.3.1.1 Description of Analysis and Example. The first step is computing the effects. Let A, B, C, D refer to the four factors under analysis: T, R, Θ, and HMM type, respectively. Table 4 lists the 16 combinations of the factor values, each considered a run of the experiment. The run labels correspond to the factors set to the high level in the run. For example, the run label "abc" refers to the run where factors A, B and C are at high levels and D is at its low level.

Equation (41) indicates the effect for the appropriate factor or higher order interaction:

$$AB\cdots D = \frac{2}{n2^4}(Contrast_{AB\cdots D}) \tag{41}$$

The variable n is the number of replications of the experiment and "Contrast" is a computation from the result of each run that provides a measure of how the factor under analysis contrasts with the others. Contrast is computed by expanding the right side of Equation (42) where the symbol "1" replaces a value of 1 in the expansion and subtraction is performed

for factors at the high level and addition at the low level:

$$Contrast_{AB\ldots D} = (a \pm 1)(b \pm 1)(c \pm 1)(d \pm 1). \tag{42}$$

For example, the contrast for the two-factor interaction AB is

$$\begin{aligned} Contrast_{AB} &= (a-1)(b-1)(c+1)(d+1) \\ &= (ab-a-b+1)(c+d+cd+1) \\ &= (1)-a-b+ab+c-ac-bc+abc+ \\ &\quad d-ad-bd+abd+cd-acd-bcd+abcd, \end{aligned} \tag{43}$$

where the measured result (Average MSE in this case) replaces the run label.

Let $\mathbf{y}$ be a 16 element vector of the measured results and $\mathbf{C}$ a 16×15 matrix of multiplicative constants (± 1) defined in Table 5 as the expansion of Equation (42) for each factor/interaction. Then the effects $\mathbf{e}$ are

$$\mathbf{e} = \frac{1}{8}\mathbf{C^T y} \tag{44}$$

for an unreplicated four factor experiment.

A normal probability plot of the effects $\mathbf{e}$ will show the insignificant effects along a straight line with approximately zero mean and variance σ^2. The significant effects are separated from those and not on the line. The normal probability plot is generated by sorting the values of $\mathbf{e}$ in ascending order and assigning a probability to the j^{th} effect of $(j-0.5)/15$. Table 6 lists an example (data taken from an example in [48]) of the sorting and probability assignment, and Figure 17 shows the plot.

From the example, the effects for A, C, D, AC and BC appear significant. Plots of the average measurements at each level of the factors with significant effects identify the appropriate levels for each to produce the desired measurement (i.e. to maximize or minimize). Figures 18(a) and (b) show the average measurements for the main significant effects (a) and the interaction plots (b). The value for each is computed by averaging the

result of all runs with the factor of interest at the corresponding high or low level. For example, the plot for factor A shows the average of all of the odd numbered runs (A at the low level) for the low level and the average of all the even numbered runs (A at the high level). If the desire is to maximize the response, interpretation of the main factor plots suggests using factors A, C and D at the high levels. However, the interaction plot for the AC interaction suggests the use of C at its low level when A is at its high level. Additionally, the AD interaction plot shows that the response is maximized when A and D are at their high levels. The bottom line is that factors A and D should be set to the high levels and C to the low level to maximize the response. Factor B has no significant effect on the response, so either level is appropriate.

The following six sections present this analysis applied to the results for each of the six trisemes. The normal probability plots and main effect plots for the analysis in these sections are in Appendix C.

4.3.1.2 Triseme L-E+F. The normal probability plots indicate no statistically significant effects for features 1, 2, 10, 17, 18, 19, 20, 21, or 24 since the effects fall nearly on a straight line. All of the plots identifying significant effects with the exception of feature 28 identify factor C as having a significant effect. Factor B had a significant effect for features 3, 6, 11, and 23. The normal probability plots indicate higher order interaction between both B and C for features 3, 5, 7, 9, 23 and 28. Other higher order interactions also exist for feature 28 (edge of lower teeth) for this triseme (BD, CD and BCD).

For factor C, in all 18 cases, the main effect plots show that the average MSE for the triseme was minimized when the factor C (dynamic window length, Θ) was at its low level (1 frame) for the 18 features.

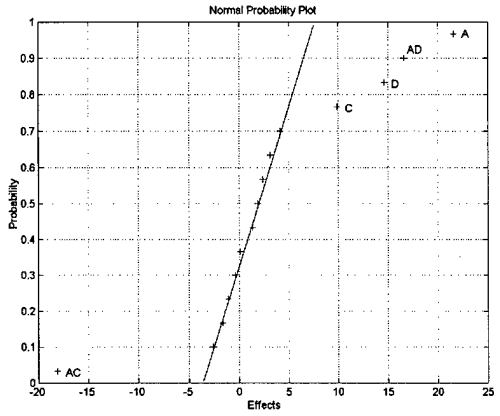


Figure 17. An example of a normal probability plot of the data in Table 6

Table 4. Experimental run labels for a 2^4 factorial experiment

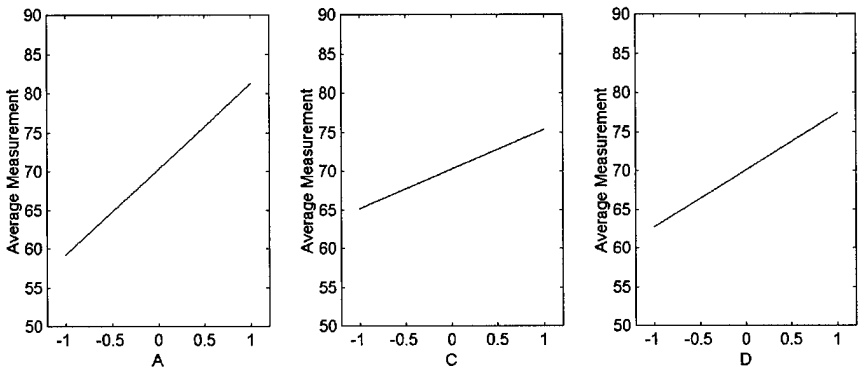
Run	Factor				Run
Number	A	B	C	D	Label
1	-	-	-	-	(1)
2	+	-	-	-	a
3	-	+	-	-	b
4	+	+	-	-	ab
5	-	-	+	-	c
6	+	-	+	-	ac
7	-	+	+	-	bc
8	+	+	+	-	abc
9	-	-	-	+	d
10	+	-	-	+	ad
11	-	+	-	+	bd
12	+	+	-	+	abd
13	-	-	+	+	cd
14	+	-	+	+	acd
15	-	+	+	+	bcd
16	+	+	+	+	abcd

Table 5. Contrast multiplicative constants from expansion of Equation 42

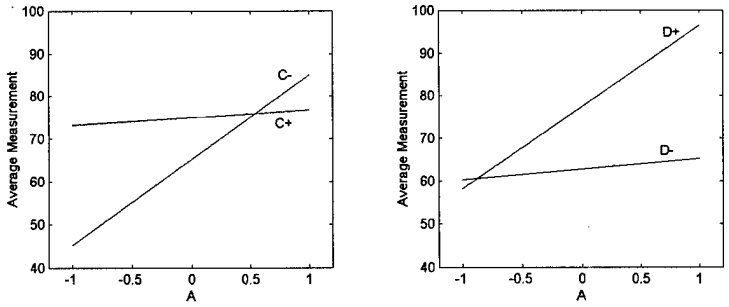
	A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	ABCD
(1)	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1
a	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
b	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1
ab	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
c	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1
ac	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
bc	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
abc	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
d	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1
ad	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
bd	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1
abd	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
cd	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1
acd	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
bcd	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
abcd	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 6. Example of assigning normal probability for plotting to identify significant effects

j	$\mathbf{e}_j$	$(j-0.5)/15$	Factor/Interaction
1	-18.125	0.033	AC
2	-2.625	0.100	BCD
3	-1.625	0.167	ACD
4	-1.125	0.233	CD
5	-0.375	0.300	BD
6	0.125	0.367	AB
7	1.375	0.433	ABCD
8	1.875	0.500	ABC
9	2.375	0.567	BC
10	3.125	0.633	B
11	4.125	0.700	ABD
12	9.875	0.767	C
13	14.625	0.833	D
14	16.625	0.900	AD
15	21.625	0.967	A



(a) Main Effect Plots



(b) Interaction Plots

Figure 18. Main effect and interaction plots for measurement data used to determine effects in Table 6.

The main effect plots for features 6, 11 and 23 indicate the average MSE for the triseme was lowest when factor B was at the high level (20). While the plot for feature 3 shows the average MSE lower when B is at the low level (10), the average MSE at the high level is only 2.5 pixels^2. Therefore, the conclusion from the plots is: use the high level of the outlier threshold to minimize the average MSE in general for this triseme. However, the analysis includes averages from factor C at the high level, and several features showed signs of significant BC interaction. It is possible that these runs perturb the main effect plots. The impact of the runs with C at its high level is apparent when comparing the BC interaction plots for features 3 and 23 with the main effect plots for factor B for the same features. For the average MSE with C at the low level, the MSE averaged across all 6 of the features is 11 and 12.75 pixels^2 for B at the low and high levels, respectively. From these values it is difficult to determine which level to choose for factor B. However, the MSE averaged across only features 3, 5, 7, 9 and 23 is 4.7 and 2.5 pixels^2 for the low and high levels of B, respectively. This result suggests that feature 28 is likely an anomaly. This is also supported by the normal probability plots where the BC effect is on the opposite side of the mean (-) in feature 28 whereas the effect is positive in the other five features. Thus the conclusion from the B main effect plots holds, and B set to the high level minimizes the average MSE in most cases for this triseme.

The main effect plot for D suggests that a three-state left-to-right model with no skips minimizes the average MSE for this triseme and feature. The other three interaction plots also support this suggestion. The BD, CD and BCD interaction plots indicate using B at the high level and C and D at the low level, which agrees with the BC interaction for this feature as previously discussed.

In summary, with the exception of feature 28, the analysis indicates that using a three-state left-to-right no skip HMM trained with an outlier threshold at the high level (20) and dynamic window length at the low level (1) generally minimizes the average MSE for this triseme. Additionally, since no significant effect or interaction exists for the minimum occupancy threshold (factor A), either level is appropriate for this triseme.

4.3.1.3 Triseme K-D+C. There are no statistically significant effects for features 2, 10, 15, 21, or 22. All of the plots identifying significant effects with the exception of feature 12 identify factor C as having a significant effect. Factor B has a significant effect for features 9 and 16 while features 13 and 14 indicate a significant effect for factor A. For this triseme, factor D has a significant effect for feature 27 (edge of upper teeth). The normal probability plots again indicate higher order interaction between B and C, this time for features 9, 16 and 20. Other higher order interactions also exist. The two factor AB interaction is significant for feature 12, AC in feature 14 and the CD interaction is significant in feature 27. Two three level interactions have significant effects in this triseme: ABC in feature 12, and BCD in feature 5.

As with triseme L-E+F, all of the main effect plots for factor C show minimization of the average MSE when the dynamic feature window length is 1 frame. For features 13 and 14 the runs with the log-likelihood threshold (factor A) at the low level (20) had the smaller average MSE. The average MSE for both feature 9 and feature 16 is smaller for the runs with factor B (outlier threshold) at the high level (20), which is consistent with the results from triseme L-E+F. Again, as in triseme L-E+F, a no skip model type results in a lower average MSE for this feature and triseme.

The two factor interaction plots for the BC interaction indicate that significance is due mainly to the inclusion of runs with factor C at the high level. With C at the low level, there is no significant differences in the average MSE between the two levels of factor B. This is the same result as the AC interaction plot shows. Interpretation of the AB interaction plot shows a slightly lower average MSE when both factors A and B are at their low levels. However, with both at the high levels, the average MSE is less than 0.2 pixels^2 different than with both at the low levels. Because of the small magnitude of the average MSE, it is difficult to draw a firm conclusion from the AB interaction plot. The CD interaction plot reveals that with C at the low level, runs with factor D also at the low level result in a lower average MSE for feature 27 in this triseme.

The ABC interaction plot shows runs with factors A and B at the low level and C at the high level produced a lower average MSE for feature 12. However, with factor B at the

high level and C at the low level, the average MSE is less than 1 pixel^2 higher and is only 1.5 pixels^2 with A at the high level. The BCD interaction plot shows a lower average MSE when all three factors are at the low levels for feature 5. However, runs with B at the high level and C and D at the low level had an average MSE only 1.3 pixels^2 higher.

In summary, the main effect plots for the factors with significant effects along with the interaction plots for significant interactions indicate that a no skip HMM trained with a log-likelihood increase threshold at the low level, outlier threshold at the high level, and a dynamic feature window length of 1 frame minimizes the average MSE for triseme K-D+C for most features. These results are consistent with the results from triseme L-E+F.

4.3.1.4 Triseme K-E+G. There are no statistically significant effects for features 2, 3, or 9. All of the normal probability plots identifying significant effects identify factor C as having a significant effect along with a significant BC interaction. Factor B has a significant effect for features 1, 4, 5, 6, 7, 13, 17, 18, 20, 21, and 23 while feature 13 also indicates a significant effect for factor A. For this triseme, factor D has a significant effect for features 7 and 21. Other higher order interactions also exist. The two factor AB interaction is significant for features 21 and 28 and the AC interaction in feature 13. One three level interaction has a significant effect in this triseme: ABC in feature 28.

Once again, the main effect plots for factor C indicate that runs with a dynamic feature window length of 1 frame result in lower average MSE. An interesting phenomenon appears when viewing the BC interaction plots and comparing them to the main effect plots for factor B. The main effect plots suggest that runs with factor B at the high level result in a lower average MSE for the features in which its effect was significant. However, the BC interaction plots suggest that runs with B and C at their low levels result in a lower average MSE. The disparity arises because the main effect plots average in the runs with C at both levels whereas the BC interaction plots differentiate the different levels of factor C. Thus, interpretation of the two sets of plots (factor B and BC interaction) suggests the use of factors B and C at their respective low levels reduce the average MSE for this triseme.

The only feature where the effect of factor A is significant is feature 13. Runs on average with factor A at the low level result in a much lower average MSE than those at the high level, a result consistent with triseme K-D+C.

The main effect plot for factor D seems to indicate a conflicting result for factor D in features 7 and 21, but the difference in average MSE between runs with D at the low level and those at the high level for feature 21 is less than 0.35 pixels^2. Thus, the conclusion drawn from the main effect plots for factor D is that a no skip model results in lower average MSE.

The AB interaction plots for features 21 and 28 seem to suggest runs with factor A at the low level and factor B at the high level result in a lower average MSE. However, as with the main effect plots for factor B, each case includes runs averaged in with factor C at the high level which is a significant source of error. The three factor interaction plot for ABC for feature 28 also illustrates this fact. Runs with factors A, B and C at their low levels result in lower average MSE for feature 28.

Thus, the analysis of the effects and interactions for triseme K-E+G suggest again that a no skip HMM trained with the log-likelihood threshold at the low level (20 in this case), outlier threshold at the low level (10), and a dynamic window length of 1 frame results in a lower average MSE for this triseme. The conclusion of using the low outlier threshold for this triseme conflicts with the conclusions from the previous two trisemes.

4.3.1.5 Triseme A-U+C. The normal probability plots of the factor effects for the 28 features of triseme A-U+C indicate that there are no statistically significant effects for features 4, 9, 17, 23, 25 and 26. For this triseme several features have factors with identical effect magnitudes of zero or a very small number. Several effects identified as significant have very small magnitudes as well (features 3, 6-8, 17, 19, 22 and 24). Factor C has a significant effect in this triseme for features 1-3, 5, 6, 8, 11-19, 21, 22, 24, 27 and 28. Factor D has significant effects in 5 features (1, 2, 7, 10 and 18). There are no indications that factor A has a significant effect, and B has significant effects in only four features (10, 14, 20 and 28). Feature 10 also shows indication of higher order interactions BC, BD and

BCD. The two factor interaction BC also has a significant effect in features 14, 20 and 28, and CD has a significant effect in features 14, 18 and 20.

The main effect plots for factor C show varied results. The average MSE is lower with C at its high level in 10 of the 19 features where the C effect is significant. However, the mean average MSE per feature for this set of 19 was approximately 4.5 pixels^2 per feature for suns with C at its low level and approximately 6 pixels^2 per feature with C at its high level. Much of the error comes from features 1 and 2 (the mouth widths). Excluding those two features, the mean is approximately 3 pixels^2 per feature for runs with C at the low level and 2.5 pixels^2 per feature for those with C at the high level. Since the mouth widths are fairly important features whose average MSE is than the others, minimizing the average MSE overall suggests using HMMs with C at the low level for this triseme.

The main effect plots for factor D in features 1 and 2 indicate that runs with a no skip model result in a lower average MSE, but the plots for the other three features indicate otherwise. However, the difference in average MSE between runs at the low and high levels for the last three features is small relative to the differences for features 1 and 2. Thus, the interpretation of these plots is that runs with a no skip HMM result in lower average MSE overall for this triseme.

For factor B, the main effect plots indicate that runs with the outlier threshold at the high level result in lower average MSE for three of the four features (10, 14 and 28). The average MSE for all features remains small ($1.4 - 8.5$ pixels^2). The MSE averaged across all four features is approximately 3.9 pixels^2 for runs with the low level of the outlier threshold and approximately 3.4 pixels^2 for those with the factor at the high level. This closeness makes it difficult to draw a firm conclusion on the appropriate level of factor B for minimizing the average MSE from the main effect plots.

The two factor BC interaction plots support the conclusion that in most cases runs with factor B at the high level result in lower average MSE. Runs with factor C at the low level have lower average MSE in features 14, 20 and 28 with factor B at the high level, and the difference for those runs in feature 10 is less than 1 pixels^2.

The BD and BCD interaction plots also support the conclusion that factor B should be used at the high level. In the BD interaction plot, runs with factor B at the high level result in lower average MSE for all runs regardless of the level of factor D. The plot suggests, however, that runs with the HMM with a skip state allowed result in lower average MSE for feature 10, which is consistent with the main effect plots for factor D. For runs with factors C and D at their low levels, the BCD interaction plot shows no difference in average MSE between runs with factor B at its low and high levels. The lowest average MSE results in runs with factor B at the high level, factor C at the high level, and either model type (factor D).

The plots for the CD interaction again indicate conflicting conclusions. The average MSE is lowest for runs with factor C at the high level and factor D at the low level for features 18 and 20, but in feature 14 the lowest average MSE resulted when factor C was at the low level and D at the high level.

Given the conflicting results between the various features, it is not possible to draw a general conclusion on the set of factor levels that minimize the average MSE for this triseme. Since the magnitude of the error is modest in most cases shown here, the grouping of factor levels identified as appropriate in other trisemes is likely to be effective for A-U+C.

4.3.1.6 Triseme F-K+D. Only a few of the 28 features show factors with significant effects for this triseme. Factor C has significant effects in features 2, 3, 9, 23, 24, 27 and 28, and Factor B has a significant effect only in feature 10. Only feature 27 has any other factors with significant effects (A, a CD interaction, and possibly D).

The main effect plots for factor C show that for the seven features for which C has a significant effect, runs with a dynamic window length of 3 frames result in a lower average MSE. This conclusion is in contrast with the results from the previous three trisemes.

The main effects plots for factors A and D and the CD interaction plot indicate that for this triseme and feature, runs with a log likelihood threshold at the high level (100), dynamic window length of 3 frames and a no skip HMM result in lower average MSE.

Finally, the main effect plot for factor B indicates that, at least for feature 10, runs with an outlier threshold (factor B) at the high level (20) result in lower average MSE.

The results for this triseme differ somewhat from those in the previously analyzed trisemes. The analysis here indicates that experimental runs with the longer dynamic feature window result in lower average MSE for this triseme (instead of the shorter window that analysis of the other trisemes indicated). The results for the other factors/interactions with significant effect tend to indicate that runs with factor A at the high level, B at the high level, and D at the low level result in the lowest average MSE. However, that conclusion is indicated for only two of the 28 features, one of which was the location of the teeth (not any lip contour features).

4.3.1.7 Triseme R-P+G. The final triseme analyzed here is R-P+G. The normal probability plots for each of the 28 features for this triseme indicate significant effects for factor C in several features (1, 5, 7-9, 18-21 and 25-28). Additionally, factor A has significant effects in features 1, 2, and 18, and factor B has significant effects in features 21, 25 and 26. The plots show higher order interactions in features 5 (BC), 14-16 (AB), 18 and 20 (AC), 25 (ABC)and 26 (BD and CD).

The main effect plots for factor C for features where C has a significant effect show that runs with factor C at the low level result in lower average MSE than those with C at the high level, with the exception of features 9 and 27.

The main effect plots for factors A and B for those features where they have significant effects show that runs with factor A at the low level and factor B at the high level result in lower average MSE.

The BC interaction plot supports the earlier conclusion that runs with factor B at the high level and C at its low level result in lower average MSE. However, the AC interaction plots for features 18 and 20 indicate slightly lower average MSE for experimental runs with factor A at the high level (contrary to the main effect plots). However, the difference between runs with A at the low level and A at the high level (with a 1 frame dynamic feature window) is less than 1 pixel^2. The AB interaction plots, however, do concur with the main effect plots

Table 7. Summary of Results from Individual Triseme Analysis

	Log-Likelihood Threshold (A)	Outlier Threshold (B)	Dynamic Feature Window Length (C)	Model Type (D)
L-E+F	20 or 100	20	1	No Skip
K-D+C	20	20	1	No Skip
K-E+G	20	10	1	No Skip
A-U+C	20 or 100	10 or 20	1 or 3	No Skip or Skip Allowed
F-K+D	100	20	3	No Skip
R-P+G	100	20	1	No Skip

in that the runs with factor A at the low level and B at the high level resulted in lower average MSE. The discrepancy is likely due to the inclusion of runs with factor C at the high level in computing the average MSE for the AB interaction. As shown by the AC plots, with both factors A and C at the high level, the average MSE is much higher. One plot which illuminates this problem is the ABC interaction plot for feature 25. Runs with factor B at the high level and C at the low level result in slightly lower average MSE with factor A at the high level than those with factor A at the low level. Therefore, the main effect plots are likely biased by the inclusion of runs with factor C at the high level, and runs with factors A and B at the high level and C at the low level have lower average MSE for this triseme. The only remaining factor that requires determination of an appropriate level is factor D. The BD and CD interaction plots for feature 26 make this determination. It is clear from the two plots in Figure 128 that with factor B at the high level and C at the low level, runs with factor D at its low level (a no skip model) result in lower average MSE.

Thus, for triseme R-P+G experimental runs with factors A and B at the high level and C and D at the low level generally result in lower average MSE than other combinations of factor levels.

4.3.1.8 Summary of Individual Feature Analysis. Table 7 summarizes the conclusions from the individual triseme analysis. It is clear that runs with a 3 state left-to-right HMM with no skips results in lower average MSE. Four of the six trisemes indicate that using a dynamic feature window length of one frame results in lower average MSE, while triseme A-U+C can use either level. With the exception of triseme K-E+G, experimental

runs with an outlier threshold of 20 produce a lower average MSE. Finally, a log-likelihood threshold at either level is appropriate since there is no significant effect in L-E+F or A-U+C and the results are equally varied for the other four trisemes. Although the results differ slightly from triseme to triseme, Table 7 indicates that using either level of log-likelihood threshold, an outlier threshold of 20, a dynamic feature window length of one frame, and a no skip HMM result in lower average MSE in a majority of trials. The next section conducts the same analysis but with the MSE averaged over all features for each triseme.

4.3.2 Average Feature Analysis by Triseme. This section uses the same analysis and procedure as the previous section but for MSE averaged over all features for a given triseme.

Figure 19 presents the normal probability plots for the effects/interactions in each of the six trisemes. Note that triseme F-K+D has no significant effects or interactions. Factor C has a significant effect in all 5 of the other trisemes. The plot for K-E+G shows an interaction between factors B and C. Factor D has a significant effect in triseme A-U+C, and factors A and B are significant in R-P+G.

Figure 20 shows the main effect and BC interaction plots for each of the above significant effects. The main effect plots for factor C (Figures 20(a)-(e)) show that runs with a one frame dynamic feature window length result in lower average MSE than those with a three frame window. Figure 20(f) confirms this result in the BC interaction for triseme K-E+G. It also suggests that for K-E+G runs with an outlier threshold of 10 have lower average MSE than those with a threshold of 20. Figure 20(g) reflects the main effect plot for factor D in triseme A-U+C and shows that runs for this triseme with a no skip HMM have lower average MSE. The remaining two plots (Figures 20(h) and (i)) are the main effect plots for factors A and B in R-P+G. Note that the plot for factor B suggests using an outlier threshold of 20 which contrasts with the BC interaction plot for K-E+G. This contrast is likely due to the inclusion of runs with factor C at its high level in the plots. Finally, the plot for factor A suggests that runs for R-P+G have a lower average MSE when the log-likelihood threshold is 20 than those with the threshold at 100.

In summary, the effect and interaction plots for this analysis (where the measurement was the total MSE per frame averaged over the triseme) suggest that using a parameter combination of a log-likelihood (A) threshold of 20, outlier threshold (B) of 10, dynamic feature window length (C) of one frame, and models without a skip state generally result in lower average MSE. The only difference between this result and the result from the individual feature analysis is the outlier threshold. Although the conclusion of using a lower outlier threshold here was drawn from the BC interaction in K-E+G, a lower outlier threshold is also suggested in the individual feature analysis for the same triseme.

4.4 Analysis of results with blocking

The individual triseme analysis in the previous section indicates variation from triseme to triseme. This section investigates the effects of the four factors in a more general sense by using the trisemes as blocks in a 2^4 factorial analysis of variance (ANOVA).

4.4.1 Individual feature analysis. Table 8 lists the computations performed in the ANOVA [48]. The factor notations A, B, C, and D refer to the four factors in Table 3 and the small letters $a, b, c, d = 2$ refer to the number of levels of each factor. The "Blocks" source refers to the $n = 6$ trisemes. The variable of interest, y, is the MSE for the given analysis. Each factor is visualized as an orthogonal dimension with "Blocks" being a fifth dimension. A dot $(\cdot)$ means that the MSE for all elements along the appropriate dimension are totaled. For example, the SS_A is computed by summing all MSEs where A is at level 1, squaring that value, and adding it to the square of the sum of all MSEs where A is at level 2. This result is then divided by $(abcn) = 48$. The square of the sum of all MSE values ($y^2_{....}$) is subtracted from the result to give the SS_A. Mean Square values (MS) are computed by dividing the corresponding SS value by the degrees of freedom. Also, F_0 represents the critical value in the F distribution for the corresponding factor with the corresponding degrees of freedom. The area under the F distribution from F_0 to ∞ is the statistical significance level α. If α is smaller than or equal to a desired significance threshold, the factor has an effect on the MSE that is statistically significant at a level α.

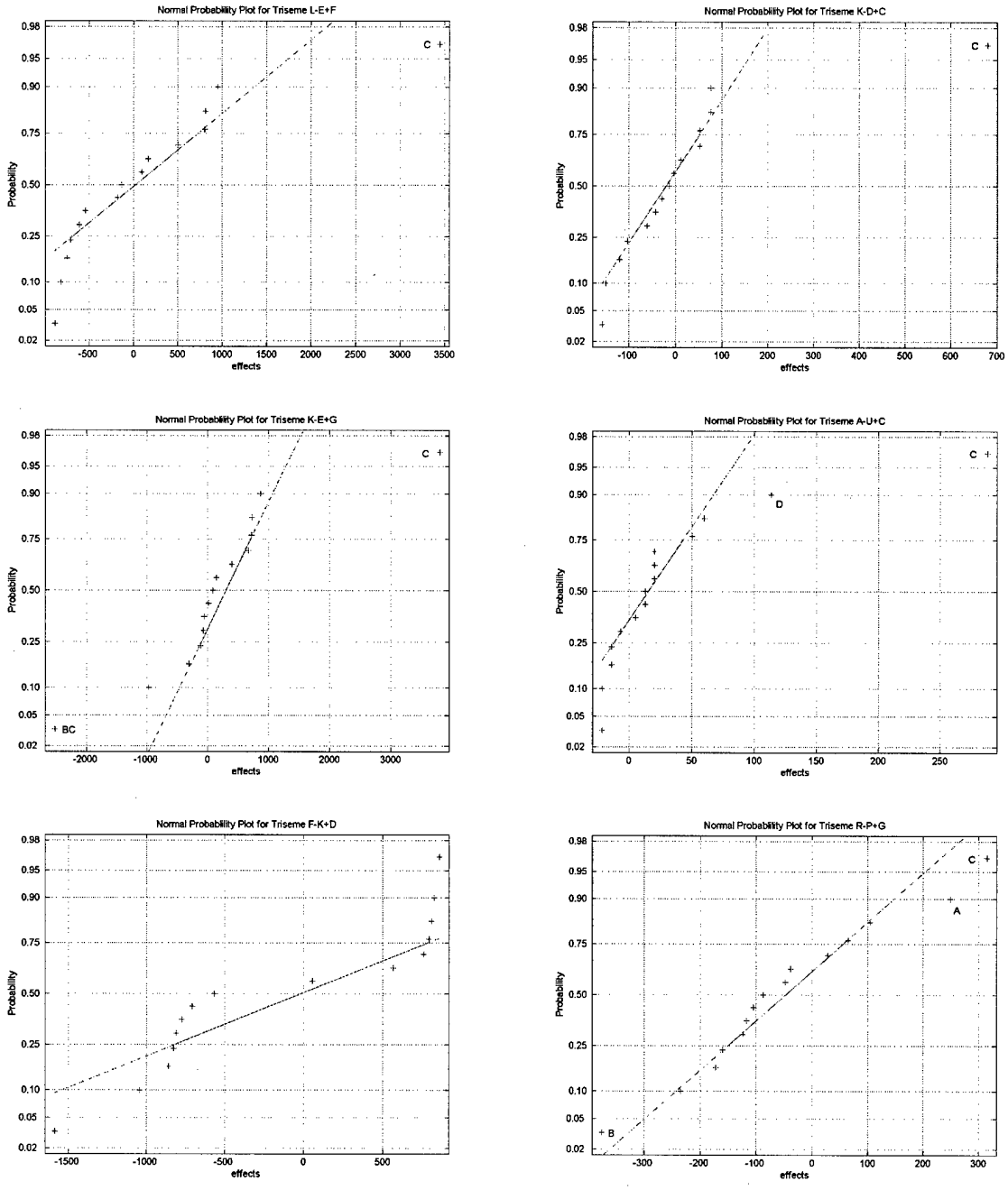


Figure 19. Normal probability plots for each triseme with the MSE totaled for each frame in the triseme.

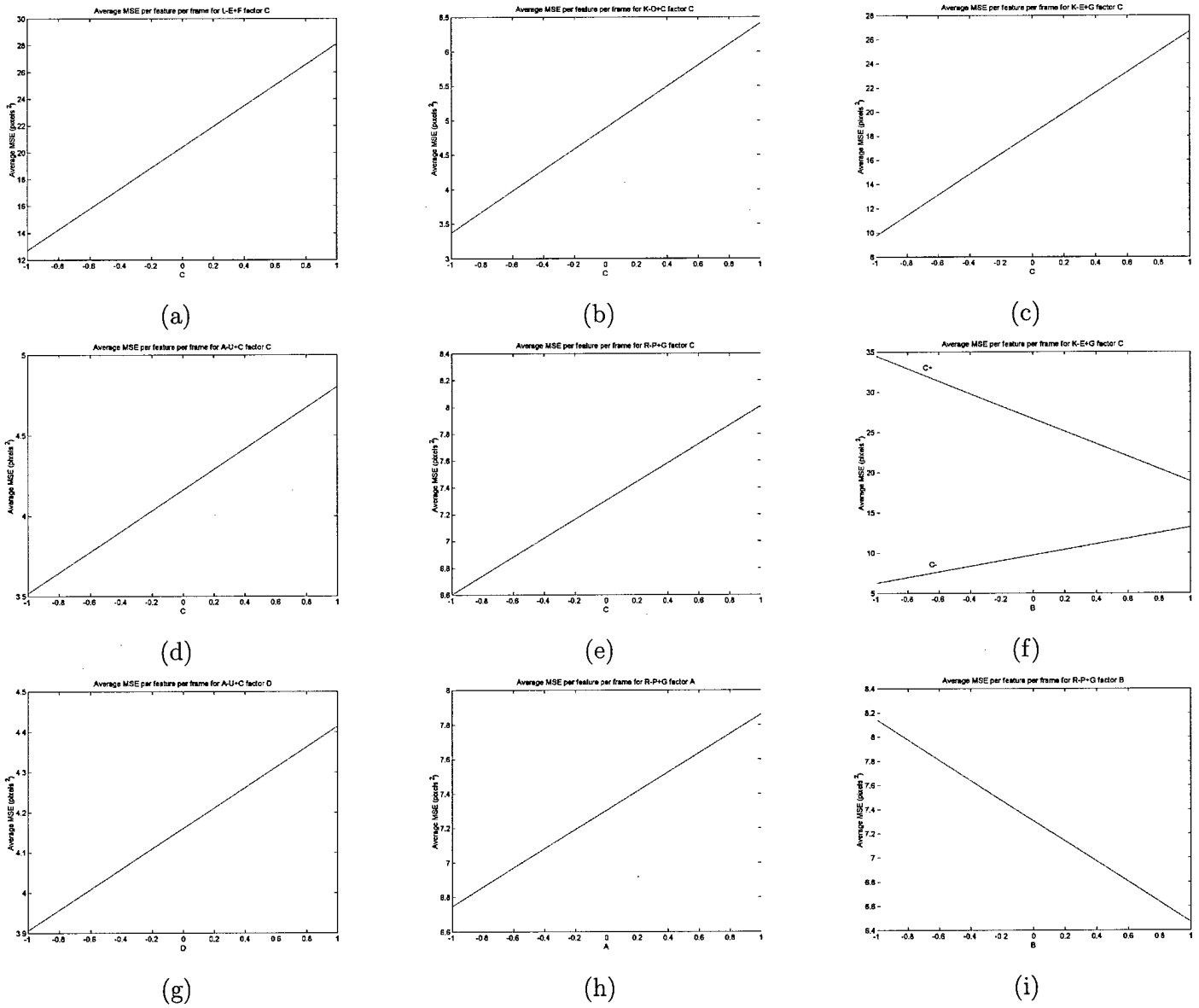


Figure 20. Main effect and interaction plots for each triseme using average MSE per feature per triseme.

Table 8. Analysis of Variance for a Four-Factor Factorial Design in a Randomized Complete Block

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0
Blocks	$\frac{1}{abcd}\sum_{m} y^2_{....m} - \frac{y^2_{....}}{abcdn}$	$n-1$	
A	$\frac{1}{bcdn}\sum_{i} y^2_{i....} - \frac{y^2_{....}}{abcdn}$	$a-1$	$\frac{MS_A}{MS_{ERROR}}$
B	$\frac{1}{acdn}\sum_{j} y^2_{.j...} - \frac{y^2_{....}}{abcdn}$	$b-1$	$\frac{MS_B}{MS_{ERROR}}$
AB	$\frac{1}{cdn}\sum_{i}\sum_{j} y^2_{ij...} - \frac{y^2_{....}}{abcdn} - SS_A - SS_B$	$(a-1)(b-1)$	$\frac{MS_{AB}}{MS_{ERROR}}$
C	$\frac{1}{abdn}\sum_{k} y^2_{..k..} - \frac{y^2_{....}}{abcdn}$	$c-1$	$\frac{MS_C}{MS_{ERROR}}$
AC	$\frac{1}{bdn}\sum_{i}\sum_{k} y^2_{i.k..} - \frac{y^2_{....}}{abcdn} - SS_A - SS_C$	$(a-1)(c-1)$	$\frac{MS_{AC}}{MS_{ERROR}}$
BC	$\frac{1}{adn}\sum_{j}\sum_{k} y^2_{.jk..} - \frac{y^2_{....}}{abcdn} - SS_B - SS_C$	$(b-1)(c-1)$	$\frac{MS_{BC}}{MS_{ERROR}}$
ABC	$\frac{1}{dn}\sum_{i}\sum_{j}\sum_{k} y^2_{ijk..} - \frac{y^2_{....}}{abcdn} - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC}$	$(a-1)(b-1)(c-1)$	$\frac{MS_{ABC}}{MS_{ERROR}}$
D	$\frac{1}{abcn}\sum_{l} y^2_{...l.} - \frac{y^2_{....}}{abcdn}$	$d-1$	$\frac{MS_D}{MS_{ERROR}}$
AD	$\frac{1}{bcn}\sum_{i}\sum_{l} y^2_{i..l.} - \frac{y^2_{....}}{abcdn} - SS_A - SS_D$	$(a-1)(d-1)$	$\frac{MS_{AD}}{MS_{ERROR}}$
BD	$\frac{1}{acn}\sum_{j}\sum_{l} y^2_{.j.l.} - \frac{y^2_{....}}{abcdn} - SS_B - SS_D$	$(b-1)(d-1)$	$\frac{MS_{BD}}{MS_{ERROR}}$
CD	$\frac{1}{abn}\sum_{k}\sum_{l} y^2_{..kl.} - \frac{y^2_{....}}{abcdn} - SS_C - SS_D$	$(c-1)(d-1)$	$\frac{MS_{CD}}{MS_{ERROR}}$
ABD	$\frac{1}{cn}\sum_{i}\sum_{j}\sum_{l} y^2_{ij.l.} - \frac{y^2_{....}}{abcdn} - SS_A - SS_B - SS_D - SS_{AB} - SS_{AD} - SS_{BD}$	$(a-1)(b-1)(d-1)$	$\frac{MS_{ABD}}{MS_{ERROR}}$
ACD	$\frac{1}{bn}\sum_{i}\sum_{k}\sum_{l} y^2_{i.kl.} - \frac{y^2_{....}}{abcdn} - SS_A - SS_C - SS_D - SS_{AC} - SS_{AD} - SS_{CD}$	$(a-1)(c-1)(d-1)$	$\frac{MS_{ACD}}{MS_{ERROR}}$
BCD	$\frac{1}{an}\sum_{j}\sum_{k}\sum_{l} y^2_{.jkl.} - \frac{y^2_{....}}{abcdn} - SS_B - SS_C - SS_D - SS_{BC} - SS_{BD} - SS_{CD}$	$(b-1)(c-1)(d-1)$	$\frac{MS_{BCD}}{MS_{ERROR}}$
ABCD	$\frac{1}{n}\sum_{i}\sum_{j}\sum_{k}\sum_{l} y^2_{ijkl.} - \frac{y^2_{....}}{abcdn} - SS_A - SS_B - SS_C - SS_D - SS_{AB} - SS_{AC} - SS_{AD} - SS_{BC} - SS_{BD} - SS_{CD} - SS_{ABC} - SS_{ABD} - SS_{BCD}$	$(a-1)(b-1)(c-1)(d-1)$	$\frac{MS_{ABCD}}{MS_{ERROR}}$
Error	$SS_{TOTAL} - SS_A - SS_B - SS_C - SS_D - SS_{AB} - SS_{AC} - SS_{AD} - SS_{BC} - SS_{BD} - SS_{CD} - SS_{ABC} - SS_{ABD} - SS_{BCD} - SS_{ABCD}$	$(abcd-1)(n-1)$	
Total	$\sum_{i}\sum_{j}\sum_{k}\sum_{l}\sum_{m} y^2_{ijklm} - \frac{y^2_{....}}{abcdn}$	$abcdn-1$	

In the analysis, the variable y_{ijklm} is the MSE for each feature in each frame of the m^{th} triseme at the appropriate levels of each factor averaged over all frames of the triseme. The tables summarizing the ANOVA for each feature and all plots used in this analysis are provided in Appendix D.

Using a threshold of $\alpha = 0.1$, features 3 and 19 have no statistically significant effects or interactions. Significant main effects (factors B and/or C) exist in all of the remaining 26 features except feature 2. The two-factor BC interaction is significant in features 9, 14, 18, 23, 24, 25, 27 and 28, whereas CD is significant in feature 28. Feature 2 contains only high order significant interactions ACD and ABCD.

As in the previous analysis, the main effect and interaction plots are appropriate for determining appropriate levels for the factors where significant effects or interactions exist. Figures 129 and 130 show the main effect plots for Factor B in the eleven features where it has a significant effect. All of the plots suggest that runs have a lower average MSE when models are trained using an outlier threshold of 20 than with a threshold of 10.

Consistent with previous analyses, the main effect plots in Appendix D indicate that runs with a one frame dynamic feature window result in lower average MSE for all 21 of the 28 features where factor C has a significant effect.

Interpreting the BC interaction plots reveals an interesting result. In the interaction plots (with the exception of the plot for feature 18), with factor C at the low level, runs with factor B at the low level result in lower or the same average MSE than those with factor B at the high level and the difference in average MSE between the two levels for feature 18 is extremely small. This result conflicts with the earlier analysis of the main effect plots for factor B. The probable cause of the discrepancy is that the main effect plots include averaging of results of runs with factor C at the high level. The BC interaction plots show that runs with factor C at the high level and B at the low level result in a much higher average MSE than the other combinations. Averaging these runs into the main effect analysis raises the average MSE for factor B at the low level above the average MSE at the high level.

The other significant two-level interaction is the CD interaction in feature 28. For this feature, runs with both factors C and D at the low level (one frame dynamic feature window and a no skip HMM) result in lower average MSE.

Finally, feature 2 has the significant high order interactions ACD and ABCD. The interaction plots indicate that for feature 2 (inner mouth width), runs with factors A, B, and C at the low level and D at the high level (a skip-state HMM) have lower average MSE.

In summary, this section presents a look at the synthesis results and attempts to compensate for errors due to differences between trisemes. Investigation into the significant main effects and interactions shows that for features other than feature 2, a parameter combination of an outlier threshold of 10, a one frame dynamic feature window, and a no skip HMM generally results in lower average MSE. Even the MSE for this combination in feature 2 with factor A at the high level is only 2 pixels^2 higher than the best combination.

4.4.2 Average feature analysis with blocking. In the analysis in this section, the variable y_{ijklm} is the average of the squared error for each feature in each frame of the m^{th} triseme, at the appropriate levels of each factor, averaged over all frames of the triseme. Table 9 summarizes the ANOVA for this scenario.

The significance level indicated by α shows one main effect significant at the 0.0002 level (Delta Window Length), and one interaction effect significant at the 0.0527 level (Delta Window Length coupled with outlier threshold). The main effect B (outlier threshold) is significant at the 0.0928 level. These are the same three factors/interactions that are significant in most of the features in the previous analysis.

Figure 21 shows the main effect plots for factors B and C as well as the BC interaction plot. As with the individual feature analysis, the main effect plot for factor B indicates that runs with B at the high level result in lower error. However, the BC interaction plot again shows that this results from the inclusion of runs with factor C at the high level in the computation of the main effects for factor B. The interpretation of the plots is as before: experimental runs with factors B and C at the low level generally result in lower average

Table 9. ANOVA Summary for Pooled Feature Analysis.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	3.367×10^6	5		
A	418.96	1	0.0126	0.9110
B	9.653×10^4	1	2.8990	0.0928
AB	1.646×10^4	1	0.4943	0.4842
C	5.0180×10^5	1	15.0713	0.0002
AC	4.5328×10^4	1	1.3623	0.2468
BC	1.2907×10^5	1	3.8766	0.0527
ABC	445.6971	1	0.0134	0.9082
D	3.3995×10^4	1	1.0210	0.3155
AD	2.5408×10^3	1	0.0763	0.7831
BD	597.9185	1	0.0180	0.8938
CD	3.2679×10^4	1	0.9815	0.3250
ABD	267.3894	1	0.0080	0.9288
ACD	1.9673×10^3	1	0.0591	0.8086
BCD	28.4019	1	0.0009	0.9768
ABCD	2.2236×10^3	1	0.0668	0.7968
Error	2.4971×10^6	90		
Total	6.7261×10^6	95		

MSE than the other runs. It is not possible from this analysis to determine an appropriate level for factors A and D since there are no statistically significant effects or interactions.

4.5 Mouthshapes

The previous sections provided an in-depth statistical analysis of the performance of the synthesis algorithm using average MSE as the measured variable. Another method of analyzing performance is to simply look at the mouth outlines formed from the synthetic features compared to the original outlines. This section presents these comparisons using the factor combinations shown in the previous analyses to lower the error. Initially, the six trisemes used in the statistical analysis are presented followed by other examples of those six trisemes plus eight trisemes found four times each in the remaining 47 recorded sentence not to be in the 49 training sentences.

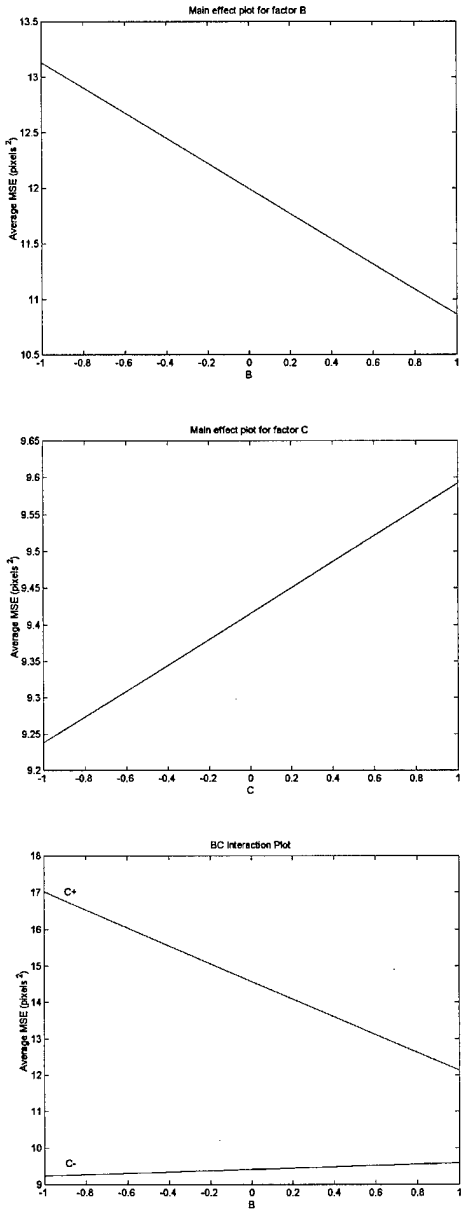


Figure 21. Main effect plots for factors B and C and the BC interaction plot for analysis with the total MSE per triseme accounting for blocking effects.

4.5.1 The Six "Analysis" Trisemes. In all four statistical analyses, it is clear that the average MSE is lower for features synthesized with a one frame dynamic window length. Likewise, a no-skip HMM prevails as the HMM structure that produces lower MSE in the majority of cases. The comparisons presented here use features synthesized using this configuration for these two factors.

Previous analysis for the outlier threshold level suggests that in three of the four cases (the individual feature analysis by triseme is the exception) that an outlier threshold of 10 results in a lower average MSE. Since individual feature analysis by triseme in Section 4.3.1 suggested a level of 20, both results are presented here.

Again, three of the four analyses (the average feature analysis by triseme is the exception) indicates that either level of the log-likelihood threshold could be used. The average feature analysis by triseme in Section 4.3.2 suggests using the lower level of 20 to reduce the average MSE. Since there is a discrepancy, both levels are presented here.

Thus, there are four combinations of possible factor levels. Figures 22-26 show the resultant mouth outlines for all synthesis factors set to their low levels. The larger views are intended to show the comparisons in more detail with the tooth locations included. The small images are versions of the corresponding frames scaled to the size they would appear in the original video in a 320×240 pixel image on a 19 inch monitor set to a resolution of 1024×768 pixels per square inch. The tooth locations are omitted from the scaled versions because the symbols obscured each other and an accurate depiction of the locations was not possible. Figures 27-31 show these outlines for the cases where factor A (log-Likelihood threshold) is at the high level (100) and the others are at the low level. Figures 32-36 are for the cases with factor A at the low level and factor B (outlier threshold) at the high level (20), and Figures 37-41 illustrate the results when both factors are at the high levels and factors C and D are low.

It appears that the outlines formed from features synthesized with a log-likelihood threshold of 100 are the same as (or closer to the original outlines than those synthesized with a threshold of 20. Also, with an outlier threshold of 20 the triseme K-E+G outlines are

much different from the originals compared to those for a threshold of 10. All other trisemes showed little to no difference in outlines for the two outlier threshold levels.

In summary, the mouth shape outlines synthesized with a log-likelihood threshold of 100, outlier threshold of 10, a one frame dynamic feature window, and a no-skip HMM overlayed with the original mouth shape outlines produces a closer approximation to the original mouth shapes than other factor configurations. This result supports the conclusions of the statistical analysis.

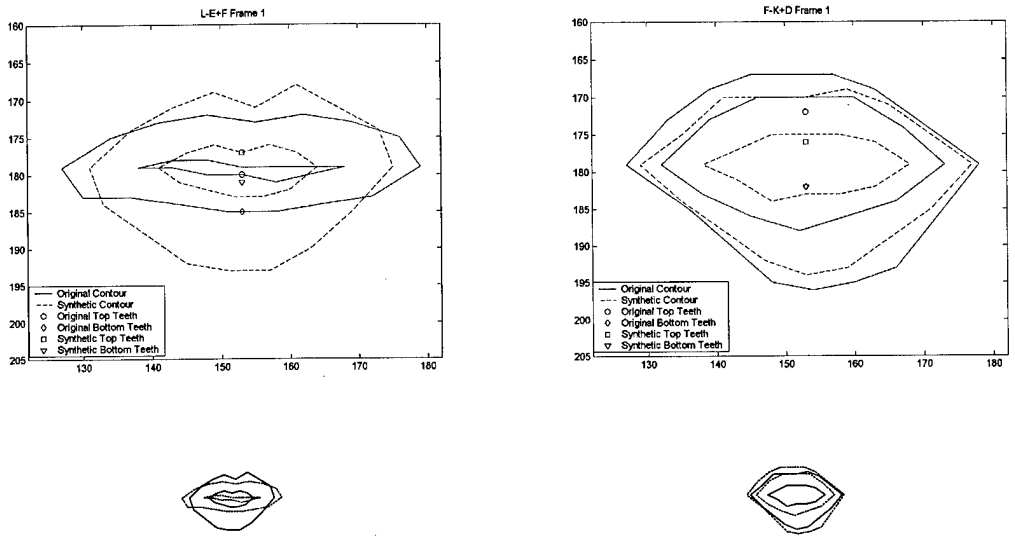


Figure 22. L-E+F (Left) and F-K+D (Right) $T = 20, R = 10, \Theta = 1$, No Skip HMM.

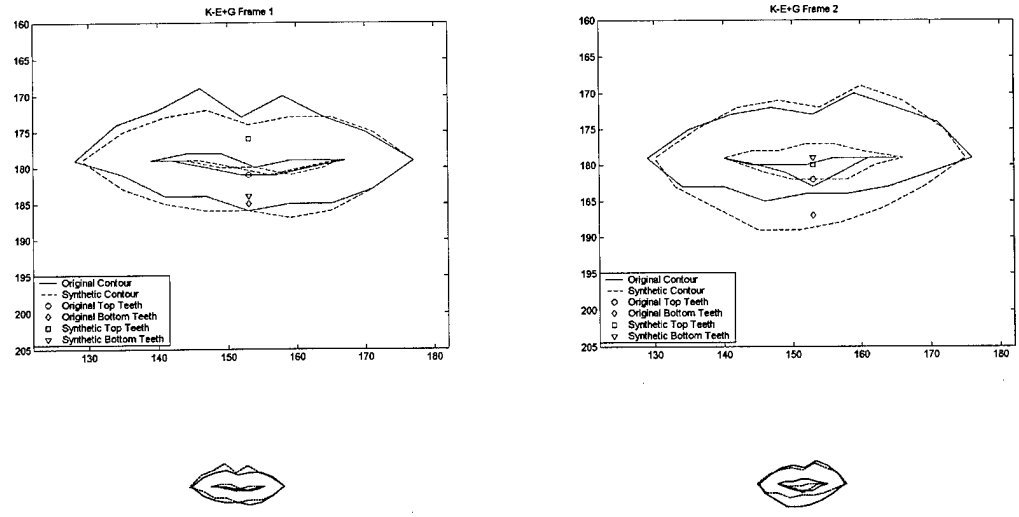


Figure 23. K-E+G $T = 20, R = 10, \Theta = 1$, No Skip HMM.

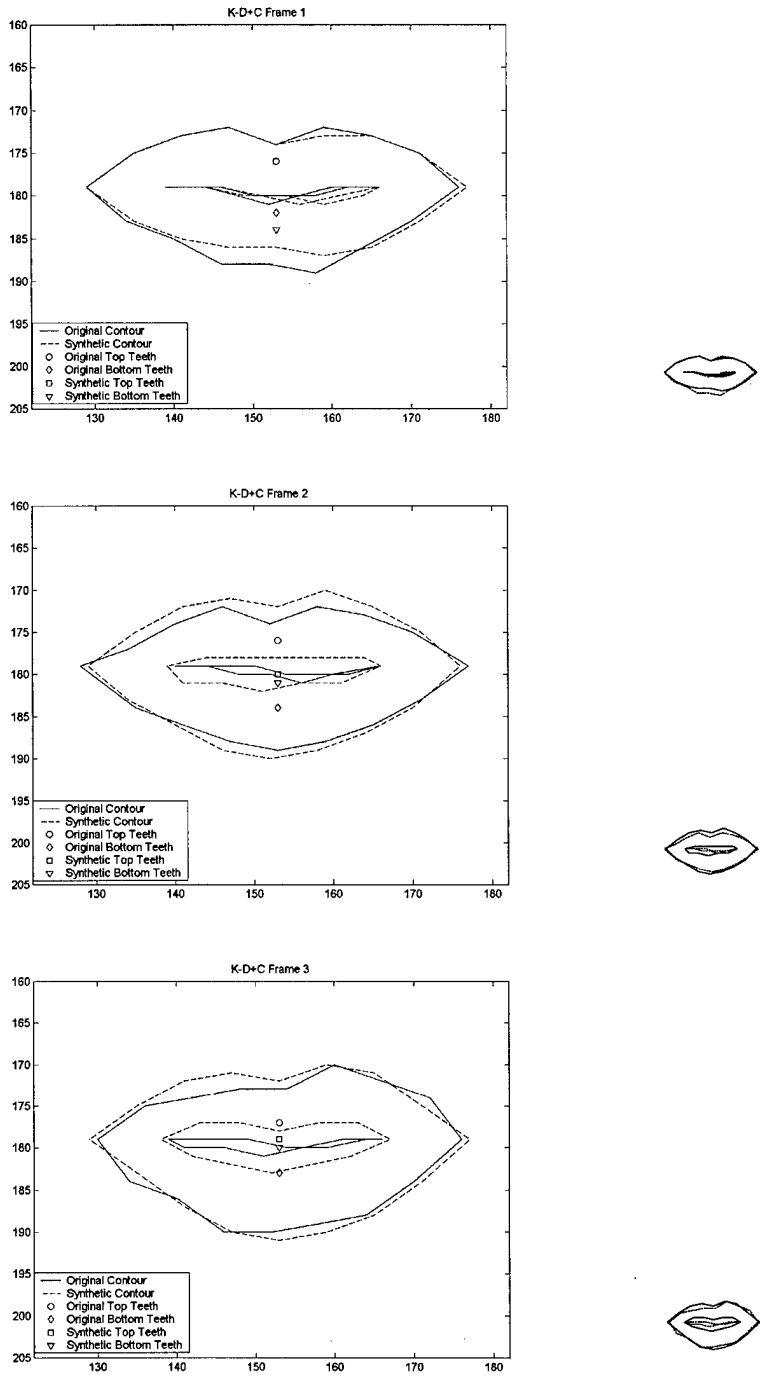
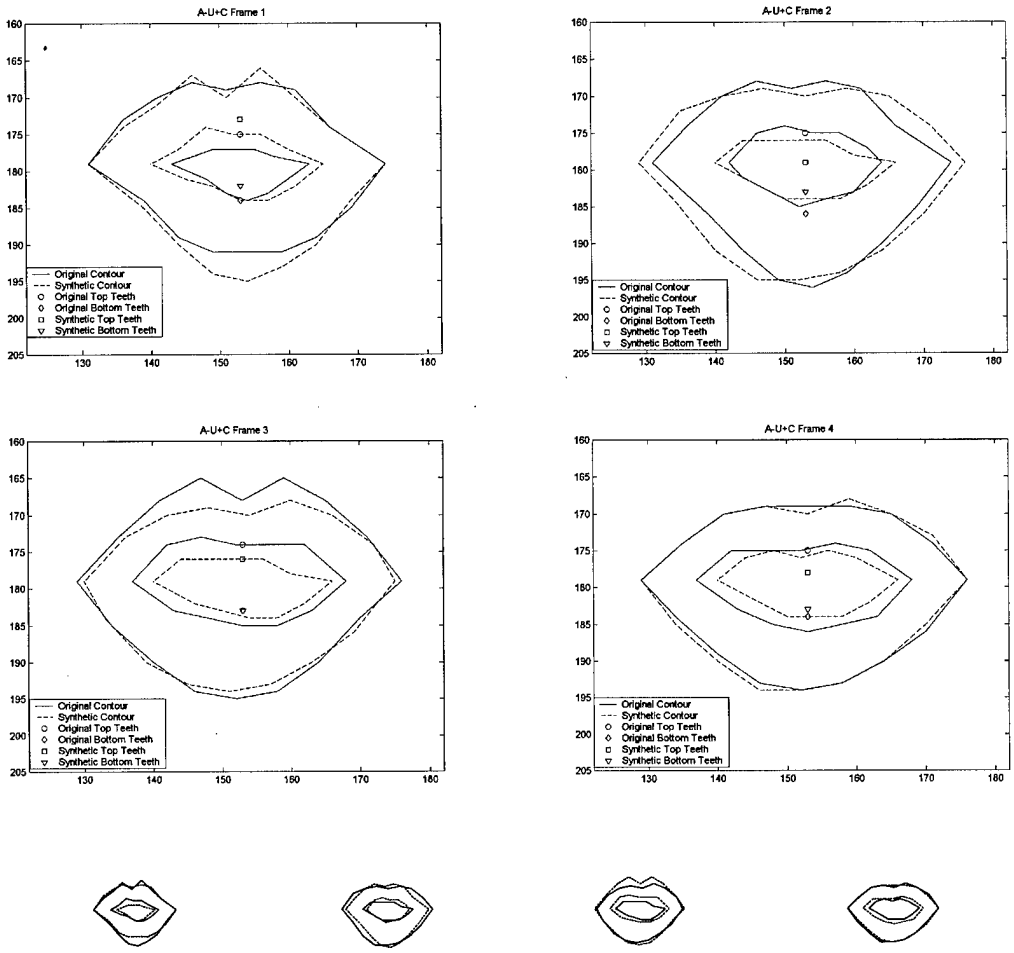
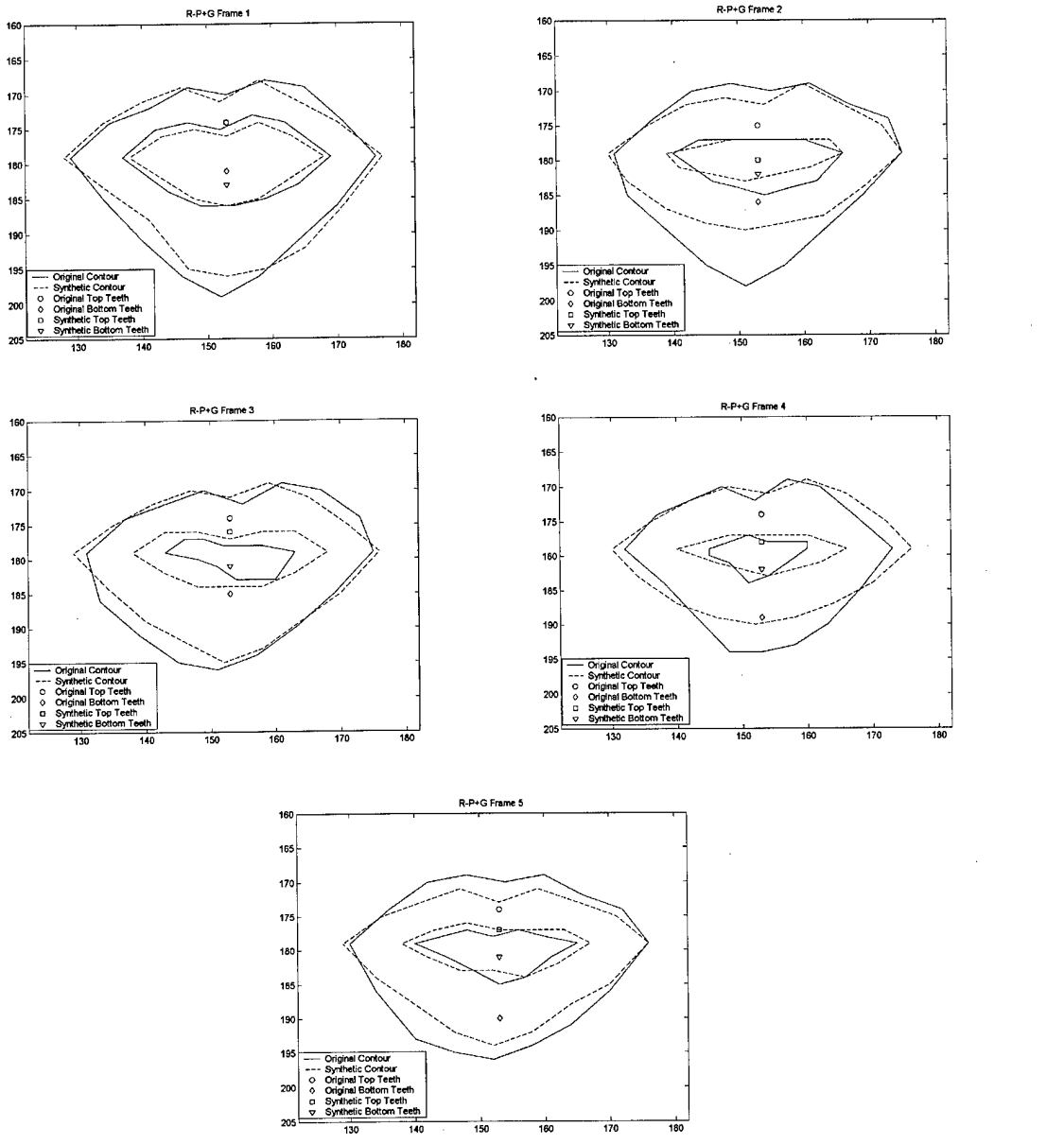


Figure 24. K-D+C $T = 20, R = 10, \Theta = 1$, No Skip HMM.



(1) (2) (3) (4)

Figure 25. A-U+C $T = 20, R = 10, \Theta = 1$, No Skip HMM.



(1) (2) (3) (4) (5)

Figure 26. R-P+G $T = 20, R = 10, \Theta = 1$, No Skip HMM.

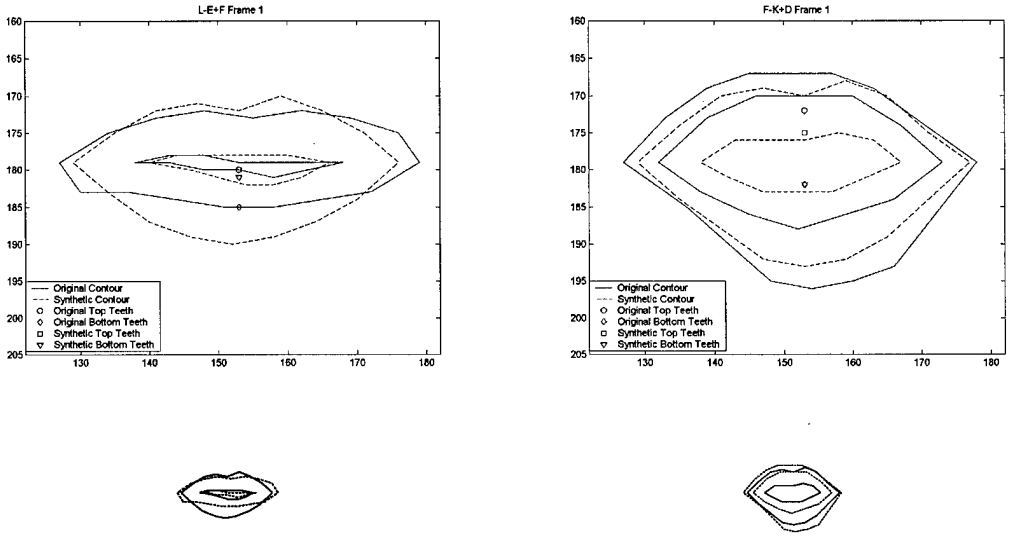


Figure 27. L-E+F (Left) and F-K+D (Right) $T = 100, R = 10, \Theta = 1$, No Skip HMM.

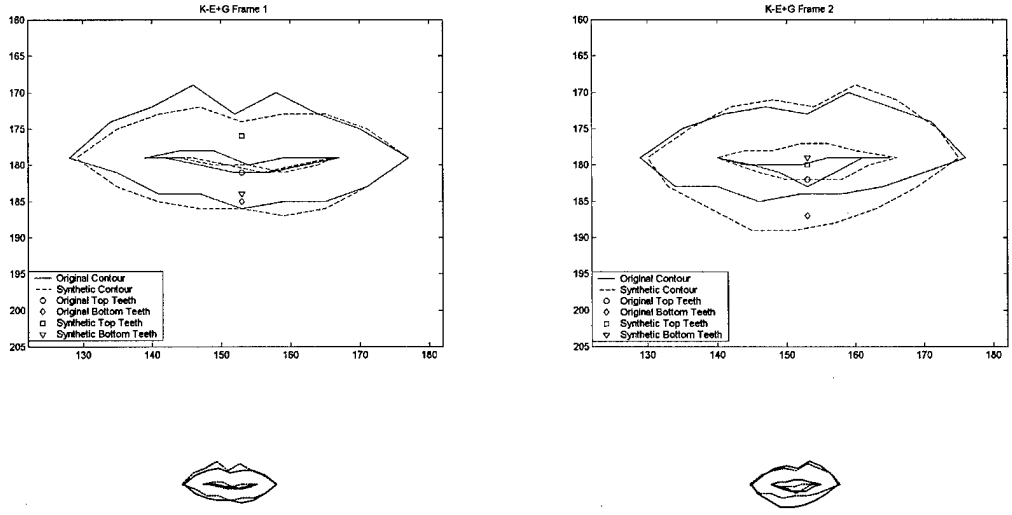


Figure 28. K-E+G $T = 100, R = 10, \Theta = 1$, No Skip HMM.

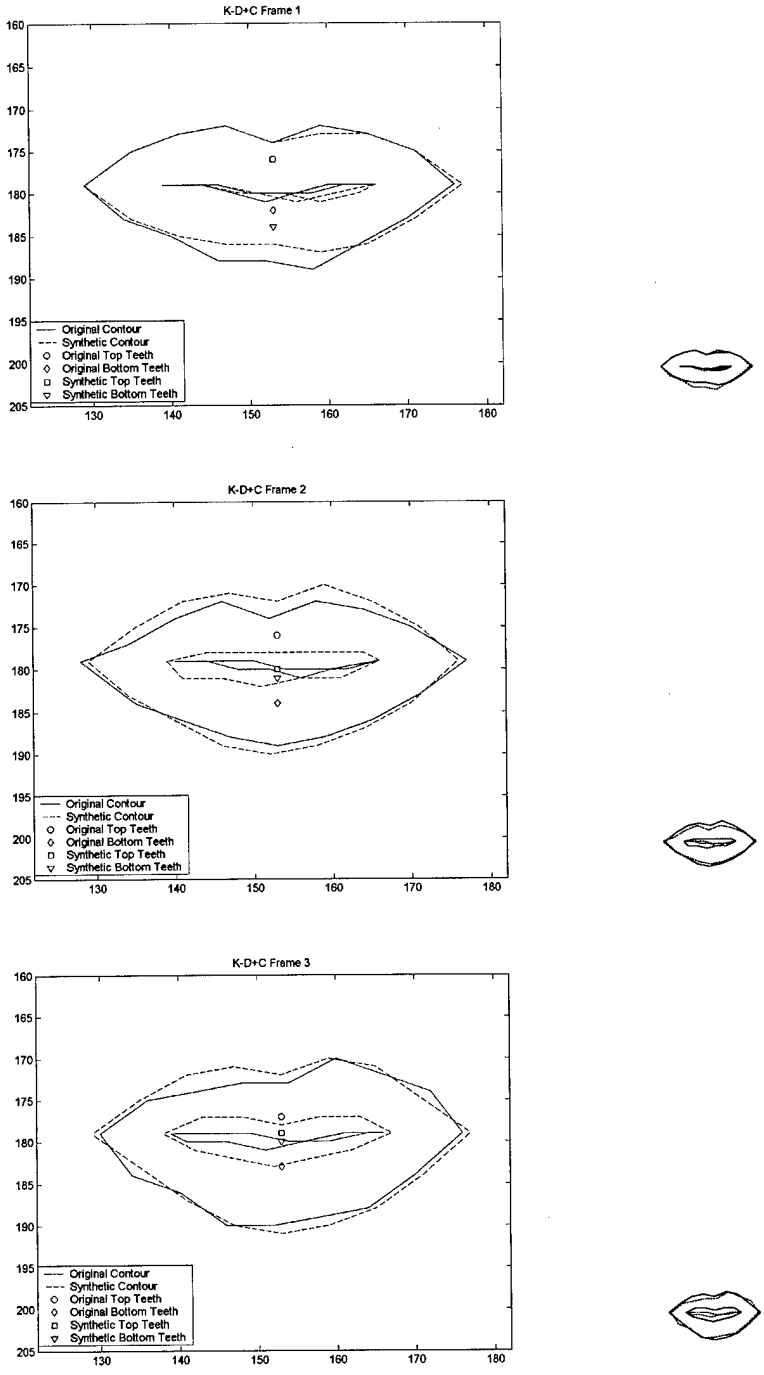
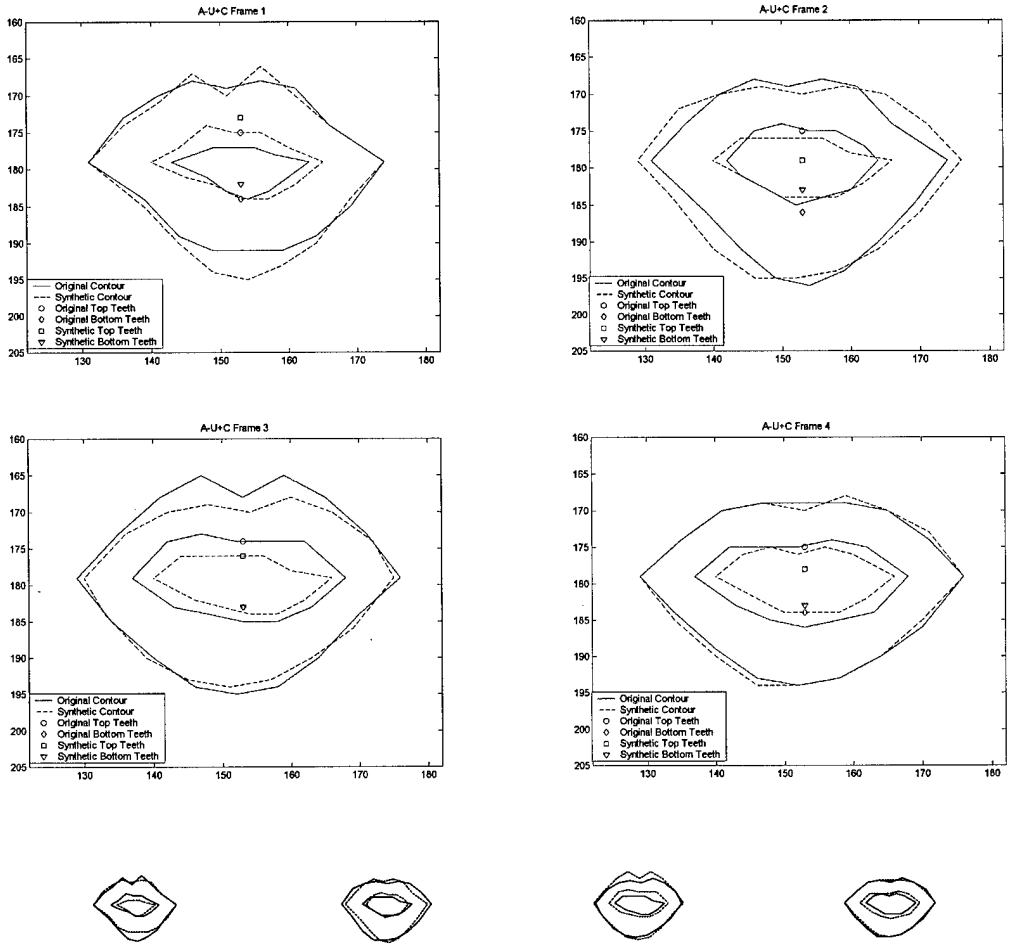
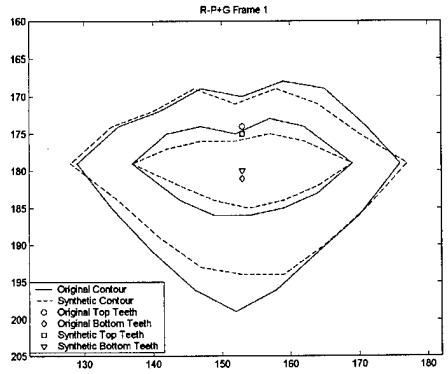


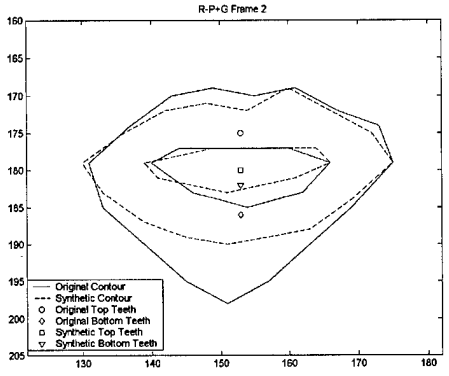
Figure 29. K-D+C $T = 100, R = 10, \Theta = 1$, No Skip HMM.

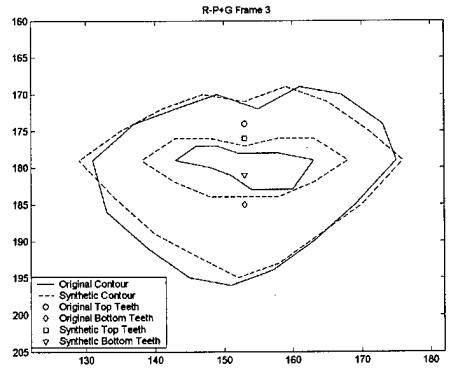


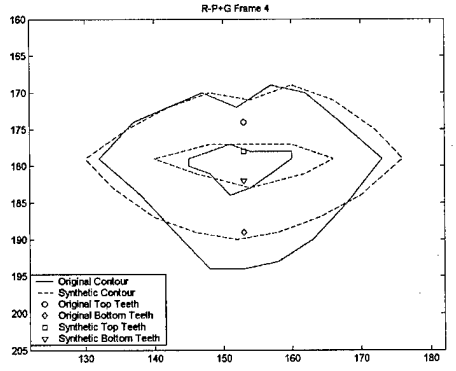
(1) (2) (3) (4)

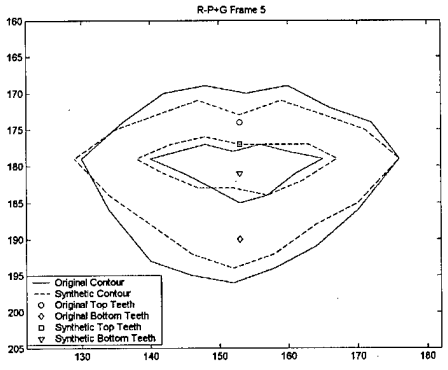
Figure 30. A-U+C $T = 100, R = 10, \Theta = 1$, No Skip HMM.











(1) (2) (3) (4) (5)

Figure 31. R-P+G $T = 100, R = 10, \Theta = 1$, No Skip HMM.

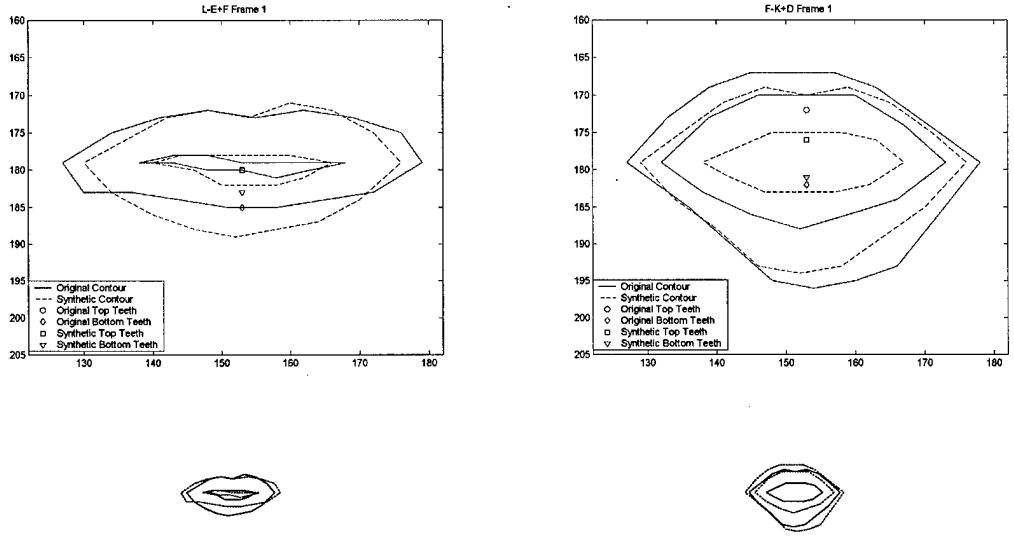


Figure 32. L-E+F (Left) and F-K+D (Right) $T = 20, R = 20, \Theta = 1$, No Skip HMM.

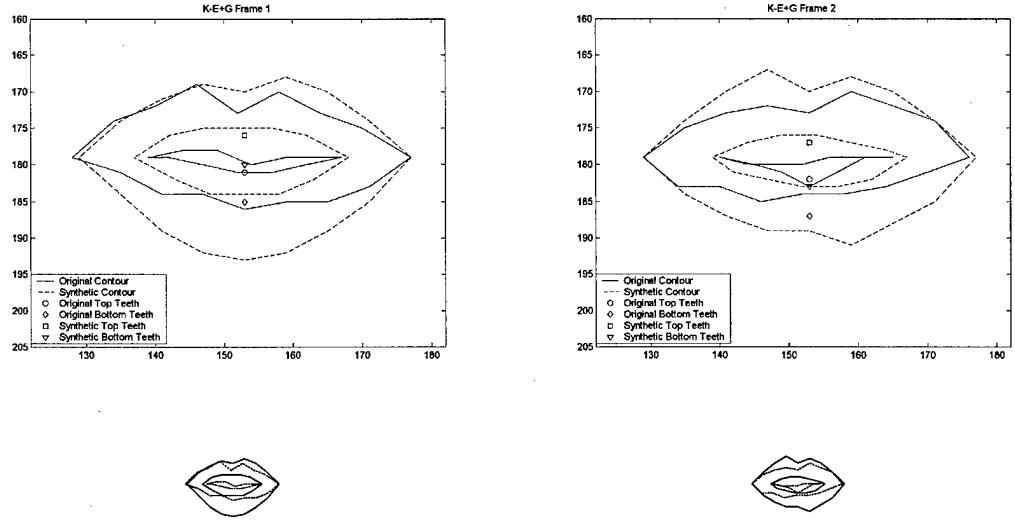


Figure 33. K-E+G $T = 20, R = 20, \Theta = 1$, No Skip HMM.

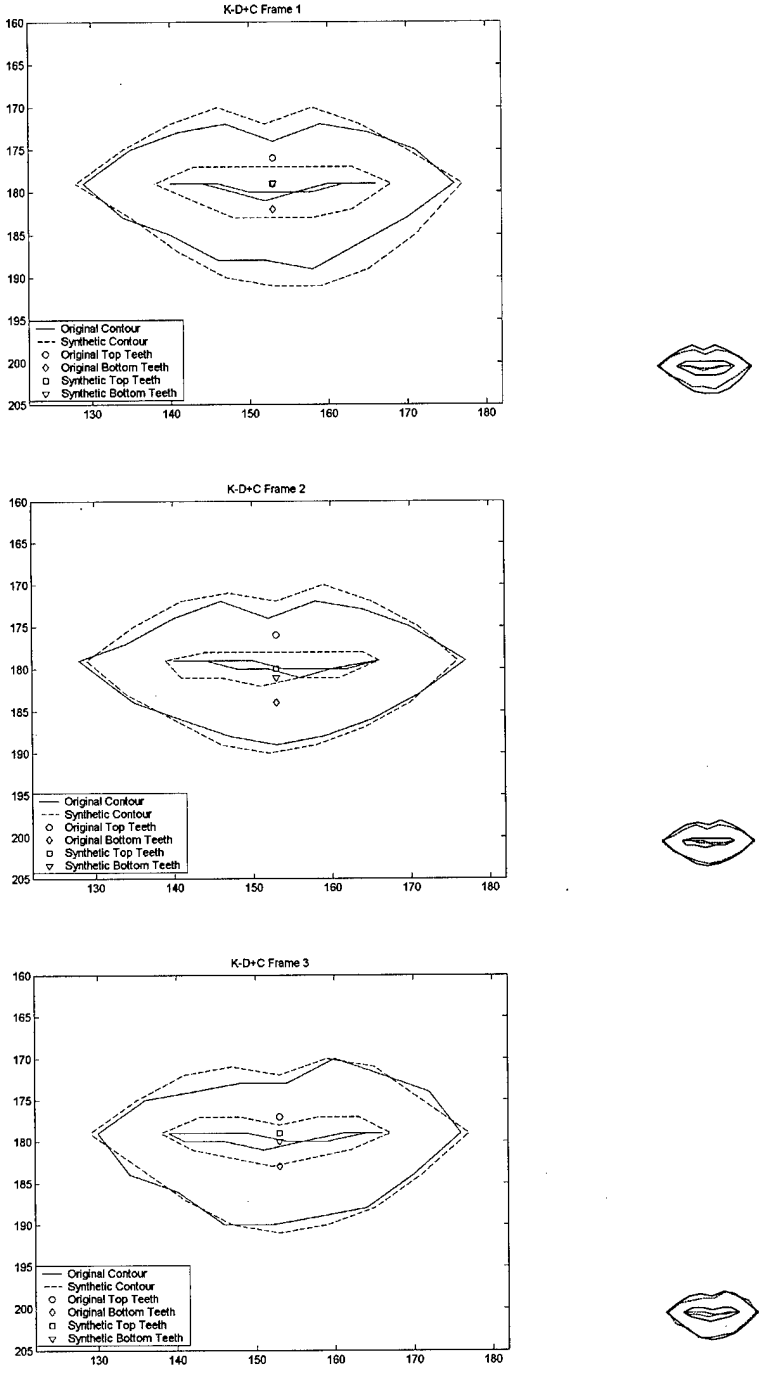
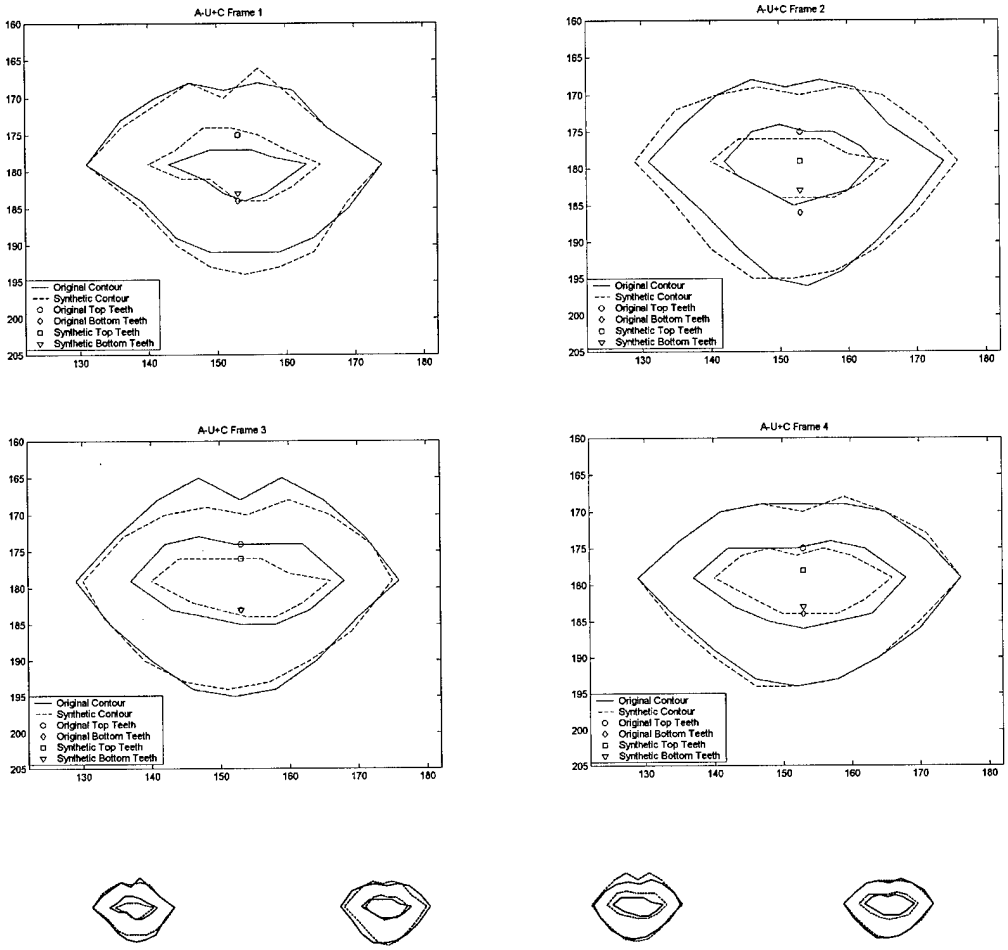
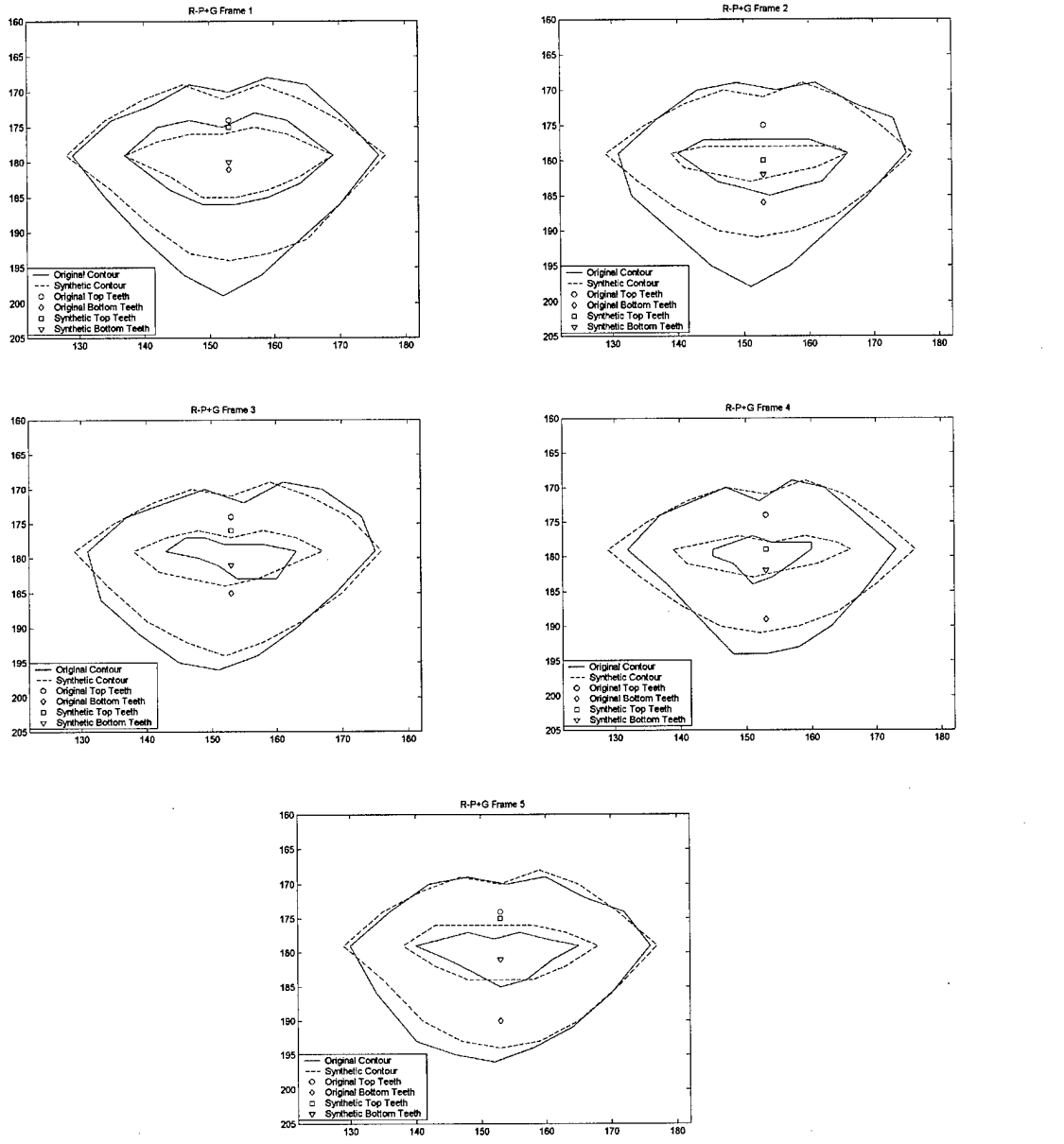


Figure 34. K-D+C $T = 20, R = 20, \Theta = 1$, No Skip HMM.



(1) (2) (3) (4)

Figure 35. A-U+C $T = 20, R = 20, \Theta = 1$, No Skip HMM.



(1) (2) (3) (4) (5)

Figure 36. R-P+G $T = 20, R = 20, \Theta = 1$, No Skip HMM.

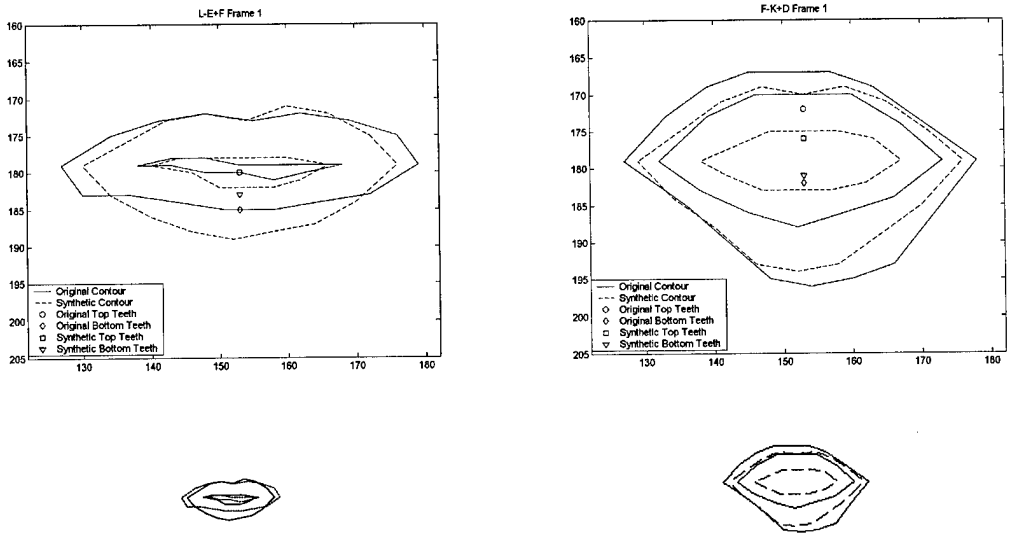


Figure 37. L-E+F (Left) and F-K+D (Right) $T = 100, R = 20, \Theta = 1$, No Skip HMM.

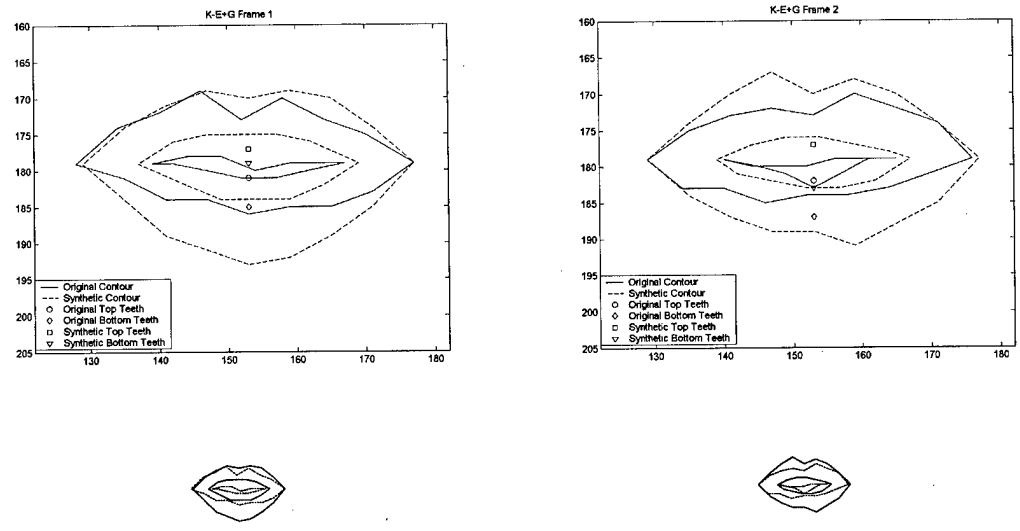


Figure 38. K-E+G $T = 100, R = 20, \Theta = 1$, No Skip HMM.

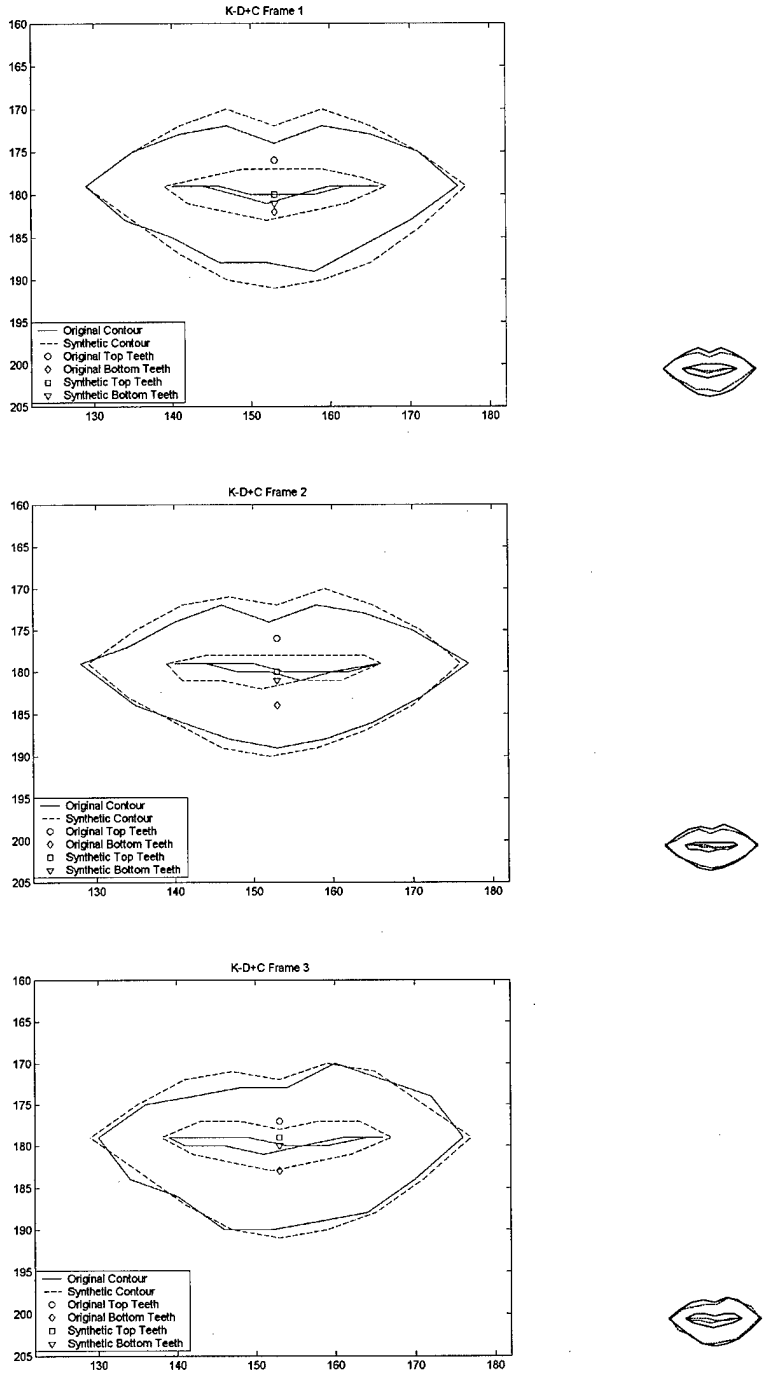
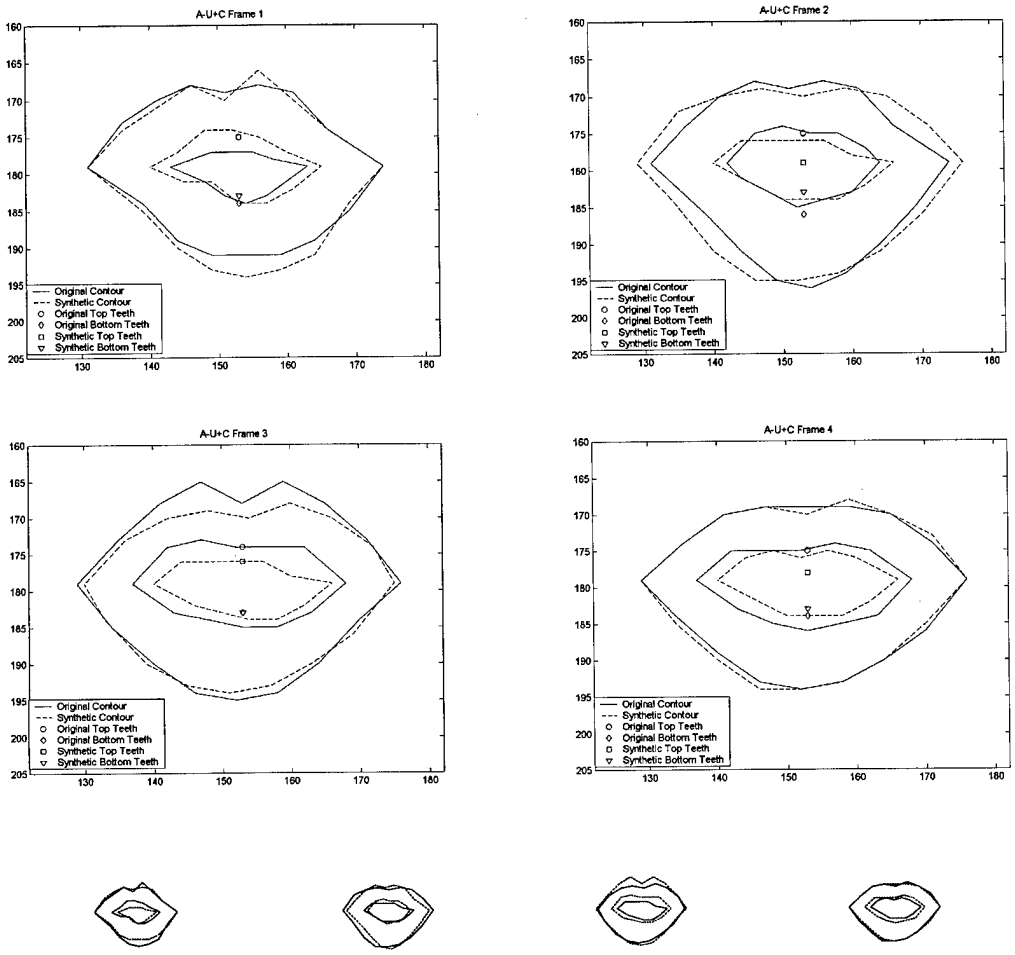
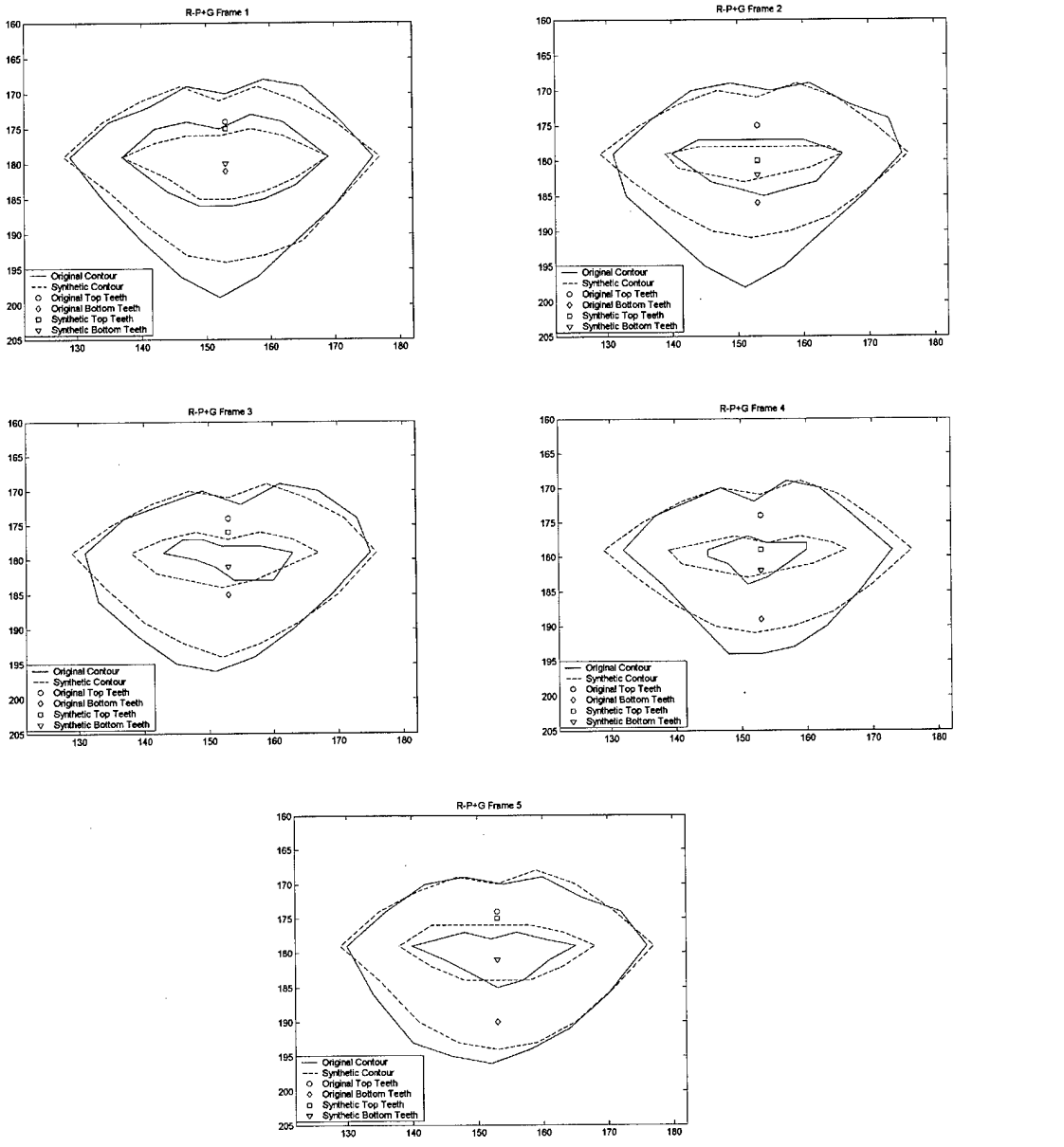


Figure 39. K-D+C $T = 100, R = 20, \Theta = 1$, No Skip HMM.



(1) (2) (3) (4)

Figure 40. A-U+C $T = 100, R = 20, \Theta = 1$, No Skip HMM.



(1) (2) (3) (4) (5)

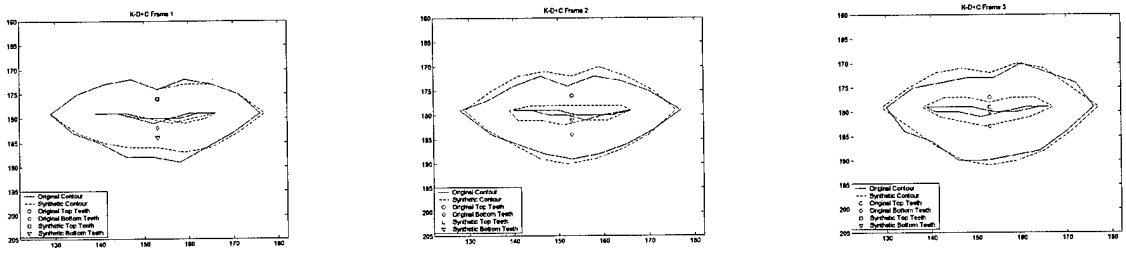
Figure 41. R-P+G $T = 100, R = 20, \Theta = 1$, No Skip HMM.

4.5.2 Additional Examples. To provide a broader sampling of performance examples, this section presents mouth outlines of additional examples of three of the six trisemes and four examples each of eight trisemes found in the remaining 47 recorded prompts that did not occur in the 49 sentence training set. All synthetic features were generated from a no-skip HMM with a log-likelihood threshold of 100, outlier threshold of 10, and a one frame dynamic feature window length. These parameter settings were chosen as a result of the conclusions of the previous section.

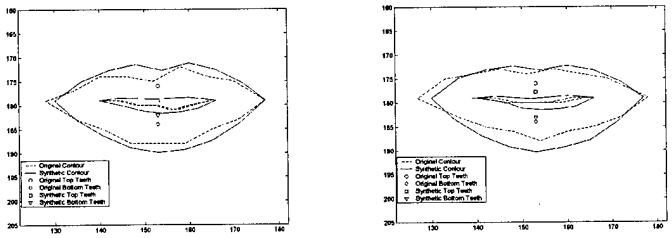
Figures 42-44 show the mouth shapes from the previous analysis along with additional examples from either the same test sentence (sx212) or from other sentences. The additional examples support the conclusion that the technique works well for generation of synthetic control points when the trisemes do not exist in the training data. The mouth shapes formed from the synthetic control points are very similar to those formed from the originally estimated control points. The worst examples are from triseme K-D+C where the mouth should be closed. The synthetic control points indicate an open mouth. Once possible cause for this is in the phonetic transcript produced by "Aligner". "Aligner" sets the beginning of the phoneme to when sound is being produced, not necessarily when the mouth is actually closed. Yet, it is clear that in most cases, the technique produces accurate control point estimates for these additional examples.

Figures 45-55 show four examples each of eight additional trisemes not occurring in the training data. Nearly all figures indicate the algorithm produced control point estimates whose resulting mouth shapes are nearly indistinguishable from those from the manually reviewed control points, especially at the smaller scale more likely to represent the scale a person would view the video at.

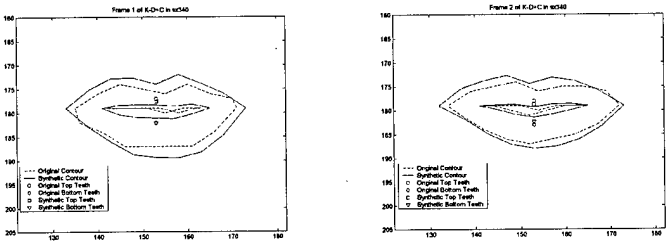
One exception is again from a triseme where the mouth should have been closed (B-D+V). Figures 46 and 47 show that the mouth shapes formed from synthetic control point estimates have an open mouth. Yet, the small scale pictures in 47 show that it is still difficult to see a large error in the mouth shape comparisons.



sx212 First Occurrence "eh-m+s"



sx212 Second Occurrence "eh-b+d"

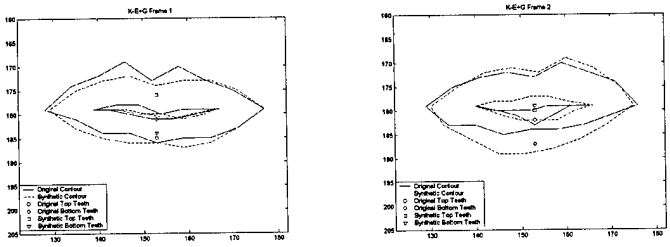


sx340 "eh-v+r"

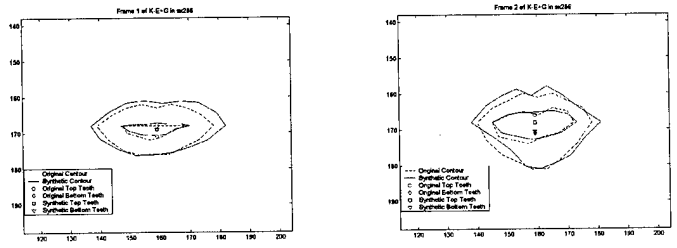
sx212 First Occurrence

sx212 Second Occurrence sx340

Figure 42. K-D+C Magnified and at Original Scale.



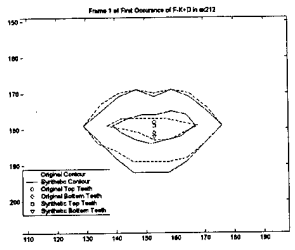
sx212 "eh-v+r"



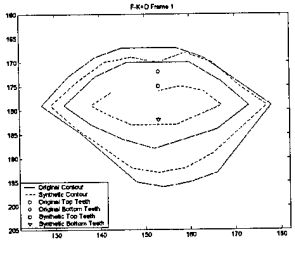
sx286 "eh-v+r"

sx212 sx286

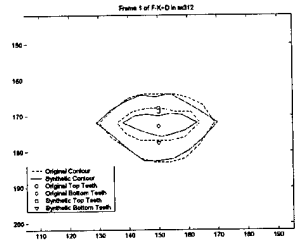
Figure 43. K-E+G Magnified and at Original Scale.

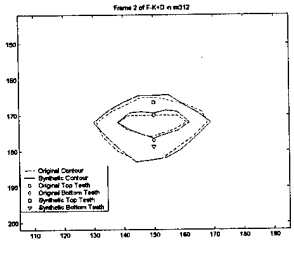


sx212 First Occurrence "th-eh+m"

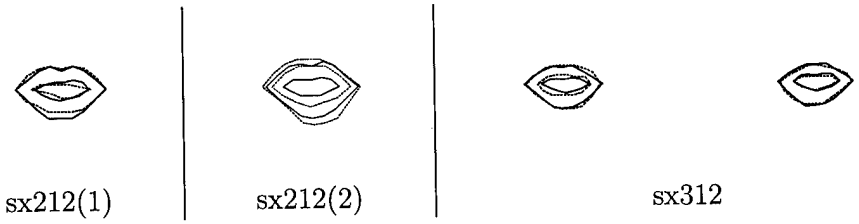


sx212 Second Occurrence "th-eh+m"





sx312 "th-eh+m"



sx212(1) sx212(2) sx312

Figure 44. F-K+D Magnified and at Original Scale.

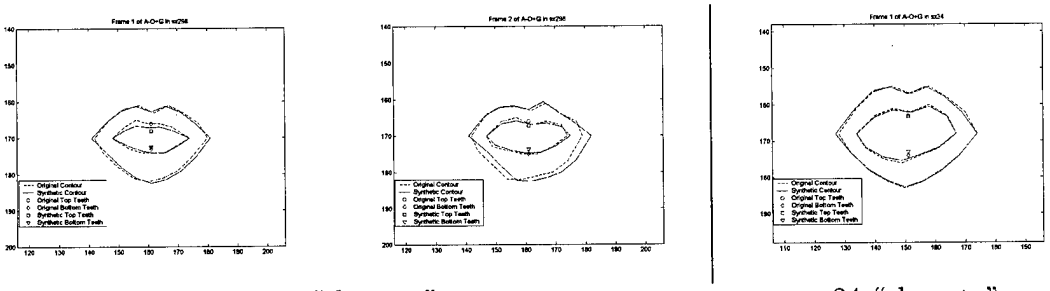
In summary, these additional examples show that decision tree-based HMMs do estimate accurate control points for trisemes not occurring in the training data by presenting mouth shape outlines for a wide array of trisemes and corresponding lip positions.

4.6 Chapter Summary

This chapter presents a detailed statistical analysis of the controllable factors in the estimation of visual speech synthesis features using HMMs. One observation is that the performance is slightly different from triseme to triseme. This result is likely due to differences in the amount of training data used to estimate the HMM as well as contextual differences. There are also differences in the level of performance between features. However, looking more generally at the statistical analysis combined with the resultant mouth outlines indicates that an appropriate combination of algorithm parameters lowers the average MSE in most cases and produces feature estimates closer to the desired feature. A larger log-likelihood threshold coupled with a smaller outlier threshold and dynamic feature window and a no-skip HMM has the potential to produce better estimates than other combinations of factors.

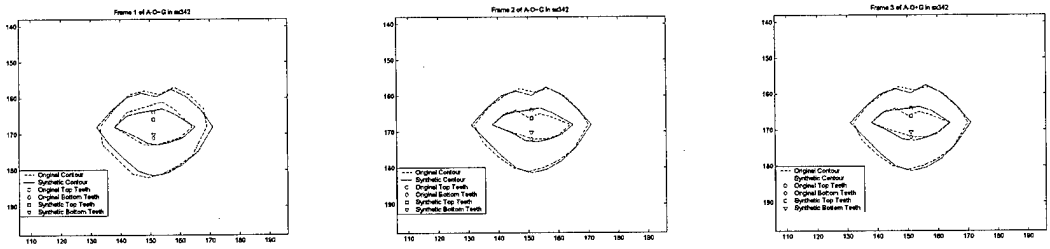
Mouth outline examples from an array of trisemes missing from the training data indicate the algorithm works extremely well in estimating accurate control points.

While this research provides insight into the performance of the algorithm and the appropriate levels for the controllable algorithm parameters, much additional research is needed in this area. Research should be accomplished into alternative features, potentially more visually relevant decision tree questions and additional video subjects and facial orientations. The next chapter concludes by summarizing the contributions reported here and by presenting new questions raised and potential areas for future work.

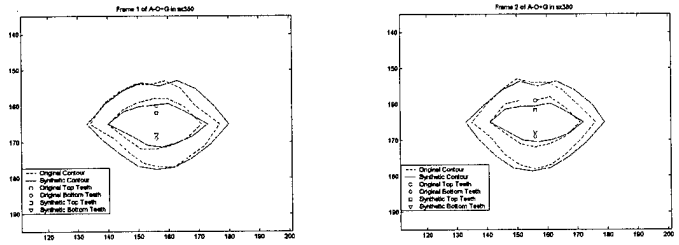


sx298 "ch-aa+r"

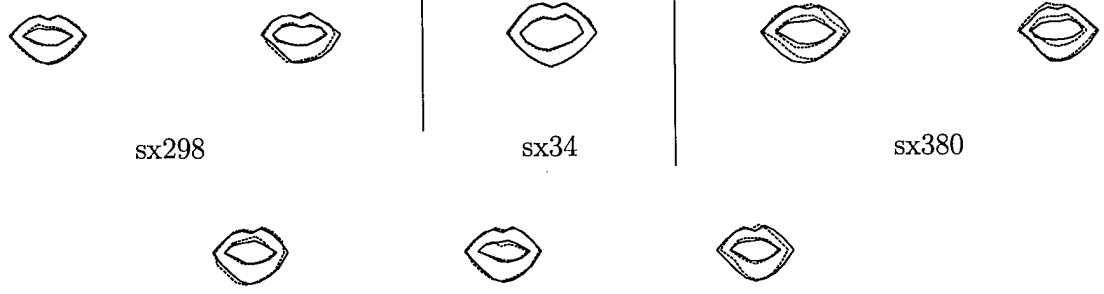
sx34 "ch-aa+r"



sx342 "ch-aa+r"



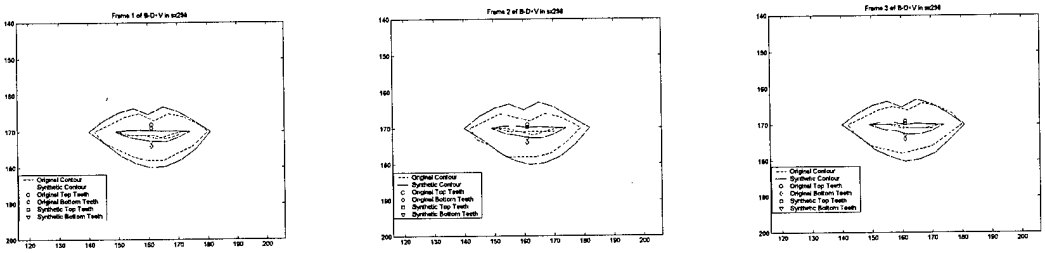
sx380 "ch-aa+r"



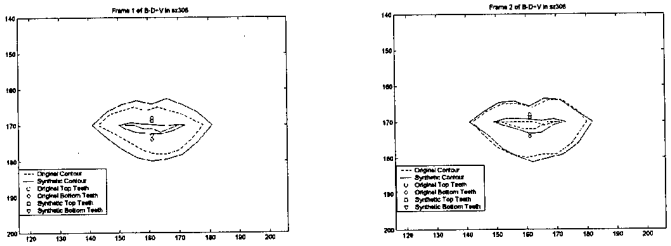
sx298 sx34 sx380

sx342

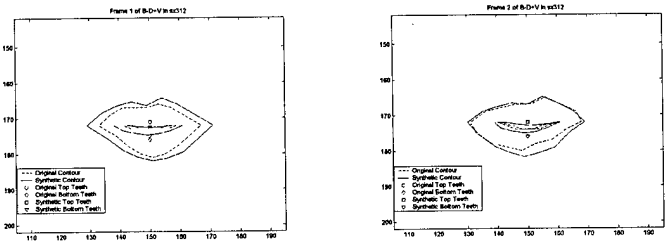
Figure 45. A-O+G Magnified and at Original Scale.



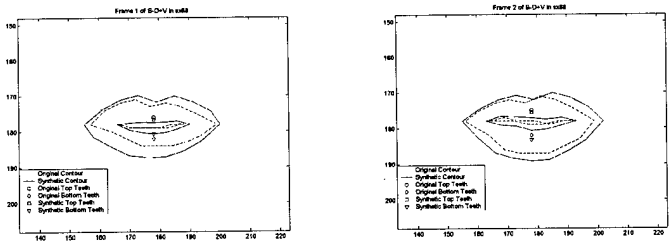
sx298 "n-b+iy"



sx306 "l-b+iy"



sx312 "l-b+iy"



sx88 "n-b+iy"

Figure 46. B-D+V Magnified.

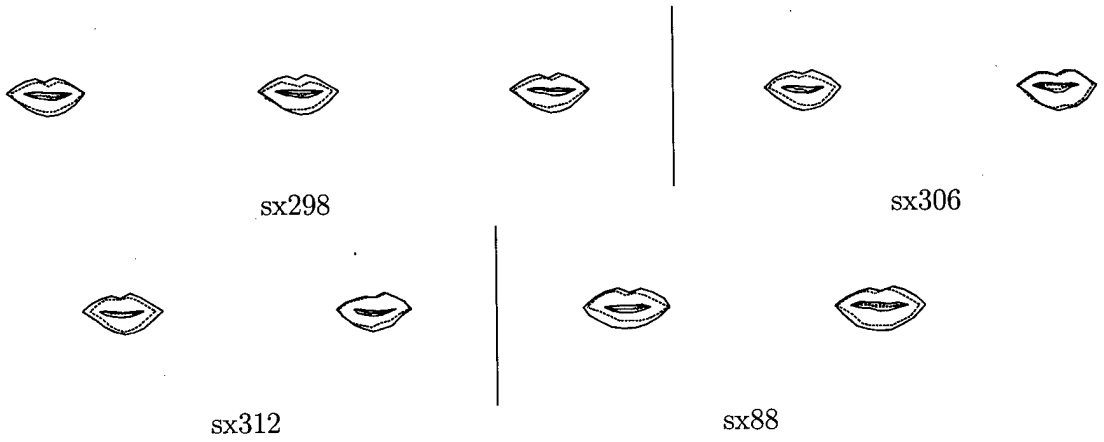
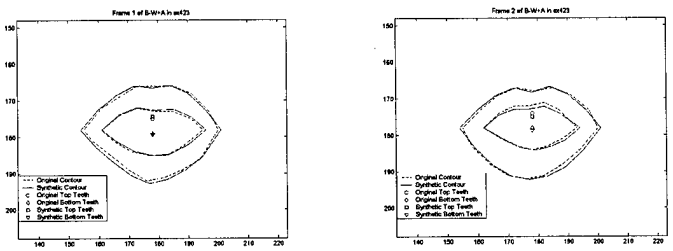
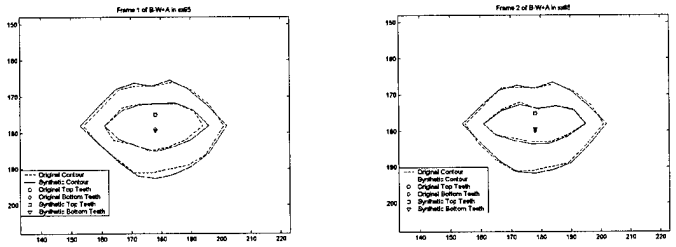


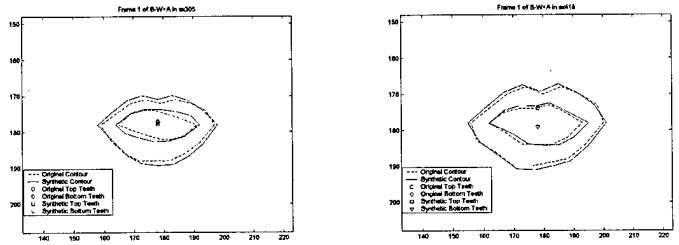
Figure 47. B-D+V Original Scale.



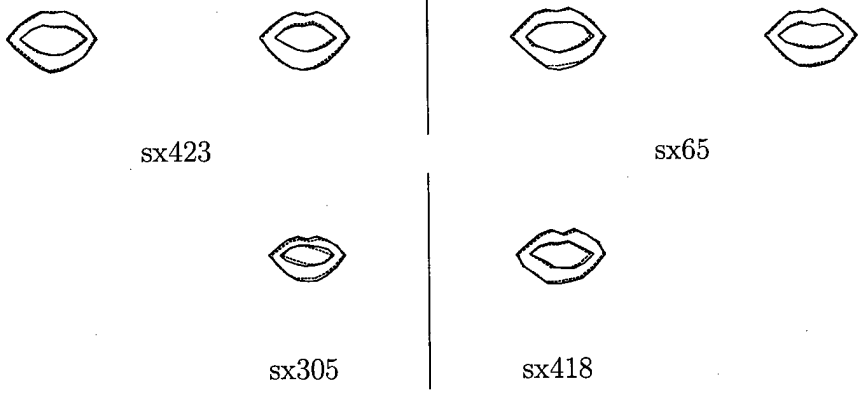
sx423 "n-ih+sh"



sx65 "l-ih+jh"



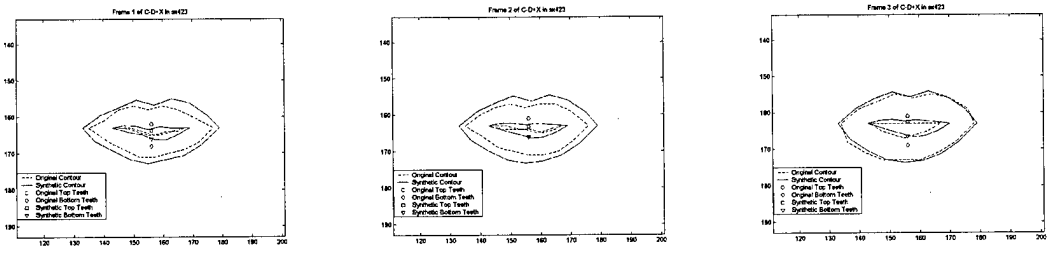
sx305 "l-ih+sh" sx418 " n-ih+sh"



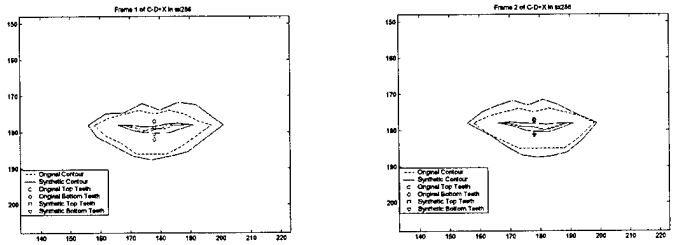
sx423 sx65

sx305 sx418

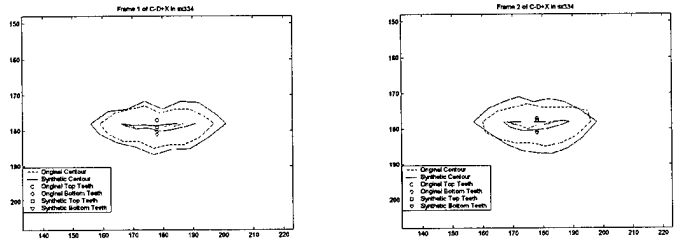
Figure 48. B-W+A Magnified and at Original Scale.



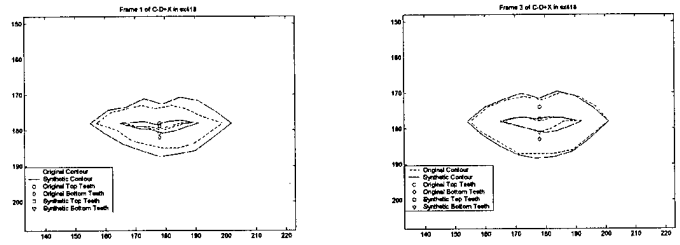
sx423 "s-m+ae"



sx286 "s-b+ae"



sx334 "s-p+ae"



sx418 "s-m+ae"

Figure 49. C-D+X Magnified.

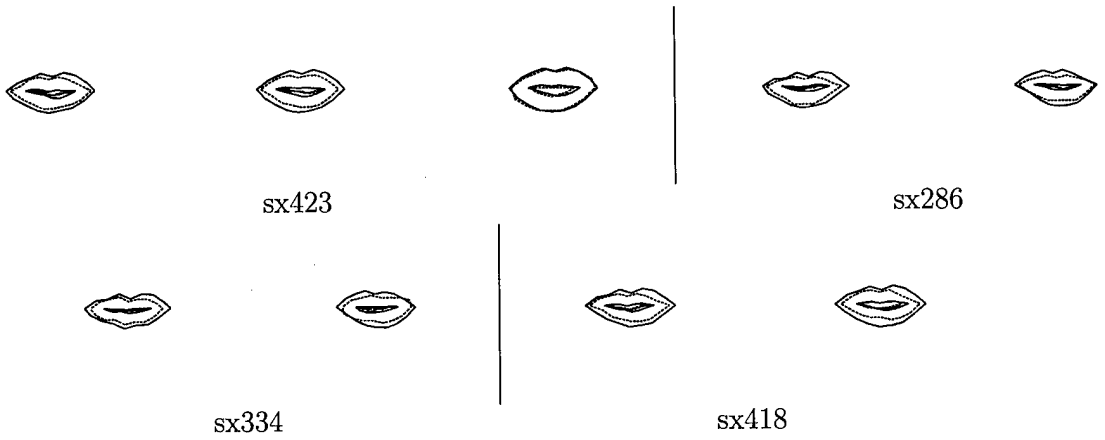
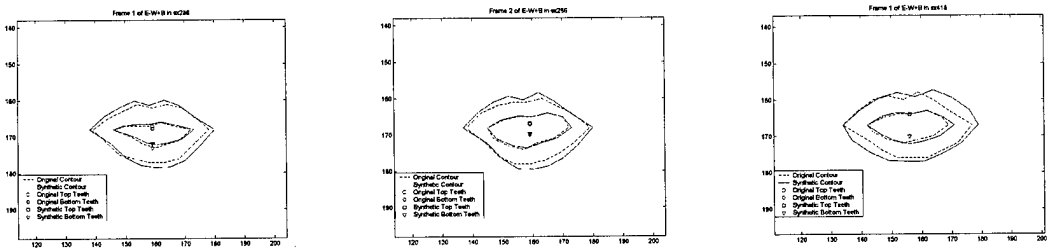
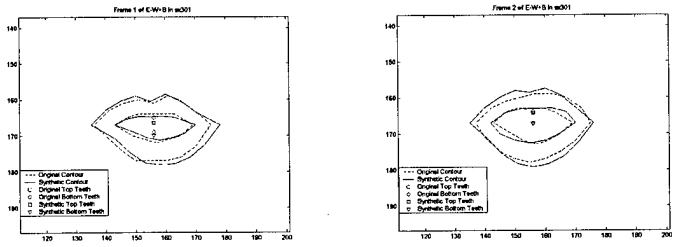


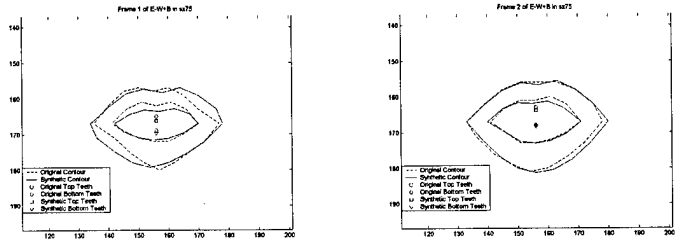
Figure 50. C-D+X at Original Scale.



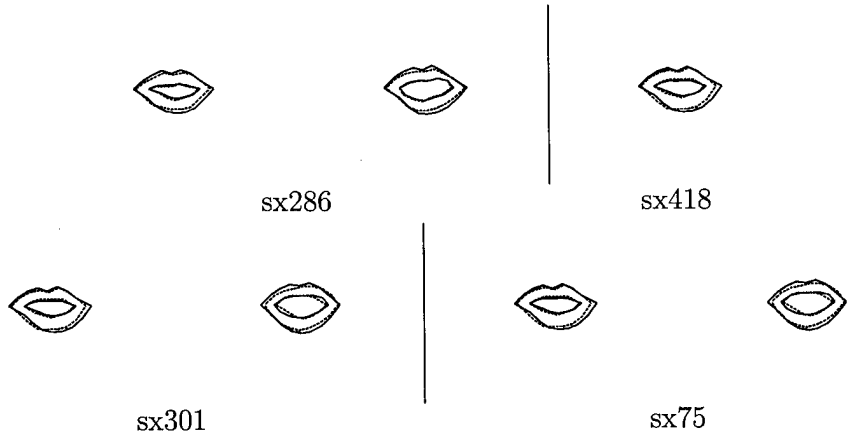
sx286 "f-ih+k" sx418 "v-ih+n"



sx301 "v-ih+k"



sx75 "f-ih+l"



sx286 sx418

sx301 sx75

Figure 51. E-W+B Magnified and at Original Scale.

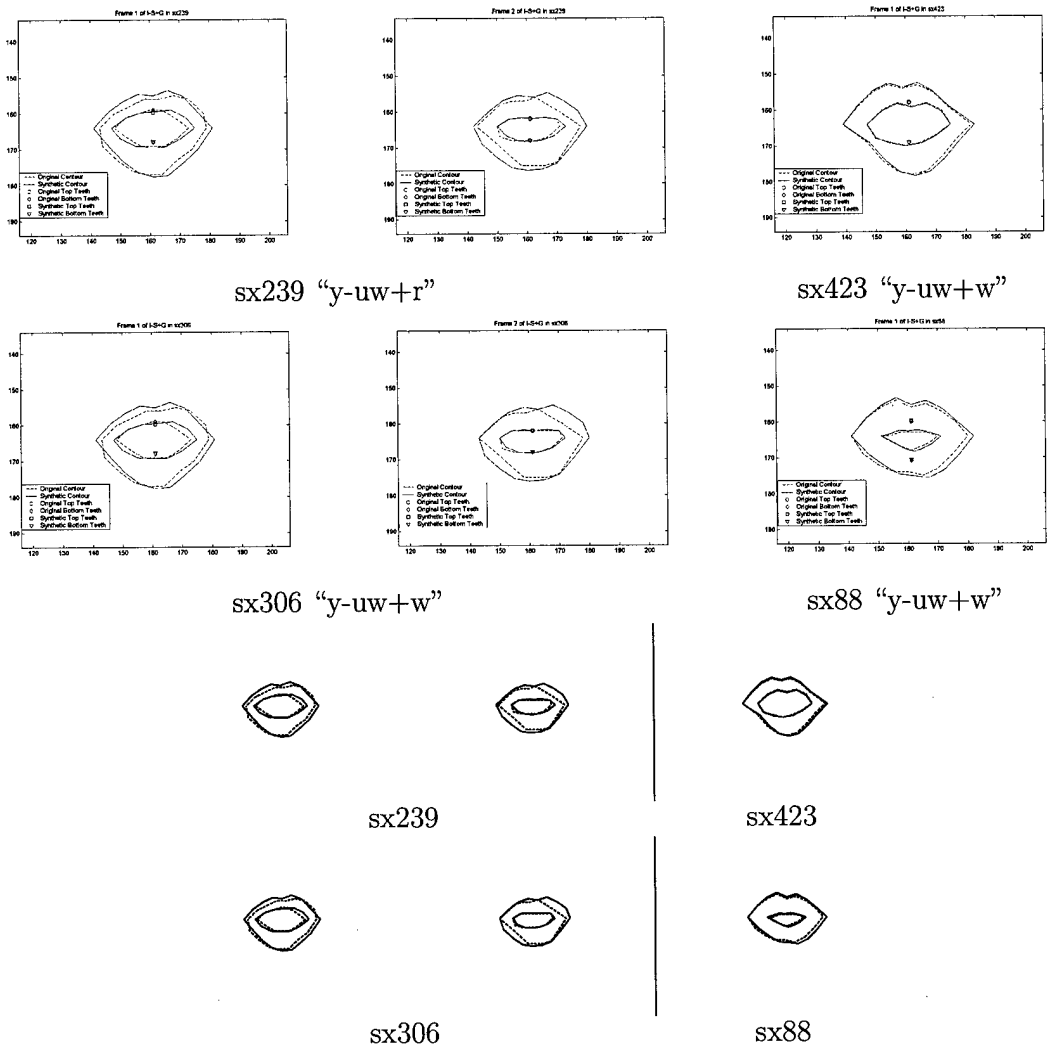
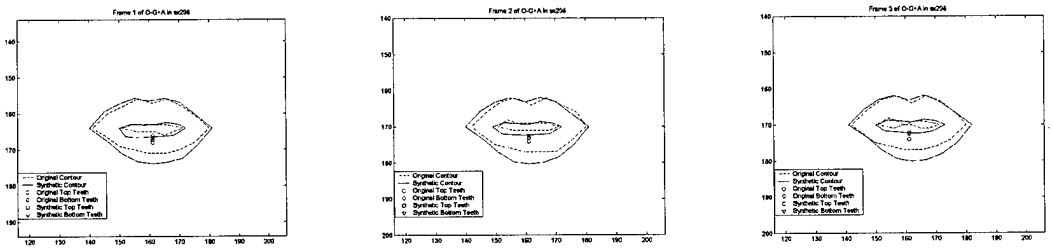
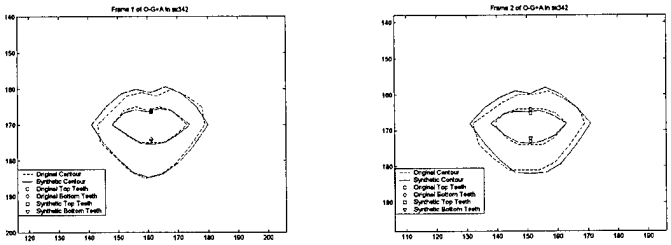


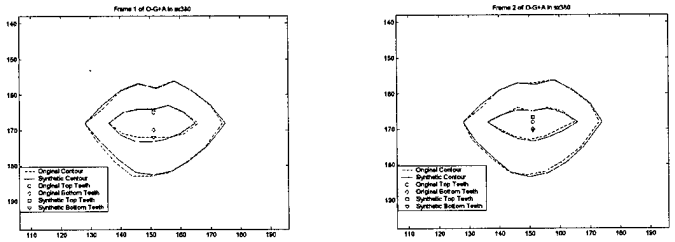
Figure 52. I-S+G Magnified and at Original Scale.



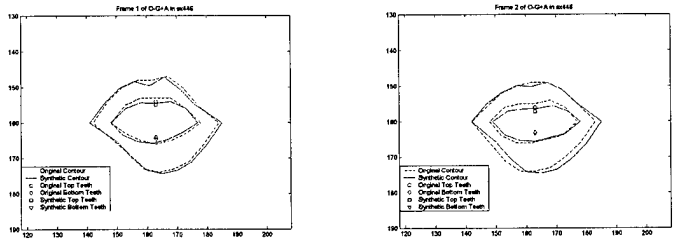
sx298 “aa-r+jh”



sx342 “aa-r+jh”



sx380 “aa-r+jh”



sx446 “aa-r+jh”

Figure 53. O-G+A Magnified.

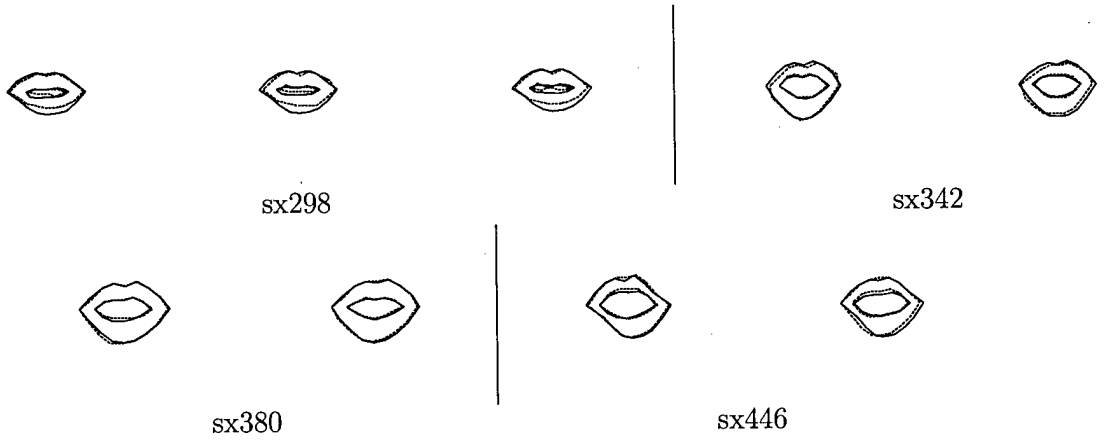


Figure 54. O-G+A at Original Scale.

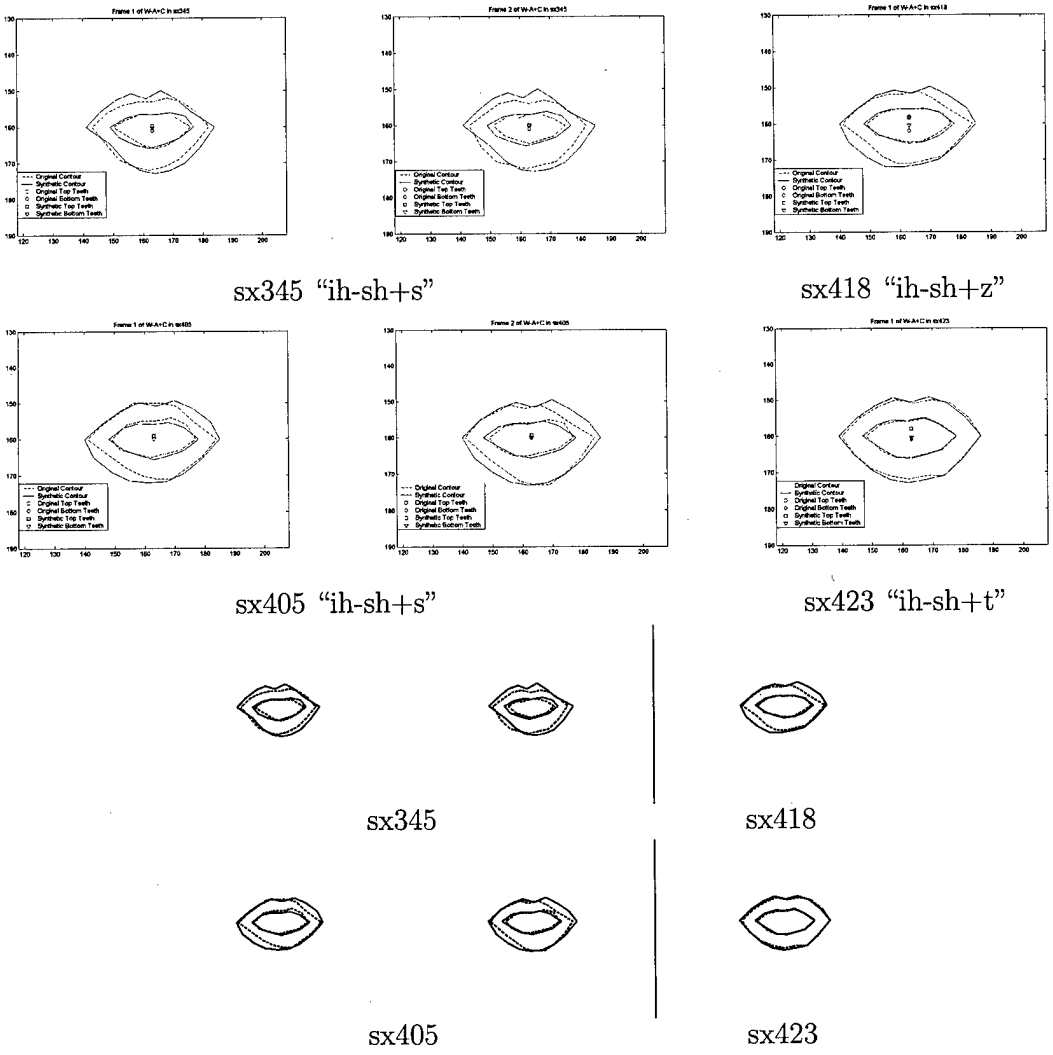


Figure 55. W-A+C Magnified and at Original Scale.

V. Conclusions and Recommendations

This chapter summarizes the conclusions made from the experimental results. It also presents recommendations for future research based on new questions motivated by the conclusions. Finally, a description of the contributions of this research is presented.

5.1 Conclusions

Visual speech synthesis is an area of technology that is rapidly expanding in commercial and military importance. Chapter II presented examples of such applications. However, the quality of synthetic visual speech produced by current "state-of-the-art" automatic synthesis systems is often visually adequate only when a large quantity of triseme examples exist. Since only a few trisemes may have representation in stored data, it is vitally important to have a method that adequately synthesizes visual speech when the available amount of stored video is limited. This research presents such a method.

Hidden Markov Models are powerful tools for characterizing the temporal and contextual properties of speech. This research showed their use in modeling the intrinsic and transient properties of visual speech as well. The mouth outlines presented in Chapter IV illustrate that for some audio-driven visual speech synthesis applications, sequences missing from the available data may be adequately estimated from HMMs.

This research also presents evidence that HMM structure, derivation of characteristic features, and state clustering parameters play significant roles in the quality of estimates of visual speech synthesis parameters. A detailed statistical analysis showed that for the synthesis algorithm presented here and the database considered, dynamic features should be computed considering only the single frames adjacent to the frame under analysis. The analysis also indicated that a common three-state left-to-right no skip HMM structure was adequate for producing high quality estimates. Investigations of the effects of varying the outlier and log-likelihood thresholds in the state clustering portion of the algorithm revealed that a high log-likelihood threshold and a lower outlier threshold resulted in lower average mean squared error in the experiments.

The results presented here show that the methods employed may be able to estimate visual speech synthesis parameters of sufficient quality despite limited data environments. However, further analysis of the algorithm could identify potential improvements. The next section details some of the possibilities.

5.2 Recommendations for Future Research

This research presented a detailed analysis of experimental results. However, it is not possible here to investigate every possible adaptation of the algorithm, and thus the following is suggested for future research:

- Further analysis should be conducted on other sentences and trisemes in the database recorded for this project. This research investigated a single test sentence and six trisemes for the statistical analysis of the effects of the process variables on the resultant MSE. The analysis should be expanded to other sentences and trisemes as well as to a different core data set.
- While the features used here produced fairly accurate mouth shapes, other kinds of features may be more appropriate. Examples include features that describe lip width versus the simple distances of the control points to the line between the mouth corners, or features that account for correlation between the control points and from the left and right sides of the mouth.
- A new audio-visual database should be recorded to include the remaining prompts and introduce new visual angles (*e.g.*, head tilt or rotation) and new faces to the video. It is likely that visual speech synthesis systems will need to estimate control points for various head rotation/tilt angles and for a wide variety of subjects.
- The eigenpoints algorithm proved to be inadequate for automatic control point estimation for this database. A more robust algorithm that accounts for more varied environmental conditions and head rotations/camera angles is required.

- This research presented the use of a decision tree derived from acoustic speech recognition/synthesis. While the questions used resulted in reasonable estimates and models, different questions may potentially further improve the results.
- Analysis of the experimental results indicated variations in the appropriate combination of factor levels for lowering the average MSE between features. An investigation into using separate HMM structures and parameter settings for each feature trained using process variable levels set to individual feature specific values should be conducted.

5.3 Contributions

- No statistical analysis previously existed on the effects of variables in the new parameter estimation process using HMMs. This research provided a detailed assessment of experimental results to investigate these effects and also provided a structure for conducting statistical analyses of future experimental results.
- The use of Hidden Markov Models with decision tree-based state tying for the estimation of audio-driven English visual speech synthesis parameters in a limited data environment is an application that has not been previously accomplished. A viable method for estimating the synthesis parameters/control points when representative examples do not exist is presented.
- The recorded audio-visual database is the only one of its kind currently available for A/V speech research.

5.4 To the Future

This research represents a key development in furthering visual speech synthesis technology. For the data of this research, the experimental results indicate that the method presented may work well in generating control points for use in synthesizing visual speech segments not available in the stored data sets. The results indicate that one may use HMMs trained with decision tree-based state tying to estimate new control points for unseen trisemes and that such a procedure may improve audio-driven visual speech technology in both military and commercial applications in the future.

Appendix A. Chosen Prompts from the TIMIT Speech Corpus

Below is a list of the 405 prompts chosen from the TIMIT Speech Corpus as described in Chapter III. The label in parentheses at the end of each prompt is the identifier for that sentence in TIMIT. Only the first 97 prompts were actually recorded and used in this research. The prompts in bold face are the 50 prompts manually reviewed for control point correction and used for the analysis in Chapter IV.

1. Why yell or worry over silly items? (sx8)
2. **A roll of wire lay near the wall. (sx16)**
3. **Alimony harms a divorced man's wealth. (sx18)**
4. Don't do Charlie's dirty dishes. (sx34)
5. Academic aptitude guarantees your diploma. (sx56)
6. The prowler wore a ski mask for disguise. (sx57)
7. We experience distress and frustration obtaining our degrees. (sx58)
8. The legislature met to judge the state of public education. (sx60)
9. Continental drift is a geological theory. (sx63)
10. Challenge each general's intelligence. (sx65)
11. Upgrade your status to reflect your wealth. (sx68)
12. Porcupines resemble sea urchins. (sx71)
13. Fill that canteen with fresh spring water. (sx75)
14. It's illegal to postdate a check. (sx80)
15. Flying standby can be practical if you want to save money. (sx88)
16. Penguins live near the icy Antarctic. (sx97)
17. **Birthday parties have cupcakes and ice cream. (sx107)**
18. **The best way to learn is to solve extra problems. (sx110)**
19. **While waiting for Chipper she crisscrossed the square many times. (sx149)**
20. **Mosquitoes exist in warm, humid climates. (sx152)**
21. **Orange juice tastes funny after toothpaste. (sx154)**
22. **The water contained too much chlorine and stung his eyes. (sx155)**
23. **Employee layoffs coincided with the company's reorganization. (sx167)**

24. **We could barely see the fjords through the snow flurries. (sx179)**
25. **Military personnel are expected to obey government orders. (sx205)**
26. **Rob sat by the pond and sketched the stray geese. (sx208)**
27. **Jeff's toy go-cart never worked! (sx210)**
28. **Only the best players enjoy popularity. (sx211)**
29. **I gave them several choices and let them set the priorities. (sx212)**
30. **The sermon emphasized the need for affirmative action. (sx215)**
31. **Try to recall the events in chronological order. (sx218)**
32. **Put the butcher block table in the garage. (sx223)**
33. **It's healthier to cook without sugar. (sx227)**
34. **Allow leeway here, but rationalize all errors. (sx230)**
35. Puree some fruit before preparing the skewers. (sx239)
36. They all agree that the essay is barely intelligible. (sx243)
37. John's brother repainted the garage door. (sx276)
38. In every major cloverleaf, traffic sometimes gets backed up. (sx286)
39. The avalanche triggered a minor earthquake. (sx291)
40. The two artists exchanged autographs. (sx296)
41. I'd rather not buy these shoes than be overcharged. (sx298)
42. Amoebas change shape constantly. (sx300)
43. We like bleu cheese but Victor prefers swiss cheese. (sx301)
44. Cottage cheese with chives is delicious. (sx305)
45. A chosen few will become Generals. (sx306)
46. Those answers will be straightforward if you think them through carefully first. (sx312)
47. Often you'll get back more than you put in. (sx334)
48. Gus saw pine trees and redwoods on his walk through Sequoia National Forest. (sx335)
49. Ducks have webbed feet and colorful feathers. (sx340)
50. Take charge of choosing her bride's maids' gowns. (sx342)
51. Trish saw hours and hours of movies Saturday. (sx345)
52. I'll have a scoop of that exotic purple and turquoise sherbet. (sx348)
53. According to my interpretation of the problem, two lines must be perpendicular. (sx352)
54. Diane may splurge and buy a turquoise necklace. (sx356)
55. George seldom watches daytime movies. (sx358)
56. I assume moisture will damage this ship's hull. (sx360)

57. I just saw Jim near the new archeological museum. (sx378)
58. Why charge money for such garbage? (sx380)
59. He picked up nine pairs of socks for each brother. (sx390)
60. The fish began to leap frantically on the surface of the small lake. (sx396)
61. Of course you can have another tunafish sandwich. (sx405)
62. There was a gigantic wasp next to Irving's big top hat. (sx406)
63. Greg buys fresh milk each weekday morning. (sx410)
64. Iris thinks this zoo has eleven Spanish zebras. (sx418)
65. Smash lightbulbs and their cash value will diminish to nothing. (sx423)
66. Please sing just the club theme. (sx436)
67. Thomas thinks a larger clamp solves the problem. (sx446)
68. **This is not a program of socialized medicine. (si458)**
69. **Day after day some new episode is reported. (si461)**
70. **It also provides for funds to clear slums and help colleges build dormitories. (si462)**
71. **Nobody really expects to evacuate. (si465)**
72. **He really crucified him; he nailed it for a yard loss. (si467)**
73. **There is definitely some ligament damage in his knee. (si468)**
74. **So, if anybody solicits by phone, make sure you mail the dough to the above. (si476)**
75. **Her position covers a number of daily tasks common to any social director. (si477)**
76. **Range was a vital detail. (si480)**
77. **We must be ready for any needed sacrifice. (si489)**
78. **Maybe it's taking longer to get things squared away than the bankers expected. (si490)**
79. **Displayed as lamps, the puppets delight the children and are decorative accent. (si497)**
80. **To create such a lamp, order a wired pedestal from any lamp shop. (si498)**
81. **This is a problem that goes considerably beyond questions of salary and tenure. (si511)**
82. **Success, for many turnpikes, has come hard. (si520)**
83. **Progress is being made, too, in improving motorists' access to many turnpikes. (si522)**
84. **Wooded stream valleys in the folds of earth would be saved. (si529)**

85. **The narrow fringe of sadness that ran around it only emphasized the pleasure. (si546)**
86. **You certainly can't expect the infield to do any better than it did last year. (si550)**
87. **This is going to be a language lesson, and you can master it in a few minutes. (si561)**
88. **Contrast trim provides other touches of color. (si566)**
89. **At the left is a pair of dressy straw pumps in a light, but crisp texture. (si567)**
90. **Teaching guides are included with each record. (si573)**
91. **He doesn't want her to look frowningly at him, or speak to him angrily. (si575)**
92. **It seems that open season upon veterans' hospitalization is once more upon us. (si582)**
93. **Maybe he will help to turn our fair city into a ghost town. (si587)**
94. **Replace it with the statue of one or another of the world's famous dictators. (si591)**
95. **The gallant half-city is dying on its feet. (si593)**
96. **On these they feel they can rely. (si596)**
97. **New self-deceiving rags are hurriedly tossed on the too-naked bones. (si598)**
98. Please take this dirty table cloth to the cleaners for me. (sx293)
99. Or certain words or rituals that child and adult go through may do the trick. (si554)
100. But the attack was made from an advance copy. (si610)
101. Her debut over, perhaps the earlier scenes will emerge equally fine. (si615)
102. He injected more vitality into the score than it has revealed in many years. (si616)
103. A tiny handful never did make the concert. (si625)
104. The works are presented chronologically. (si626)
105. He liked to nip ear lobes of unsuspecting visitors with his needle-sharp teeth. (si633)
106. Here, he is, quite persuasively, the very embodiment of meanness and slyness. (si634)
107. He enlisted a staff of loyal experts and of many zealous volunteers. (si637)
108. This theory eventually proved inexact. (si638)
109. Clever light songs were overly coy, tragic songs a little too melodramatic. (si641)
110. The armchair traveler preserves his illusions. (si643)
111. Yet it exists and has an objective reality which can be experienced and known. (si654)
112. His history is his alone, yet each man must recognize his own history in it. (si659)

113. Death reminds man of his sin, but it reminds him also of his transience. (si661)
114. Both eventualities are possible logically, but practically they are impossible. (si675)
115. He is not talking in the main about probabilities, risks and danger in general. (si679)
116. Propriety was synonymous with ritual observance, the mark of a true gentleman. (si690)
117. But this esoteric doctrine was lost in the shuffle to acquire special powers. (si691)
118. Why, said the teacher, art thou so late? (si692)
119. The new birth is immediate and instantaneous. (si695)
120. For me it has more of both elements than the majority of its competitors. (si715)
121. There are canoes ideal for fishing in protected waters or for camping trips. (si716)
122. This changes the formula to an equation. (si722)
123. The total of these three volumes is the final combustion chamber volume. (si723)
124. Keep the desserts simple; fruit does nicely. (si739)
125. In the winter, hibachi in the kitchen or grill over the logs of the fireplace. (si741)
126. Get copper or earthenware mugs that keep beer chilled or soup hot. (si744)
127. Trim excess clay from around lip of mold and set aside while assembling lid. (si764)
128. Turn shaker upside down. (si765)
129. Make lid for sugar bowl the same as jar lids, omitting design disk. (si768)
130. Vases: make same as salt and pepper shakers, leaving off top pieces. (si770)
131. Thread matching yarn in tapestry needle. (si775)
132. Again, these blocks were set in resin-saturated glass cloth and nailed. (si781)
133. The decking is quarter-inch mahogany marine plywood. (si782)
134. When you're less fatigued, things just naturally look brighter. (si798)
135. Insulate, weatherstrip, double-glaze to the maximum. (si803)
136. An area sheltered from strong winds may be highly desirable for recreation use. (si810)
137. Push-ups: push-ups are essential, but few have the strength for them at first. (si813)
138. Selecting bunks by economic comparison is usually an individual problem. (si831)
139. Computers are being used to keep branch inventories at more workable levels. (si837)
140. Complexity of complete marketing planning. (si840)
141. Are planning and strategy development emphasized sufficiently in your company? (si842)
142. A complete plan: we have made limited application of the parallel ladder plan. (si867)
143. Be careful that you keep adequate coverage, but look for places to save money. (si873)

144. Before that we lumber dealers were working almost single-handed on the problem. (si881)
145. The way is to rent a chauffeur-driven car. (si882)
146. But why is it necessary to reproduce the retinal image within the brain? (si890)
147. It was applied with a compost shredder made from a converted manure spreader. (si891)
148. New ideas for improving nutrition came with the study of soil treatment. (si894)
149. Adults take a long time to convince and you are thwarted if you try to push. (si895)
150. Sprouted grains and seeds are used in salads and dishes such as chop suey. (si900)
151. We always thought we would die with our boots on. (si908)
152. Do you love to run up a hem, sew on buttons, make neat buttonholes? (si913)
153. Differences were related to social, economic, and educational backgrounds. (si921)
154. In tradition and in poetry, the marriage bed is a place of unity and harmony. (si922)
155. To many experts, this trend was inevitable. (si923)
156. But such cases were, in the past, unusual. (si929)
157. Crooked, overlapping, twisted, or widely spaced teeth. (si935)
158. Impressions often appear in a symbolic form and cannot be taken at face value. (si939)
159. The farmer's life must be arranged to meet the demands of crops and livestock. (si941)
160. Production may fall far below expectations. (si943)
161. Advantages a farm provides: a wholesome and healthful environment for children. (si945)
162. As a rule, part-time farmers hire little help. (si946)
163. How much will it cost to do any necessary modernizing and redecorating? (si948)
164. It may be hard to decide what this labor is worth, but charge something for it. (si949)
165. We flew in rickety planes so overloaded that we wondered why they didn't crash. (si970)
166. Bathing the itching parts with kerosene gave relief and also killed the pests. (si985)
167. A kerosene shampoo seems a heroic treatment, but it did the job. (si988)
168. She had no way of knowing in advance whether an opportunity for murder existed. (si991)
169. They were a vagabond lot and considered to be shady and undesirable characters. (si992)
170. At different times he served as glee-club and choir leader and as organist. (si995)
171. Cereal grains have been used for centuries to prepare fermented beverages. (si1013)
172. Beer, generally fermented from barley, is an old alcoholic beverage. (si1014)
173. To keep 'em scattered somewhat and yet herd 'em was called loose herdin'. (si1016)

174. The cowboy's humorous name for a cow givin' milk was a milk pitcher. (si1021)
175. Her tail no longer lashed. (si1029)
176. With a whirling jump, it could get into gear. (si1035)
177. Residential associations struggle to insulate themselves against intrusions. (si1044)
178. It offered to surrender its right to exclusive trade, but asked an indemnity. (si1045)
179. Internal national responsibility, now a truism, need not be documented. (si1055)
180. Vital questions would be quickly answered according to a preprepared agenda. (si1062)
181. They were chasing a rain cloud. (si1066)
182. Perhaps this is what gives the aborigine his odd air of dignity. (si1069)
183. The theme may be the formation of a shape from which other shapes evolve. (si1084)
184. Solar activities could presumably bring long periods of flood or drought. (si1091)
185. Yet we no longer feel uneasy. (si1092)
186. His talk turns to what he calls the mess, or sometimes this buzzing confusion. (si1097)
187. Why should this be so? (si1114)
188. The record teems with romance and adventure. (si1119)
189. Let me quote him even more fully, for his analysis is important to my theme. (si1124)
190. And let me add, utopianism, also. (si1126)
191. Others invoked technology and common sense. (si1128)
192. Individual human strength is needed to pit against an inhuman condition. (si1134)
193. The stepmother, almost without exception, has been presented as a cruel ogress. (si1141)
194. We would establish no censorship. (si1144)
195. We knew that it was, as reassurance, the ironic fruit of a deeply moral nature. (si1146)
196. This is what necessitates the nonsystematic character of his astronomy. (si1156)
197. But, considered within technical astronomy, a different pattern can be traced. (si1157)
198. The continuing modernization of these forces is a costly but necessary process. (si1165)
199. In the course of its inquiry, it took testimony from only seven witnesses. (si1166)
200. He has also an extraordinary conscience. (si1170)
201. He was busy, he said, in having someone submit to a monkey-gland operation. (si1172)
202. What shall these effects be? (si1179)
203. All chance of fulfilling my destiny is over. (si1191)
204. He had collared one of his generals in public. (si1193)
205. He had accordingly cultivated eccentricity to the point of second nature. (si1196)
206. Let him become honest, and they discard him. (si1200)

207. His artistic accomplishments guaranteed him entry into any social gathering. (si1202)
208. Students' choices of ideal educational goals are not arbitrary or whimsical. (si1219)
209. They also want money, prestige, and security. (si1220)
210. Conservatism and traditionalism seem implied by what has just been said. (si1221)
211. This is nevertheless a minority view. (si1224)
212. It made no difference that most evidence points to an opposite conclusion. (si1233)
213. We may say of some unfortunates that they were never young. (si1237)
214. In many of his poems, death comes by train: a strongly evocative visual image. (si1241)
215. Ideas are the thruway to nowhere. (si1243)
216. Only rarely is attention given to accurate progress reports and evaluation. (si1251)
217. My sincere wish is that he continues to add to this record he sets here today. (si1255)
218. Ten others made no reply. (si1262)
219. This has been attributed to helium film flow in the vapor pressure thermometer. (si1271)
220. The same shelter could be built into an embankment or below ground level. (si1281)
221. Living in a shelter the radioactivity of fallout decays rapidly at first. (si1283)
222. A portable electric heater is advisable for shelters in cold climates. (si1288)
223. One of these is the solidarity and the confidential relationship of marriage. (si1290)
224. It was exposed to a high velocity gas jet. (si1304)
225. If they are not ellipsoids, the conclusions will be a reasonable approximation. (si1305)
226. Being based on so few events, these results are of dubious validity. (si1317)
227. Biological warfare is considered to be primarily a strategic weapon. (si1319)
228. This area is highly susceptible to infection. (si1320)
229. This may be of overriding importance in considering military objectives. (si1322)
230. Quite often, honeybees form a majority on the willow catkins. (si1325)
231. The lower portion provides a rating at any stage between infancy and adulthood. (si1330)
232. Of particular importance is the study of the actions of drugs in this respect. (si1339)
233. Before deriving this formula, we explain what we mean by problems of this kind. (si1342)
234. Experiment without a fixed number of trials. (si1344)
235. The population can thereby replenish itself and actually grow larger. (si1357)
236. But the information on the dynamics of population was often quite misleading. (si1358)
237. Our campus, unfortunately, owns no films. (si1366).

238. As these maladies overlap, so must the cure. (si1377)
239. Morphophonemic rules may be thought of as joining certain points in the system. (si1378)
240. There is little doubt that the students benefit from vocational education. (si1380)
241. There need be no squeamishness about admitting this. (si1384)
242. Several factors contributed to this change. (si1387)
243. Such legislation was clarified and extended from time to time thereafter. (si1388)
244. However, when labor disputes arise, its provisions come clearly into play. (si1391)
245. The elementary-school child grows gradually in his ability to work in groups. (si1395)
246. The plaintiff in school desegregation cases. (si1397)
247. Now that this is at odds with our meaning may be shown as follows. (si1400)
248. And the law virtually ignored the situation. (si1405)
249. Receiving no answer, they set the fire. (si1406)
250. We will achieve a more vivid sense of what it is by realizing what it is not. (si1410)
251. In most discussions of this phenomenon, the figures are substantially inflated. (si1413)
252. There should be no reason to misinterpret or ignore the intent of this letter. (si1417)
253. Personal predispositions tend to blunt the ear and, in turn, the voice as well. (si1418)
254. The failure to keep these two usages distinct presents hazards to the reader. (si1419)
255. There are many such competently anonymous performances among the earlier poems. (si1422)
256. The world is constantly changing; what was new yesterday is obsolescent today. (si1435)
257. A third volume remains to be published. (si1440)
258. Coverage of primary literature will follow. (si1445)
259. In earlier years, the preservation of food was essentially related to survival. (si1446)
260. Microorganisms are often responsible for the rapid spoilage of foods. (si1447)
261. Measured performance characteristics for this experimental tube will be listed. (si1461)
262. We can, however, maximize its expected value. (si1463)
263. This process is especially difficult since gyro drifting is typically random. (si1465)
264. The platform accelerometers must be slightly modified for this procedure. (si1467)
265. Poverty imposes a kind of chastity on the ambitious. (si1488)
266. Perhaps it was right; perhaps it was just. (si1490)
267. He merely said, any good decorator these days can make you a tasteful home. (si1491)
268. It was not exactly panic they gave way to, but they could not just sit there. (si1502)
269. He chuckled, the memory vivid. (si1513)

270. No one will even suspect that it is your work. (si1530)
271. Who is going to stop me? (si1531)
272. Is it because of my slovenliness that hair grows on my face? (si1534)
273. It gave the rubberized fabric a dull gleam, like metal. (si1541)
274. In the pity for them his loneliness was gone. (si1546)
275. Outside, only a handful of reporters remained. (si1552)
276. However, the litter remained, augmented by several dozen lunchroom suppers. (si1553)
277. A sleepy voice answered. (si1555)
278. He waited until they were inside the elevator and then said: now what do we do? (si1556)
279. They had vermouth, sitting in front of a cafe. (si1557)
280. Husbandry was bounded by snake-rail fences, and there were grazing cattle. (si1564)
281. Ice baths, electric shocks, lashings, wild dogs, testicle crushers. (si1571)
282. He found an empty bench, opened a newspaper, and stretched his legs before him. (si1572)
283. Rector was often curious; often tempted to ask questions but he never did. (si1584)
284. Your voice is delightful, he approved with a warm smile. (si1586)
285. No, she would not pretend modesty, but neither must she be crudely bold. (si1589)
286. The figure in the corner belched loudly, a deep, liquid eruption. (si1597)
287. The flat-bottomed boat swung slowly to the pull of the current. (si1598)
288. They moved toward the skiffs with shocking eagerness, elbowing and shoving. (si1599)
289. Intelligence jabbed at him accusingly. (si1603)
290. Nobody does, but they gave me ten last year. (si1605)
291. "Infectious hepatitis!", he shouted heartily. (si1606)
292. She smiled, and the teeth gleamed in her beautifully modeled olive face. (si1608)
293. They've never met, you know. (si1612)
294. She knew she was feeling afraid and inwardly laughed at herself. (si1622)
295. They even pay me six dollars a month. (si1644)
296. She looked at me provocatively. (si1651)
297. The knifelike pain in his groin nearly brought him down again. (si1671)
298. It snowed continuously, but quietly, evenly. (si1676)
299. She asked with a reportorial gleam in her eye. (si1684)
300. Something else distracted him, yet there was no sound, only tomblike silence. (si1688)
301. No, they could kill him just as easy right now. (si1691)

302. So somebody else knew what would happen to her father's money if she died. (si1698)
303. Well, then who brought it? (si1699)
304. Emotionally, they penetrated him not at all. (si1701)
305. They'll tell you what's cooking. (si1706)
306. Here's where luck would normally step in. (si1714)
307. Do you know what they think of me? (si1718)
308. And men also used vacuum cleaners in both rooms, sucking dust up once more. (si1726)
309. Bitter, unreasoning jealousy? (si1732)
310. The poor, unlucky – the telephone rang. (si1736)
311. No signs of these, no gross hemorrhage of lungs, heart, brain or stomach. (si1738)
312. Nevertheless it's true. (si1749)
313. Meanwhile, fishermen took advantage of them to pull up whoppers. (si1756)
314. His sarcasm was followed by a stupid grin of his thick mouth and bad teeth. (si1757)
315. Got a heck of a buy on this, dirt cheap. (si1759)
316. Hastily the boy switched on a ceiling light. (si1772)
317. There was a grunt, curiously inarticulate, like that of an animal in pain. (si1784)
318. C'mon, he whispered; floor level's about three feet down, so don't fall. (si1797)
319. If we left one we'd have to wipe it for fingerprints. (si1801)
320. They went downstairs, loaded a tray lavishly. (si1808)
321. That'll be plenty to eat by. (si1810)
322. "Presently", his water brother said breathlessly. (si1811)
323. My beloved ward, my perennial gadfly, said the whining voice. (si1817)
324. A range of bluffs hugged the horizon, the color of decaying moss. (si1820)
325. He wasn't likely to come down with hot chills and puzzling striped fever. (si1825)
326. His body processes only slowed down enough for him to dispel fatigue poisons. (si1829)
327. When she awoke, she was the ship. (si1842)
328. Hello, anyone at home? (si1844)
329. My dress needs some work on it. (si1861)
330. We'll pay you back if you'll let us. (si1862)
331. We haven't slept together since we started. (si1863)
332. Now there's nothin left of me. (si1865)
333. He rode low on the mare's neck. (si1869)
334. You'll have to try it alone. (si1878)

335. The wagons were burning fiercely. (si1884)
336. Although my shot killed his horse, he rolled off the bale on top of me. (si1890)
337. There would still be plenty of moments of regret and sadness and guilty relief. (si1900)
338. When he saw me coming he turned his radio off. (si1903)
339. He lived and breathed for the mining company. (si1910)
340. As she drove, she thought about her plan. (si1914)
341. Perhaps it was all a vividly conceived dream. (si1916)
342. Every movement she made seemed unnecessarily noisy. (si1919)
343. And his relatively small hands and feet gave him an almost delicate appearance. (si1921)
344. Or maybe you just don't feel like a cigar? (si1922)
345. Not immediately, as the deputy demanded. (si1928)
346. A rabbit punch redoubled him. (si1930)
347. We'll ride out as soon as we've had chuck. (si1936)
348. Lighted windows glowed jewel-bright through the downpour. (si1938)
349. Privately, he created and magnified an image of himself as a hired assassin. (si1945)
350. Honey, he whispered. (si1954)
351. He had not covered a hundred yards before a gun crashed from somewhere behind. (si1959)
352. A second twitched his shirtsleeve, and he felt a brief burn on his upper arm. (si1961)
353. Twelve o'clock level. (si1967)
354. If any of us miss, they can pick up the pieces. (si1969)
355. The clearly identifiable enemy continued on as if no one else were around. (si1970)
356. How's it strike you, foul or fair? (si1988)
357. Or, more rhythmically, the swarthy witches are laughing? (si1992)
358. So if all these beers was to get me in bed, man, you just spent a lot of money. (si1994)
359. The clouds bulged downward and burst suddenly into a great black funnel. (si2009)
360. Then he would realize they were really things that only he himself could think. (si2021)
361. They came upon cheetal deer at woodland pools. (si2027)
362. He took his mask from his forehead and threw it, unexpectedly, across the deck. (si2028)
363. He felt a good deal less shaky. (si2035)
364. He strolled back to the door, whistling softly, hands still clasped behind him. (si2057)
365. Thank you, she said, dusting herself off. (si2060)

366. Lifting her skirts, she climbed in, never relinquishing her grip on his arm. (si2065)
367. Yeah, seems so, don't it, the boy laughed, hugging her close. (si2067)
368. Maybe you and me will, girlie, but these two ain't goin' nowhere. (si2077)
369. Above him slid the evasive shadow of a storm cloud. (si2083)
370. He'd had no idea how unhappy his sweet peach had been. (si2088)
371. He then offered his own estimate of the weather, which was unenthusiastic. (si2093)
372. Left mine many a time, only she never knew it. (si2097)
373. It sounded silly; why go on? (si2107)
374. Don't forget whiskey; it's such a big seller. (si2108)
375. He saw the dangers, not the glories of being identified as a mutineer. (si2117)
376. His black hat with its wide brim, high crown, and fur trim rode high. (si2125)
377. Does this bother you? (si2130)
378. Here they did not need to be in ghettos. (si2133)
379. He had never felt particularly close to her. (si2160)
380. A smile pulled at the lower strip of adhesive tape. (si2167)
381. How long would it be occupied? (si2171)
382. Aw, shut up, he said. (si2174)
383. Shivering, he put on his clothes. (si2175)
384. But now she looked ugly. (si2192)
385. A profession that brought me as good an income as mine wasn't to be sneezed at. (si2196)
386. But if she wasn't interested, she'd just go back to the same life she'd left. (si2198)
387. A bullet, she answered. (si2209)
388. It was as blissful and fulfilling a night as any bride ever experienced. (si2215)
389. It's never wrong if love is real. (si2216)
390. He'd not only told me so, he'd proved it. (si2217)
391. You young men get to be my age, you won't take flu so lightly. (si2222)
392. Why not try another club? (si2240)
393. Snakes are ugly, he said again. (si2243)
394. Joking stopped and everyone gathered around. (si2267)
395. Needless to say, my art suffered drastically during this turbulent period. (si2268)
396. Samples: he walked in upon her invitation. (si2281)
397. The girl nodded understandingly. (si2285)
398. Finally he asked, do you object to petting? (si2289)

399. Those were especially the ones that all other grownups laughed at loudest. (si2293)
400. Their curiosity went happily out of bounds. (si2295)
401. When we left Washington his son Tad was ill and Mrs. Lincoln hysterical. (si2323)
402. Keep your seats, boys, I just want to put some finishing touches on this thing. (si2324)
403. I'm sorry but I shall have to search this house. (si2327)
404. Now if you'll pardon me, I won't keep him waiting. (si2331)
405. But I'm so sunburned that every move I make is agony. (si2334)

Appendix B. Decision Tree Questions

Below is the list of questions in the decision tree used for the state clustering and tying. Each is in the format used by HTK. "QS" is the command used by the HTK utility HHEd defining the rest of the line as a question. The "L_.." is the question label and the items in the curly braces give the requirements for a "Yes" answer. For example, if the triseme is "A-B+Z", the answer to the question "R_Silence" would be "Yes" since the silence viseme "Z" is to the right of the core viseme "B". If it were "A-B+C", the answer would be "No" to "R_Silence", but "Yes", for example, for "R_C".

```
QS  "R_NonBoundary"      { *+* }
QS  "R_Silence"          { *+Z }
QS  "R_Stop"             { *+D,*+C,*+B }
QS  "R_Nasal"            { *+D,*+B,*+J }
QS  "R_Fricative"        { *+C,*+A,*+E,*+F }
QS  "R_Liquid"           { *+B,*+G,*+I,*+H }
QS  "R_Vowel"            { *+K,*+W,*+P,*+O,*+S,*+Y,*+M,*+R,*+U,*+L,*+V,*+T }
QS  "R_C-Front"          { *+D,*+E,*+G }
QS  "R_C-Central"        { *+C,*+B,*+A,*+F,*+G }
QS  "R_C-Back"           { *+A,*+I,*+B,*+J,*+H }
QS  "R_V-Front"          { *+V,*+W,*+K }
QS  "R_V-Central"        { *+K,*+O,*+M,*+P }
QS  "R_V-Back"           { *+S,*+O,*+Y,*+N }
QS  "R_Front"            { *+D,*+E,*+G,*+V,*+W,*+K }
QS  "R_Central"          { *+C,*+B,*+A,*+F,*+G,*+K,*+O,*+M,*+P }
QS  "R_Back"             { *+A,*+I,*+B,*+J,*+H,*+O,*+S,*+Y,*+N }
QS  "R_Lenis"            { *+D,*+C,*+B,*+E,*+F,*+A }
QS  "R_UnFortLenis"      { *+D,*+B,*+J,*+H,*+G,*+I }
QS  "R_Coronal"          { *+C,*+B,*+F,*+A,*+G }
QS  "R_NonCoronal"       { *+D,*+B,*+J,*+E,*+H,*+I,*+G }
QS  "R_Anterior"         { *+D,*+C,*+B,*+E,*+F,*+G }
QS  "R_NonAnterior"      { *+B,*+J,*+A,*+H,*+G,*+I }
QS  "R_Continuent"       { *+D,*+B,*+J,*+E,*+F,*+C,*+A,*+H,*+G,*+I }
QS  "R_NonContinuent"    { *+D,,*+C,*+B,*+A }
QS  "R_Strident"         { *+C,*+A }
QS  "R_NonStrident"      { *+E,*+F,*+H }
QS  "R_UnStrident"       { *+D,*+C,*+B,*+J,*+G,*+I }
QS  "R_Glide"            { *+H,*+B,*+G,*+I }
QS  "R_Syllabic"         { *+B,*+D,*+M }
QS  "R_Unvoiced-Cons"    { *+D,*+C,*+B,*+A,*+E,*+F,*+H }
QS  "R_Voiced-Cons"      { *+A,*+D,*+C,*+F,*+B,*+I,*+J,*+G,*+E }
QS  "R_Unvoiced-All"     { *+D,*+C,*+B,*+A,*+E,*+F,*+H }
```

```
QS  "R_Long"            { *+V,*+O,*+T,*+P,*+S,*+B,*+D }
QS  "R_Short"           { *+K,*+L,*+O,*+W,*+R,*+U,*+Y,*+N }
QS  "R_Dipthong"        { *+L,*+R,*+U,*+O,*+M,*+B,*+D }
QS  "R_Front-Start"     { *+L,*+O,*+M }
QS  "R_Fronting"        { *+R,*+L,*+U }
QS  "R_High"            { *+W,*+S,*+O,*+Y,*+V }
QS  "R_Medium"          { *+L,*+M,*+O,*+Y,*+K,*+B,*+D }
QS  "R_Low"             { *+K,*+R,*+O,*+Q,*+P,*+U }
QS  "R_Rounded"         { *+P,*+S,*+O,*+Y,*+U,*+G }
QS  "R_Unrounded"       { *+K,*+W,*+O,*+M,*+R,*+L,*+V,*+Q,*+Y,*+B,*+D,*+H,*+G,*+I }
QS  "R_NonAffricate"    { *+C,*+A,*+E,*+F }
QS  "R_Affricate"       { *+A }
QS  "R_IVowel"          { *+W,*+V }
QS  "R_EVowel"          { *+K,*+L }
QS  "R_AVowel"          { *+K,*+O,*+M,*+R,*+Q }
QS  "R_OVowel"          { *+P,*+U,*+O }
QS  "R_UVowel"          { *+O,*+Y,*+B,*+D,*+S }
QS  "R_Voiced-Stop"     { *+D,*+C,*+B }
QS  "R_Front-Stop"      { *+D }
QS  "R_Central-Stop"    { *+C }
QS  "R_Back-Stop"       { *+B }
QS  "R_Voiced-Fric"     { *+C,*+A,*+F,*+E }
QS  "R_Front-Fric"      { *+E }
QS  "R_Central-Fric"    { *+C,*+F }
QS  "R_Back-Fric"       { *+A }
QS  "R_A"               { *+A }
QS  "R_B"               { *+B }
QS  "R_C"               { *+C }
QS  "R_D"               { *+D }
QS  "R_E"               { *+E }
QS  "R_F"               { *+F }
QS  "R_G"               { *+G }
QS  "R_H"               { *+H }
QS  "R_I"               { *+I }
QS  "R_J"               { *+J }
QS  "R_K"               { *+K }
QS  "R_L"               { *+L }
QS  "R_M"               { *+M }
QS  "R_N"               { *+N }
QS  "R_O"               { *+O }
QS  "R_P"               { *+P }
QS  "R_Q"               { *+Q }
QS  "R_R"               { *+R }
QS  "R_S"               { *+S }
QS  "R_T"               { *+T }
QS  "R_U"               { *+U }
QS  "R_V"               { *+V }
```

```
QS "R_W"                { *+W }
QS "R_X"                { *+X }
QS "R_Y"                { *+Y }
QS "L_NonBoundary"      { *-* }
QS "L_Silence"          { Z-* }
QS "L_Stop"             { D-*,C-*,B-* }
QS "L_Nasal"            { D-*,B-*J-* }
QS "L_Fricative"        { C-*,A-*,E-*,F-* }
QS "L_Liquid"           { B-*,G-*,I-*,H-* }
QS "L_Vowel"            { K-*,W-*,P-*,O-*,S-*,Y-*,M-*,R-*,U-*,L-*,V-*,T-* }
QS "L_C-Front"          { D-*,E-*,G-* }
QS "L_C-Central"        { C-*,B-*,A-*,F-*,G-* }
QS "L_C-Back"           { A-*,I-*,B-*,J-*,H-* }
QS "L_V-Front"          { V-*,W-*,K-* }
QS "L_V-Central"        { K-*,O-*,M-*,P-* }
QS "L_V-Back"           { S-*,O-*,Y-*,N-* }
QS "L_Front"            { D-*,E-*,G-*,V-*,W-*,K-* }
QS "L_Central"          { C-*,B-*,A-*,F-*,G-*,K-*,O-*,M-*,P-* }
QS "L_Back"             { A-*,I-*,B-*,C-*,B-*,J-*,H-*,O-*,S-*,Y-*,N-* }
QS "L_Fortis"           { D-*,B-*,E-*,F-*,C-*,A-* }
QS "L_UnFortLenis"      { D-*,B-*,J-*,H-*,B-*,B-*,G-*,I-*,G-* }
QS "L_Coronal"          { C-*,F-*,A-*,B-*,G-* }
QS "L_NonCoronal"       { D-*,B-*,J-*,E-*,H-*,I-*,G-* }
QS "L_Anterior"         { D-*,E-*,F-*,C-*,B-*,G-* }
QS "L_NonAnterior"      { B-*,J-*,A-*,H-*,G-*,I-* }
QS "L_Continuent"       { D-*,J-*,E-*,F-*,C-*,A-*,H-*,B-*,G-*,I-* }
QS "L_NonContinuent"    { D-*,C-*,B-*,A-* }
QS "L_Strident"         { C-*,A-* }
QS "L_NonStrident"      { E-*,F-*,H-* }
QS "L_UnStrident"       { D-*,C-*,J-*,B-*,G-*,I-* }
QS "L_Glide"            { H-*,B-*,G-*,I-* }
QS "L_Syllabic"         { B-*,D-*,M-* }
QS "L_UnvoiceC-Cons"    { D-*,B-*,C-*,A-*,E-*,F-*,H-* }
QS "L_VoiceC-Cons"      { A-*,D-*,C-*,F-*,B-*,I-*,J-*,G-*,E-* }
QS "L_UnvoiceC-All"     { D-*,B-*,C-*,A-*,E-*,F-*,H-*,A-* }
QS "L_Long"             { V-*,O-*,T-*,P-*,S-*,B-*,D-* }
QS "L_Short"            { K-*,L-*,O-*,W-*,R-*,U-*,Y-*,N-* }
QS "L_Dipthong"         { L-*,R-*,U-*,O-*,M-*,B-*,D-* }
QS "L_Front-Start"      { L-*,O-*,M-* }
QS "L_Fronting"         { R-*,L-*,U-* }
QS "L_High"             { W-*,S-*,O-*,Y-*,V-* }
QS "L_Medium"           { L-*,M-*,O-*,Y-*,K-*,B-*,D-* }
QS "L_Low"              { K-*,R-*,O-*,Q-*,P-*,U-* }
QS "L_Rounded"          { P-*,S-*,O-*,Y-*,U-*,G-* }
QS "L_Unrounded"        { K-*,W-*,O-*,M-*,R-*,L-*,V-*,Q-*,Y-*,B-*,D-*,H-*,G-*,I-* }
QS "L_NonAffricate"     { C-*,A-*,E-*,F-* }
QS "L_Affricate"        { A-* }
```

```
QS  "L_IVowel"          { W-*,V-* }
QS  "L_EVowel"          { K-*,L-* }
QS  "L_AVowel"          { K-*,O-*,M-*,R-*,Q-* }
QS  "L_OVowel"          { P-*,U-*,O-* }
QS  "L_UVowel"          { O-*,Y-*,B-*,D-*,B-*,S-* }
QS  "L_Voiced-Stop"     { D-*,C-*,B-* }
QS  "L_Front-Stop"      { D-* }
QS  "L_Central-Stop"    { C-* }
QS  "L_Back-Stop"       { B-* }
QS  "L_Voiced-Fric"     { C-*,A-*,F-*,E-* }
QS  "L_Front-Fric"      { E-* }
QS  "L_Central-Fric"    { C-*,F-* }
QS  "L_Back-Fric"       { A-* }
QS  "L_A"               { A-* }
QS  "L_B"               { B-* }
QS  "L_C"               { C-* }
QS  "L_D"               { D-* }
QS  "L_E"               { E-* }
QS  "L_F"               { F-* }
QS  "L_G"               { G-* }
QS  "L_H"               { H-* }
QS  "L_I"               { I-* }
QS  "L_J"               { J-* }
QS  "L_K"               { K-* }
QS  "L_L"               { L-* }
QS  "L_M"               { M-* }
QS  "L_N"               { N-* }
QS  "L_O"               { O-* }
QS  "L_P"               { P-* }
QS  "L_Q"               { Q-* }
QS  "L_R"               { R-* }
QS  "L_S"               { S-* }
QS  "L_T"               { T-* }
QS  "L_U"               { U-* }
QS  "L_V"               { V-* }
QS  "L_W"               { W-* }
QS  "L_X"               { X-* }
QS  "L_Y"               { Y-* }
```

Appendix C. Normal Probability Plots, Main Effect Plots, and Interaction Plots for Individual Triseme Analysis

C.1 Triseme L-E+F

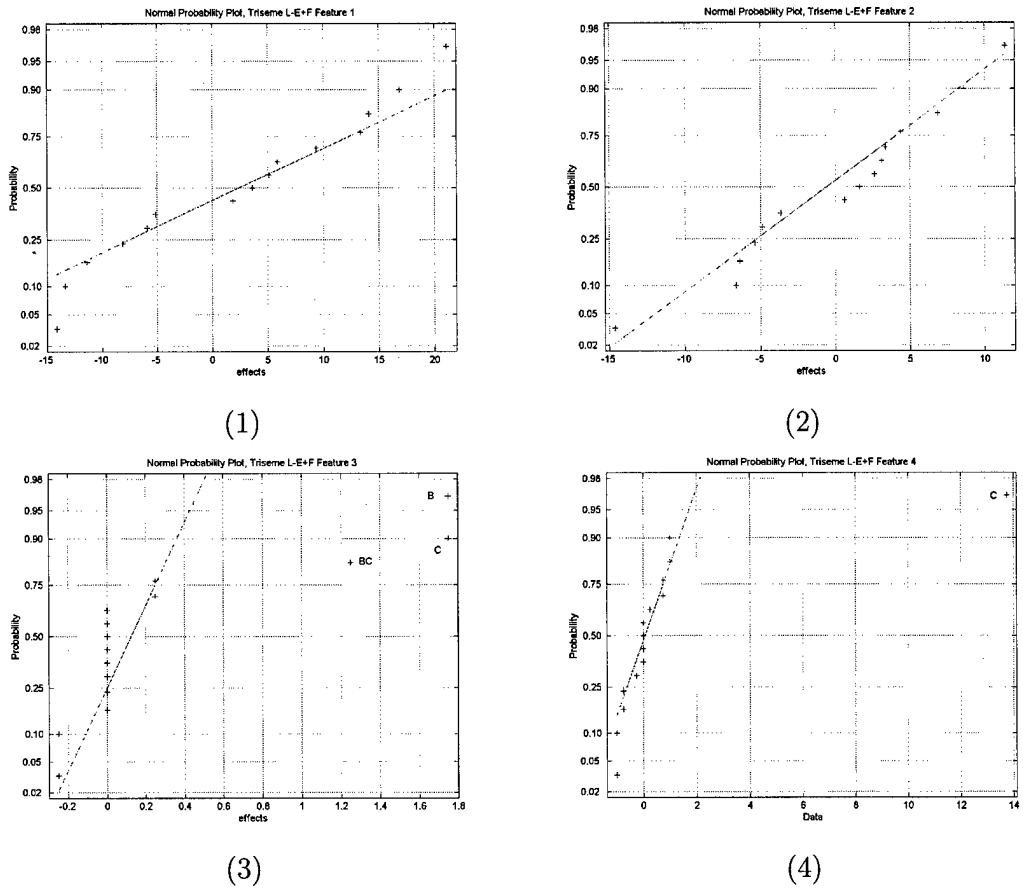
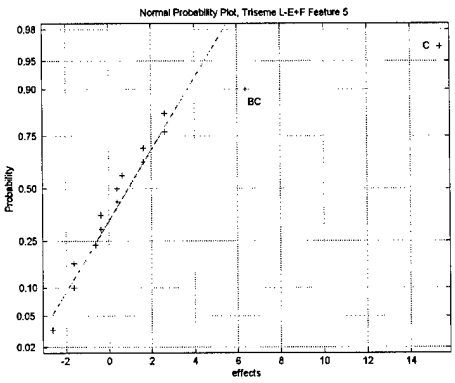
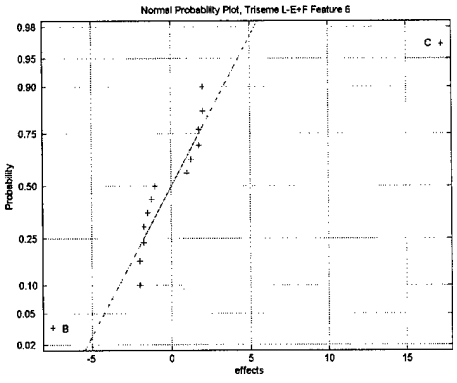


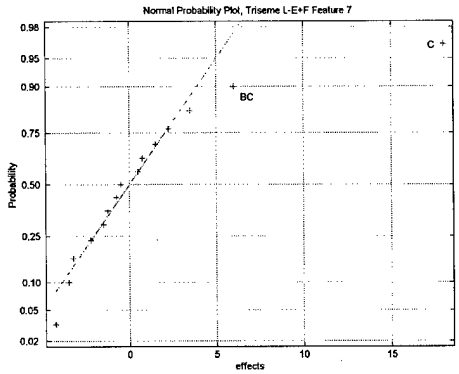
Figure 56. Normal probability plots of effects for triseme L-E+F features 1-4.



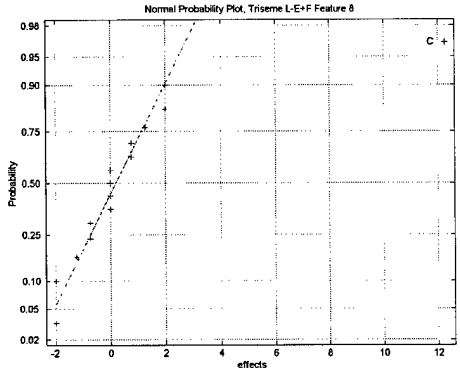
(5)



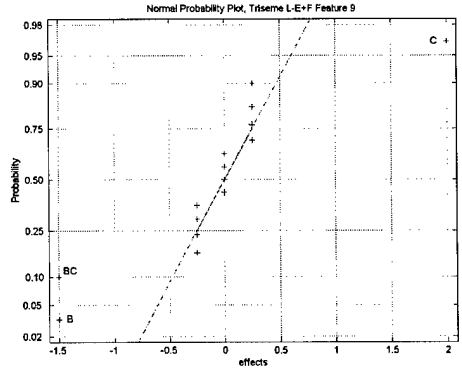
(6)



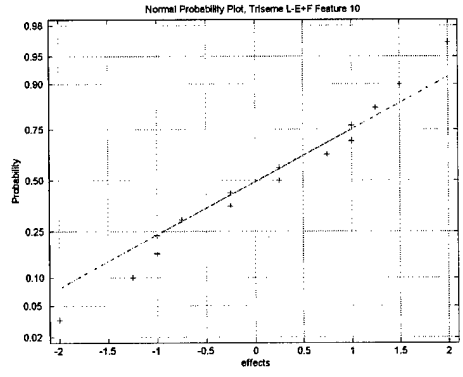
(7)



(8)

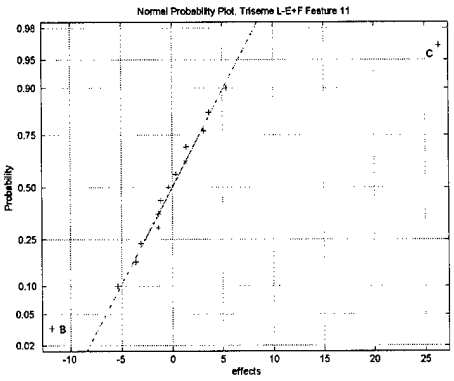


(9)

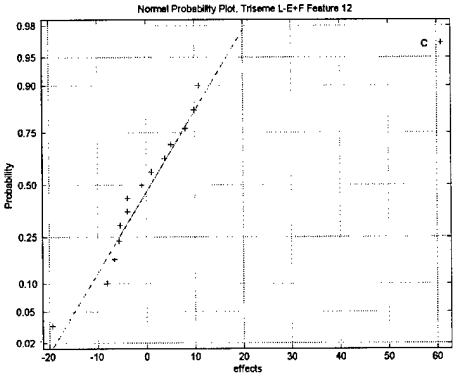


(10)

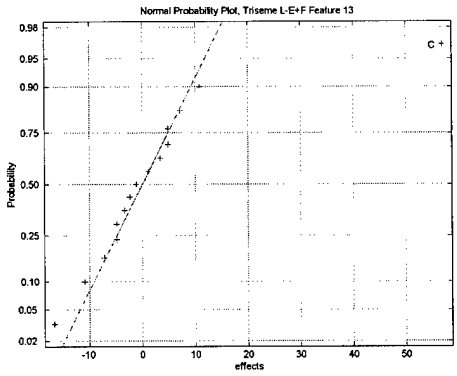
Figure 57. Normal probability plots of effects for triseme L-E+F features 5-10.



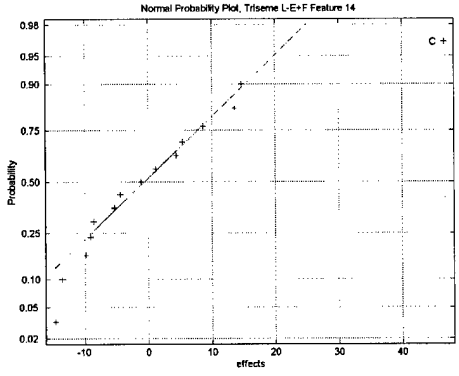
(11)



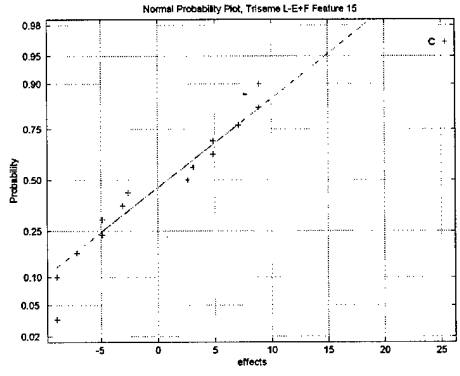
(12)



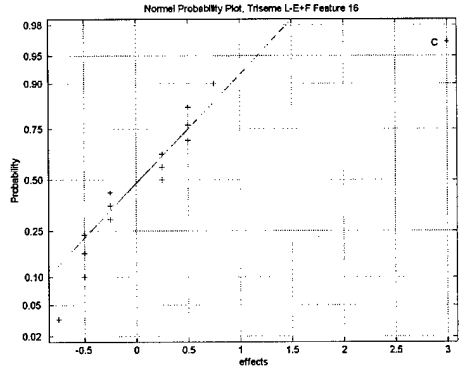
(13)



(14)

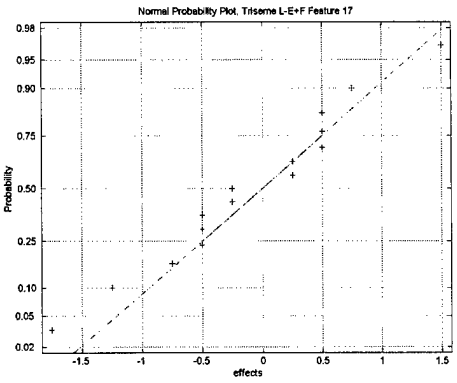


(15)

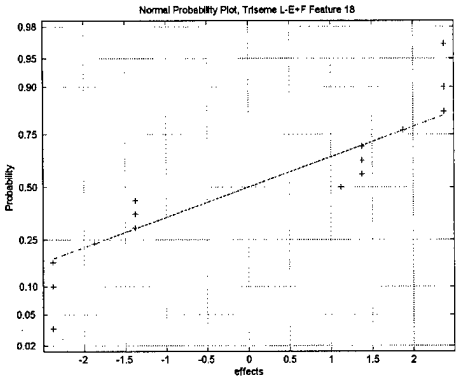


(16)

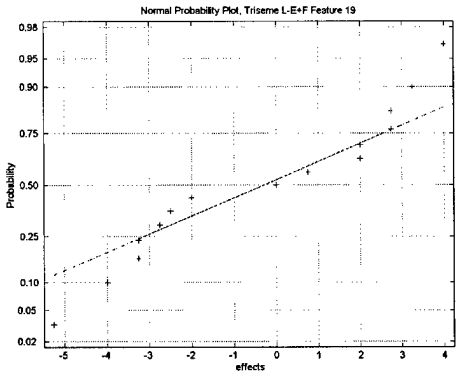
Figure 58. Normal probability plots of effects for triseme L-E+F features 11-16.



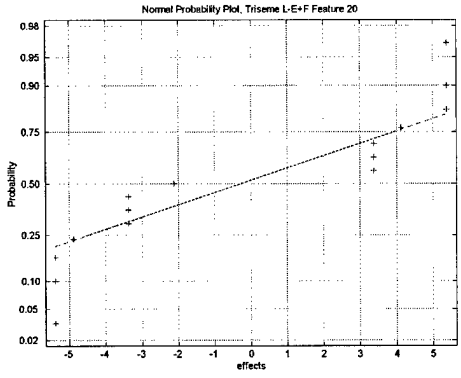
(17)



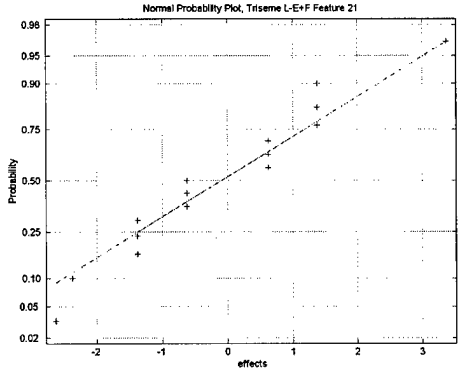
(18)



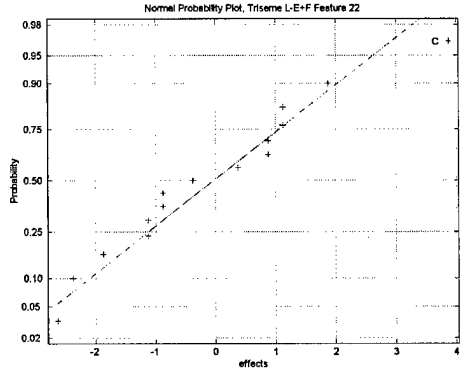
(19)



(20)

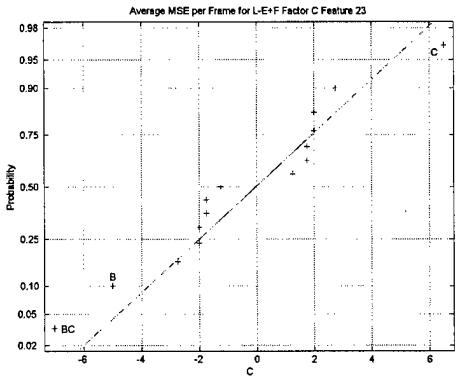


(21)

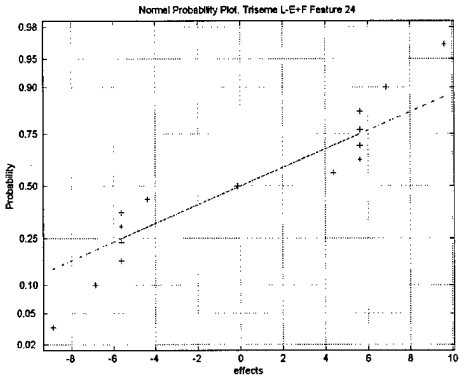


(22)

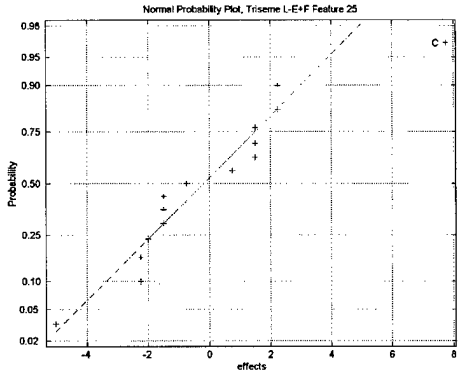
Figure 59. Normal probability plots of effects for triseme L-E+F features 17-22.



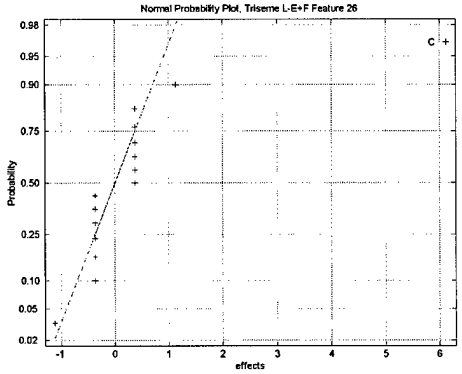
(23)



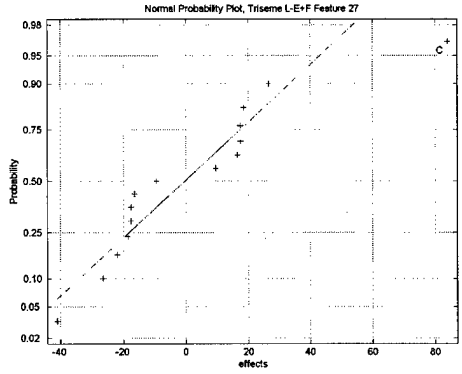
(24)



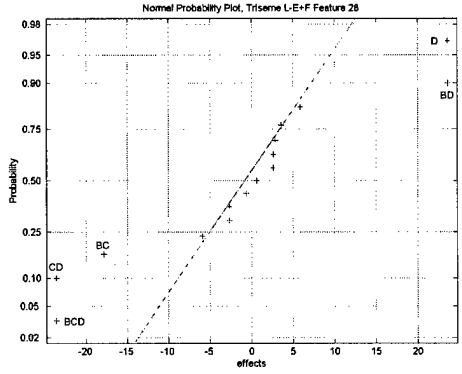
(25)



(26)



(27)



(28)

Figure 60. Normal probability plots of effects for triseme L-E+F features 23-28.

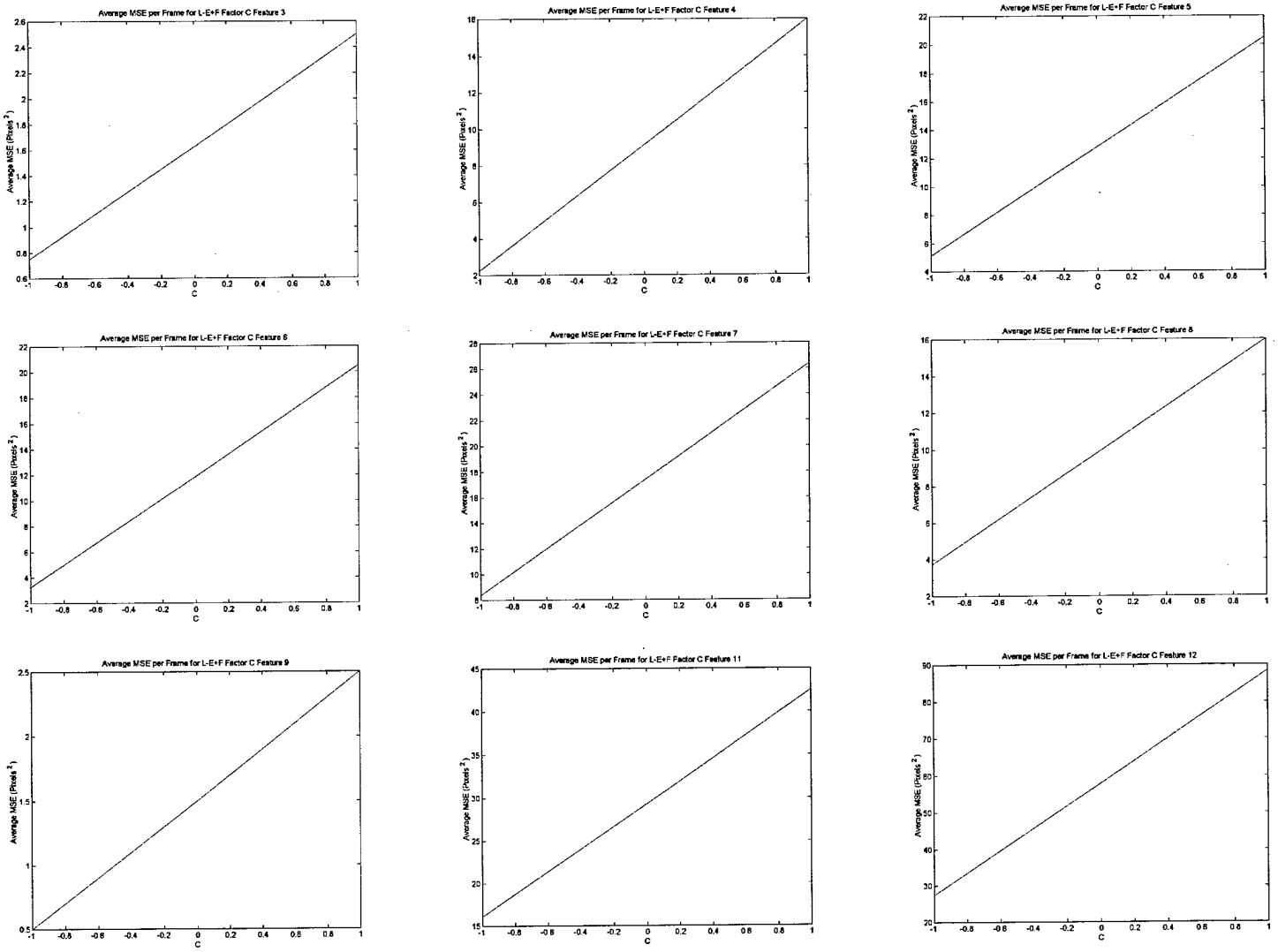


Figure 61. Main effect plots for factor C in triseme L-E+F for first 9 features where the effect of C was significant.

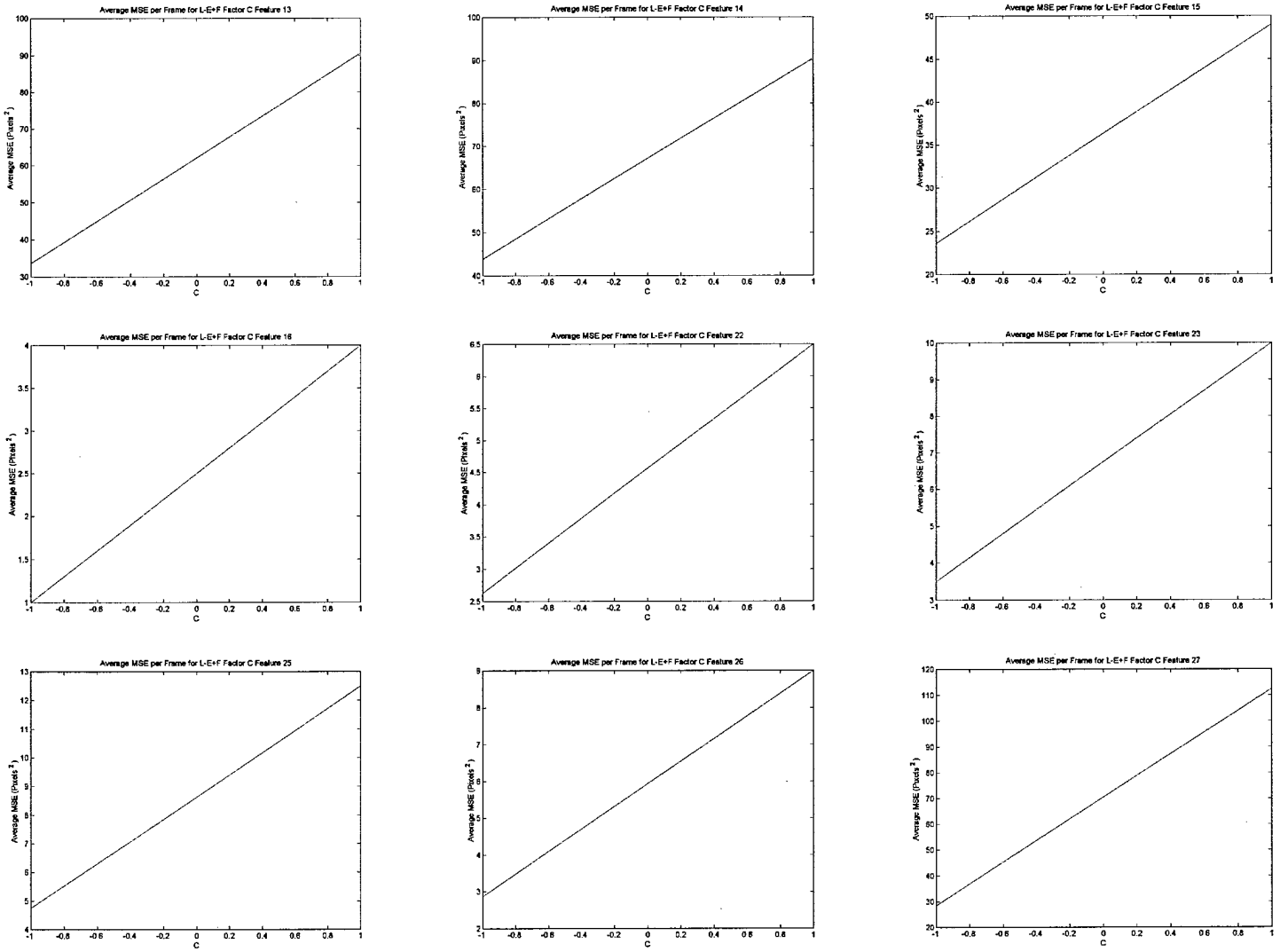


Figure 62. Main effect plots for factor C in triseme L-E+F for last 9 features where the effect of C was significant.

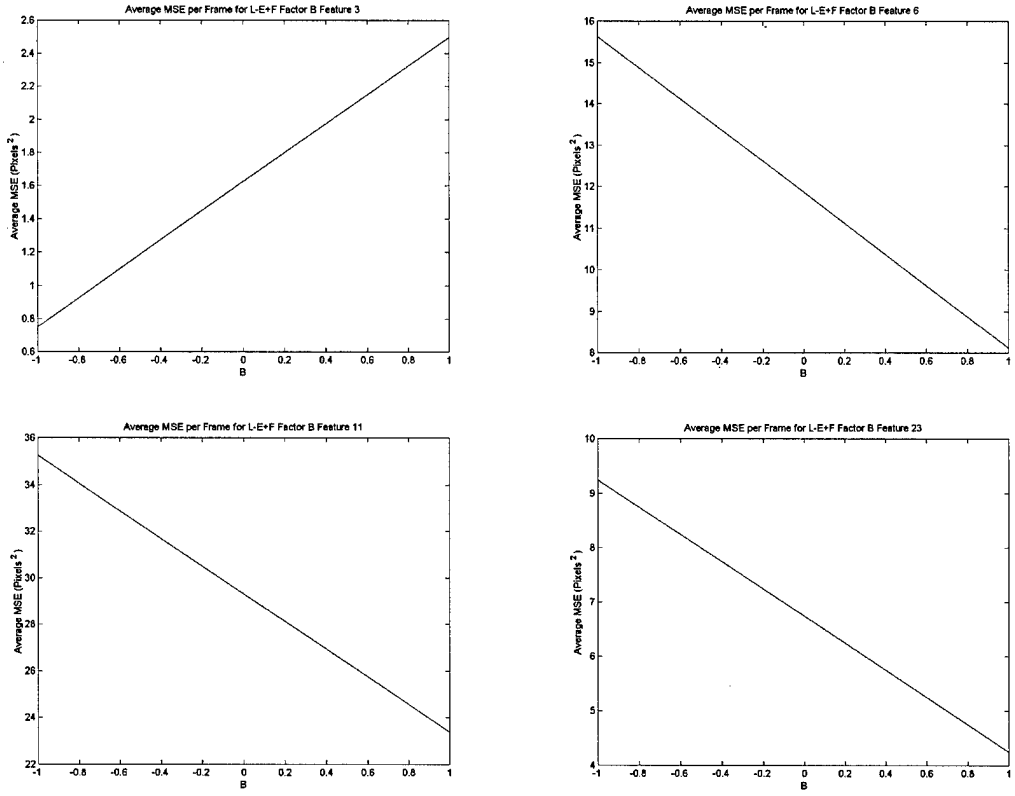


Figure 63. Main effect plots for factor B in triseme L-E+F for features 3, 6, 11 and 23 where the effect of B was significant.

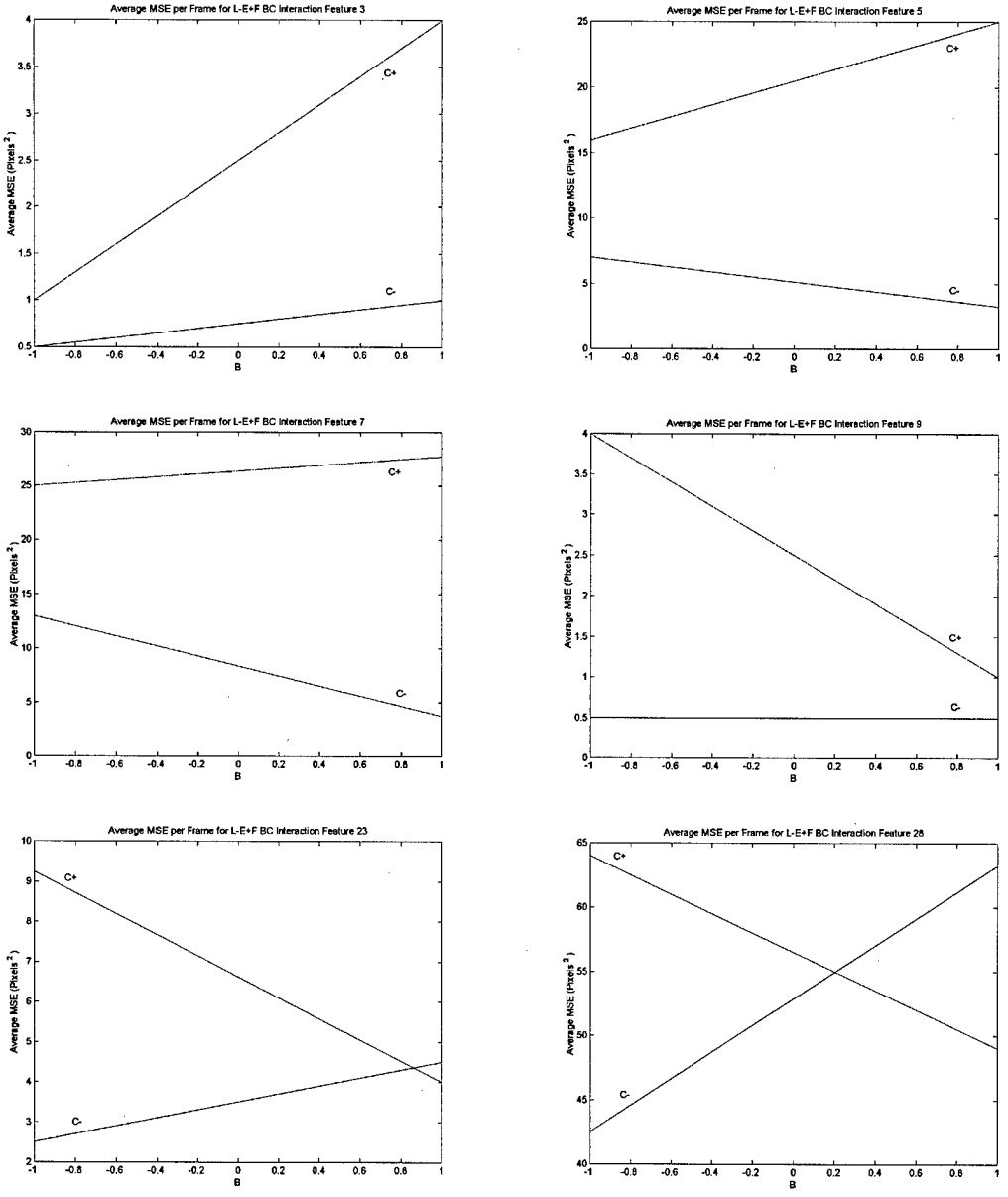


Figure 64. Interaction plots for factors B and C in triseme L-E+F for features 3, 5, 7, 9, 23 and 28 where BC interaction was significant.

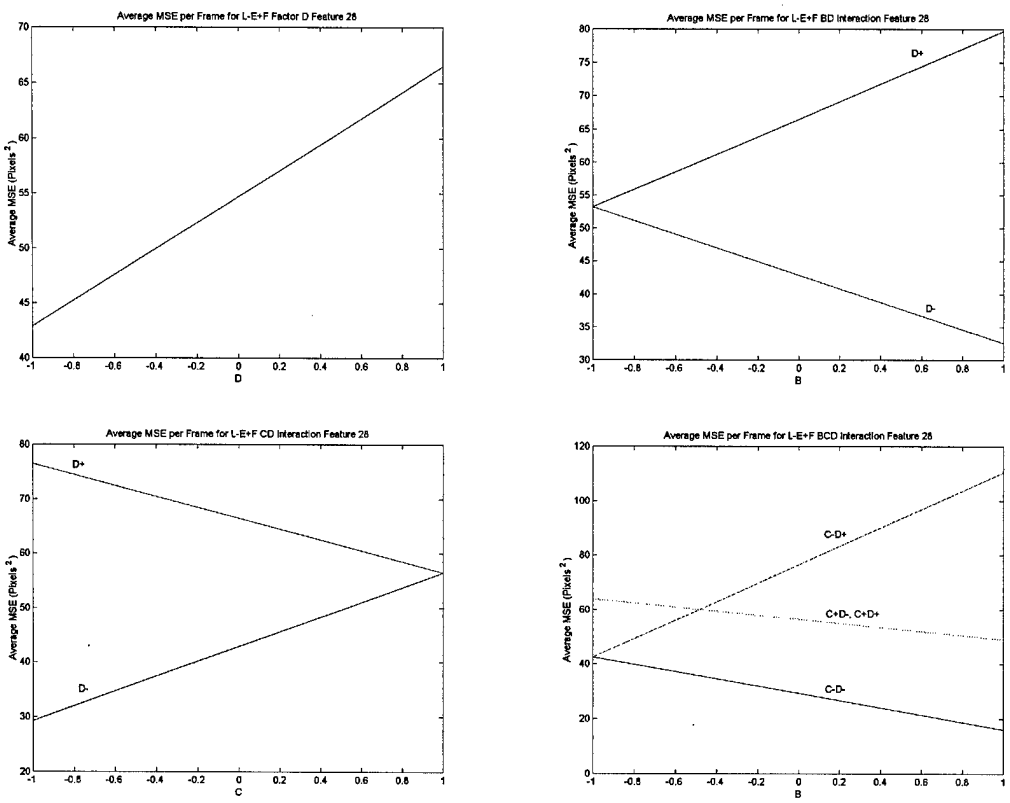


Figure 65. Main effect and interaction plots for factor D, BD, CD and BCD for feature 28 in triseme L-E+F.

C.2 *Triseme K-D+C*

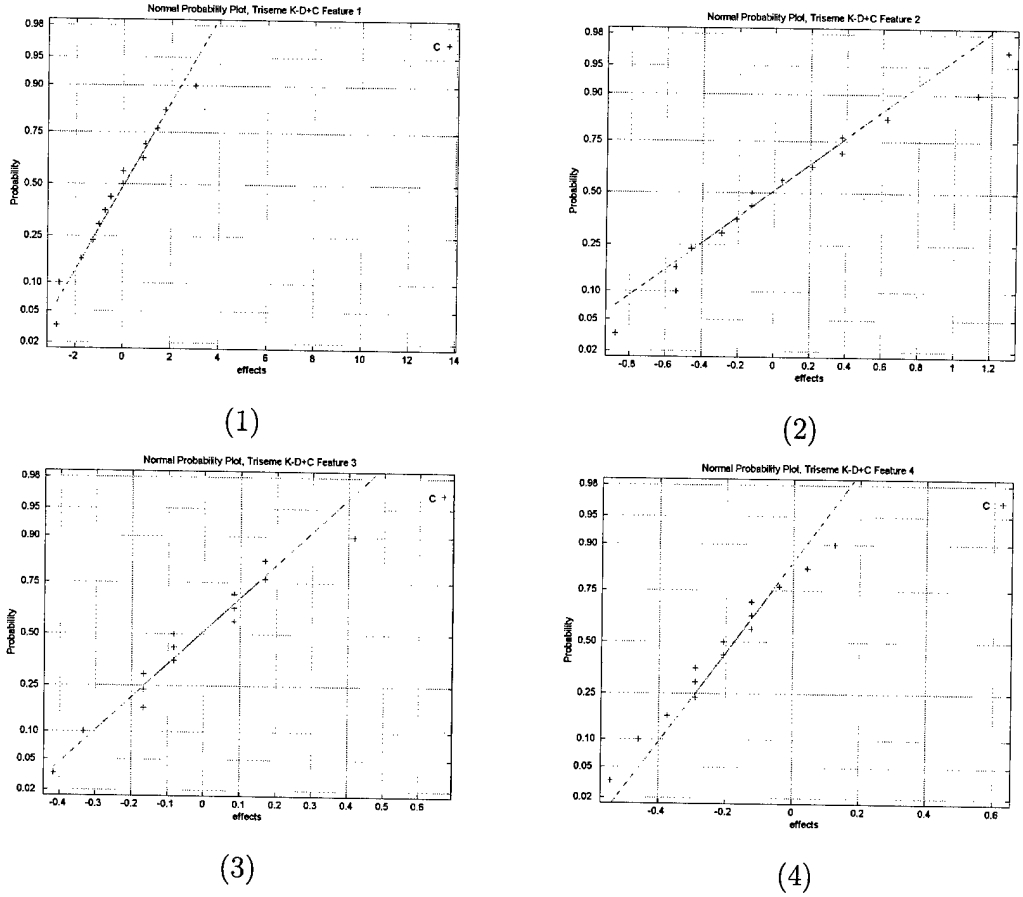
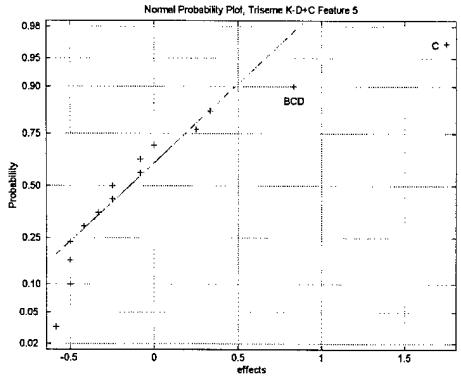
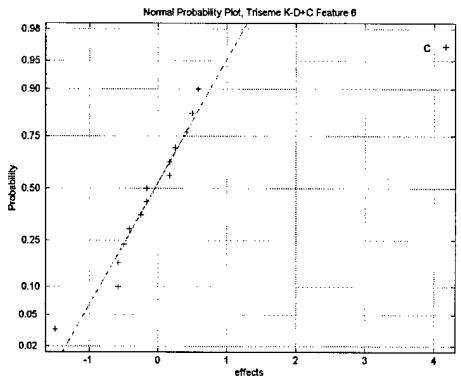


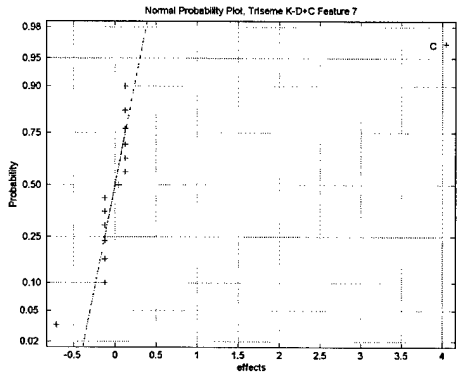
Figure 66. Normal probability plots of effects for triseme K-D+C features 1-4.



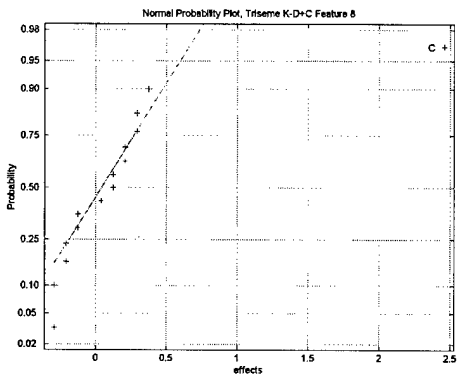
(5)



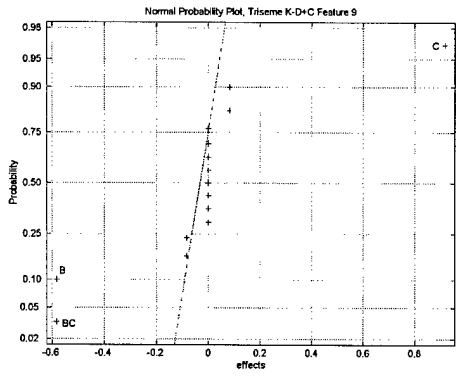
(6)



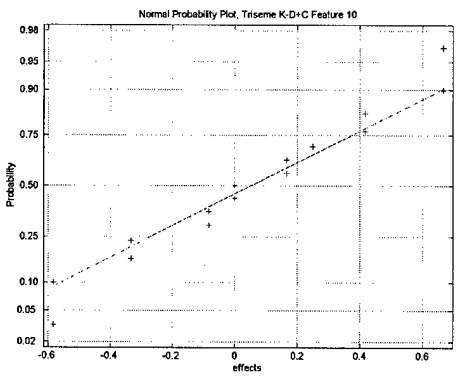
(7)



(8)

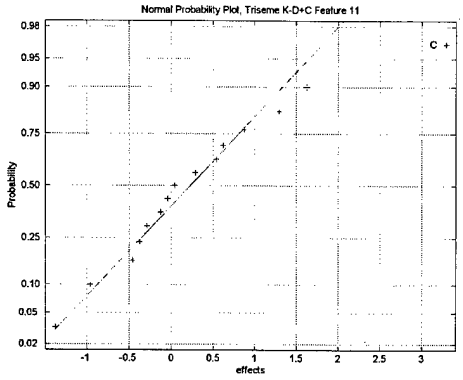


(9)

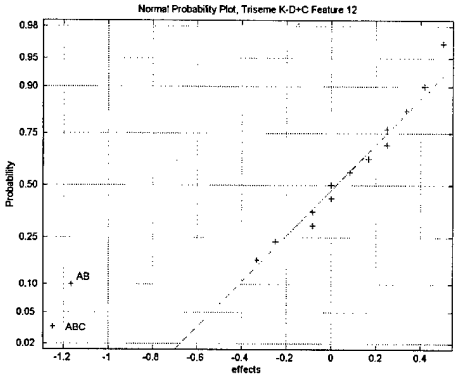


(10)

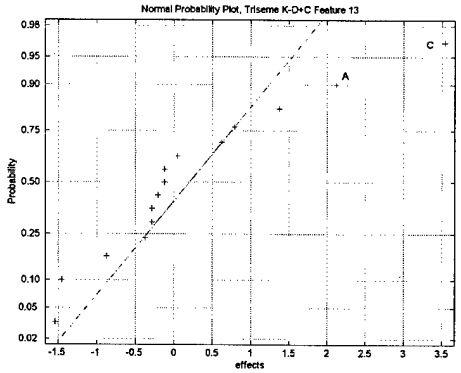
Figure 67. Normal probability plots of effects for triseme K-D+C features 5-10.



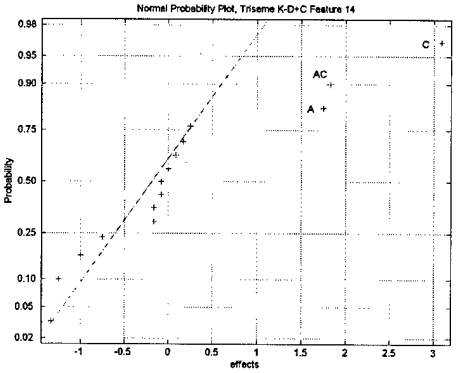
(11)



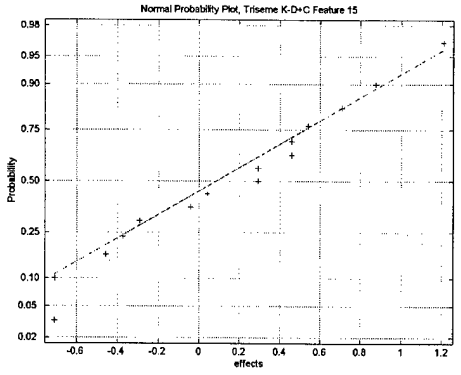
(12)



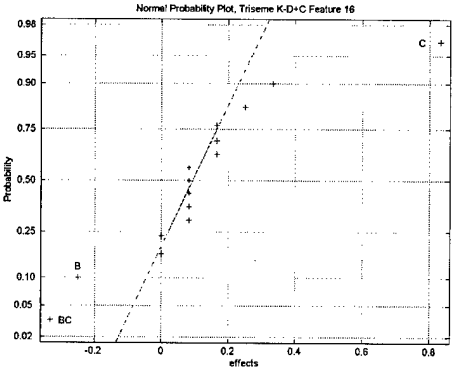
(13)



(14)

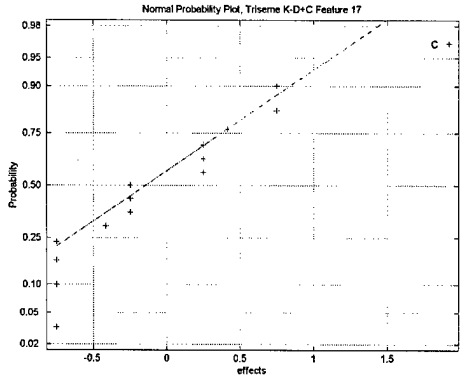


(15)

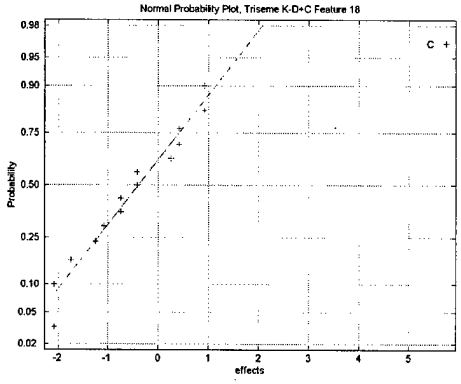


(16)

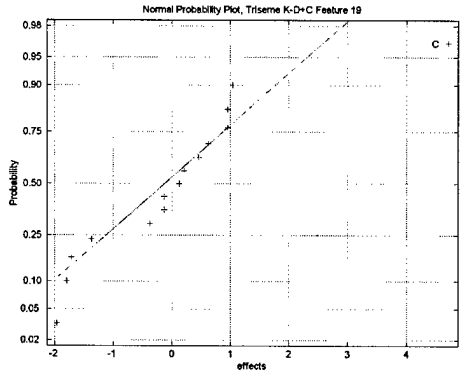
Figure 68. Normal probability plots of effects for triseme K-D+C features 11-16.



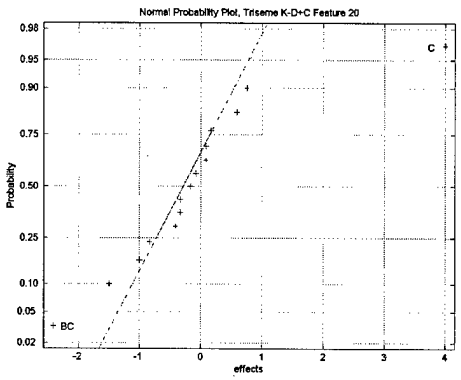
(17)



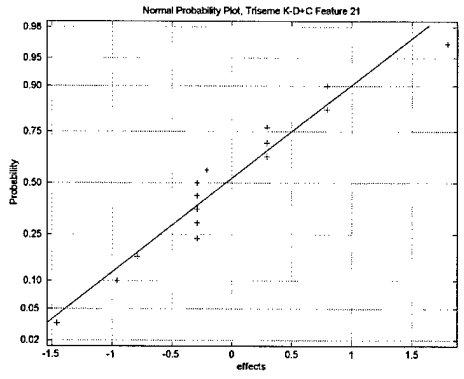
(18)



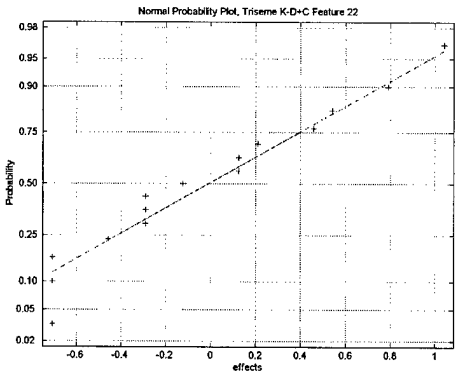
(19)



(20)

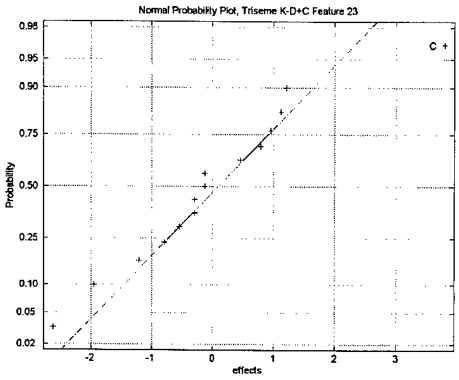


(21)

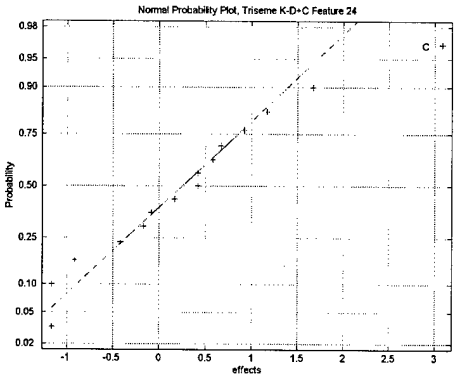


(22)

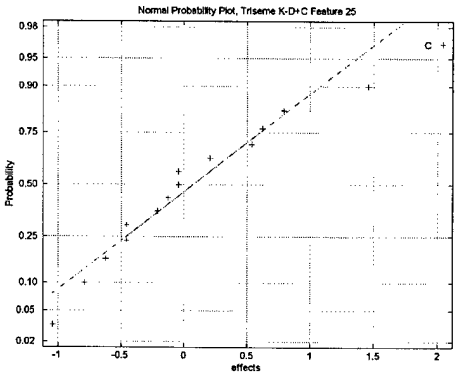
Figure 69. Normal probability plots of effects for triseme K-D+C features 17-22.



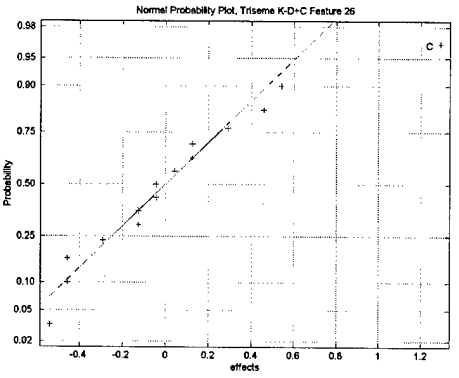
(23)



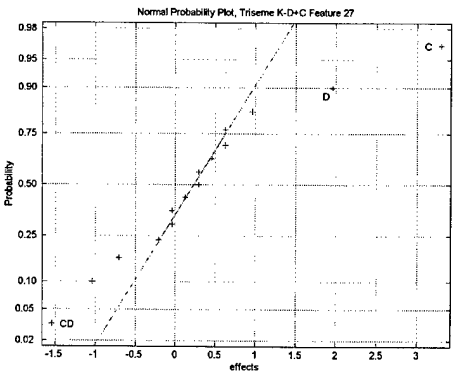
(24)



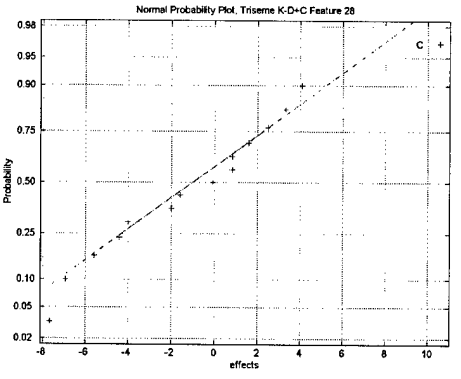
(25)



(26)



(27)



(28)

Figure 70. Normal probability plots of effects for triseme K-D+C features 23-28.

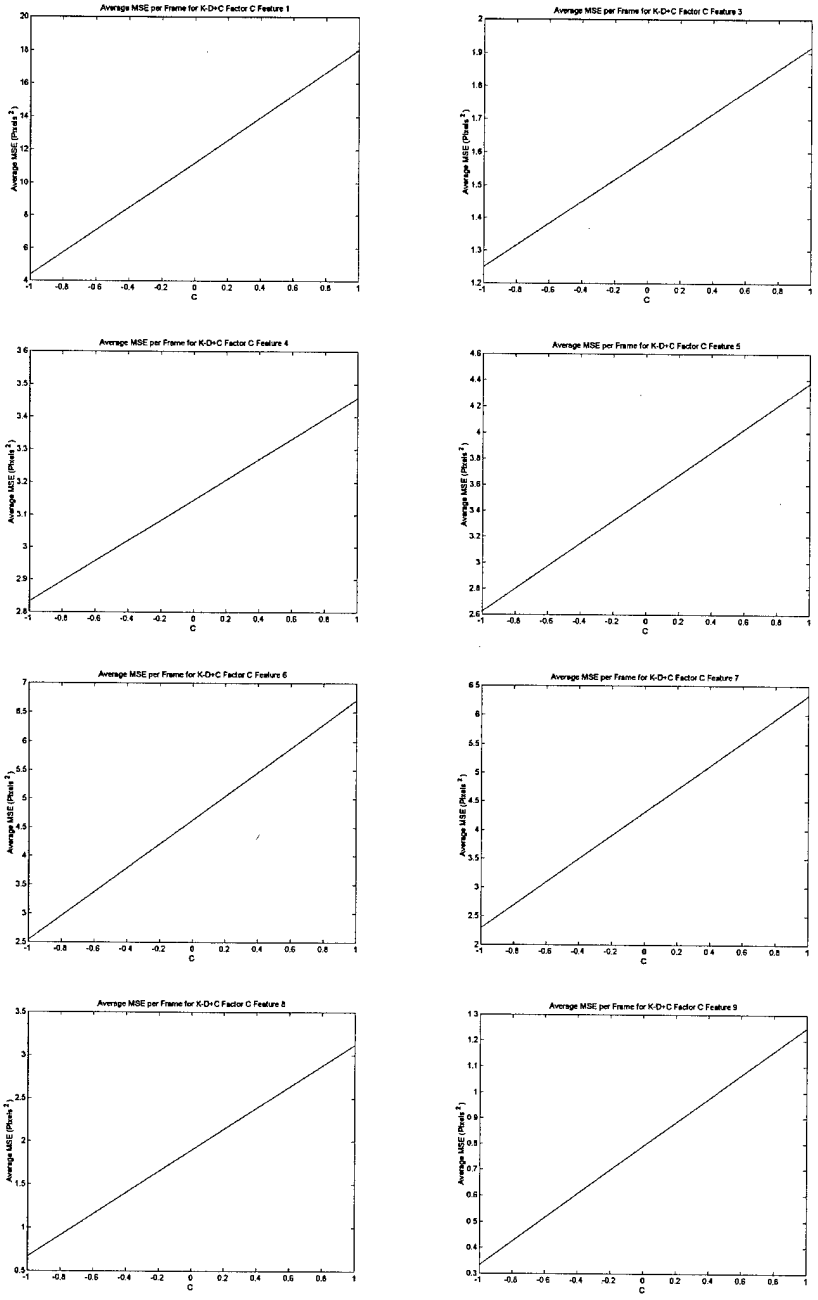


Figure 71. Main effect plots for factor C in triseme K-D+C for features 1 and 3-9.

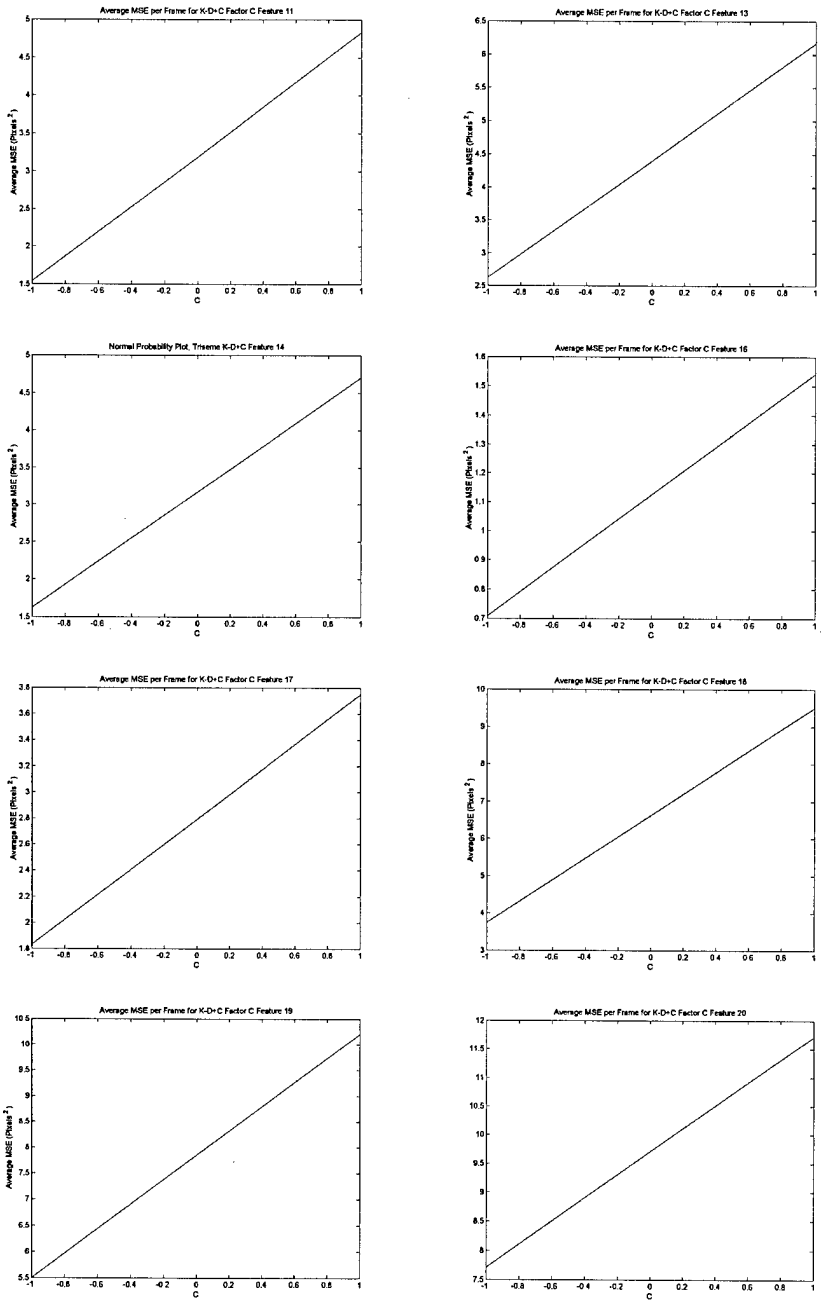


Figure 72. Main effect plots for factor C in triseme K-D+C for features 11, 13, 14 and 16-20.

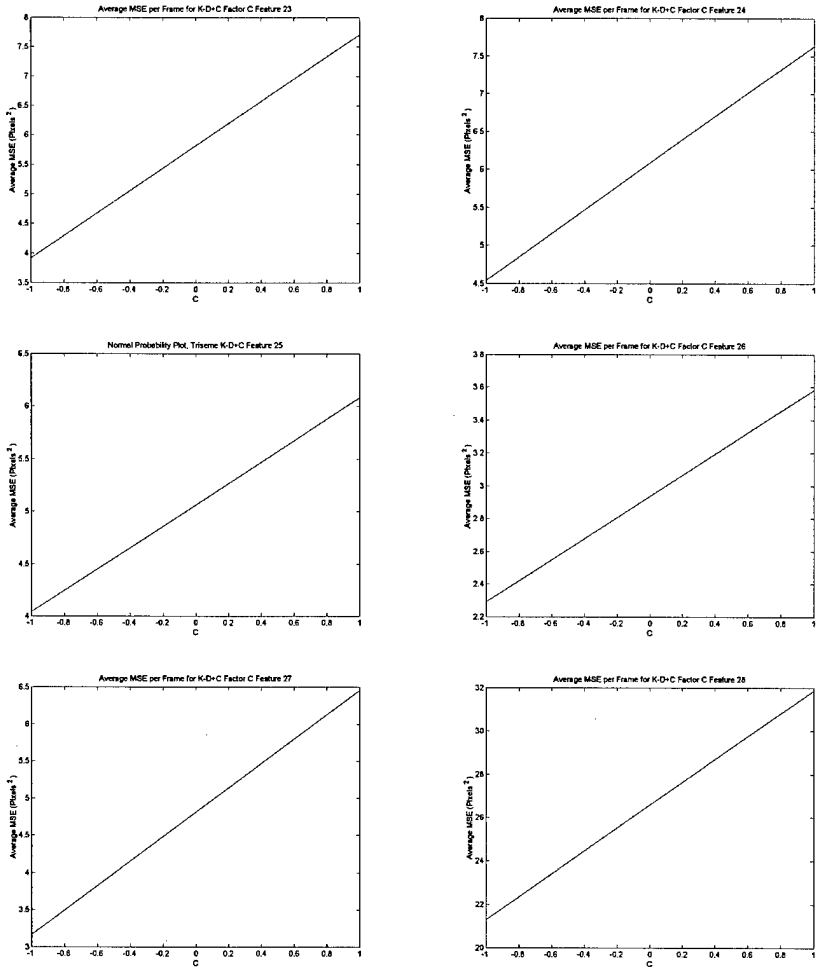


Figure 73. Main effect plots for factor C in triseme K-D+C for features 23-28.

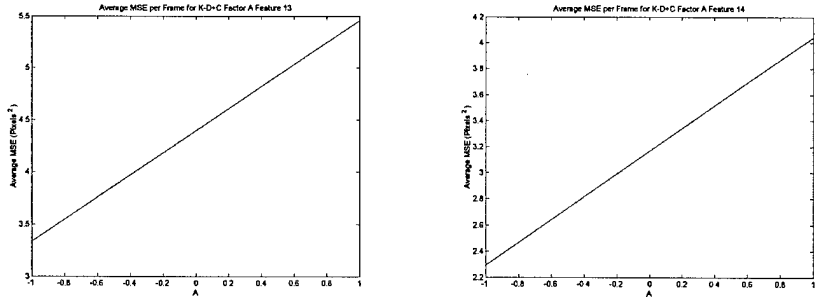


Figure 74. Main effects plots for factor A in features 13 and 14 of triseme K-D+C.

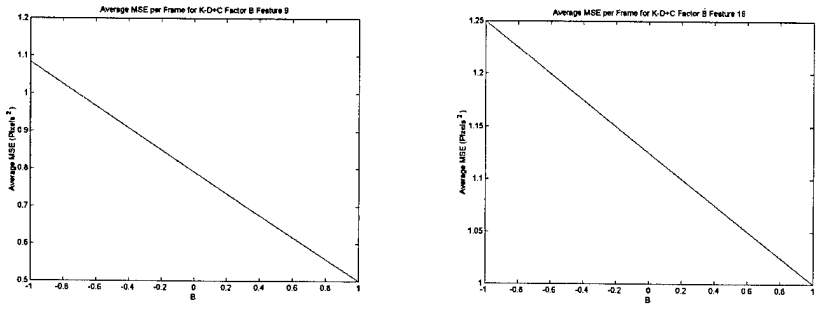


Figure 75. Main effect plots for factor B in features 9 and 16 of triseme K-D+C.

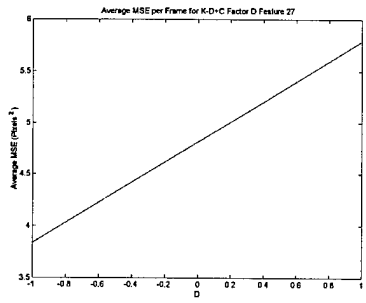


Figure 76. Main effect plot for factor D in feature 27 of triseme K-D+C.

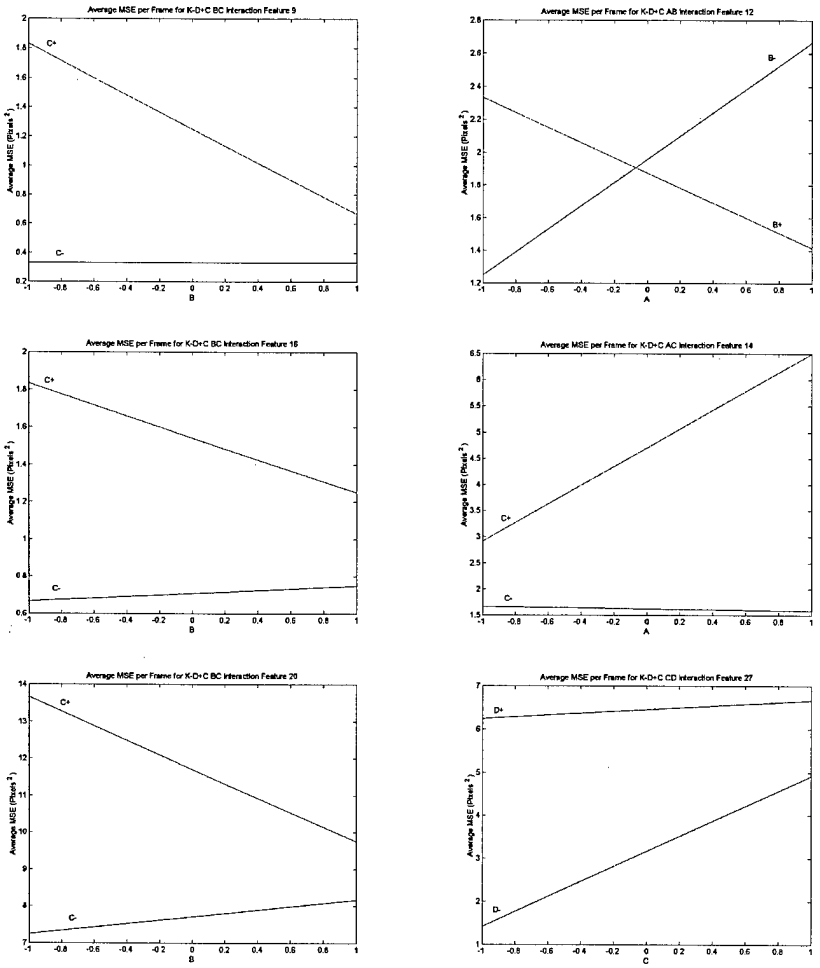


Figure 77. Plots of for the two factor interactions: BC (the 3 left plots), AB (top right), AC (middle right) and CD (bottom right) for triseme K-D+C.

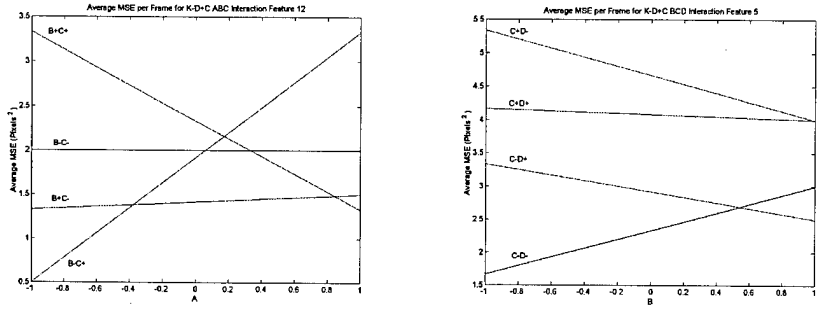


Figure 78. Three factor interaction plots for ABC (left) and BCD (right) interactions in features 12 and 5 respectively of triseme K-D+C.

C.3 Triseme K-E+G

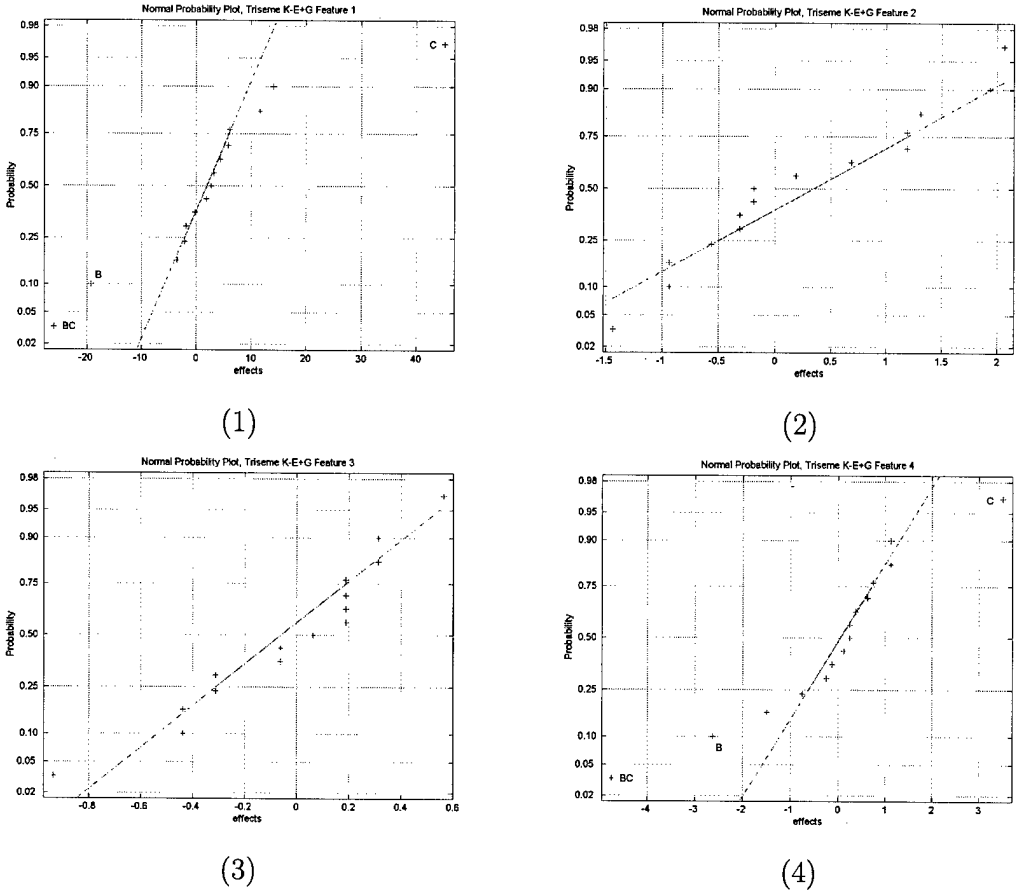


Figure 79. Normal probability plots of effects for triseme K-E+G features 1-4.

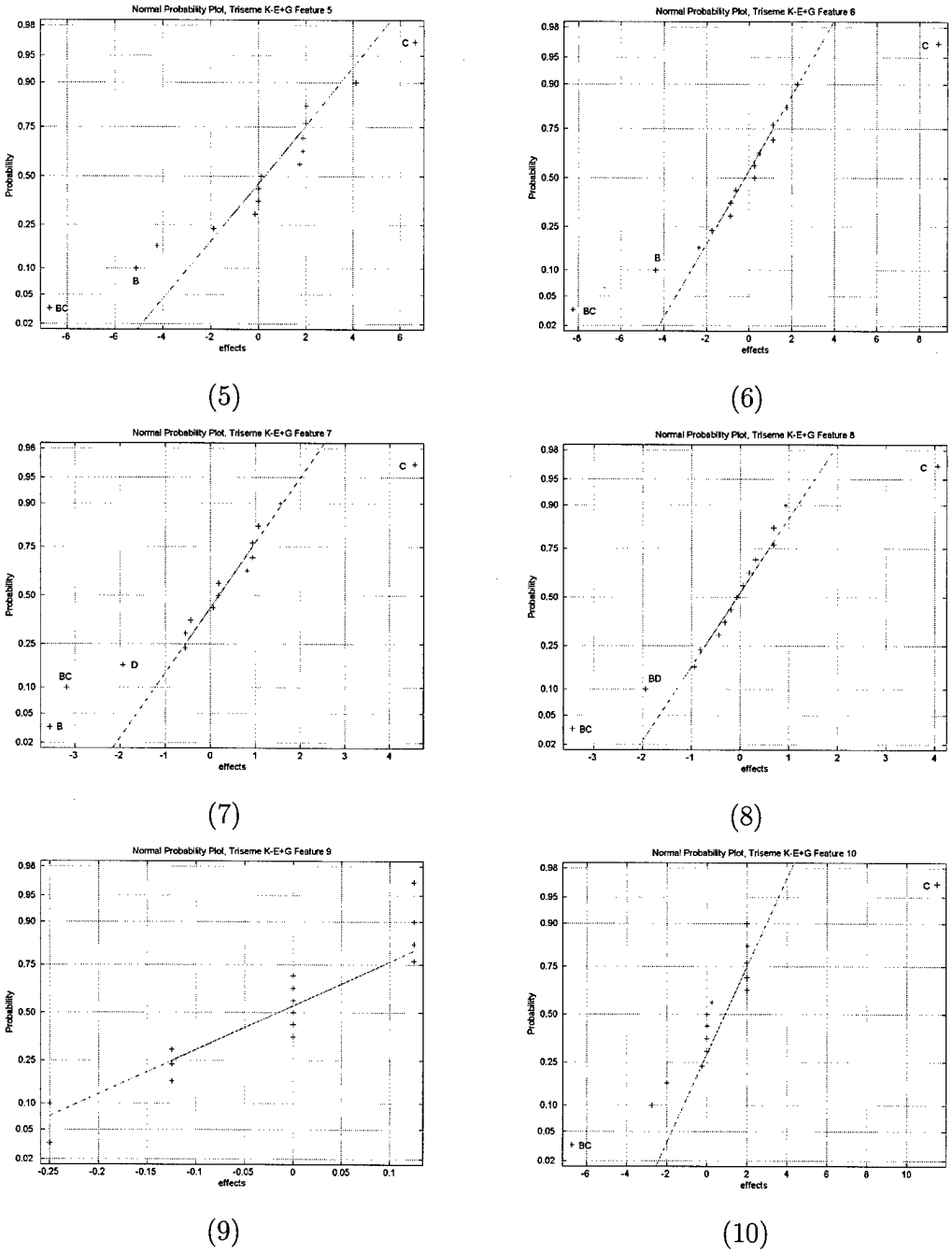
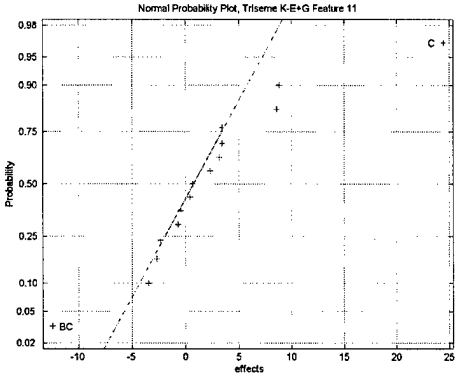
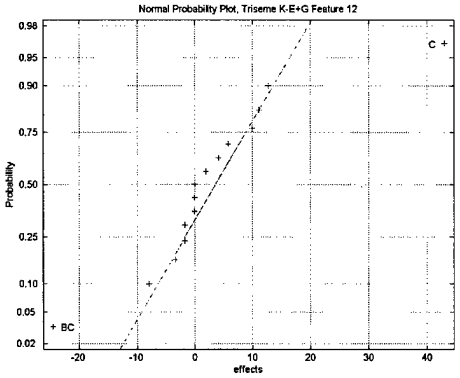


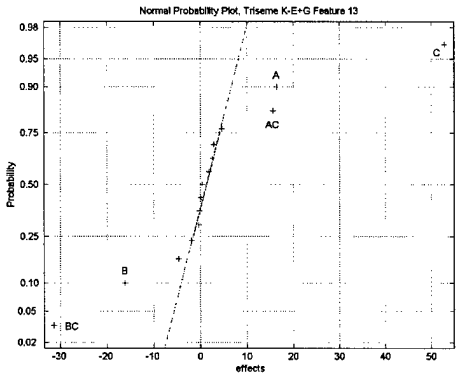
Figure 80. Normal probability plots of effects for triseme K-E+G features 5-10.



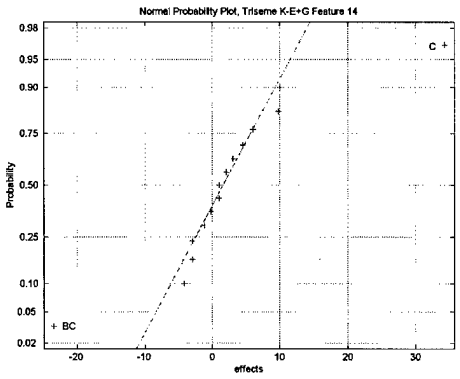
(11)



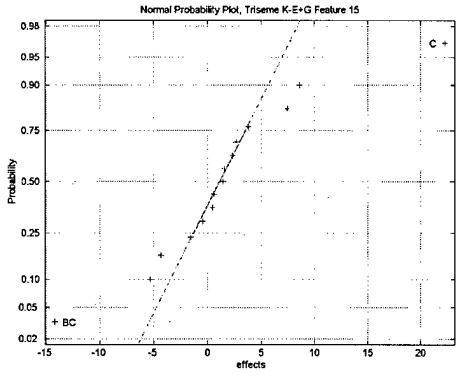
(12)



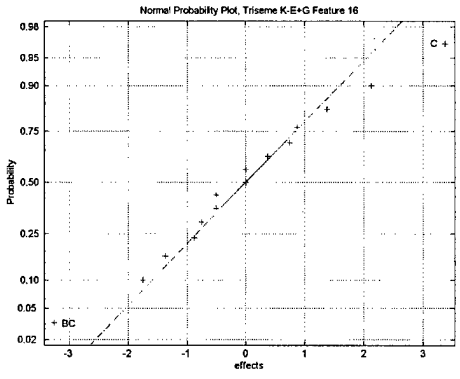
(13)



(14)

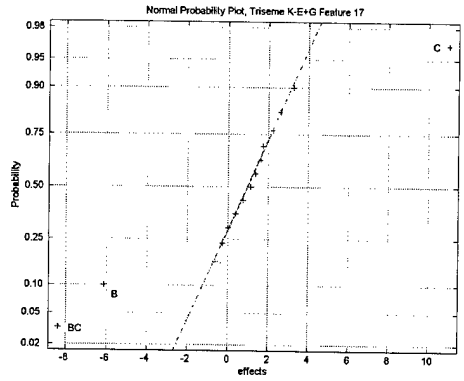


(15)

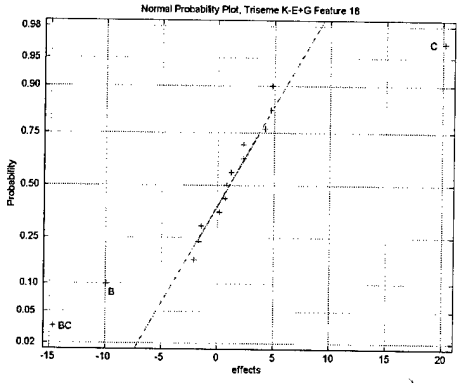


(16)

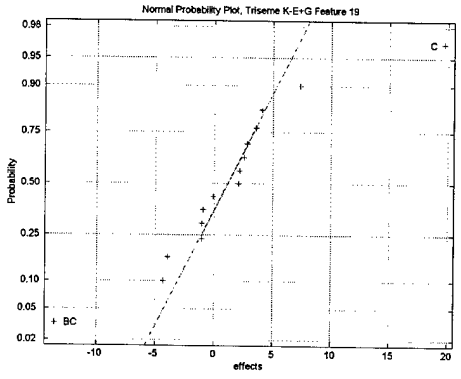
Figure 81. Normal probability plots of effects for triseme K-E+G features 11-16.



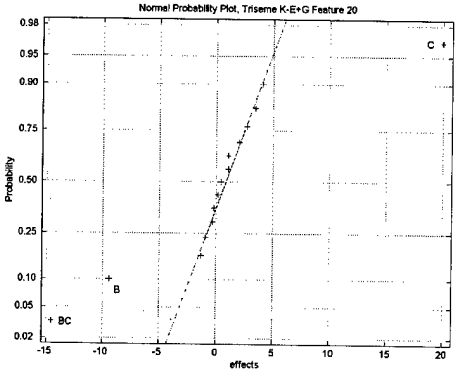
(17)



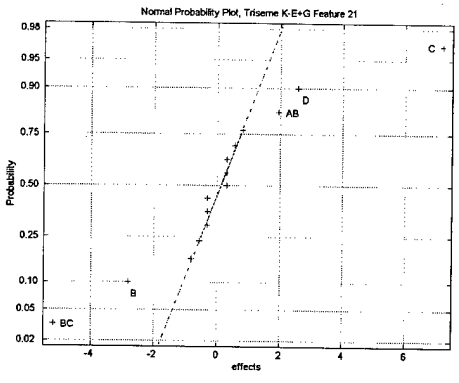
(18)



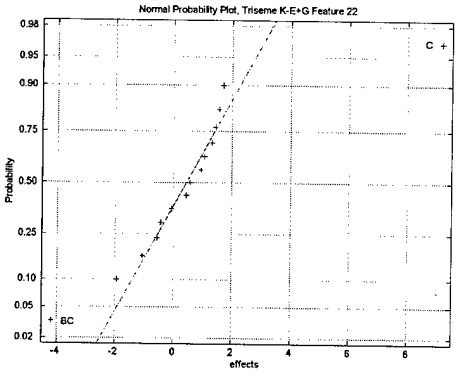
(19)



(20)

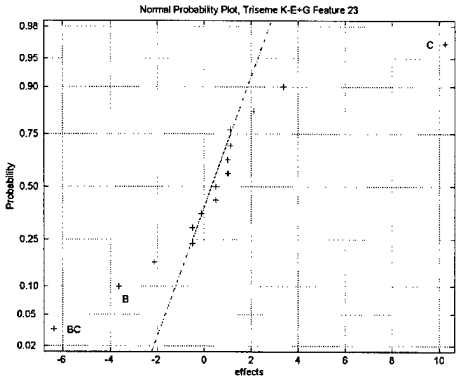


(21)

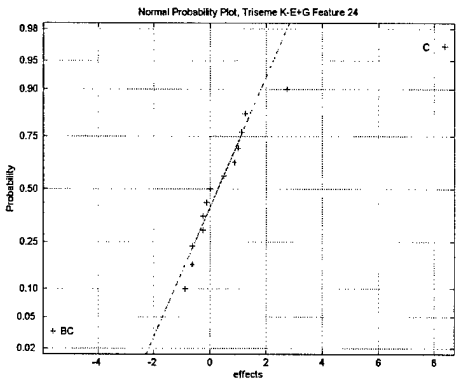


(22)

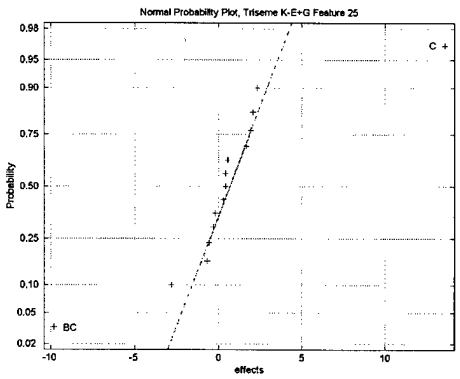
Figure 82. Normal probability plots of effects for triseme K-E+G features 17-22.



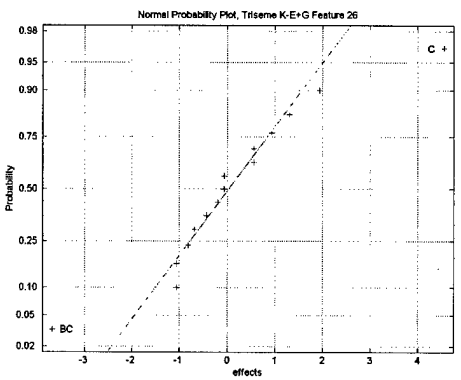
(23)



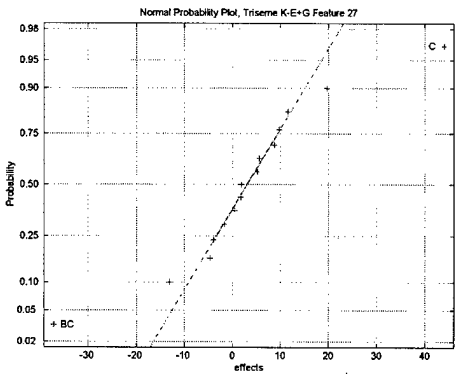
(24)



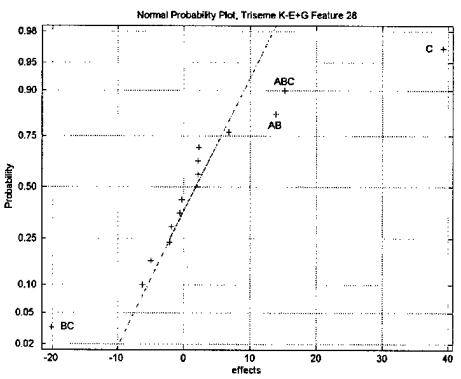
(25)



(26)



(27)



(28)

Figure 83. Normal probability plots of effects for triseme K-E+G features 23-28.

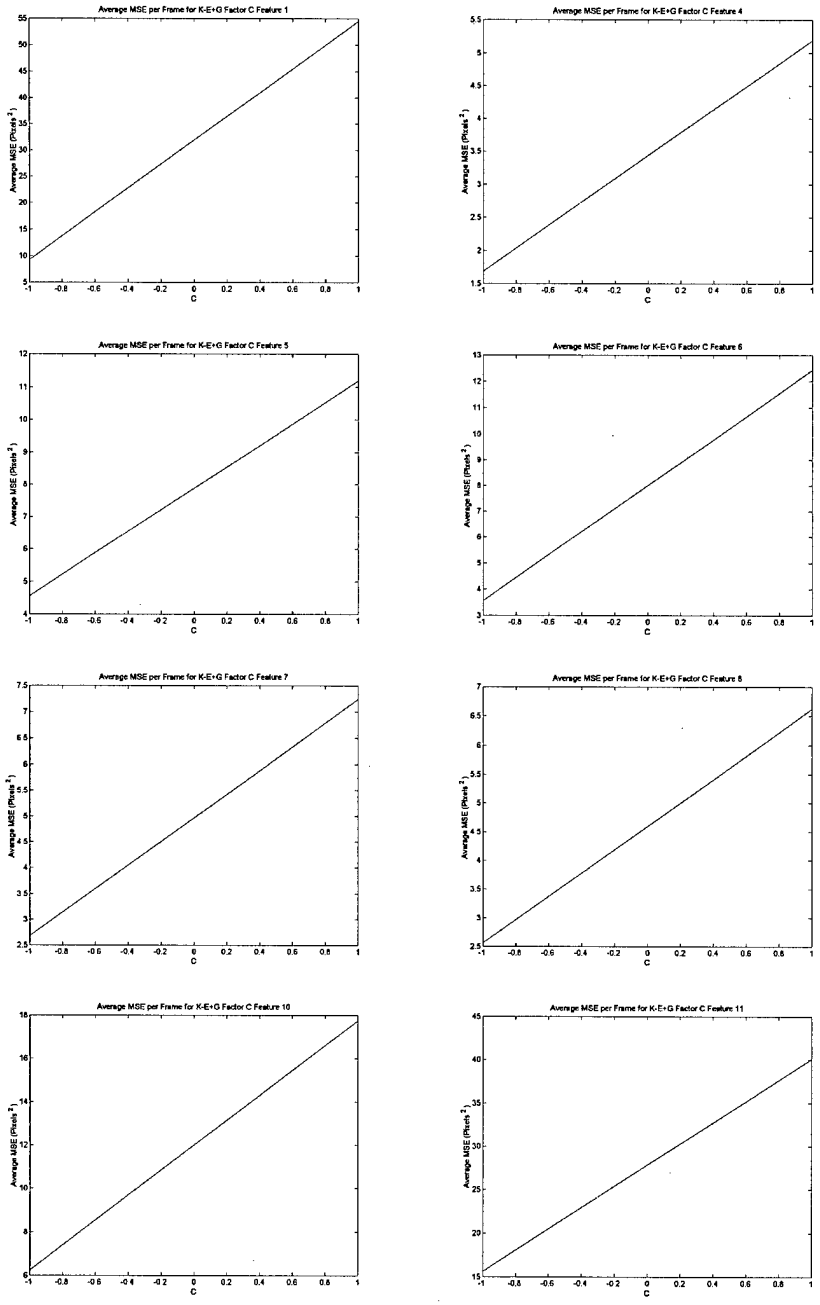


Figure 84. Main effect plots for factor C in triseme K-E+G for features 1, 4-8 and 10 and 11.

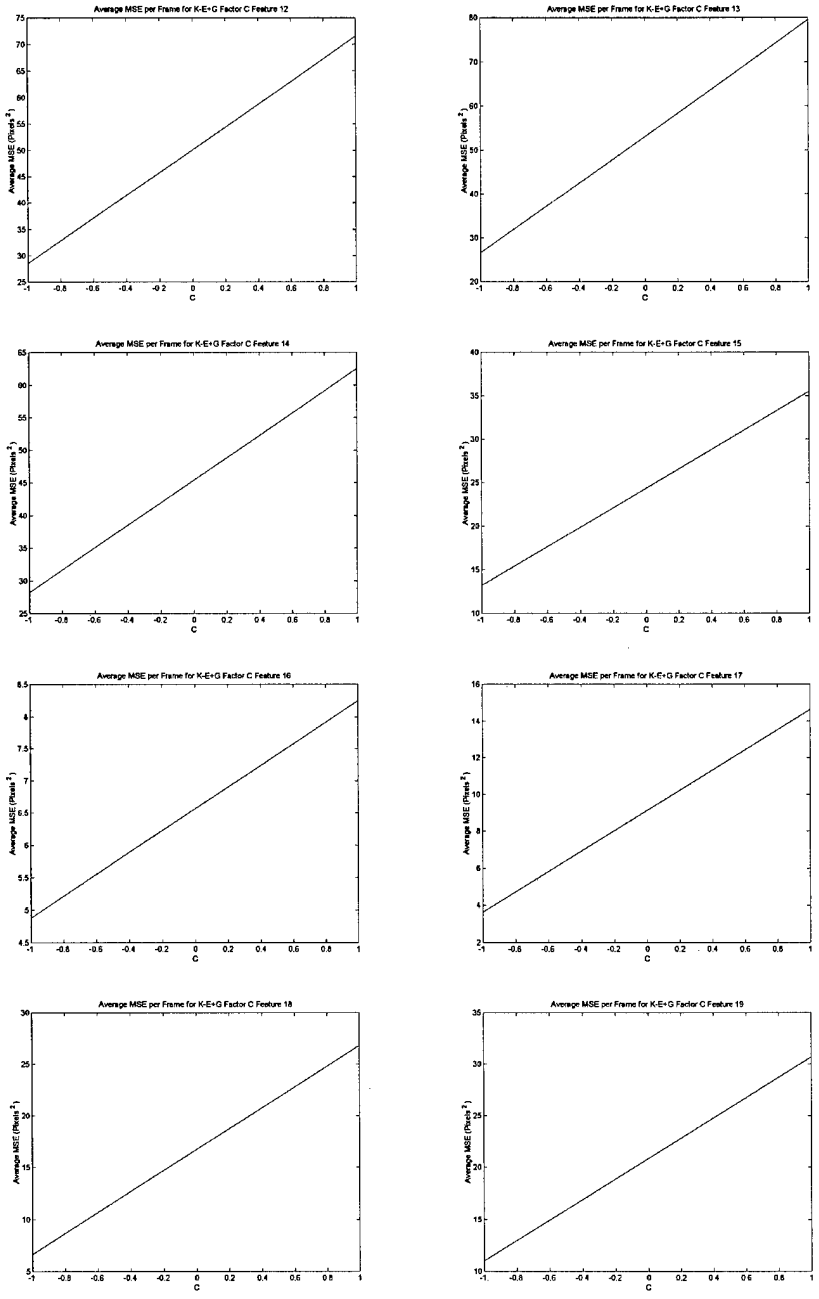


Figure 85. Main effect plots for factor C in triseme K-E+G for features 12-19.

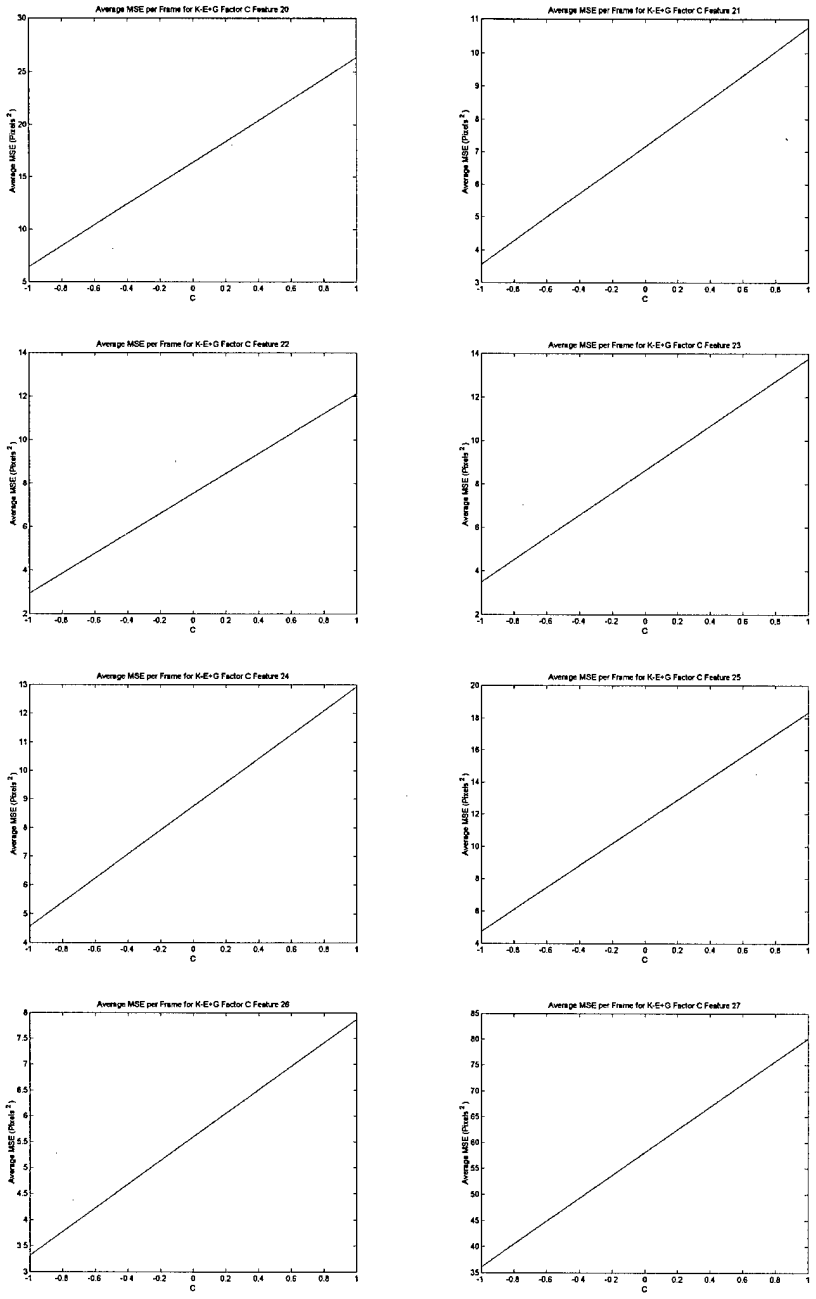


Figure 86. Main effect plots for factor C in triseme K-E+G for features 20-27.

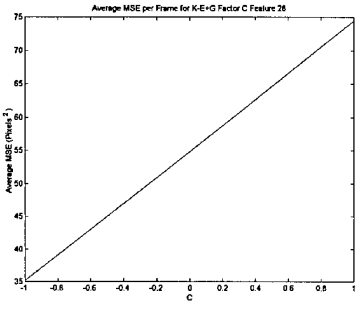


Figure 87. Main effect plot for factor C in triseme K-E+G for feature 28.

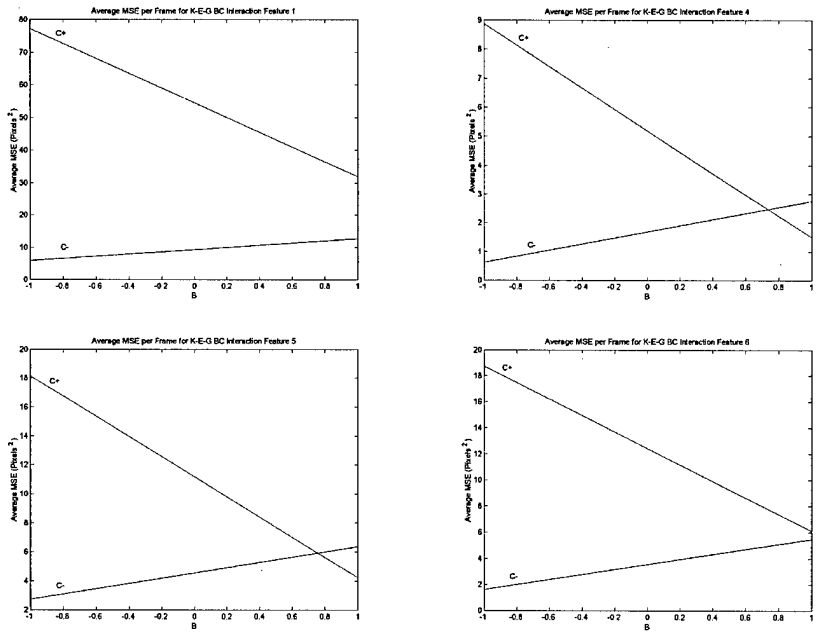


Figure 88. Plots for BC interaction in triseme K-E+G for features 1, and 4-6.

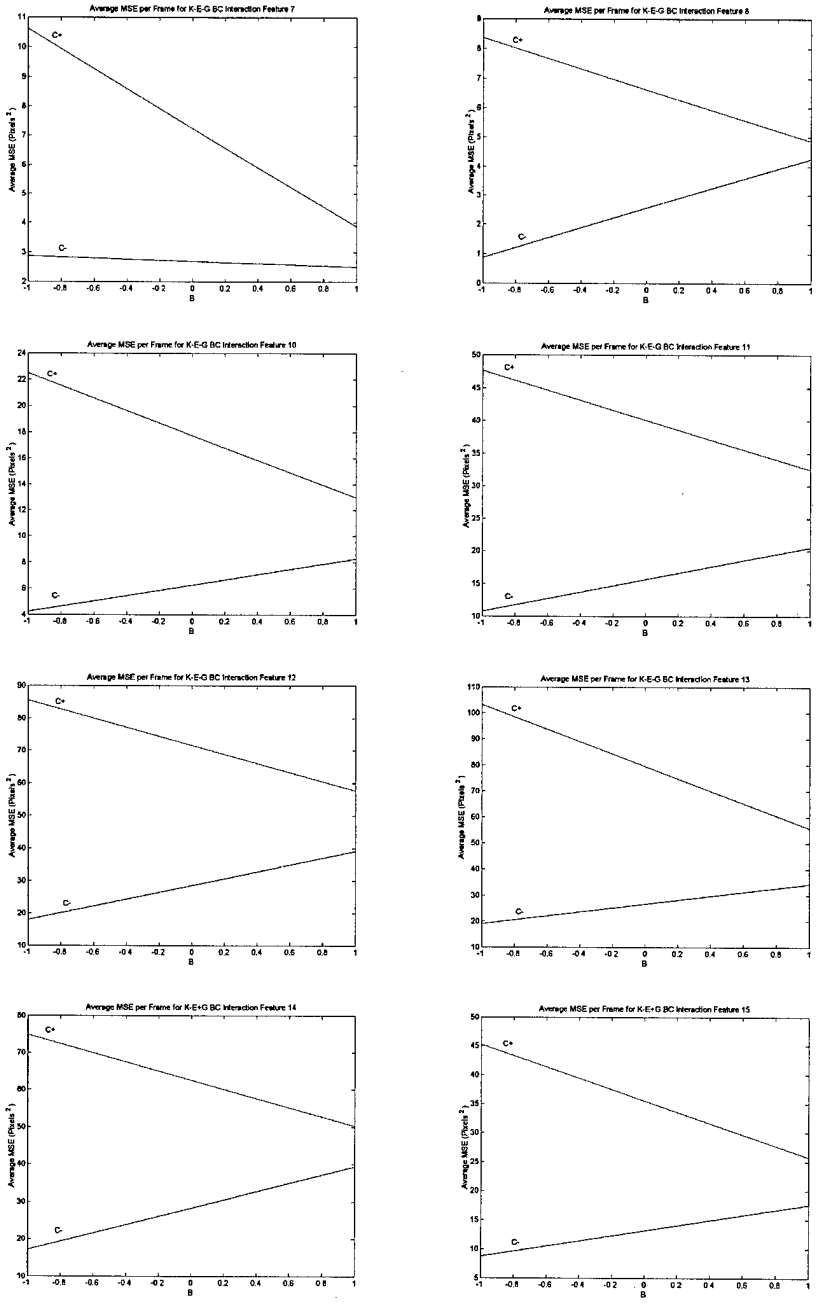


Figure 89. Plots for BC interaction in triseme K-E+G for features 7,8 and 10-15.

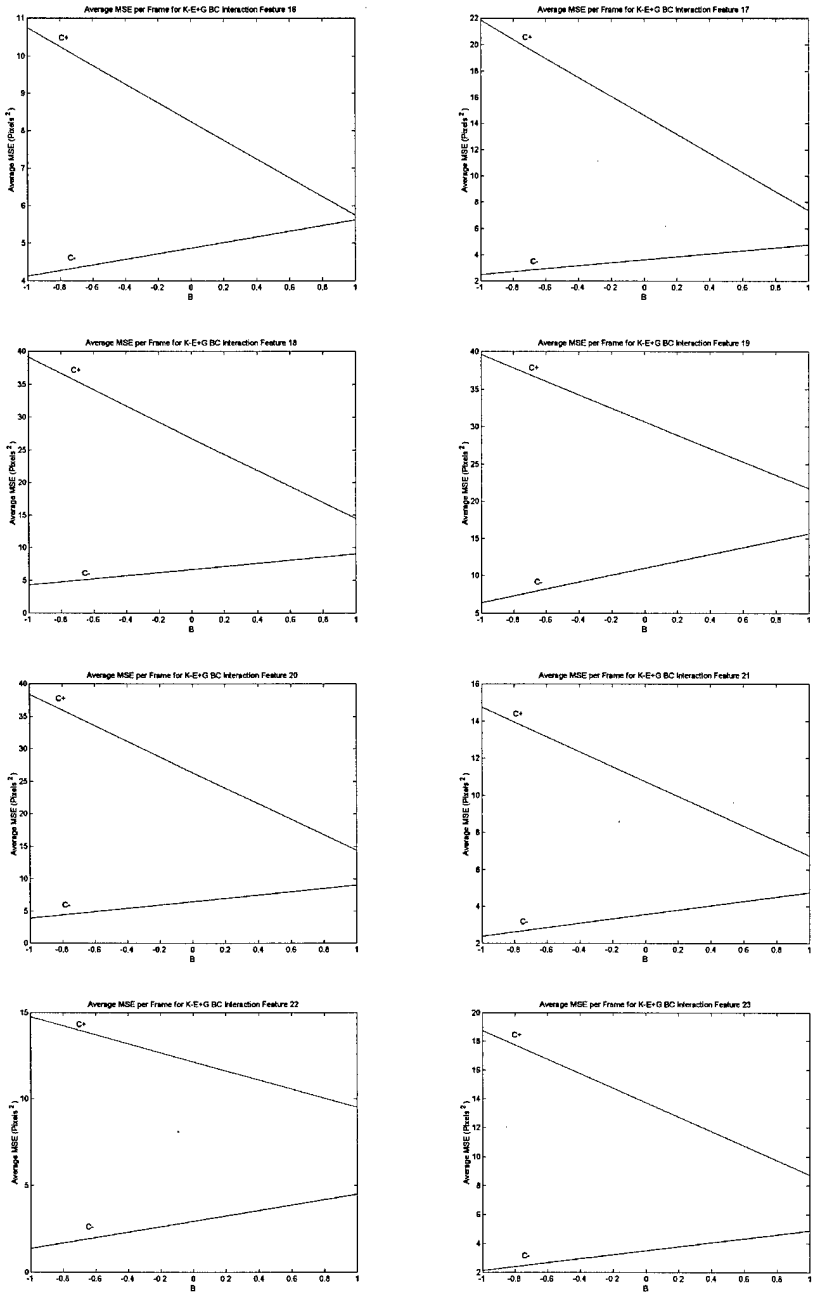


Figure 90. Plots for BC interaction in triseme K-E+G for features 16-23.

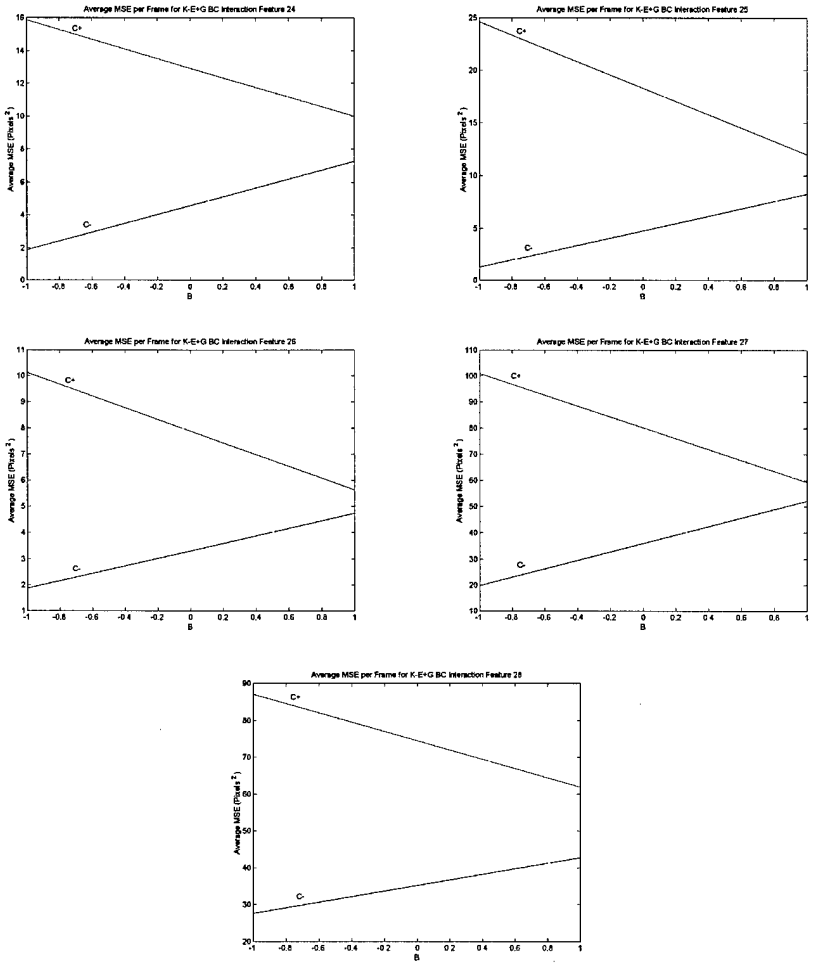


Figure 91. Plots for BC interaction in triseme K-E+G for features 24-28.

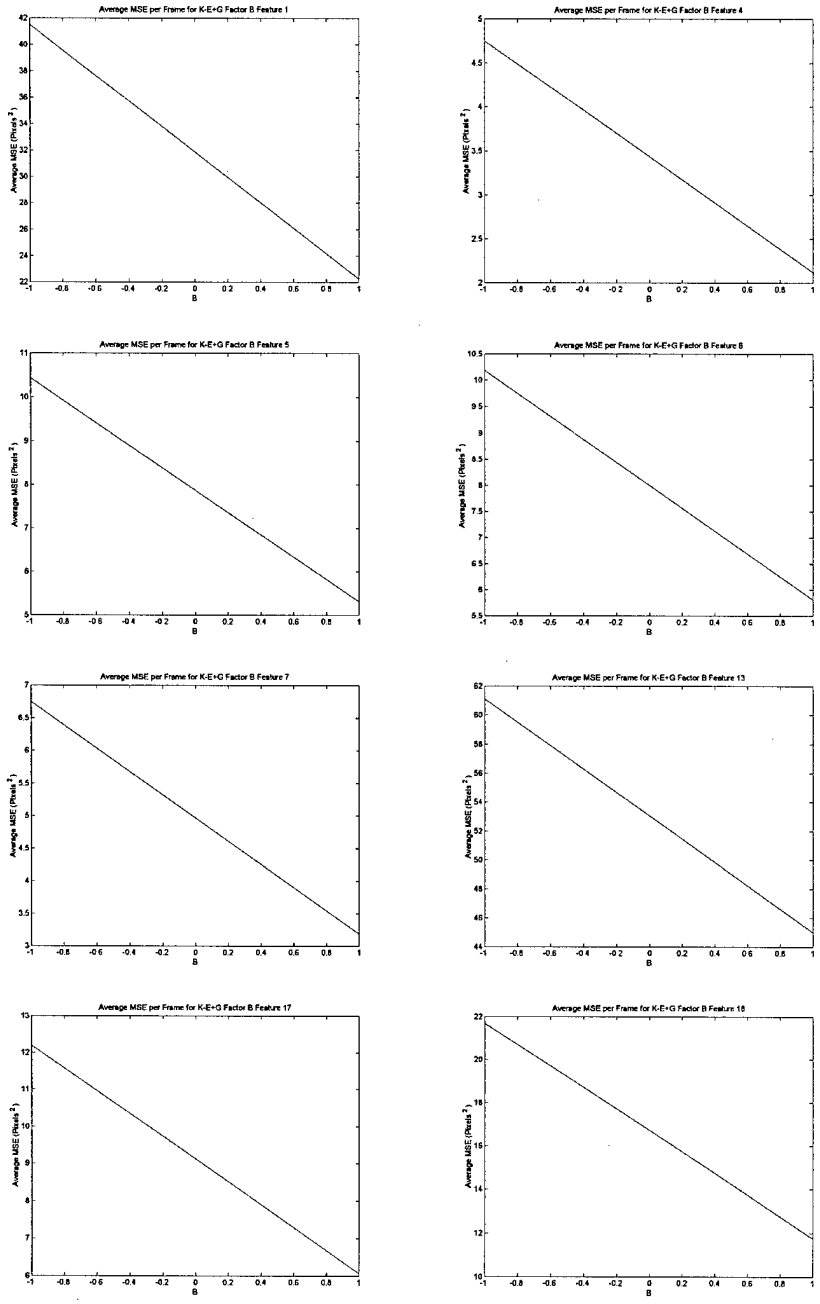


Figure 92. Main effect plots for factor B in triseme K-E+G for features 1, 4-7, 13, 17 and 18.

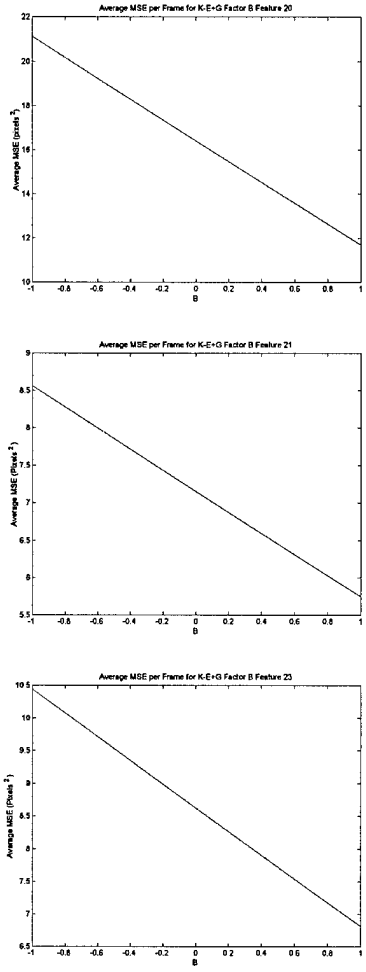


Figure 93. Main effect plots for factor B in triseme K-E+G for features 20, 21 and 23.

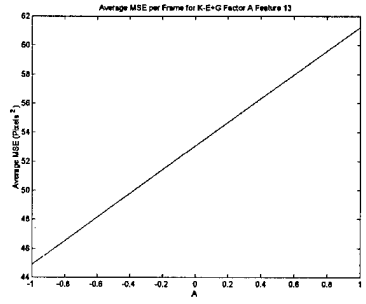


Figure 94. Main effect plot for factor A in triseme K-E+G for feature 13.

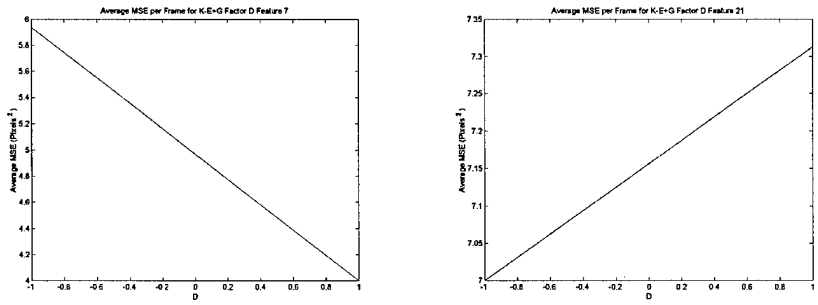


Figure 95. Main effect plots for factor D in triseme K-E+G for features 7 and 21.

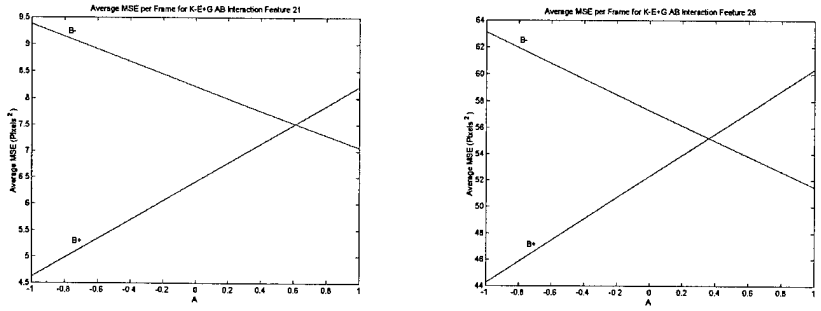
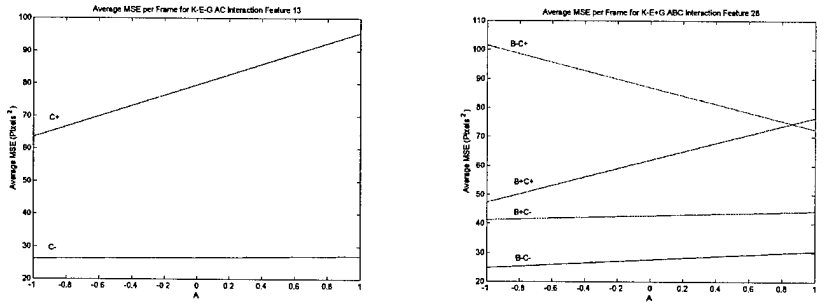


Figure 96. Plots of the AB interaction in triseme K-E+G for features 21 and 28.



(a) (b)

Figure 97. Plots of the AC interaction in feature 13 (a) and ABC interaction in triseme K-E+G for feature 28 (b).

C.4 *Triseme A-U+C*

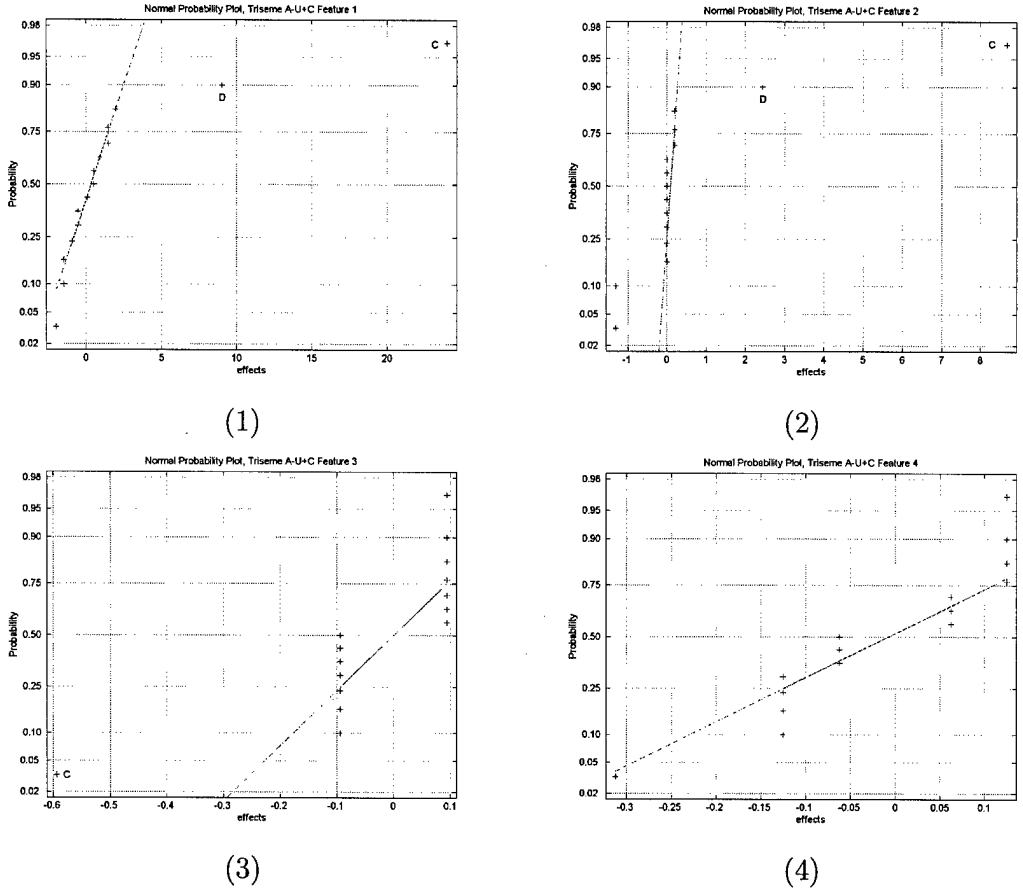


Figure 98. Normal probability plots of effects for triseme A-U+C features 1-4.

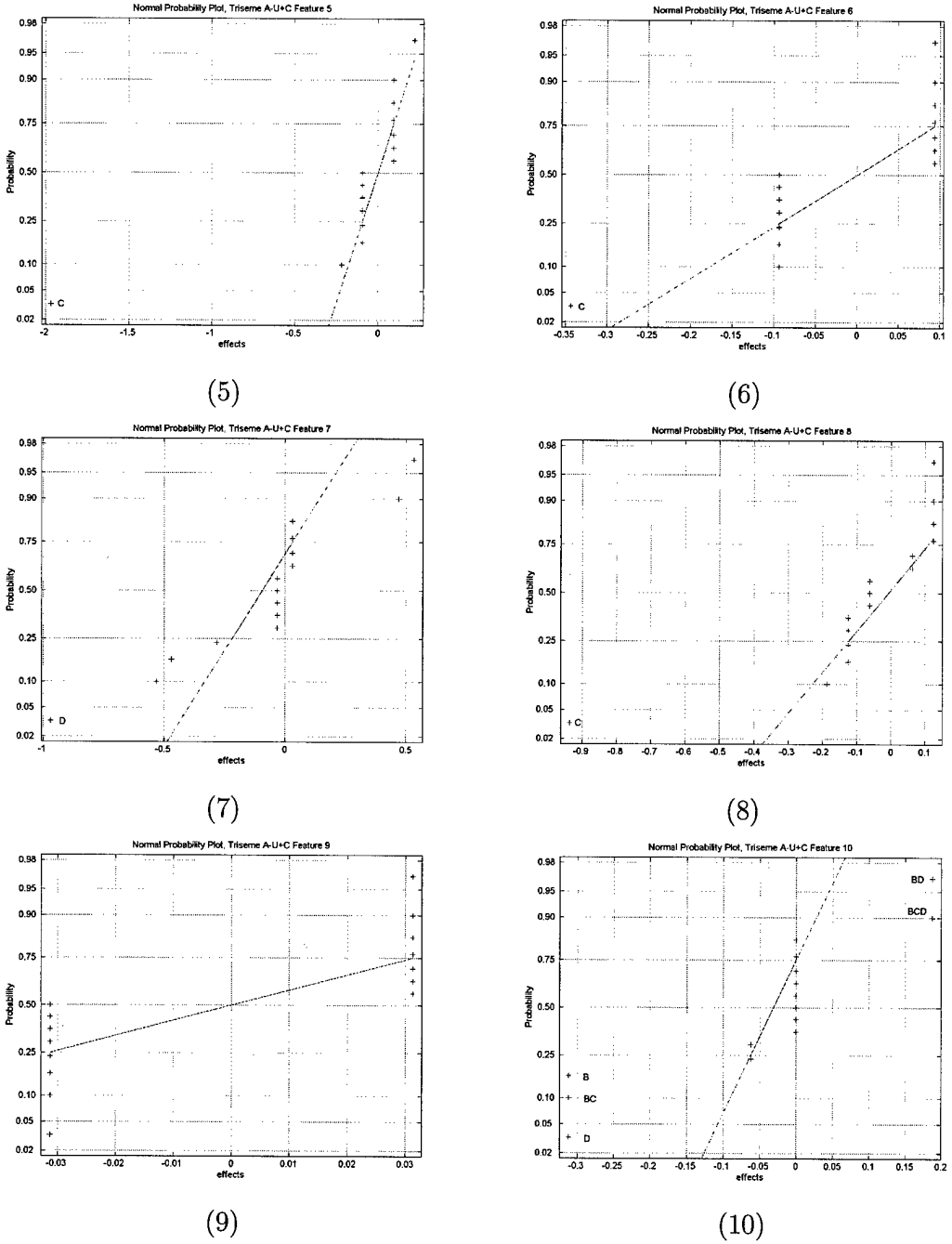
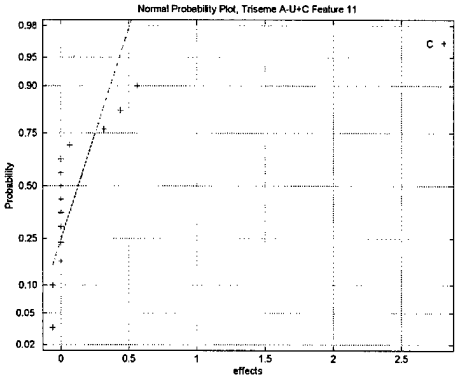
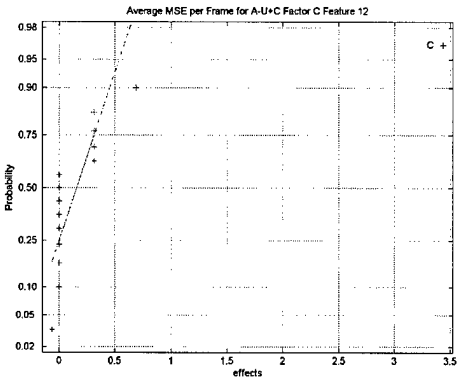


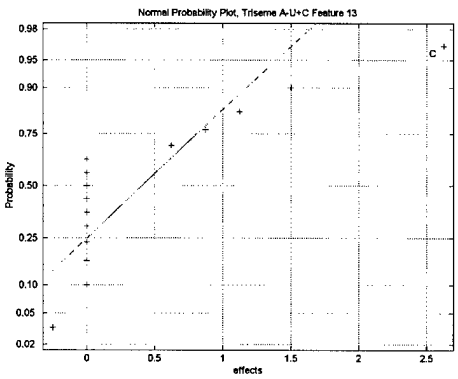
Figure 99. Normal probability plots of effects for triseme A-U+C features 5-10.



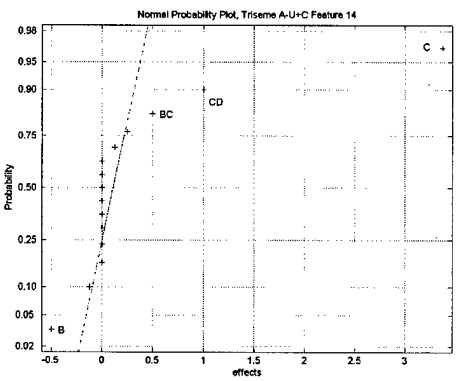
(11)



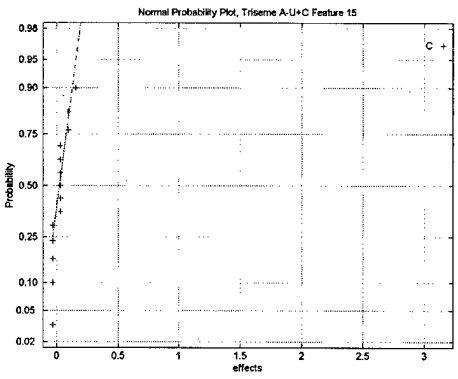
(12)



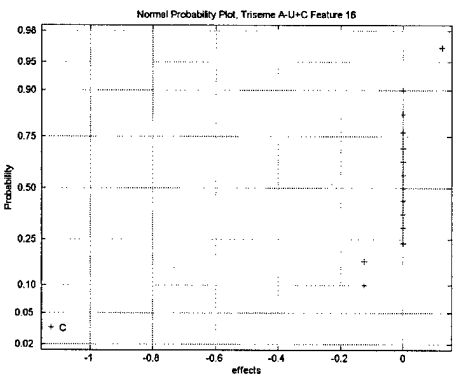
(13)



(14)

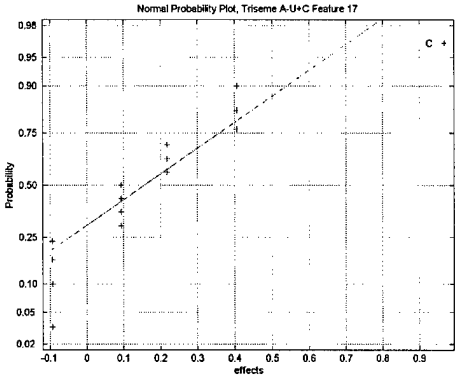


(15)

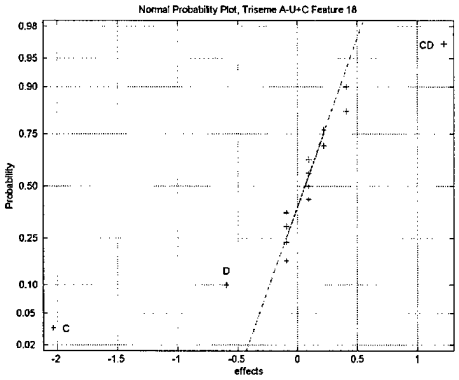


(16)

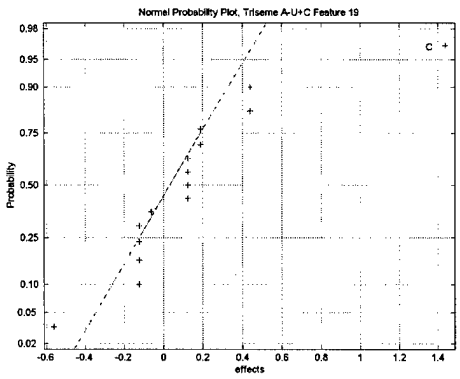
Figure 100. Normal probability plots of effects for triseme A-U+C features 11-16.



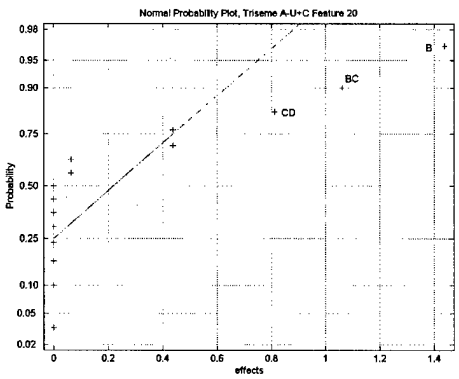
(17)



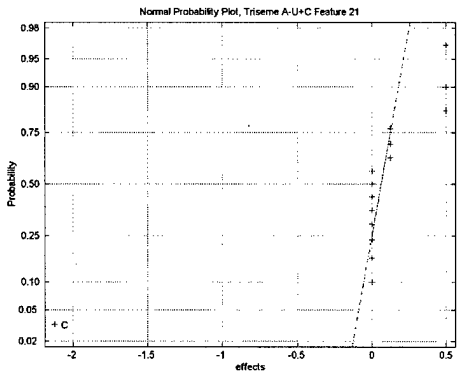
(18)



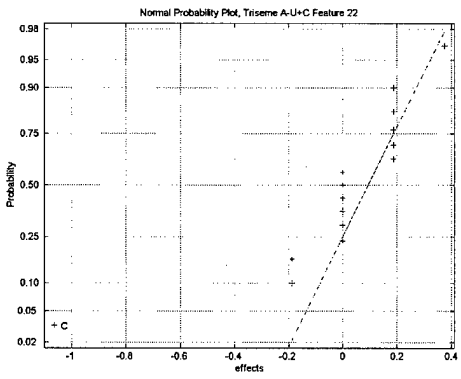
(19)



(20)

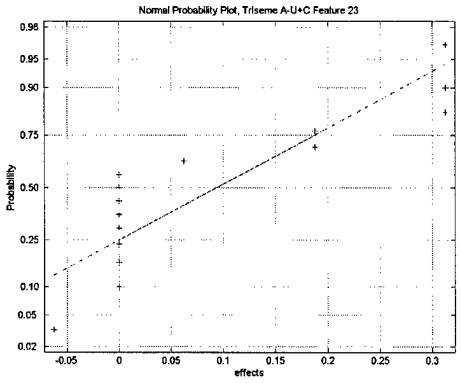


(21)

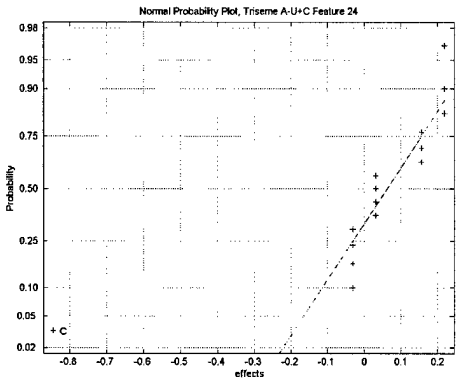


(22)

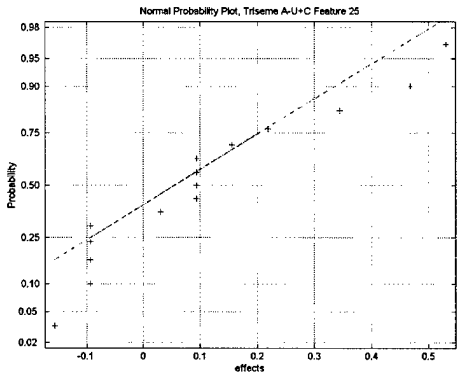
Figure 101. Normal probability plots of effects for triseme A-U+C features 17-22.



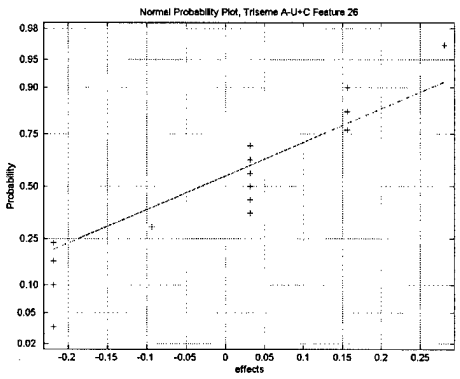
(23)



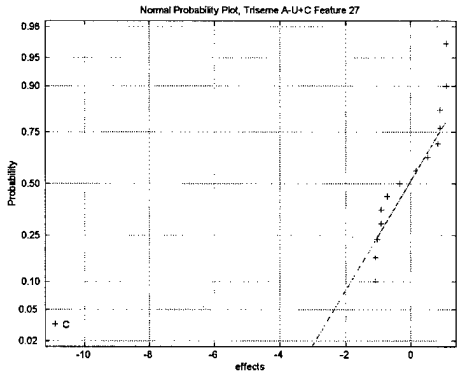
(24)



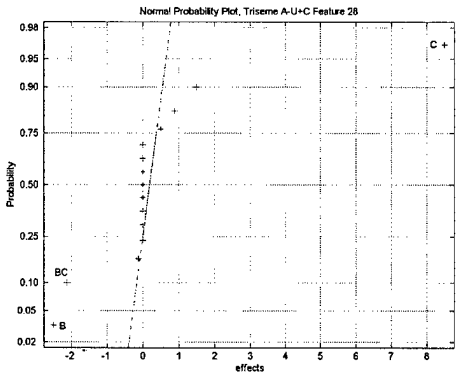
(25)



(26)



(27)



(28)

Figure 102. Normal probability plots of effects for triseme A-U+C features 23-28.

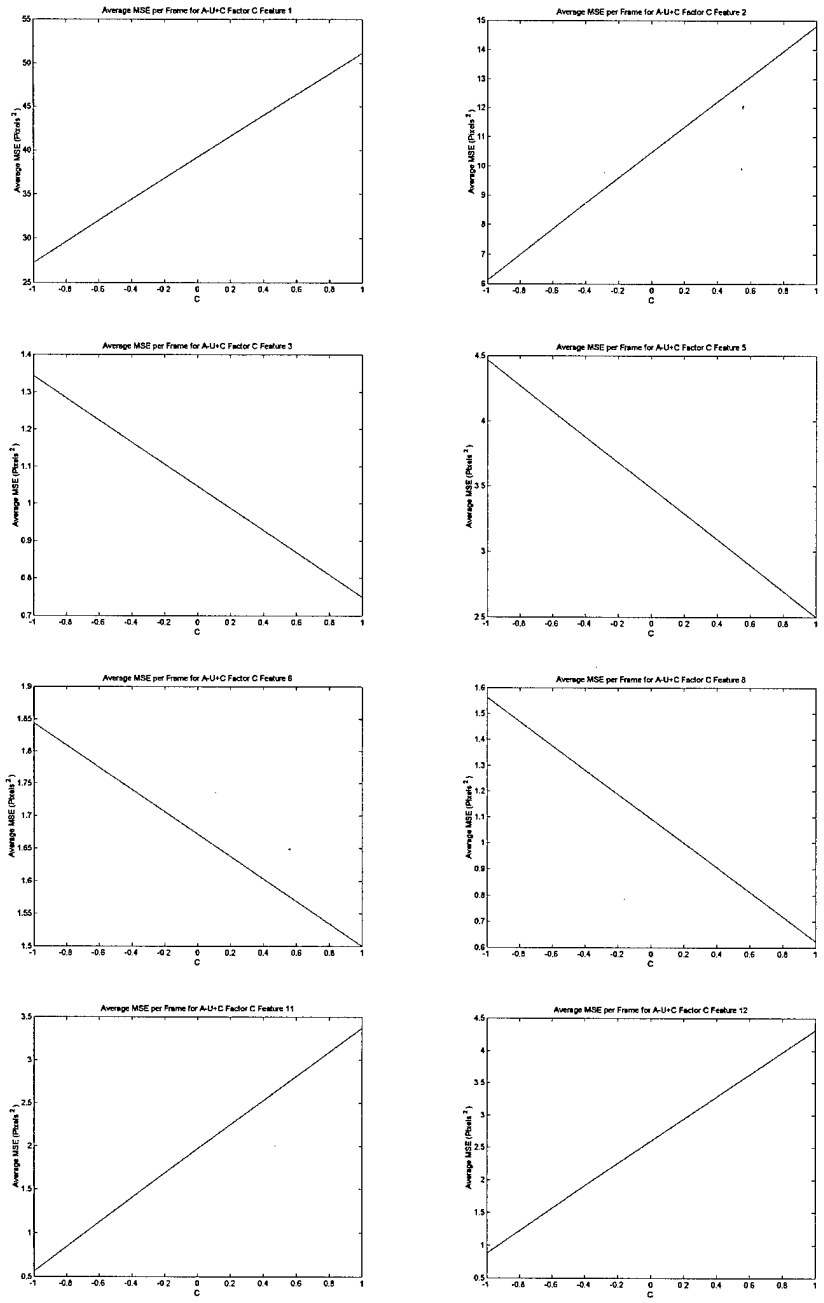


Figure 103. Main effect plots for factor C in triseme A-U+C for features 1-3, 5, 6, 8, 11 and 12.

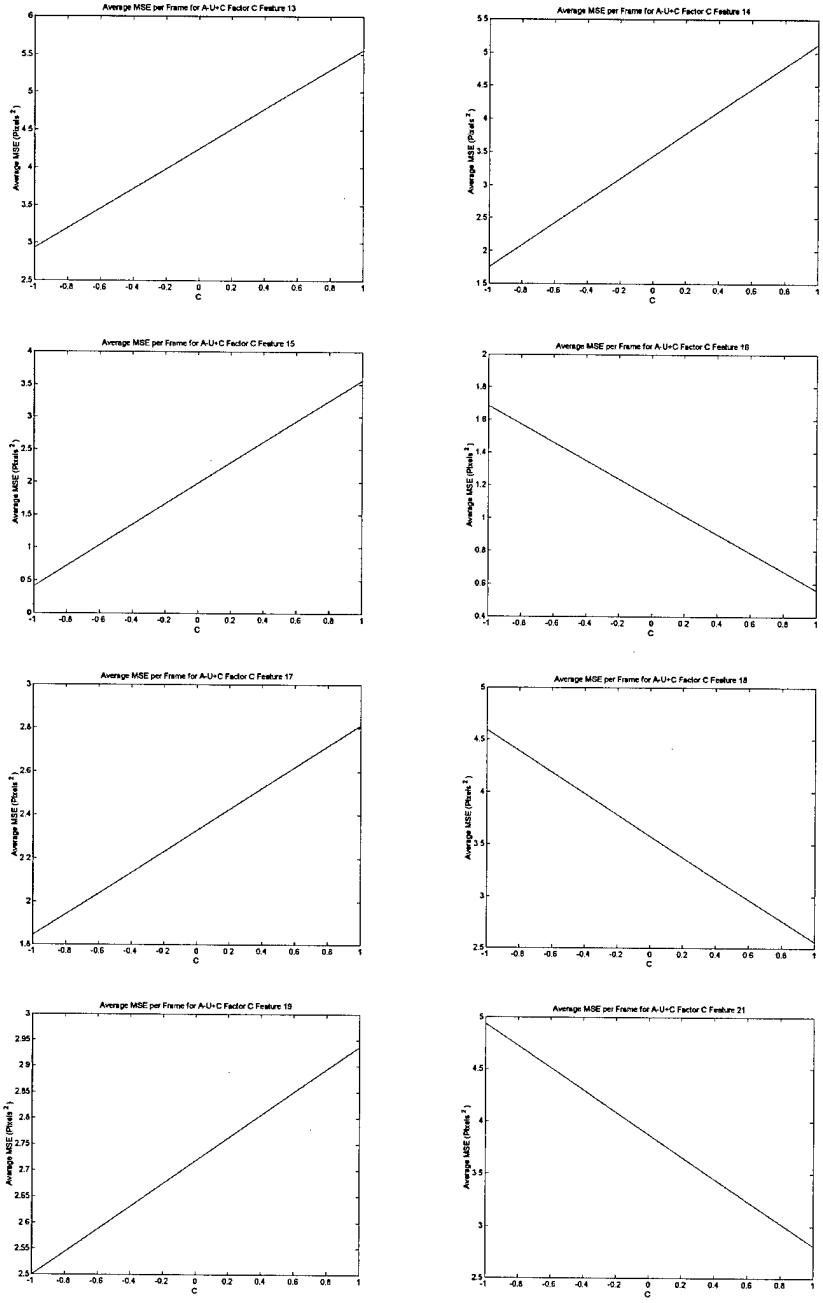


Figure 104. Main effect plots for factor C in triseme A-U+C for features 13-19 and 21.

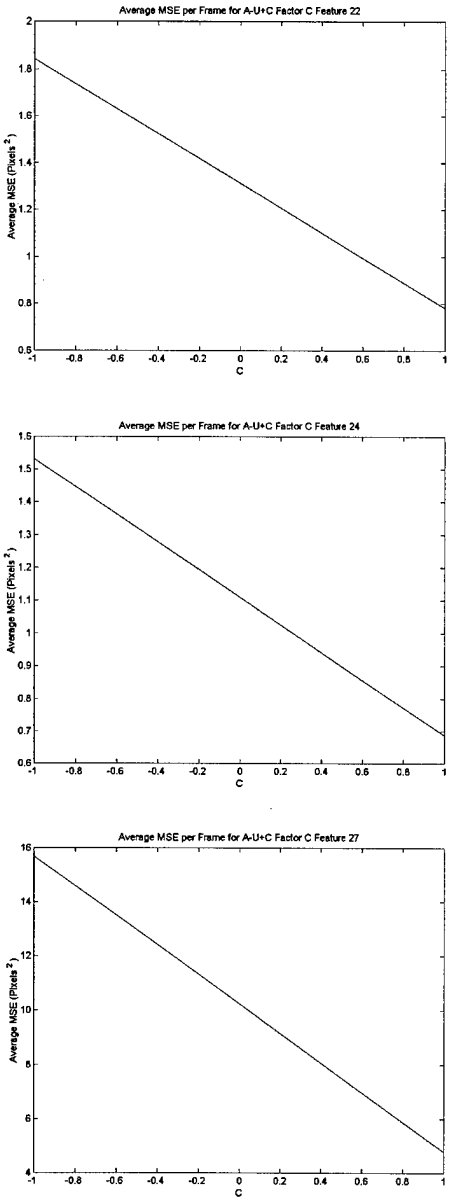


Figure 105. Main effect plots for factor C in triseme A-U+C for features 23, 24 and 27.

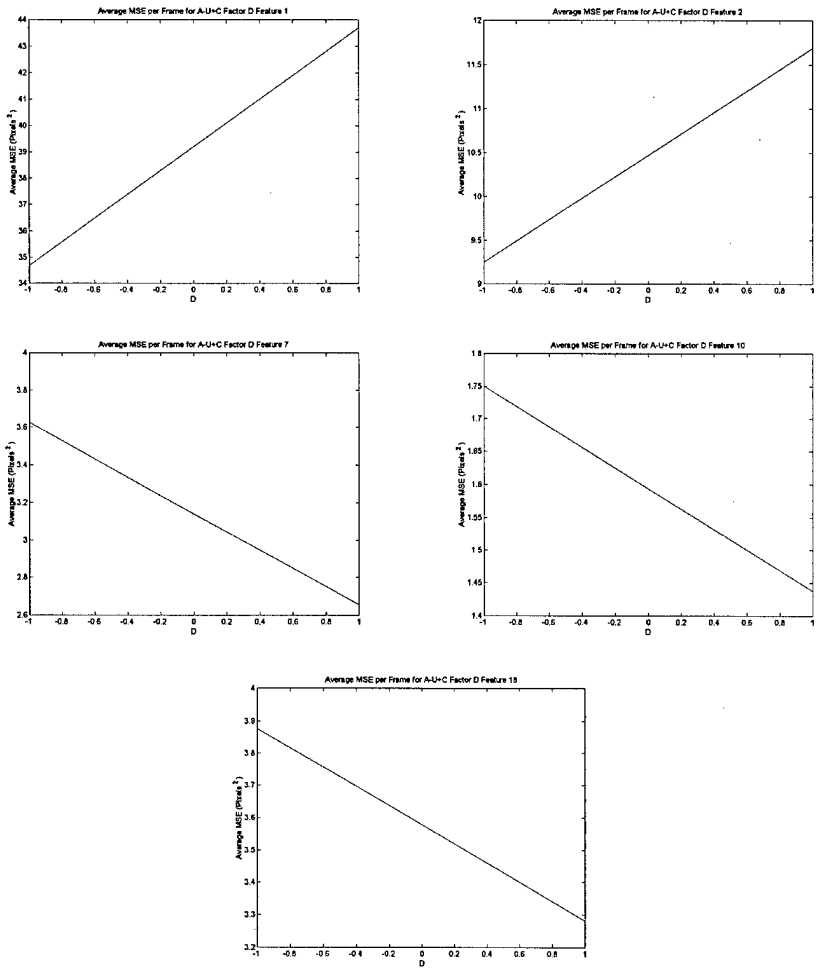


Figure 106. Main effect plots for factor D in triseme A-U+C for features 1, 2, 7, 10 and 18.

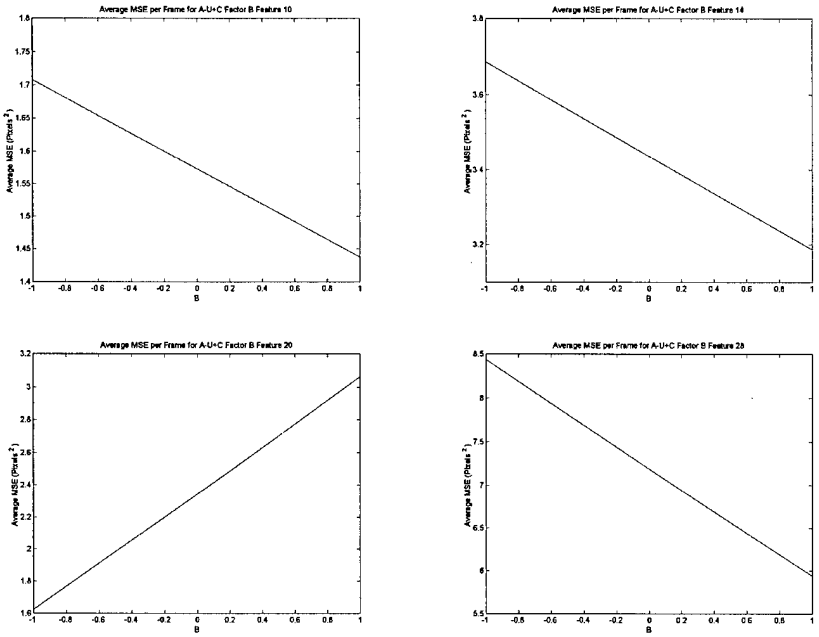


Figure 107. Main effect plots for factor B in triseme A-U+C for features 10, 14, 20 and 28.

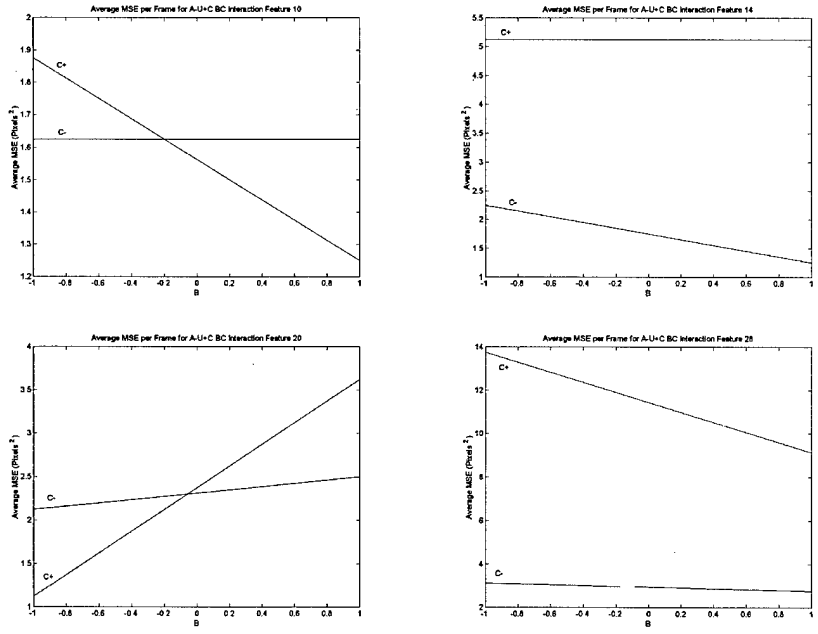


Figure 108. Plots of the BC interaction in triseme A-U+C for features 10, 14, 20 and 28.

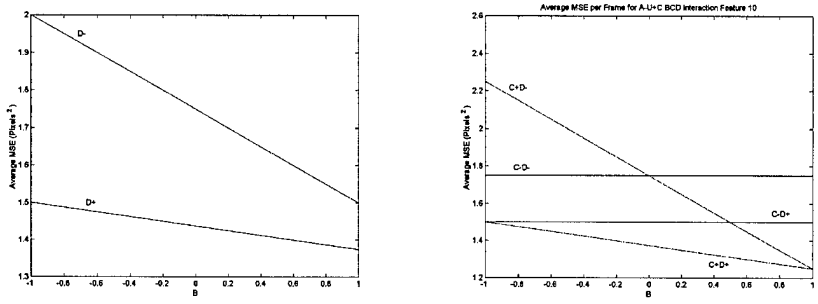


Figure 109. Plots of the BD and BCD interactions in triseme A-U+C for feature 10.

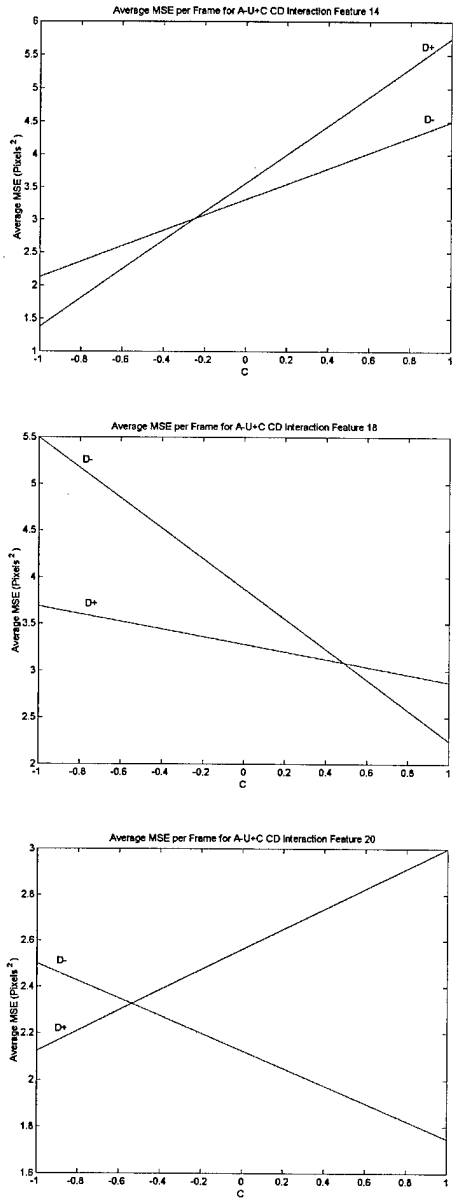


Figure 110. Plots of the CD interaction in triseme A-U+C for features 14, 18 and 20.

C.5 *Triseme F-K+D*

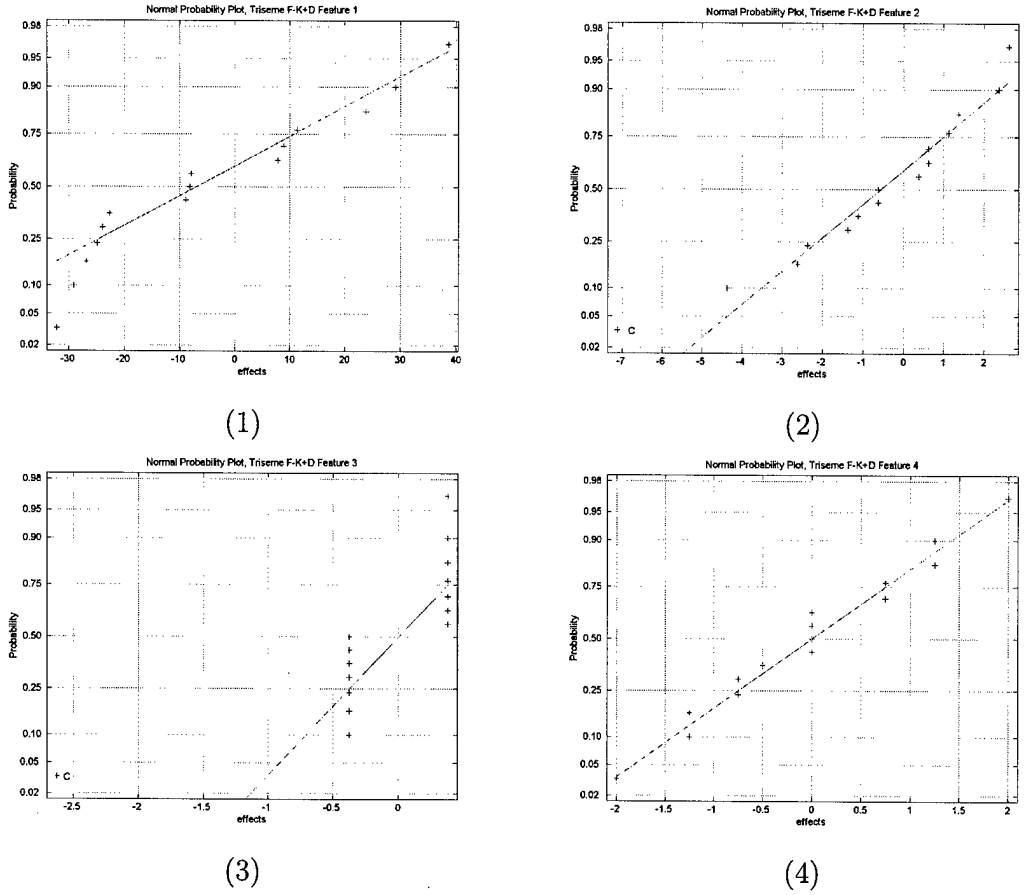
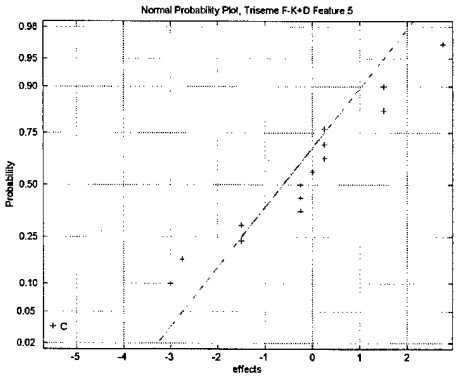
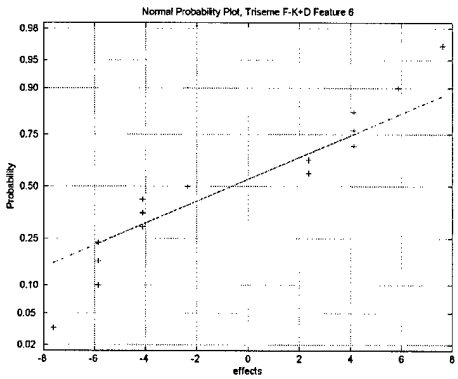


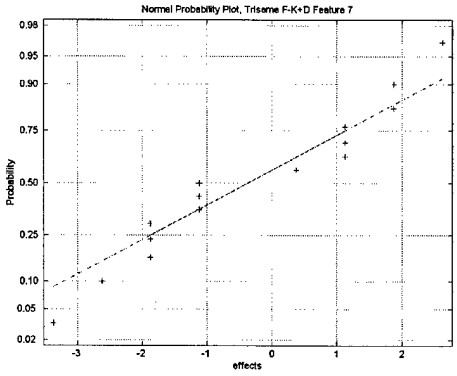
Figure 111. Normal probability plots of effects for triseme F-K+D features 1-4.



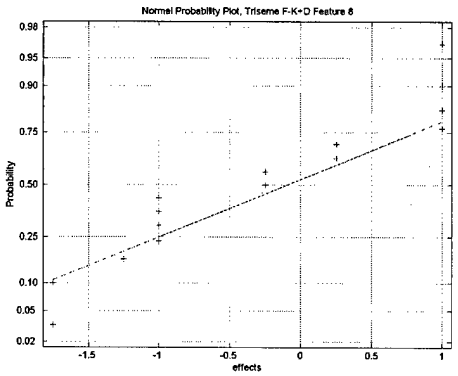
(5)



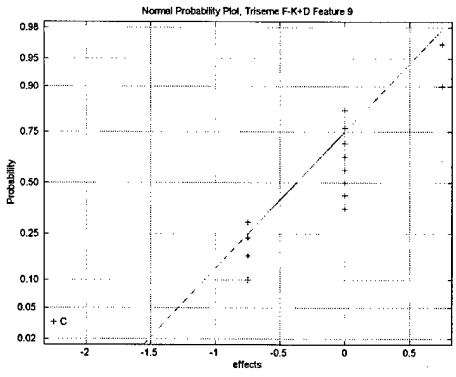
(6)



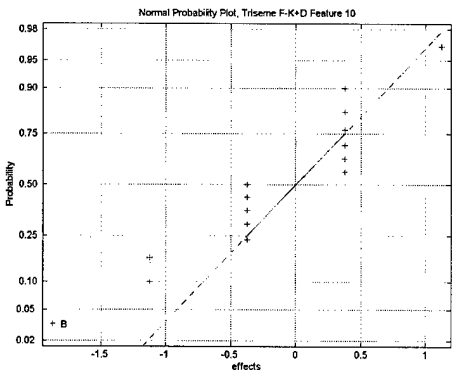
(7)



(8)

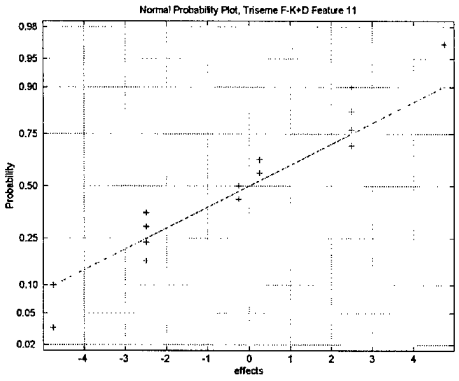


(9)

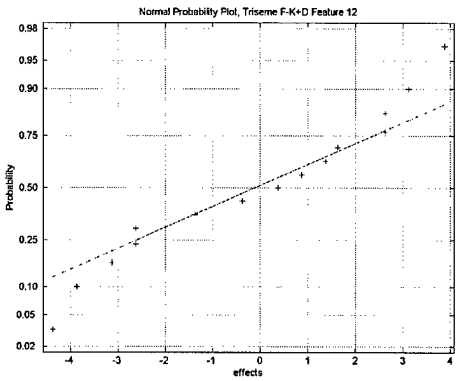


(10)

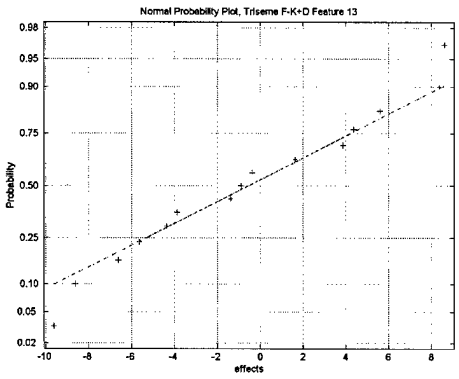
Figure 112. Normal probability plots of effects for triseme F-K+D features 5-10.



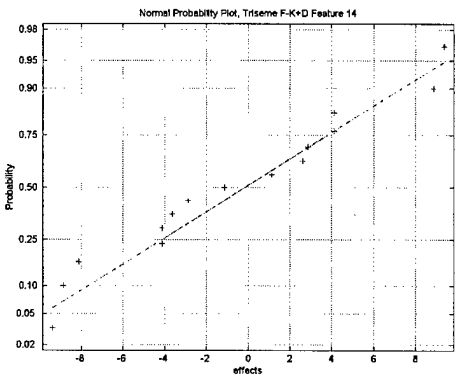
(11)



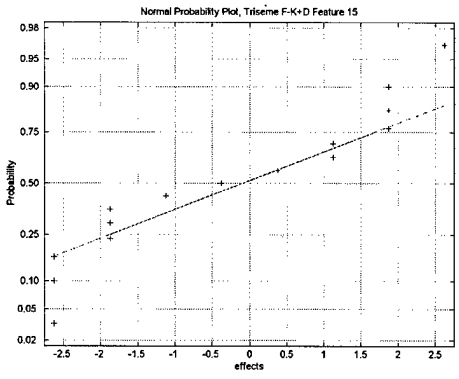
(12)



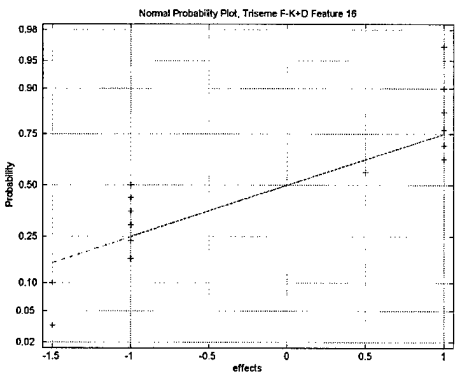
(13)



(14)

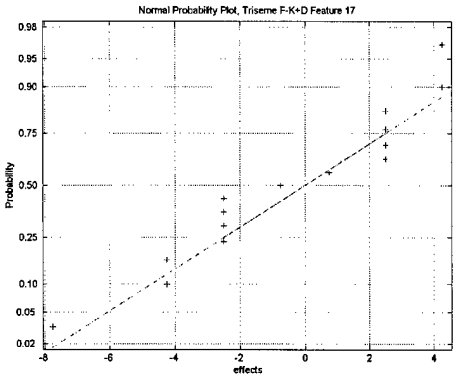


(15)

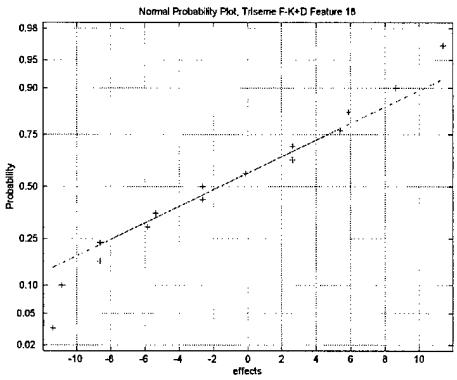


(16)

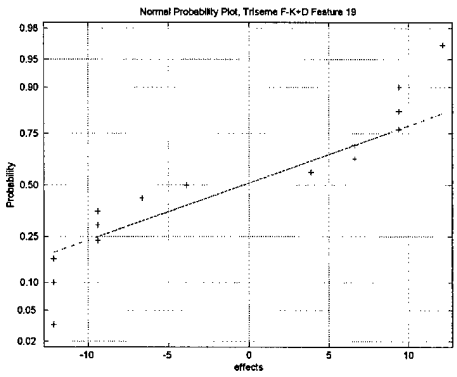
Figure 113. Normal probability plots of effects for triseme F-K+D features 11-16.



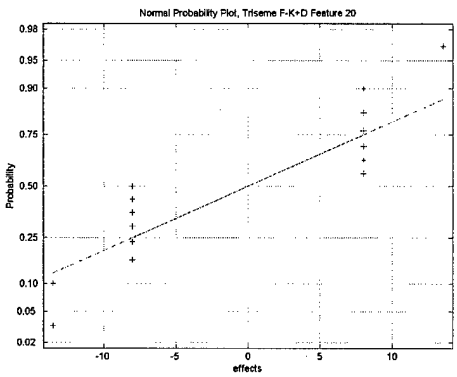
(17)



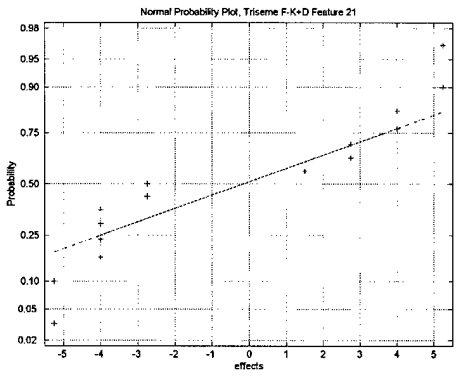
(18)



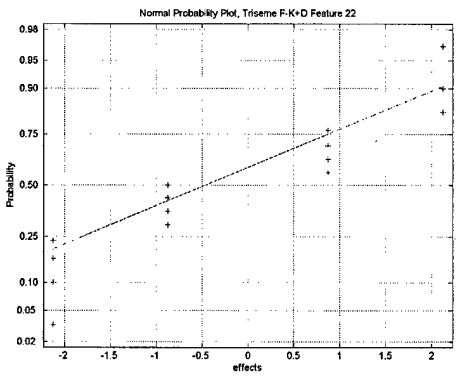
(19)



(20)

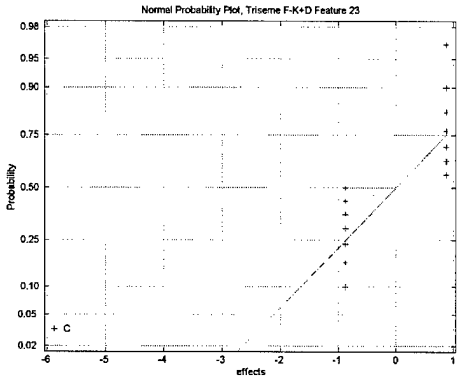


(21)

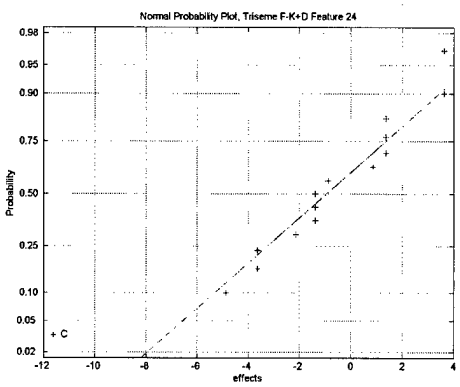


(22)

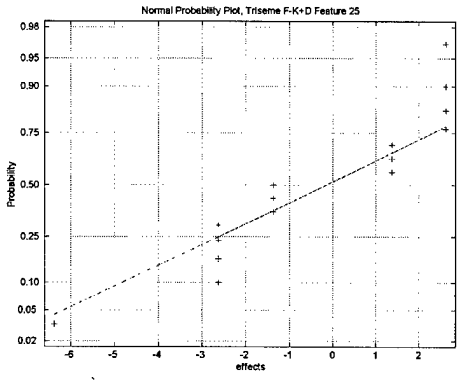
Figure 114. Normal probability plots of effects for triseme F-K+D features 17-22.



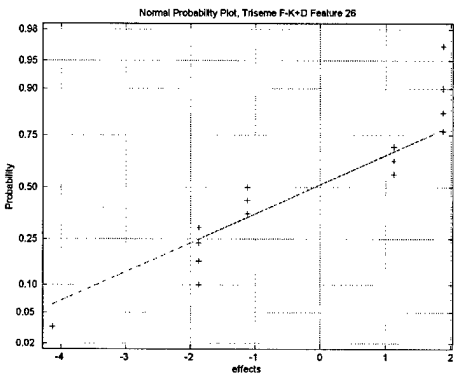
(23)



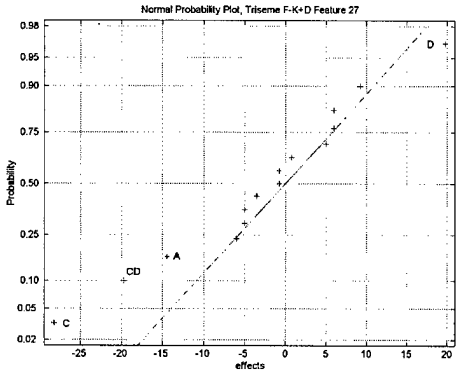
(24)



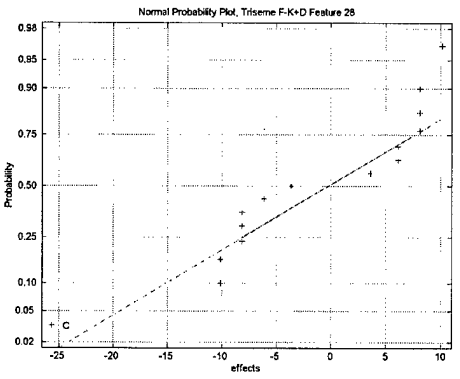
(25)



(26)



(27)



(28)

Figure 115. Normal probability plots of effects for triseme F-K+D features 23-28.

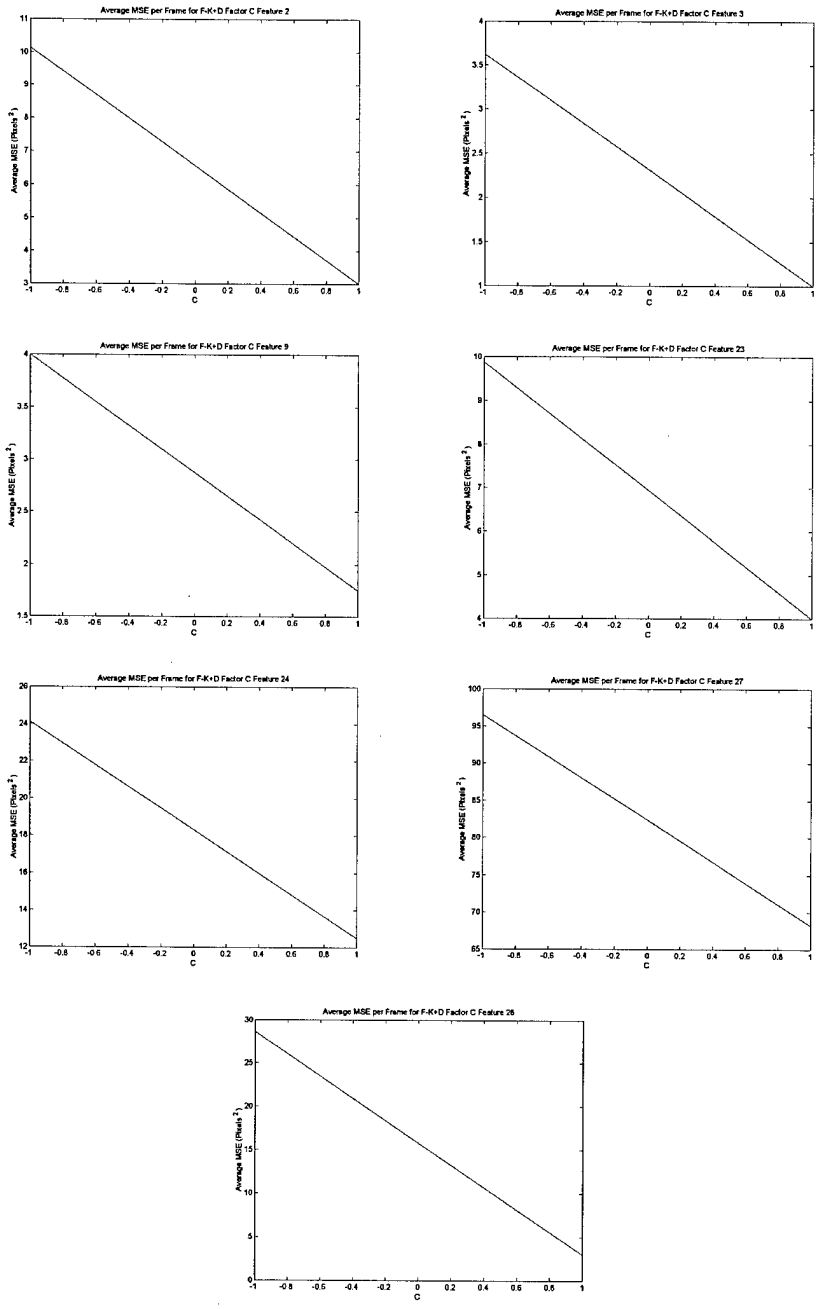


Figure 116. Main effect plots for factor C in triseme F-K+D for features 2, 3, 9, 23, 24, 27 and 28.

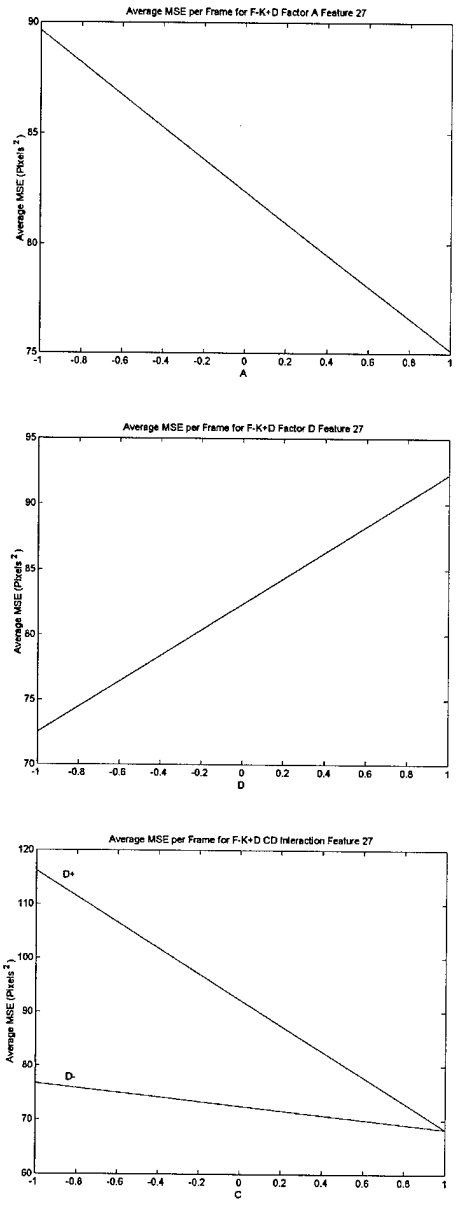


Figure 117. Main effect plots for factors A and D and interaction plot for the CD interaction in triseme F-K+D feature 27.

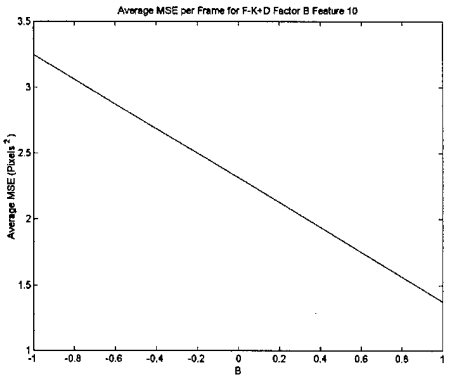


Figure 118. Main effect plot for factor B in triseme F-K+D for feature 10.

C.6 *Triseme R-P+G*

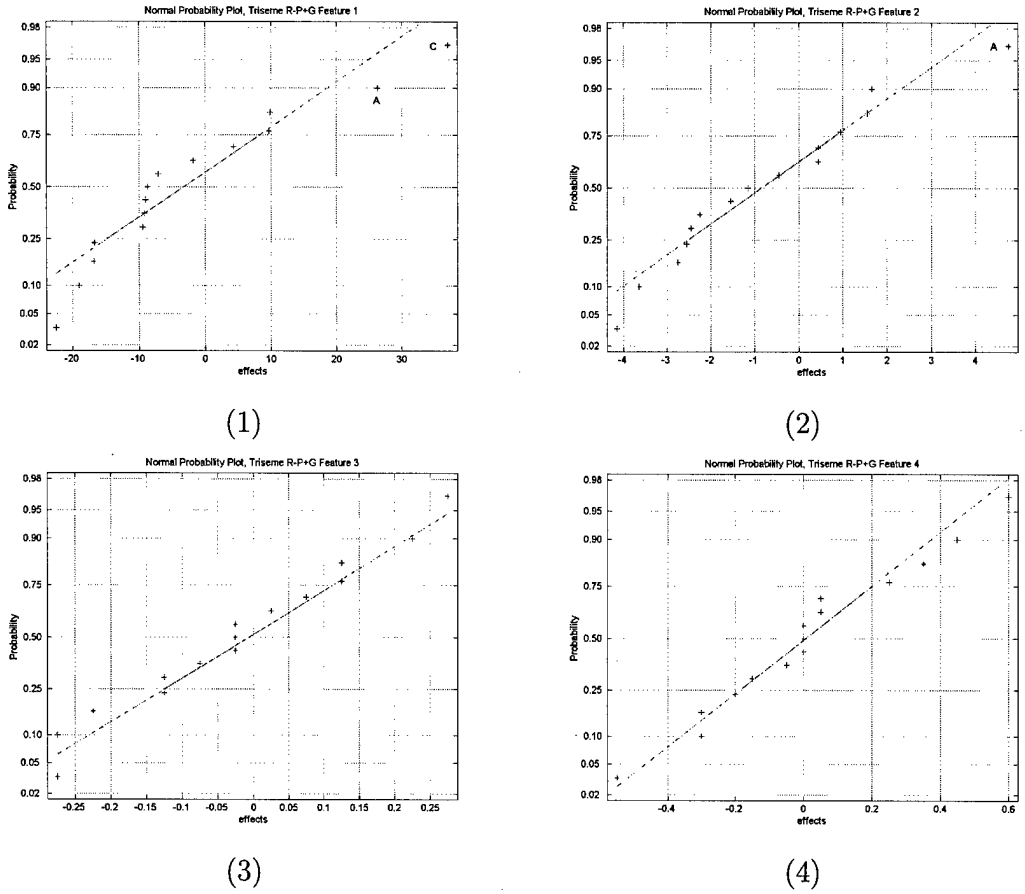
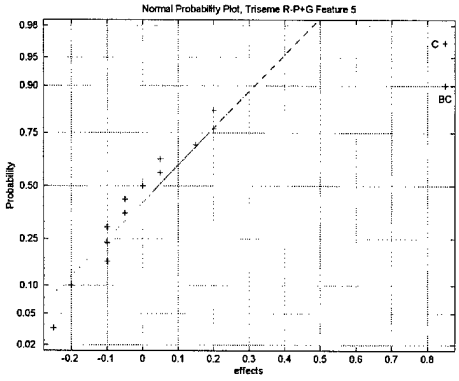
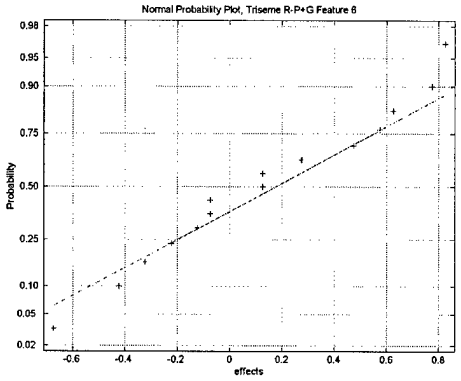


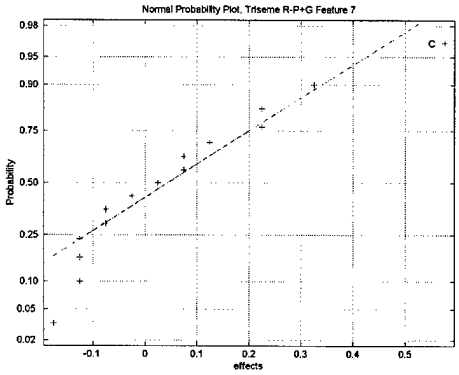
Figure 119. Normal probability plots of effects for triseme R-P+G features 1-4.



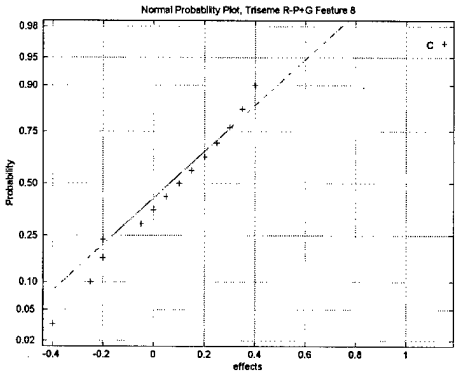
(5)



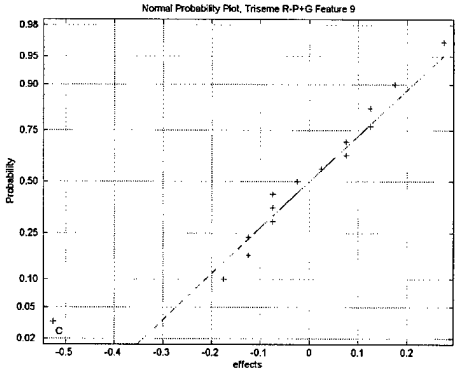
(6)



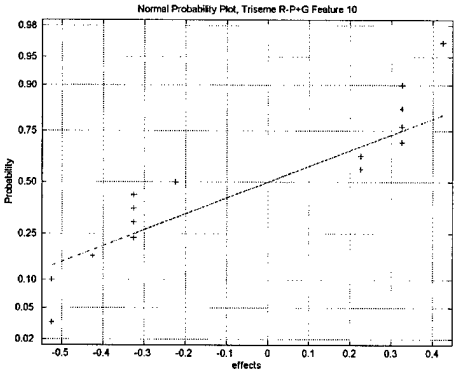
(7)



(8)

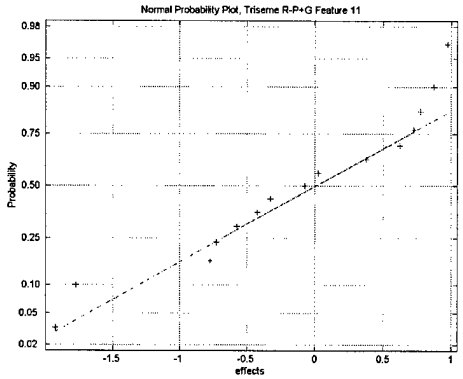


(9)

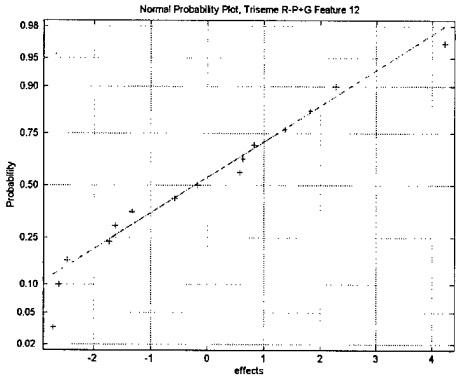


(10)

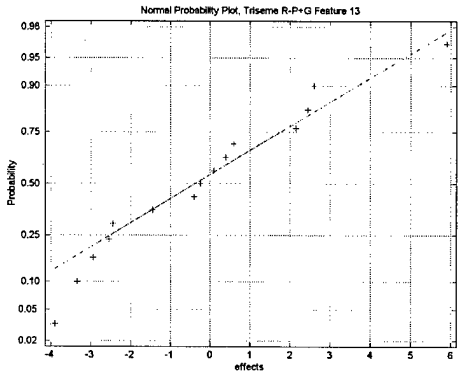
Figure 120. Normal probability plots of effects for triseme R-P+G features 5-10.



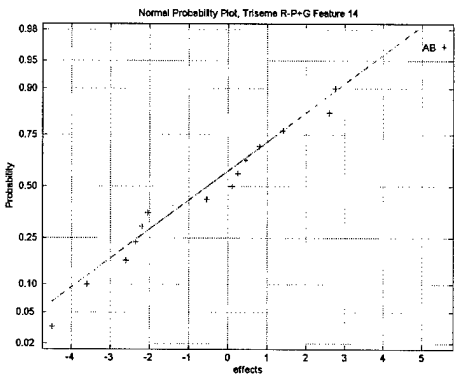
(11)



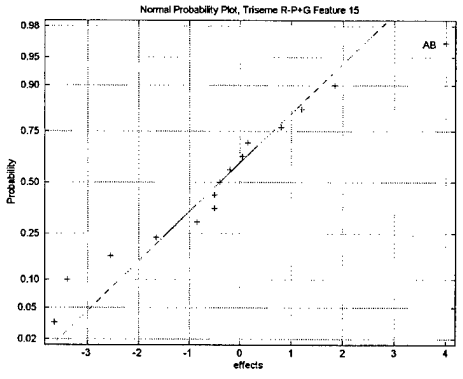
(12)



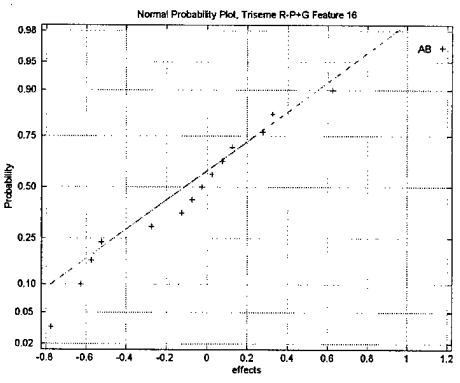
(13)



(14)

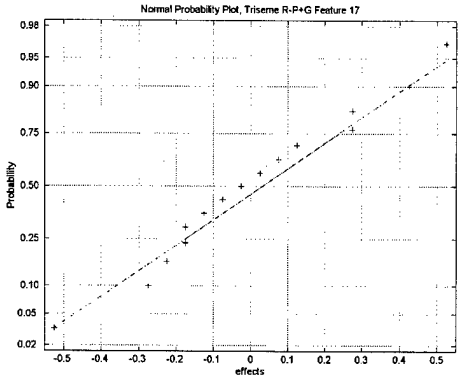


(15)

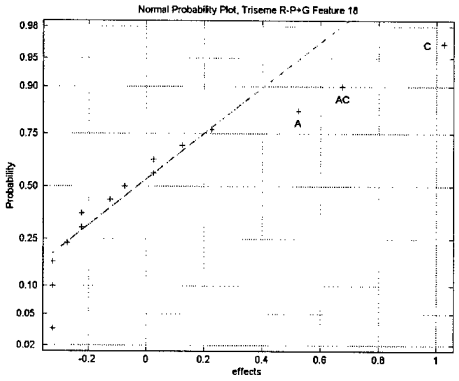


(16)

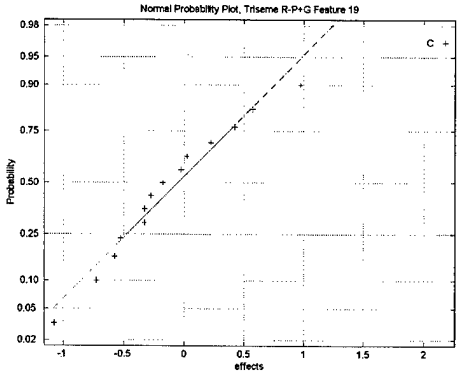
Figure 121. Normal probability plots of effects for triseme R-P+G features 11-16.



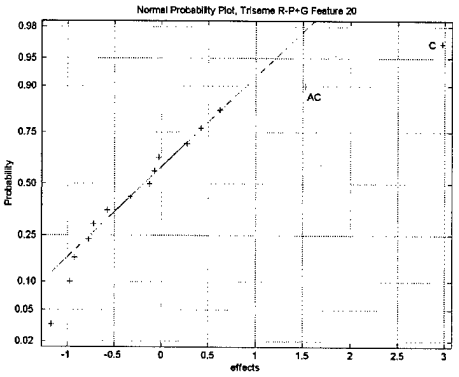
(17)



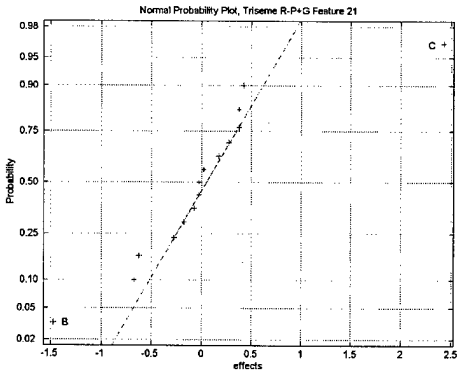
(18)



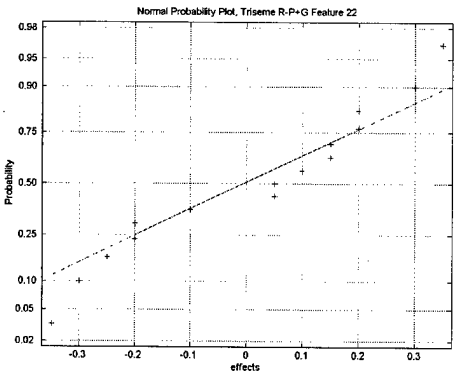
(19)



(20)

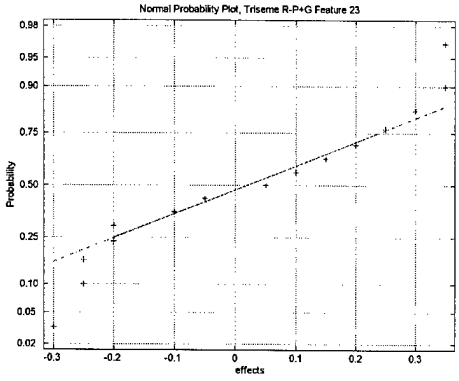


(21)

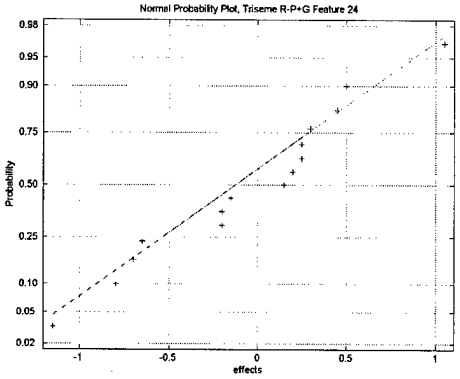


(22)

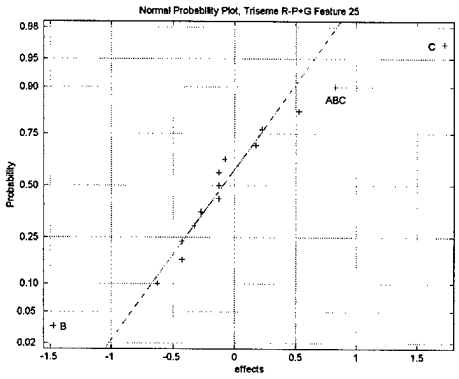
Figure 122. Normal probability plots of effects for triseme R-P+G features 17-22.



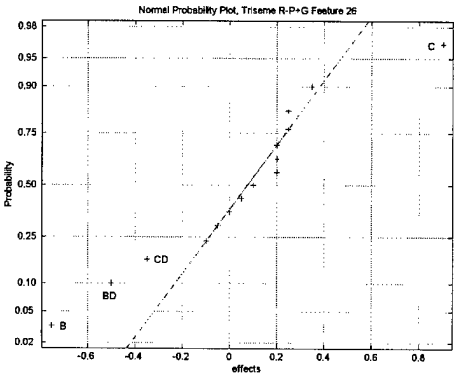
(23)



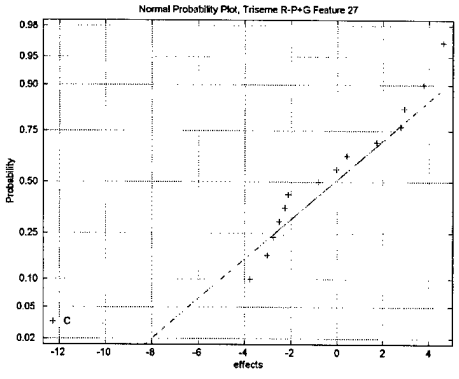
(24)



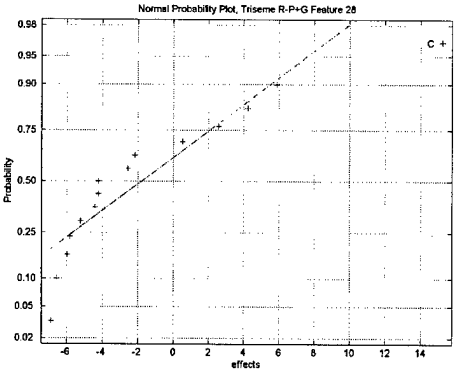
(25)



(26)



(27)



(28)

Figure 123. Normal probability plots of effects for triseme R-P+G features 23-28.

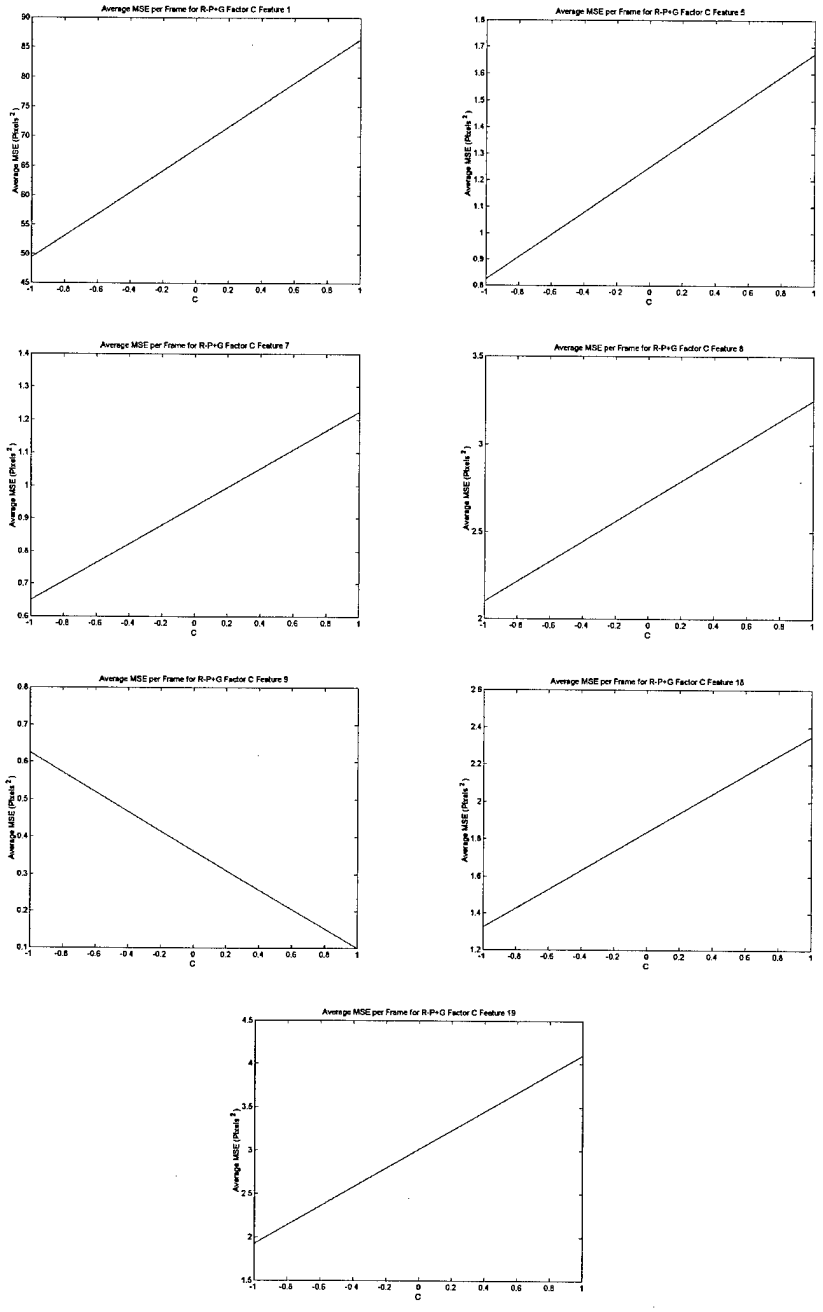


Figure 124. Main effect plots for factor C in triseme R-P+G for features 1, 5, 7-9, 18 and 19.

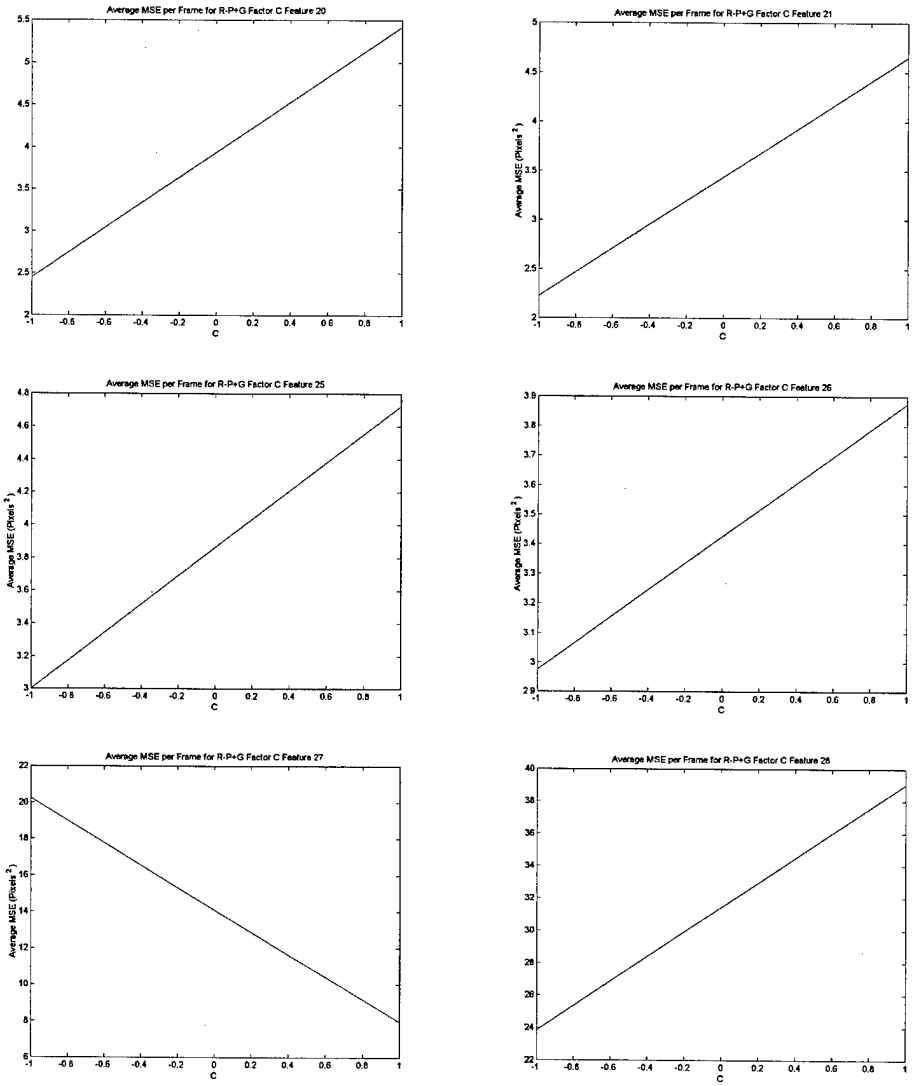
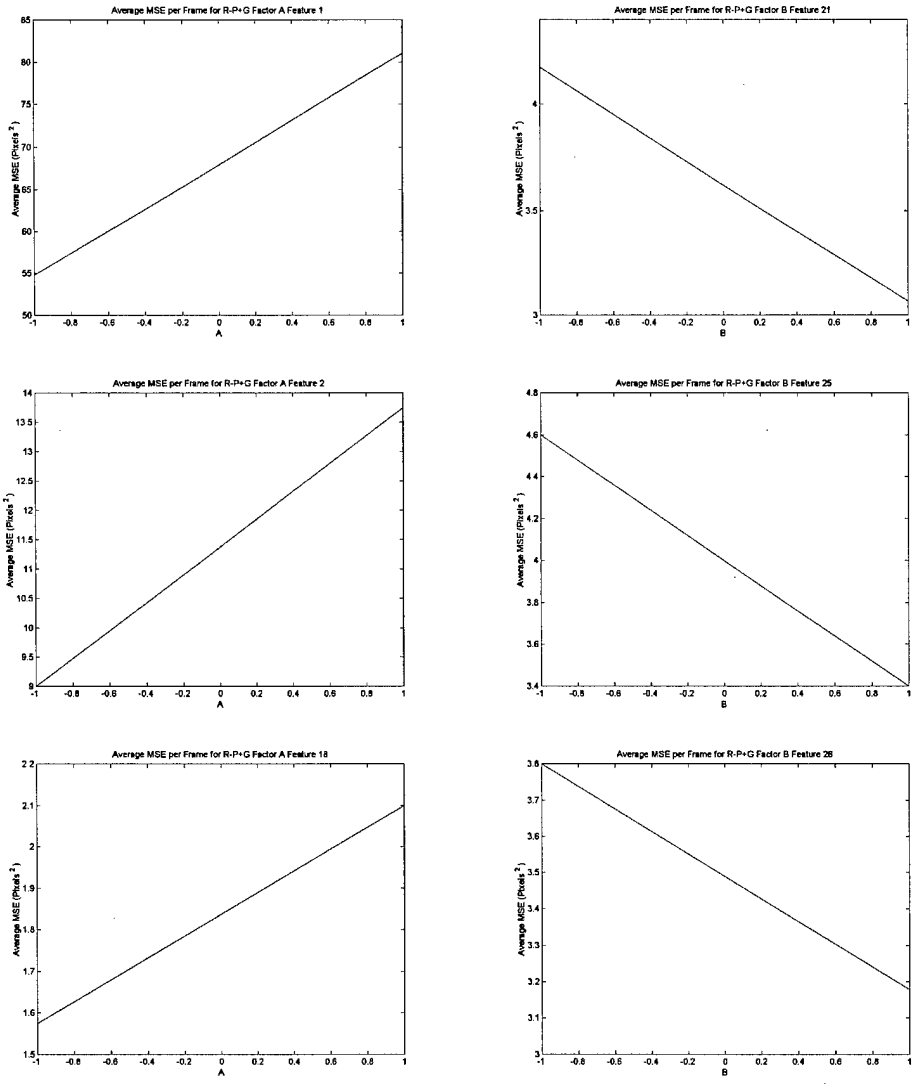


Figure 125. Main effect plots for factor C in triseme R-P+G for features 20, 21 and 25-28.



(a) (b)

Figure 126. Main effects plots for factors A in features 1, 2 and 18 (a) and B in features 21, 25 and 26 (b) for triseme R-P+G.

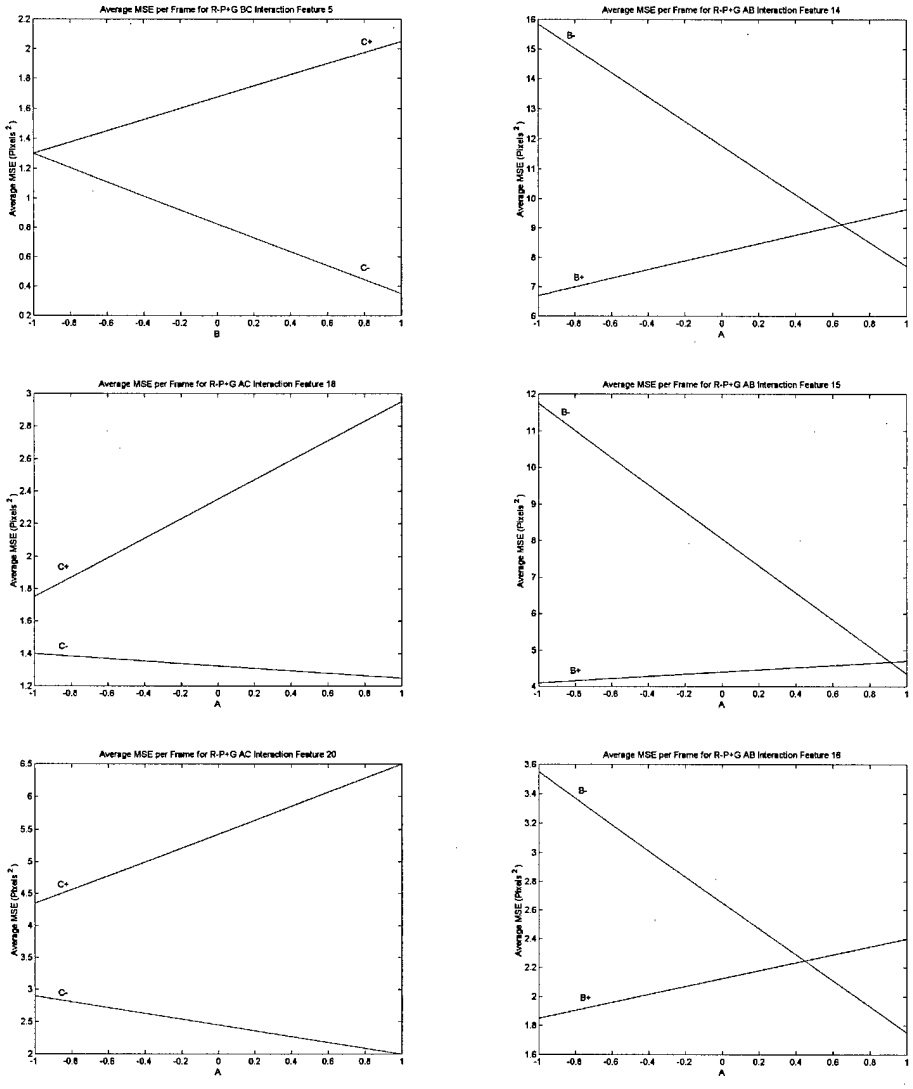


Figure 127. Interaction plots for triseme R-P+G for the BC interaction in feature 5 (top left), AC interactions in features 18 and 20 (left center and lower left), and AB interactions in features 14-16 (right top to bottom).

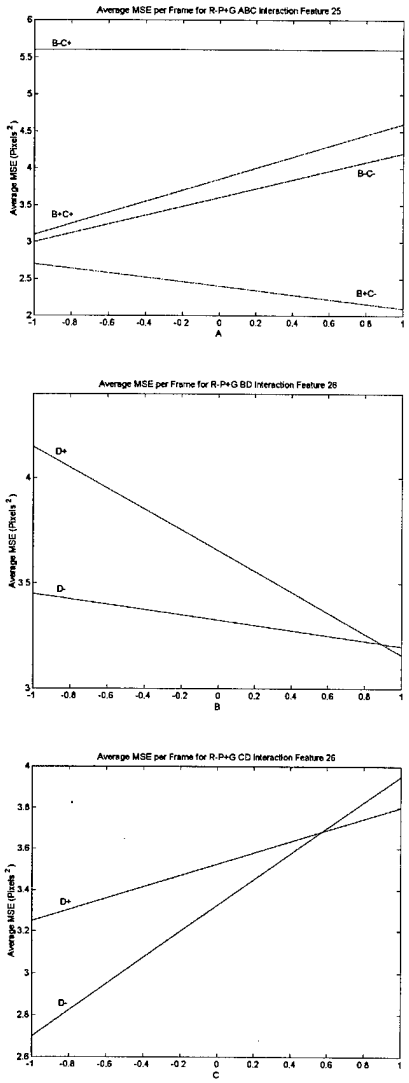


Figure 128. Interaction plots for triseme R-P+G for the ABC interaction in feature 25 (top), and the BD and CD interactions in feature 26 (middle/bottom).

Appendix D. ANOVA Tables, Main Effect Plots and Interaction Plots for Individual Triseme Analysis with Blocking

Table 10. ANOVA Summary for Feature 1 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	114756.6	5		
A	856.5163	1	1.0832	0.3013
B	1064.5566	1	1.3463	0.2496
AB	146.8913	1	0.1858	0.6677
C	4222.5590	1	5.3402	0.0236
AC	1617.4521	1	2.0456	0.1568
BC	2014.0703	1	2.5471	0.1147
ABC	1900.5950	1	2.4036	0.1253
D	1.9694	1	0.0025	0.9603
AD	3.3938	1	0.0043	0.9479
BD	15.9115	1	0.0201	0.8876
ABD	750.1221	1	0.9487	0.3332
CD	0.0846	1	0.0001	0.9918
ACD	20.2125	1	0.0256	0.8734
BCD	65.8635	1	0.0833	0.7737
ABCD	602.1013	1	0.7615	0.3857
Error	59303.6	90		
Total	187322.3	95		

Table 11. ANOVA Summary for Feature 2 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	11779.8	5		
A	31.7017	1	0.7148	0.4005
B	1.6537	1	0.0373	0.8474
AB	1.6276	1	0.0367	0.8486
C	61.1204	1	1.3782	0.2441
AC	0.8500	1	0.0192	0.8903
BC	35.4456	1	0.7993	0.3742
ABC	72.3958	1	1.6325	0.2053
D	28.7474	1	0.6482	0.4233
AD	86.7033	1	1.9551	0.1662
BD	0.1780	1	0.0040	0.9497
ABD	18.5797	1	0.4190	0.5194
CD	47.6017	1	1.0734	0.3035
ACD	137.5209	1	3.1010	0.0823
BCD	7.2600	1	0.1637	0.6869
ABCD	139.5335	1	3.1464	0.0802
Error	3326.1	90		
Total	15639.2	95		

Table 12. ANOVA Summary for Feature 3 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	33.4	5		
A	0.3897	1	0.3692	0.5453
B	0.2835	1	0.2685	0.6059
AB	1.2873	1	1.2194	0.2730
C	0.1765	1	0.1672	0.6838
AC	0.1123	1	0.1064	0.7452
BC	0.0005	1	0.0005	0.9829
ABC	0.1571	1	0.1488	0.7008
D	0.1338	1	0.1267	0.7229
AD	0.3855	1	0.3652	0.5475
BD	0.2587	1	0.2450	0.6220
ABD	0.1571	1	0.1488	0.7008
CD	0.0695	1	0.0658	0.7982
ACD	0.5226	1	0.4951	0.4839
BCD	0.1653	1	0.1566	0.6935
ABCD	0.7710	1	0.7304	0.3955
Error	79.2	90		
Total	116.9	95		

Table 13. ANOVA Summary for Feature 4 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	662.3	5		
A	2.4172	1	0.2191	0.6411
B	9.8389	1	0.8919	0.3480
AB	1.3380	1	0.1213	0.7286
C	194.0859	1	17.5945	0.0001
AC	1.0908	1	0.0989	0.7540
BC	11.4817	1	1.0409	0.3109
ABC	0.1612	1	0.0146	0.9041
D	0.8626	1	0.0782	0.7805
AD	0.3545	1	0.0321	0.8582
BD	1.1267	1	0.1021	0.7502
ABD	0.0167	1	0.0015	0.9691
CD	1.4259	1	0.1293	0.7202
ACD	0.4864	1	0.0441	0.8342
BCD	1.9837	1	0.1798	0.6727
ABCD	0.6316	1	0.0573	0.8115
Error	827.3	90		
Total	1716.4	95		

Table 14. ANOVA Summary for Feature 5 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	1361.2	5		
A	1.5547	1	0.0557	0.8140
B	24.5532	1	0.8799	0.3512
AB	9.8069	1	0.3515	0.5551
C	195.6532	1	7.0117	0.0099
AC	15.0549	1	0.5395	0.4649
BC	2.1750	1	0.0779	0.7809
ABC	0.4200	1	0.0151	0.9027
D	0.8344	1	0.0299	0.8632
AD	0.1288	1	0.0046	0.9460
BD	1.8013	1	0.0646	0.8001
ABD	1.1521	1	0.0413	0.8395
CD	2.3490	1	0.0842	0.7725
ACD	0.0319	1	0.0011	0.9731
BCD	3.5331	1	0.1266	0.7230
ABCD	0.2382	1	0.0085	0.9266
Error	2092.8	90		
Total	3713.2	95		

Table 15. ANOVA Summary for Feature 6 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	2046.3	5		
A	27.3244	1	0.7290	0.3959
B	243.2597	1	6.4900	0.0129
AB	38.7816	1	1.0347	0.3123
C	413.2708	1	11.0258	0.0014
AC	38.6123	1	1.0301	0.3134
BC	44.3496	1	1.1832	0.2802
ABC	20.0599	1	0.5352	0.4667
D	47.3438	1	1.2631	0.2646
AD	31.9127	1	0.8514	0.3591
BD	44.3043	1	1.1820	0.2804
ABD	23.1575	1	0.6178	0.4343
CD	41.1035	1	1.0966	0.2984
ACD	20.8600	1	0.5565	0.4580
BCD	34.8205	1	0.9290	0.3382
ABCD	42.6628	1	1.1382	0.2895
Error	2811.2	90		
Total	5948.5	95		

Table 16. ANOVA Summary for Feature 7 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	2769.9	5		
A	5.4865	1	0.2404	0.6254
B	57.0674	1	2.5001	0.1180
AB	20.8600	1	0.9139	0.3422
C	368.8851	1	16.1606	0.0001
AC	23.3544	1	1.0231	0.3150
BC	3.0400	1	0.1332	0.7162
ABC	0.5475	1	0.0240	0.8773
D	0.2982	1	0.0131	0.9093
AD	1.8013	1	0.0789	0.7795
BD	2.6500	1	0.1161	0.7343
ABD	5.7282	1	0.2509	0.6179
CD	18.5065	1	0.8108	0.3708
ACD	0.6750	1	0.0296	0.8639
BCD	0.2763	1	0.0121	0.9127
ABCD	1.3007	1	0.0570	0.8120
Error	1712.0	90		
Total	4991.6	95		

Table 17. ANOVA Summary for Feature 8 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	797.6	5		
A	0.6947	1	0.0760	0.7836
B	6.7557	1	0.7387	0.3928
AB	0.2236	1	0.0245	0.8762
C	197.9919	1	21.6478	0.0000
AC	5.3676	1	0.5869	0.4460
BC	3.8400	1	0.4199	0.5190
ABC	1.8058	1	0.1974	0.6581
D	4.2224	1	0.4617	0.4989
AD	1.0626	1	0.1162	0.7342
BD	15.7896	1	1.7264	0.1929
ABD	0.0214	1	0.0023	0.9615
CD	2.6667	1	0.2916	0.5908
ACD	0.1089	1	0.0119	0.9134
BCD	6.2017	1	0.6781	0.4129
ABCD	0.3456	1	0.0378	0.8464
Error	686.0	90		
Total	1730.5	95		

Table 18. ANOVA Summary for Feature 9 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	71.3	5		
A	0.1123	1	0.1265	0.7231
B	5.6632	1	6.3771	0.0137
AB	0.0642	1	0.0723	0.7887
C	0.0369	1	0.0416	0.8389
AC	0.2587	1	0.2913	0.5910
BC	5.0493	1	5.6858	0.0196
ABC	0.6976	1	0.7855	0.3783
D	0.0969	1	0.1091	0.7421
AD	0.0059	1	0.0066	0.9355
BD	0.1169	1	0.1316	0.7178
ABD	0.0163	1	0.0183	0.8927
CD	0.0969	1	0.1091	0.7421
ACD	0.0059	1	0.0066	0.9355
BCD	0.1169	1	0.1316	0.7178
ABCD	0.0221	1	0.0249	0.8750
Error	66.6	90		
Total	150.3	95		

Table 19. ANOVA Summary for Feature 10 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	1431.1	5		
A	1.1557	1	0.1150	0.7355
B	30.7889	1	3.0641	0.0841
AB	6.0669	1	0.6038	0.4396
C	122.4770	1	12.1887	0.0008
AC	5.4785	1	0.5452	0.4626
BC	21.0001	1	2.0899	0.1524
ABC	0.0007	1	0.0001	0.9932
D	1.9361	1	0.1927	0.6620
AD	0.4091	1	0.0407	0.8406
BD	6.8801	1	0.6847	0.4106
ABD	1.7424	1	0.1734	0.6783
CD	2.0709	1	0.2061	0.6512
ACD	0.0896	1	0.0089	0.9250
BCD	10.1617	1	1.0113	0.3178
ABCD	2.8459	1	0.2832	0.5962
Error	753.6	90		
Total	2397.8	95		

Table 20. ANOVA Summary for Feature 11 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	13567.9	5		
A	6.9876	1	0.0920	0.7625
B	245.1204	1	3.2270	0.0765
AB	72.2801	1	0.9516	0.3325
C	2008.7300	1	26.4448	0.0000
AC	101.7503	1	1.3395	0.2508
BC	91.7807	1	1.2083	0.2752
ABC	1.8611	1	0.0245	0.8760
D	22.2980	1	0.2936	0.5896
AD	3.0936	1	0.0407	0.8406
BD	17.7389	1	0.2335	0.6303
ABD	0.0111	1	0.0001	0.9904
CD	13.7512	1	0.1810	0.6717
ACD	6.2526	1	0.0823	0.7750
BCD	12.9067	1	0.1699	0.6814
ABCD	6.5910	1	0.0868	0.7691
Error	5696.9	90		
Total	21869.8	95		

Table 21. ANOVA Summary for Feature 12 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	51395.8	5		
A	0.5104	1	0.0017	0.9669
B	584.9292	1	1.9922	0.1622
AB	113.6800	1	0.3872	0.5357
C	7615.4375	1	25.9373	0.0000
AC	113.8252	1	0.3877	0.5354
BC	454.5751	1	1.5482	0.2173
ABC	0.0600	1	0.0002	0.9886
D	34.8406	1	0.1187	0.7315
AD	17.5674	1	0.0598	0.8074
BD	0.2709	1	0.0009	0.9758
ABD	4.1667	1	0.0142	0.9055
CD	0.0051	1	0.0000	0.9967
ACD	4.5938	1	0.0156	0.9008
BCD	31.2436	1	0.1064	0.7452
ABCD	20.2216	1	0.0689	0.7937
Error	22020.7	90		
Total	82407.8	95		

Table 22. ANOVA Summary for Feature 13 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	51807.2	5		
A	13.2017	1	0.0379	0.8461
B	1349.5000	1	3.8787	0.0526
AB	141.4585	1	0.4066	0.5257
C	8234.9800	1	23.6690	0.0000
AC	196.0817	1	0.5636	0.4552
BC	672.0417	1	1.9316	0.1687
ABC	13.1030	1	0.0377	0.8467
D	92.8267	1	0.2668	0.6070
AD	1.7067	1	0.0049	0.9443
BD	0.9467	1	0.0027	0.9585
ABD	5.2267	1	0.0150	0.9028
CD	74.3189	1	0.2136	0.6453
ACD	0.8067	1	0.0023	0.9617
BCD	7.1141	1	0.0204	0.8867
ABCD	4.8630	1	0.0140	0.9062
Error	26094.2	90		
Total	88708.8	95		

Table 23. ANOVA Summary for Feature 14 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	55769.7	5		
A	20.4426	1	0.0862	0.7699
B	369.3426	1	1.5567	0.2160
AB	212.7126	1	0.8965	0.3468
C	4189.2433	1	17.6568	0.0001
AC	127.8047	1	0.5387	0.4653
BC	627.8151	1	2.6461	0.1080
ABC	14.2861	1	0.0602	0.8068
D	466.2547	1	1.9652	0.1651
AD	0.4770	1	0.0020	0.9644
BD	58.7501	1	0.2476	0.6202
ABD	0.0042	1	0.0000	0.9967
CD	353.3058	1	1.4891	0.2262
ACD	3.2881	1	0.0139	0.9066
BCD	72.9759	1	0.3076	0.5808
ABCD	5.4381	1	0.0229	0.8801
Error	17794.4	90		
Total	80083.0	95		

Table 24. ANOVA Sumary for Feature 15 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	16840.8	5		
A	1.4875	1	0.0165	0.8981
B	41.7165	1	0.4630	0.4983
AB	15.5070	1	0.1721	0.6794
C	1602.4366	1	17.7836	0.0001
AC	87.5608	1	0.9717	0.3274
BC	178.5331	1	1.9813	0.1634
ABC	42.0688	1	0.4669	0.4965
D	59.7189	1	0.6628	0.4182
AD	15.8573	1	0.1760	0.6760
BD	3.7538	1	0.0417	0.8388
ABD	6.8712	1	0.0763	0.7832
CD	43.0899	1	0.4782	0.4914
ACD	6.3123	1	0.0701	0.7920
BCD	13.1893	1	0.1464	0.7031
ABCD	22.3871	1	0.2484	0.6196
Error	6758.1	90		
Total	25733.0	95		

Table 25. ANOVA Summary for Feature 16 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	349.8	5		
A	0.0741	1	0.0243	0.8766
B	14.4150	1	4.7210	0.0330
AB	3.1538	1	1.0329	0.3128
C	9.6689	1	3.1666	0.0792
AC	1.2757	1	0.4178	0.5200
BC	4.7112	1	1.5429	0.2181
ABC	1.1123	1	0.3643	0.5480
D	7.5563	1	2.4747	0.1199
AD	0.0150	1	0.0049	0.9443
BD	2.8935	1	0.9476	0.3335
ABD	1.7245	1	0.5648	0.4547
CD	5.1956	1	1.7016	0.1961
ACD	0.1067	1	0.0349	0.8522
BCD	3.8667	1	1.2664	0.2640
ABCD	1.3367	1	0.4378	0.5102
Error	229.0	90		
Total	635.8	95		

Table 26. ANOVA Summary for Feature 17 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	1795.1	5		
A	0.9302	1	0.0443	0.8338
B	74.6419	1	3.5558	0.0632
AB	22.3784	1	1.0661	0.3052
C	43.3135	1	2.0634	0.1550
AC	29.2052	1	1.3913	0.2419
BC	48.0959	1	2.2912	0.1343
ABC	7.0146	1	0.3342	0.5650
D	5.2500	1	0.2501	0.6185
AD	0.0438	1	0.0021	0.9637
BD	11.6088	1	0.5530	0.4594
ABD	6.6940	1	0.3189	0.5740
CD	2.1152	1	0.1008	0.7518
ACD	1.4875	1	0.0709	0.7908
BCD	27.7171	1	1.3204	0.2542
ABCD	6.2357	1	0.2971	0.5874
Error	1574.4	90		
Total	3654.7	95		

Table 27. ANOVA Summary for Feature 18 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	13039.0	5		
A	0.9970	1	0.0134	0.9081
B	276.8170	1	3.7282	0.0573
AB	38.9513	1	0.5246	0.4711
C	202.5657	1	2.7282	0.1028
AC	76.1188	1	1.0252	0.3145
BC	218.1554	1	2.9382	0.0906
ABC	13.2140	1	0.1780	0.6743
D	69.3883	1	0.9345	0.3368
AD	32.1438	1	0.4329	0.5126
BD	8.3721	1	0.1128	0.7380
ABD	37.6042	1	0.5065	0.4789
CD	77.0715	1	1.0380	0.3116
ACD	26.6177	1	0.3585	0.5511
BCD	7.0146	1	0.0945	0.7594
ABCD	73.4945	1	0.9898	0.3230
Error	5568.7	90		
Total	19739.6	95		

Table 28. ANOVA Summary for Feature 19 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	12226.8	5		
A	5.8181	1	0.0593	0.8083
B	154.7030	1	1.5771	0.2131
AB	22.9126	1	0.2336	0.6303
C	190.0313	1	1.9373	0.1681
AC	185.0926	1	1.8869	0.1736
BC	96.2669	1	0.9814	0.3250
ABC	8.9833	1	0.0916	0.7630
D	141.6204	1	1.4438	0.2333
AD	102.1625	1	1.0415	0.3108
BD	106.8223	1	1.0890	0.3000
ABD	43.4256	1	0.4427	0.5079
CD	151.5038	1	1.5445	0.2178
ACD	84.1876	1	0.8583	0.3572
BCD	27.3778	1	0.2791	0.5988
ABCD	104.0802	1	1.0611	0.3063
Error	7356.9	90		
Total	20924.4	95		

Table 29. ANOVA Summary for Feature 20 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	9188.9	5		
A	65.2850	1	0.6171	0.4346
B	282.9067	1	2.6740	0.1062
AB	31.9320	1	0.3018	0.5844
C	206.5067	1	1.9519	0.1665
AC	205.2375	1	1.9399	0.1678
BC	130.0452	1	1.2292	0.2711
ABC	1.5420	1	0.0146	0.9042
D	220.6245	1	2.0853	0.1529
AD	86.9570	1	0.8219	0.3675
BD	20.7204	1	0.1958	0.6594
ABD	1.3936	1	0.0132	0.9089
CD	249.4000	1	2.3573	0.1289
ACD	101.2020	1	0.9565	0.3312
BCD	27.5204	1	0.2601	0.6115
ABCD	108.4436	1	1.0250	0.3146
Error	7935.0	90		
Total	18762.4	95		

Table 30. ANOVA Summary for Feature 21 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	244.5	5		
A	11.9239	1	0.7151	0.4004
B	86.1972	1	5.1694	0.0258
AB	12.2075	1	0.7321	0.3949
C	49.9297	1	2.9944	0.0877
AC	27.6992	1	1.6612	0.2014
BC	42.8892	1	2.5722	0.1130
ABC	0.6611	1	0.0396	0.8427
D	41.1253	1	2.4664	0.1205
AD	14.2861	1	0.8568	0.3576
BD	6.0167	1	0.3608	0.5499
ABD	3.9339	1	0.2359	0.6286
CD	32.1631	1	1.9289	0.1690
ACD	10.9125	1	0.6544	0.4211
BCD	5.0264	1	0.3014	0.5846
ABCD	13.1839	1	0.7907	0.3767
Error	1250.6	90		
Total	1842.3	95		

Table 31. ANOVA Summary for Feature 22 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	439.8	5		
A	0.7764	1	0.0846	0.7719
B	16.7780	1	1.8291	0.1803
AB	0.2745	1	0.0299	0.8631
C	70.9557	1	7.7354	0.0068
AC	11.1612	1	1.2168	0.2735
BC	12.2551	1	1.3360	0.2514
ABC	0.4311	1	0.0470	0.8290
D	3.7869	1	0.4128	0.5225
AD	7.5563	1	0.8238	0.3670
BD	4.6670	1	0.5088	0.4779
ABD	0.8626	1	0.0940	0.7600
CD	0.1751	1	0.0191	0.8905
ACD	2.5895	1	0.2823	0.5968
BCD	4.2224	1	0.4603	0.4996
ABCD	2.7791	1	0.3030	0.5837
Error	688.0	90		
Total	1264.5	95		

Table 32. ANOVA Summary for Feature 23 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	871.5	5		
A	0.7585	1	0.0419	0.8384
B	62.4575	1	3.4487	0.0672
AB	0.1300	1	0.0072	0.9327
C	136.4061	1	7.5319	0.0076
AC	19.0817	1	1.0536	0.3080
BC	96.2001	1	5.3118	0.0239
ABC	0.0938	1	0.0052	0.9428
D	13.4750	1	0.7440	0.3911
AD	5.2267	1	0.2886	0.5927
BD	4.0700	1	0.2247	0.6368
ABD	0.1400	1	0.0077	0.9302
CD	4.8750	1	0.2692	0.6054
ACD	1.2757	1	0.0704	0.7914
BCD	1.3617	1	0.0752	0.7847
ABCD	1.8880	1	0.1042	0.7477
Error	1358.3	90		
Total	2576.0	95		

Table 33. ANOVA Summary for Feature 24 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	4063.2	5		
A	23.9167	1	0.5499	0.4607
B	11.3552	1	0.2611	0.6109
AB	2.3074	1	0.0531	0.8185
C	37.1467	1	0.8541	0.3584
AC	62.6997	1	1.4417	0.2336
BC	148.7941	1	3.4212	0.0683
ABC	0.0826	1	0.0019	0.9653
D	50.2910	1	1.1563	0.2857
AD	32.4144	1	0.7453	0.3907
BD	9.0754	1	0.2087	0.6491
ABD	1.3577	1	0.0312	0.8602
CD	25.6439	1	0.5896	0.4450
ACD	26.5826	1	0.6112	0.4368
BCD	8.7504	1	0.2012	0.6550
ABCD	33.3472	1	0.7668	0.3840
Error	3261.8	90		
Total	7772.3	95		

Table 34. ANOVA Summary for Feature 25 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	1032.4	5		
A	0.1078	1	0.0047	0.9458
B	23.8835	1	1.0318	0.3130
AB	0.4840	1	0.0209	0.8854
C	245.0672	1	10.5876	0.0017
AC	31.3769	1	1.3556	0.2480
BC	124.7844	1	5.3910	0.0230
ABC	1.9031	1	0.0822	0.7751
D	14.5834	1	0.6300	0.4298
AD	15.1077	1	0.6527	0.4217
BD	8.6500	1	0.3737	0.5428
ABD	0.7263	1	0.0314	0.8599
CD	10.7223	1	0.4632	0.4982
ACD	5.8757	1	0.2538	0.6159
BCD	1.8938	1	0.0818	0.7756
ABCD	5.9180	1	0.2557	0.6146
Error	1736.0	90		
Total	3253.6	95		

Table 35. ANOVA Summary for Feature 26 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	269.7	5		
A	0.0452	1	0.0071	0.9330
B	6.2952	1	0.9918	0.3225
AB	0.2622	1	0.0413	0.8395
C	52.9304	1	8.3391	0.0051
AC	10.0643	1	1.5856	0.2119
BC	4.4706	1	0.7043	0.4040
ABC	4.8525	1	0.7645	0.3847
D	8.7705	1	1.3818	0.2435
AD	1.4138	1	0.2227	0.6383
BD	5.0035	1	0.7883	0.3775
ABD	0.8721	1	0.1374	0.7119
CD	5.3755	1	0.8469	0.3604
ACD	0.2982	1	0.0470	0.8290
BCD	3.4567	1	0.5446	0.4628
ABCD	0.6076	1	0.0957	0.7579
Error	476.0	90		
Total	850.1	95		

Table 36. ANOVA Summary for Feature 27 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	93582.1	5		
A	284.7982	1	0.3315	0.5665
B	673.5418	1	0.7841	0.3787
AB	32.5695	1	0.0379	0.8461
C	4258.4484	1	4.9573	0.0290
AC	501.3442	1	0.5836	0.4473
BC	4596.5869	1	5.3510	0.0235
ABC	739.1675	1	0.8605	0.3566
D	1591.0174	1	1.8521	0.1776
AD	46.3658	1	0.0540	0.8169
BD	169.4688	1	0.1973	0.6582
ABD	418.2654	1	0.4869	0.4875
CD	1812.9368	1	2.1105	0.1505
ACD	97.1032	1	0.1130	0.7376
BCD	112.4862	1	0.1309	0.7185
ABCD	413.2641	1	0.4811	0.4901
Error	64426.5	90		
Total	173658.8	95		

Table 37. ANOVA Summary for Feature 28 Analysis

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F_0	α
Blocks	31073.9	5		
A	88.8067	1	0.2661	0.6075
B	287.2723	1	0.8607	0.3565
AB	181.1335	1	0.5427	0.4636
C	1772.1745	1	5.3098	0.0240
AC	0.1350	1	0.0004	0.9840
BC	1322.6400	1	3.9629	0.0502
ABC	101.8189	1	0.3051	0.5824
D	495.3445	1	1.4842	0.2269
AD	4.9807	1	0.0149	0.9031
BD	105.9800	1	0.3175	0.5748
ABD	11.0704	1	0.0332	0.8560
CD	932.0912	1	2.7928	0.0989
ACD	7.1504	1	0.0214	0.8840
BCD	209.8445	1	0.6287	0.4303
ABCD	7.2484	1	0.0217	0.8832
Error	25031.5	90		
Total	61626.0	95		

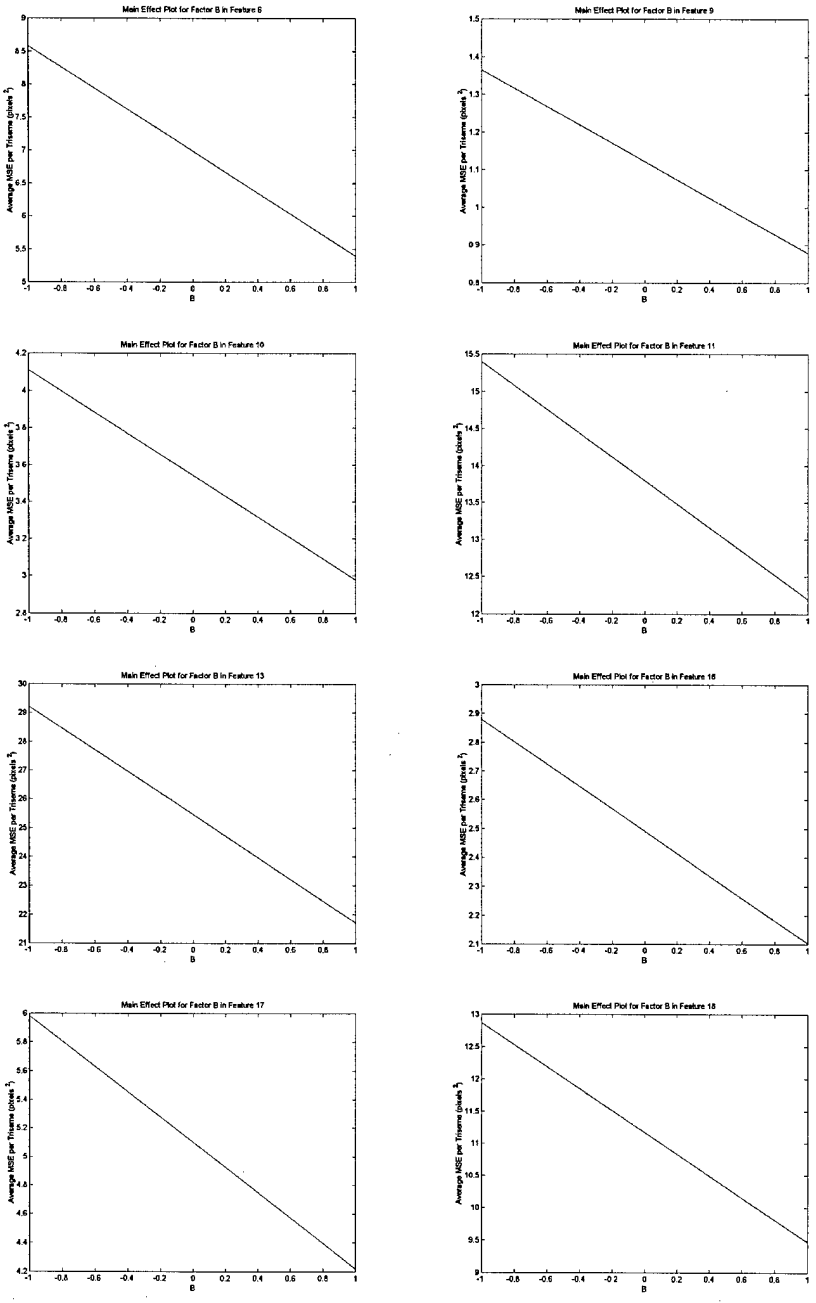


Figure 129. Main effect plots for factor B in features 6, 9, 10, 11, 13, 16-18 using block analysis

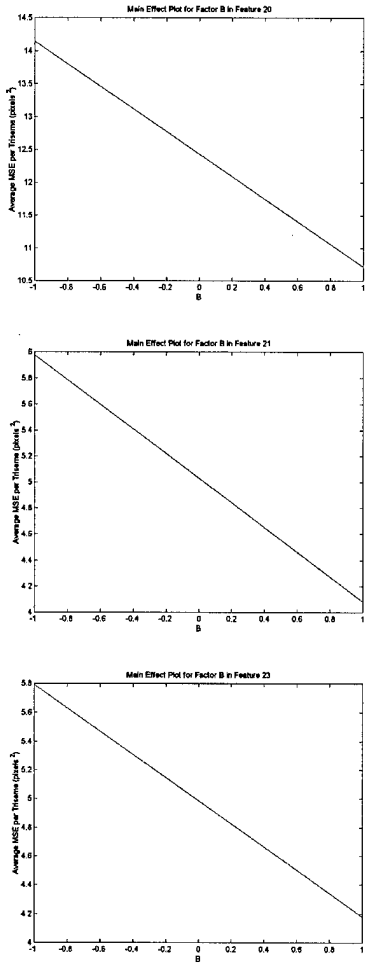


Figure 130. Main effect plots for factor B in features 20, 21 and 23 using block analysis

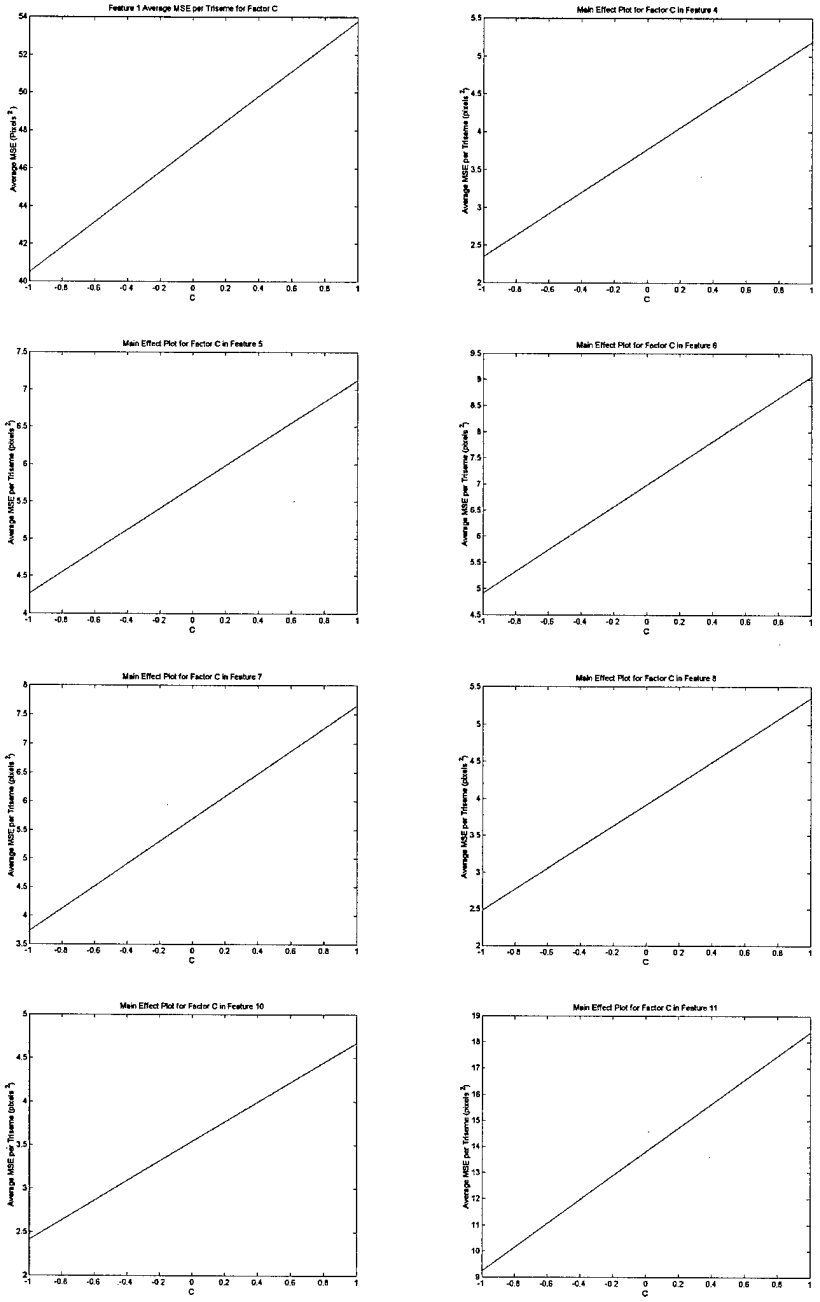


Figure 131. Main effect plots for factor C in features 1, 4-8, 10 and 11 using block analysis

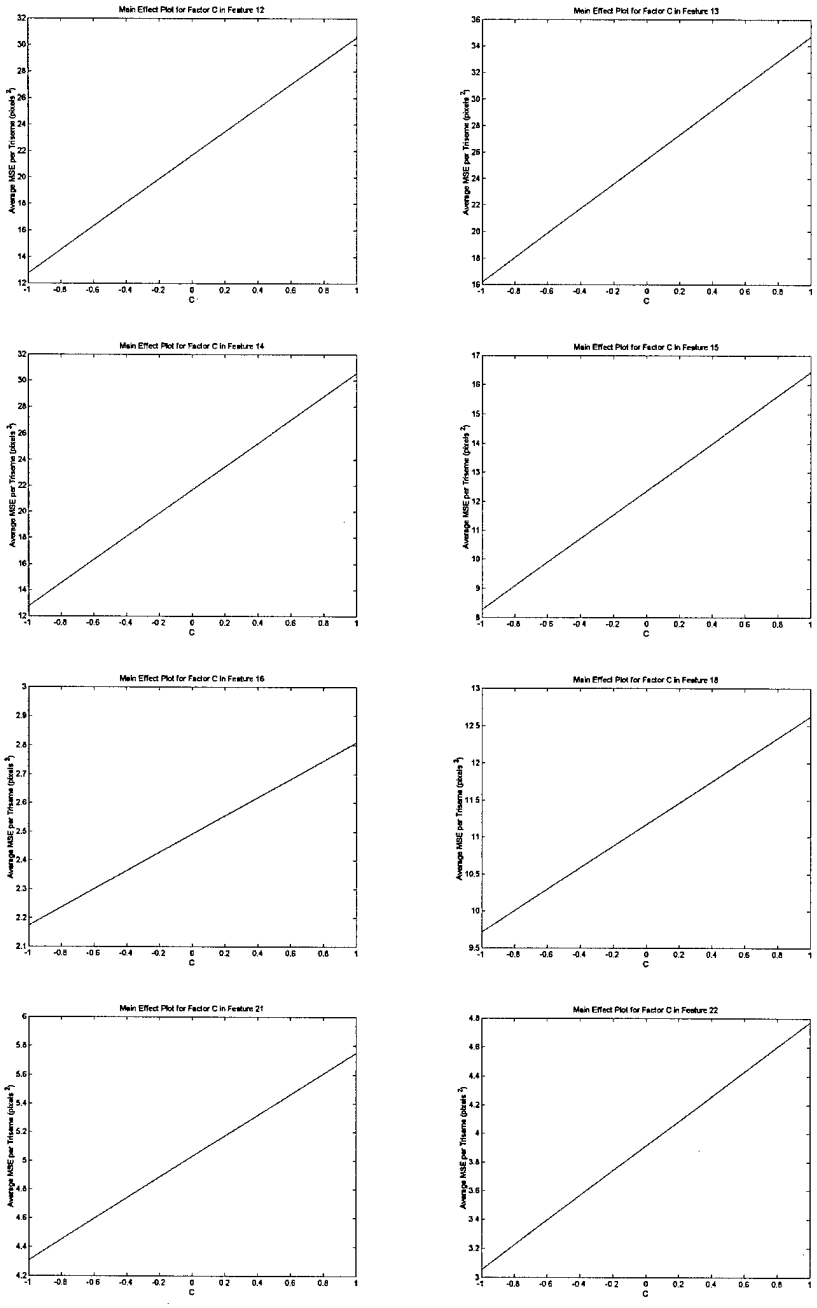


Figure 132. Main effect plots for factor C in features 12-16, 18, 21 and 22 using block analysis

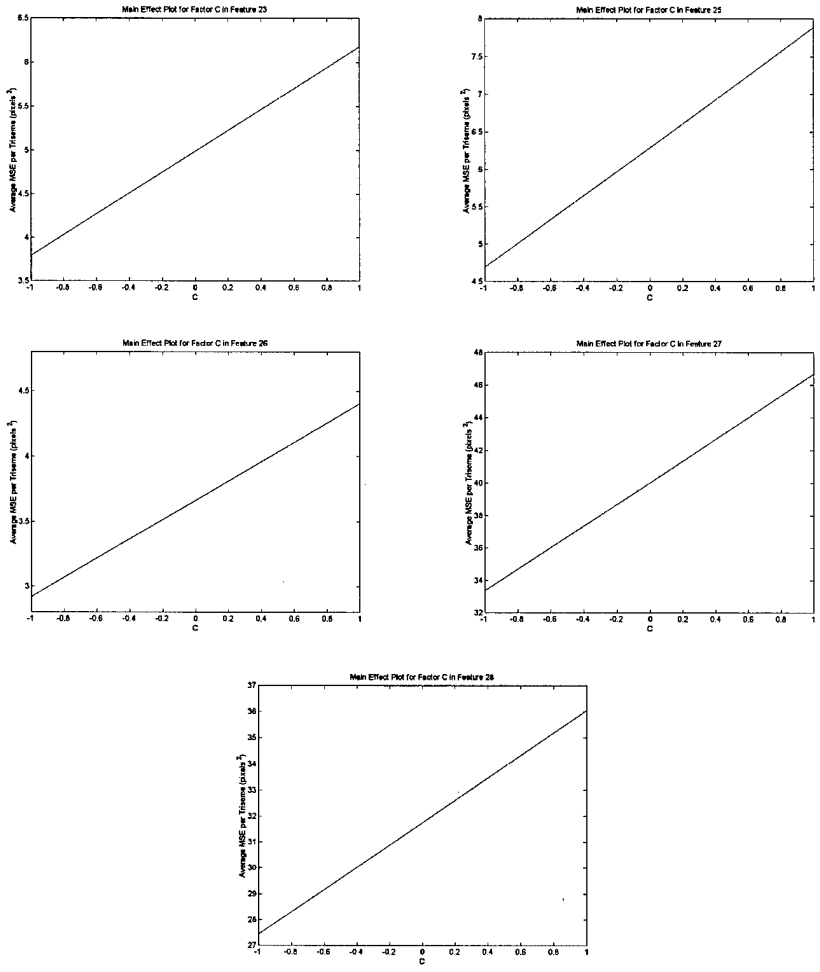


Figure 133. Main effect plots for factor C in features 23 and 25-28 using block analysis

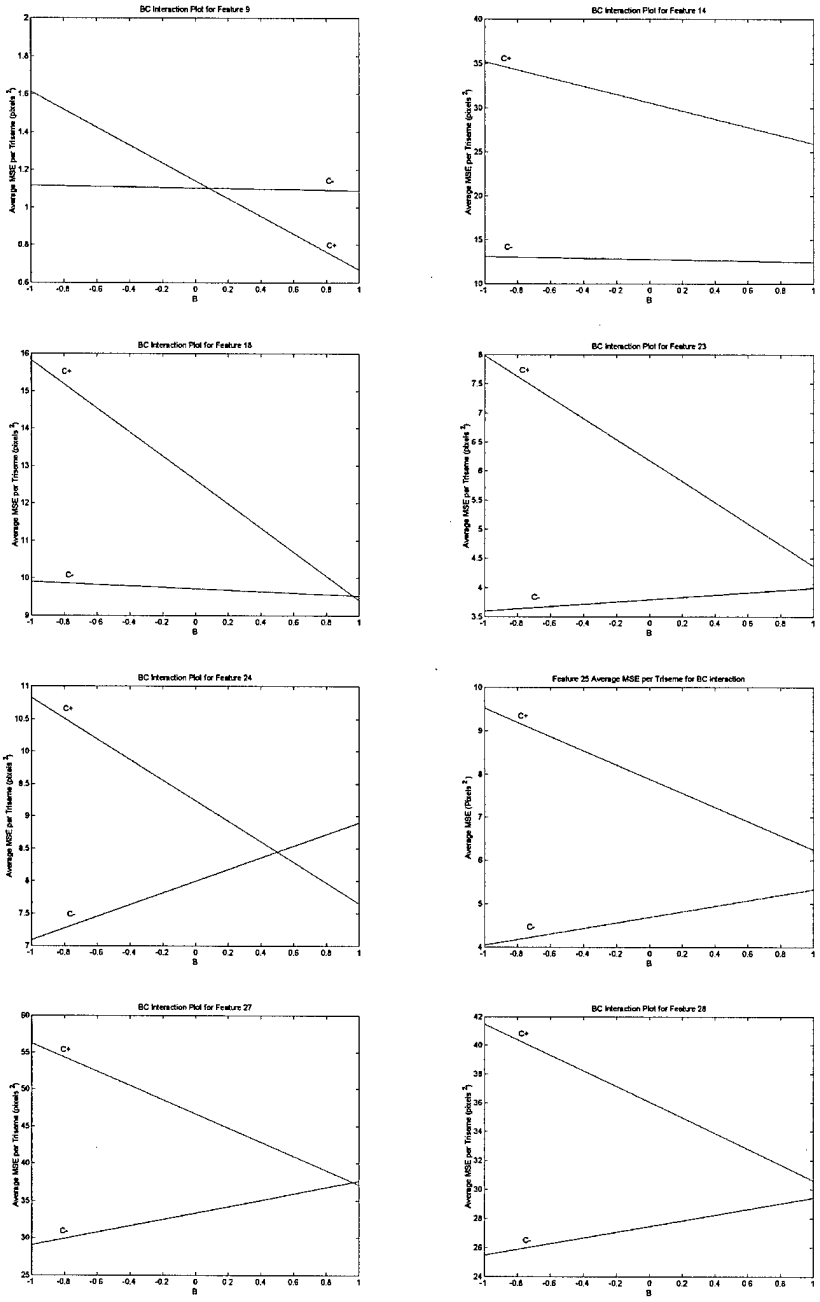


Figure 134. BC Interaction plots for features 9, 14, 18, 23-25, 27 and 28 using block analysis

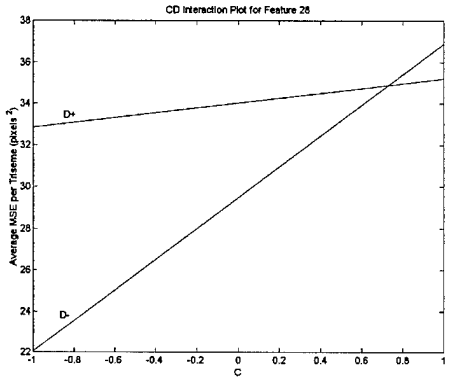


Figure 135. CD interaction plot for feature 28 using block analysis

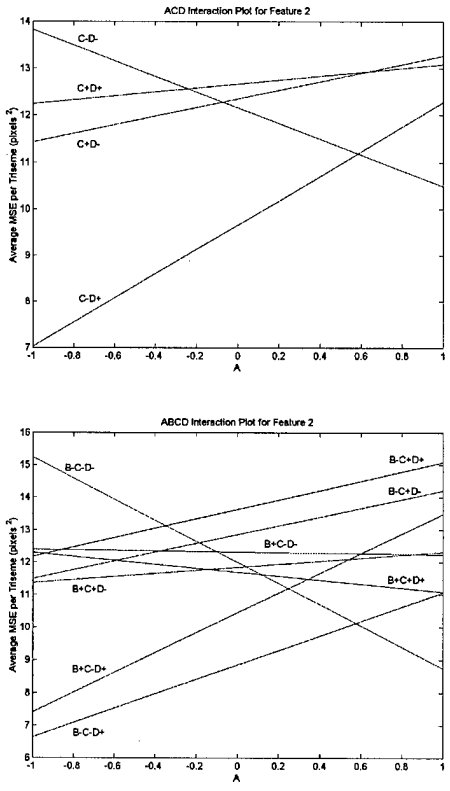


Figure 136. ACD and ABCD interaction plots for feature 2 using block analysis

Bibliography

1. Beier, T. and S. Neely. "Feature-Based Image Metamorphosis." *SIGGRAPH 92 Computer Graphics Proceedings*. 35–42. 1992.

2. Bergen, J.R. and Others. "Transparent-motion Analysis." *Computer Vision - ECCV'90, First European Conference on Computer Vision*. 566–569. May 1990.

3. Bergen, J.R. and Others. "Hierarchical model-based motion estimation." *Computer Vision - ECCV'92, Second European Conference on Computer Vision*. 237–252. May 1992.

4. Beymer, D., A. Shashua and T. Poggio. *Example Based Image Analysis and Synthesis*. A.I. Memo 1431, Massachusetts Institute of Technology, November 1993.

5. Bregler, C. and Yochai Konig. ""Eigenlips" for Robust Speech Recognition." *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*. Adelaide, Australia: IEEE, 1994. Accessible on WWW at http://mambo.ucsc.edu/psl/Bregler/icassp94.final.ps.

6. Bregler, C. and S. Omohundro. "Surface Learning with Applications to Lipreading." *Advances in Neural Information Processing Systems* edited by G. Tesauro J.D. Cowan and J. Alspector, San Francisco, CA: Morgan Kaufmann, 1994.

7. Bregler, C., M. Covell and M. Slaney. "Video Rewrite: Driving Visual Speech with Audio." *SIGGRAPH 97 Computer Graphics Proceedings*. Annual Conference Series. 353–360. 1997. Available at http://graphics.stanford.edu/ bregler/videorewrite/.

8. Brook, N.M. and E.D. Petajan. "Seeing Speech: Investigations Into the Synthesis and Recognition of Visible Speech Movements Using Automatic Image Processing and Computer Graphics." *Proceedings of the 1986 International Conference on Speech Input/Output; Techniques and Applications*. 104–109. March 1986.

9. Chen, T., H.P. Graf and K. Wang. "Speech-Assisted Video Processing: Interpolation and Low-Bitrate Coding." *Conference Record of The Twenty-Eighth Asilomar Conference on Signals, Systems & Computers2*. 975–979. November 1994.

10. Chen, T. and others. "Lip Synchronization in Talking Head Video Utilizing Speech Information." *Proceedings of the SPIE-The International Society for Optical Engineering2501*. 1690–1701. May 1995. pt. 3.

11. Chen, T. and others. "Speech-Assisted Lip Synchronization in Audio-Visual Communications." *Proceedings of the International Conference on Image Processing*. 579–582. October 1995.

12. Chen, T. and R.R. Rao. "Audio-Visual Interaction in Multimedia Communication." *Proceedings of the 1997 International Conference on Acoustics, Speech, and Signal Processing*. 179–182. April 1997.

13. Chou, W. and H. Chen. "Speech Recognition for Image Animation and Coding." *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing4*. 2253–2256. May 1995.

14. Cohen, M., J. Beskow and D. Massaro. "Recent Developments In Facial Animation: An Inside View." *AVSP'98 Workshop*. 201–206. December 1998. Available on WWW at http://mambo.ucsc.edu/psl/avsp98/11.doc.

15. Cohen, M. and D. Massaro. "Modeling Coarticulation in Synthetic Visual Speech." *Proceedings of Computer Animation 93*, edited by Magnenat-Thalmann & Thalmann. 1993.

16. Cohen, M. and D. Massaro. "Speech for Virtual Humans." *Proceedings of the Virtual Humans 2 Conference*. June 1997.

17. Cosatto, E. and H.P. Graf. "Photo-Realistic Talking-Heads from Image Samples," *IEEE Transactions on Multimedia*, *2*(3):152–163 (September 2000).

18. Covell, M. "Eigenpoints: Control-point Location using Principal Component Analysis." *Proceedings of the second International Conference on Automatic Face and Gesture Recognition*. Killington, VT: IEEE, October 1996.

19. DARPA, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc CD1-1.1." Produced on CD-ROM by the National Institute of Standards and Technology (NIST), October 1990.

20. Donovan, R.E. and P.C. Woodland. "Automatic Speech Synthesizer Parameter Estimation Using HMMs." *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing1*. 640–643. May 1995.

21. Dupont, S. and J. Luettin. "Audio-Visual Speech Modeling for Continuous Speech Recognition," *IEEE Transactions on Multimedia*, *2*(3):141–151 (September 2000).

22. Ezzat, T. and T. Poggio. "Facial Analysis and Synthesis Using Image-Based Models." *Proceedings of the Workshop on the Algorithmic Foundations of Robotics*. August 1996. Available on WWW at http://cuneus.ai.mit.edu:8000/personal/publications.html.

23. Ezzat, T. and T. Poggio. "Videorealistic Talking Faces: A Morphing Approach." *AVSP'97 Workshop*. September 1997. Available on WWW at http://cuneus.ai.mit.edu:8000/personal/publications.html.

24. Ezzat, T. and T. Poggio. "MikeTalk: A Talking Facial Display Based on Morphing Visemes." *Proceedings of the Computer Animation Conference*. June 1998. Available on WWW at http://cuneus.ai.mit.edu:8000/personal/publications.html.

25. Ezzat, T. and T. Poggio. *Visual Speech Synthesis by Morphing Visemes*. A.I. Memo 1658, Massachusetts Institute of Technology, May 1999.

26. Gagne, J.P., D. Dinon and J. Parsons. "An Evaluation of CAST: A Computer-Aided Speechreading Training Program," *Journal of Speech and Hearing Research*, *34*:213–221 (February 1991).

27. Goldschen, A.J., O.N. Garcia and E. Petajan. "Continuous Optical Automatic Speech Recognition by Lipreading." *Conference Record of The Twenty-Eighth Asilomar Conference on Signals, Systems & Computers*. 572–577. November 1994.

28. Golub, G. and C. Van Loan. *Matrix Computations* (Second Edition). Baltimore, MD: Johns Hopkins University Press, 1989.

29. Gray, M., J. Movellan and T. Sejnowski. "Dynamic Features for visual speechreading: A systematic comparison." *Advances in Neural Information Processing Systems9*. 751–757. 1997.

30. Horn, B.K.P. and B.G. Schunck. "Determining Optical Flow," *Artificial Intelligence*, *17*:185–203 (1981).

31. Kober, R., U. Harz and J. Schiffers. "Fusion of Visual and Acoustic Signals for Command-Word Recognition." *Proceedings of the 1997 International Conference on Acoustics, Speech, and Signal Processing2*. 1495–1497. 1997.

32. Lavagetto, F. and others. "Lip Motion Modeling and Speech Driven Estimation." *Proceedings of the 1997 International Conference on Acoustics, Speech, and Signal Processing1*. 183–186. March 1997.

33. Le Goff, B. and C. Benoit. "A Text-to-Audiovisual-Speech Synthesizer for French." *Proceedings of the Fourth International Conference on Spoken Language Processing*. October 1996.

34. Le Goff, B. and others. "Real-Time Analysis-Synthesis and Intelligibility of Talking Faces." *Proceedings of the 2nd International Conference on Speech Synthesis*. September 1994.

35. Leuttin, J., N. Thacker and S. Beet. "Visual Speech Recognition Using Active Shape Models and Hidden Markov Models." *Proceedings of the 1996 International Conference on Acoustics, Speech, and Signal Processing1*. 817–820. May 1996.

36. Ljolje, A and F. Fallside. "Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, *ASSP-34*(5):1074–1079 (October 1997).

37. Luo, S. H. and R. W. King. "A Novel Approach for Classifying Continuous Speech Into Visible Mouth-Shape Related Classes." *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing1*. 464–467. 1994.

38. Massaro, D. and M. Cohen. "Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables," *Speech Communication*, *13*:127–134 (1993).

39. Massaro, D., M. Cohen and P. Smeele. "Perception of asynchronous and conflicting visual and auditory speech," *Journal of the Acoustical Society of America*, *100*(3):1777–1786 (September 1996).

40. Masuko, T and others. "Speech Synthesis Using HMMs with Dynamic Features." *Proceedings of the 1996 International Conference on Acoustics, Speech, and Signal Processing1*. 398–392. May 1996.

41. Masuko, T and others. "Voice Characteristics Conversion for HMM-Based Speech Synthesis System." *Proceedings of the 1997 International Conference on Acoustics, Speech, and Signal Processing*. 1611–1614. April 1997.

42. Masuko, T and others. "Text-To-Visual Speech Synthesis Based on Parameter Generation from HMM." *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing*. 3745–3748. May 1998.

43. Matthews, I., J.A. Bangaham and S. Cox. "Audiovisual Speech recognition Using Multiscale Nonlinear Image Decomposition." *Proceedings of the Fourth International Conference on Spoken Language Processing*. October 1996.

44. McGurk, H. and J. MacDonald. "Hearing Lips and Seeing Voices," *Nature*, *264*:746–748 (December 1976).

45. Meier, U., R. Stiefelhagen and J. Yang. "Preprocessing of Visual Speech Under Real World Conditions." *AVSP'97 Workshop*. September 1997. Available on WWW at http://www.cs.cmu.edu/ uwem/work/papers/97.avsp.ps.gz.

46. Miramax Films. *Life is Beautiful*. Internet Advertisement. November 2, 1999. Available on WWW at http://www.miramax.com/lifeisbeautiful/.

47. Moghaddam, B. and A. Pentland. *Probabilistic Visual Learning for Object Detection*. M.I.T. Media Laboratory Technical Report 326, Massachusetts Institute of Technology, June 1995.

48. Montgomery, Douglas C. *Design and Analysis of Experiments* (Fourth Edition). John Wiley and Sons, 1997.

49. Movellan, J.R. "Visual Speech Recognition with Stochastic Networks." *Advances in Neural Information Processing Systems 7*, edited by D. Toruetzky Tesauro, G. and T. Leen, MIT Press, 1995.

50. Neter, J. and Others. *Applied Linear Regression Models* (Third Edition). Irwin, 1996.

51. Odell, J.J., P.C. Woodland and S.J. Young. "Tree-Based State Clustering for Large Vocabulary Speech Recognition." *Proceedings of the 1994 International Symposium on Speech, Image Processing and Neural Networks*. 690–693. April 1994.

52. Parke, F.I. *A Parametric Model for Human Faces.* PhD dissertation, University of Utah, 1974.

53. Parke, F.I. "A Model for Human Faces That Allows Speech Synchronized Animation," *Journal of Computers and Graphics*, *1*(1):3–4 (May 1975).

54. Pelachaud, C., N. Badler and M. Viaud, "Final Report to NSF of the Standards for Facial Animation Workshop," October 1994.

55. Petajan, E. *Automatic Lipreading to Enhance Speech Recognition.* PhD dissertation, University of Illinois, 1984.

56. Petajan, E. "Automatic Lipreading to Enhance Speech Recognition." *Proceedings of the 1985 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 40–47. June 1985.

57. Petajan, E. and H.P. Graf. "Robust Face Feature Analysis for Automatic Speechreading and Character Animation." *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition.* 357–362. October 1996.

58. Pichora-Fuller, M. K. and A-P. Benguerel. "The Design of CAST (Computer-Aided Speechreading Training)," *Journal of Speech and Hearing Research*, *34*:202–212 (February 1991).

59. Pixar Animation Studios. *RenderMan.* Brochure. 1999. Available on WWW at http://www.pixar.com/products/renderman/brochure99/rman.html.

60. Poritz, Alan B. "Hidden Markov Models: A Guided Tour." *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing1.* 7–13. 1988.

61. Potamianos, G. and Others. "Speaker independent audio-visual database for bimodal ASR." *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing: Cognitive and Computational Approaches.* 65–68. September 1997.

62. Prasad, K.V., D. Stork and G. Wolff. *Preprocessing Video Images for Neural Learning of Lipreading.* Technical Report CRC-TR-93-26, Ricoh California Research Center, 1993.

63. Rabiner, L. "A Tutorial of Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE77.* 257–286. February 1989.

64. Rabiner, L. and B.H. Juang. "An Introduction to Hidden Markov Models," *IEEE Acoustics, Speech and Signal Processing Magazine*, *3*(1):4–16 (January 1986).

65. Rabiner, L. and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

66. Reveret, L. and C. Benoit. "A Viseme-based Approach to Labiometrics for Automatic Lipreading." *Proceedings of First International Conference on Audio and Video Based Biometric Person Authentication.* 335–342. March 1997.

67. Robert-Ribes, Jordi and Others. "Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise," *Journal of the Acoustical Society of America*, *103*(6):3677–3689 (June 1998).

68. Rogozan, A. and P. Deléglise. "Continuous Visual Speech Recognition Using Geometric Lip-Shape Models and Neural Networks." *Proceedings of the Fifth Conference on Speech Communication Technology*. 1999–2000. September 1997.

69. Rogozan, A. and P. Deléglise. "Hybrid Hidden Markov Model/Neural Network Models for Speechreading." *Proceedings of the European Symposium on Artificial Neural Networks*. 377–382. April 1998.

70. Rogozan, A. and P. Deléglise. "Visible Speech Modelling and Hybrid Hidden Markov Model/Neural Network Models Based Learning for Lipreading." *Proceedings of the IEEE Symposia on Intelligence and Systems*. 336–342. May 1998.

71. Rogozan, A., P. Deléglise and M. Alissali. "Adaptive Determination of Audio and Visual Weights for Automatic Speech Recognition." *Proceedings of the Audio and Visual Speech Processing Workshop*. 61–64. September 1997.

72. San Filippo, Michael, "Dubbing is Beautiful?." Internet article at http://italian.about.com/education/languages/italian/library/weekly/aa990901.htm, September 1, 1999.

73. Scott, K.C. and others. "Synthesis of Speaker Facial Movement to Match Selected Speech Sequences." *Proceedings of the Fifth Australian Conference on Speech Science and Technology2*. 620–625. December 1994.

74. Sekiyama, K. and Y. Tohkura. "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," *Journal of the Acoustical Society of America*, *90*(4):1797–1805 (October 1991).

75. Shashua, A. *Correspondence and Affine Shape from two Orthographic Views: Motion and Recognition*. A.I. Memo 1327, Massachusetts Institute of Technology, December 1991.

76. Shashua, A. *Geometry and Photometry in 3D Visual Recognition*. PhD dissertation, Massachusetts Institute of Technology, 1992.

77. Silsbee, P. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD dissertation, University of Texas at Austin, 1993.

78. Tanenhaus, Michael and Others. "Integration of Visual and Linguistic Information in Spoken Language Comprehension," *Science*, *268*:1632–1634 (June 1995).

79. Tokuda, K. and others. "A Very Low Bit Rate Speech Coder Using HMM-Based Speech Recognition/Synthesis Techniques." *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing*. 609–612. May 1998.

80. Tokuda, K, T. Kobayashi and S. Imai. "Speech Parameter Generation from HMM Using Dynamic Features." *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing1*. 660–663. May 1995.

81. Van Loan, C. "Computing the CS and the Generalized Singular Value Decompositions," *Numer. Math.*, *46*:479–491 (1985).

82. Van Son, N. and others. "Viseme Classifications of Dutch Consonants and Vowels," *Journal of the Acoustical Society of America*, *96*(3):1341–1355 (September 1994).

83. Watson, Charles S. and Others. "Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition," *Journal of the Acoustical Society of America*, *100*(2):1153–1162 (August 1996).

84. Watson, S. and Others. "An Advanced Morphing Algorithm for Interpolating Phoneme Images to Simulate Speech." Unpublished Technical Report.

85. Williams, J.J. and Others. "Frame rate and viseme analysis for multimedia applications." *1997 IEEE First Workshop on Multimedia Signal Processing*. June 1997.

86. Wu, L. and others. "Multimodal Integration-A Statistical View," *IEEE Transactions on Multimedia*, *1*(4):334–341 (December 1999).

87. Young, S.J., J.J. Odell and P.C. Woodland. "Tree-Based State Tying for High Accuracy Acoustic Modeling." *Proceedings of the ARPA Workshop on Human Language Technology*. 307–312. March 1994.

88. Young, S. and others. *The HTK Book* (HTK Version 2.2 Edition). Entropic Ltd., 1999.

89. Yuhas, B., M. Goldstein Jr. and T. Sejnowski. "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Communications Magazine*, *ASSP-34*:65–71 (November 1997).

90. Zhong, J, W. Chou and E. Petajan. "Acoustic Driven Viseme Identification for Face Animation." *Proceedings of the First Signal Processing Society Workshop on Multimedia Signal Processing*. 7–12. June 1997.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED
20-03-2001	Dissertation	Jan 1997 - Mar 2001

4. TITLE AND SUBTITLE

Hidden Markov Models for Visual Speech Synthesis in Limited Data Environments

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Harold Allan Arb

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
AFIT/ENG
Building 640
2950 P Street
Wright-Patterson AFB, OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/DS/ENG/01-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/HECA
Attn: Dr. Ray Slyh
2255 H Street
Wright-Patterson AFB, OH 45433-7001
DSN 758-9248

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This work presents a new approach for estimating control points (facial locations that control movement) to allow the artificial generation of video with apparent mouth movement (visual speech) time-synced with recorded audio. First, Hidden Markov Models (HMMs) are estimated for each visual speech category (viseme) present in stored video data, where a category is defined as the mouth movement corresponding to a given sound and where the visemes are further categorized as trisemes (a viseme in the context of previous and following visemes). Next, a decision tree is used to cluster and relate states in the HMMs that are similar in a contextual and statistical sense. The tree is also used to estimate HMMs that generate sequences of visual speech control points for trisemes not occurring in the stored data. An experiment is described that evaluates the effect of several algorithm variables, and a statistical analysis is presented that establishes appropriate levels for each variable by minimizing the error between the desired and estimated control points. Also, comparisons of mouth shapes generated from the artificial control points and the true control points indicate that the process provides accurate estimates for most trisemes tested in this work. The research presented here thus establishes a useful method for synthesizing realistic audio-synchronized video facial features.

15. SUBJECT TERMS

Visual Speech Synthesis, Hidden Markov Models, Synthetic Video, Image Motion, Audiovisual Databases

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Steven C. Gustafson
U	U	U	UU	237	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636 x4598