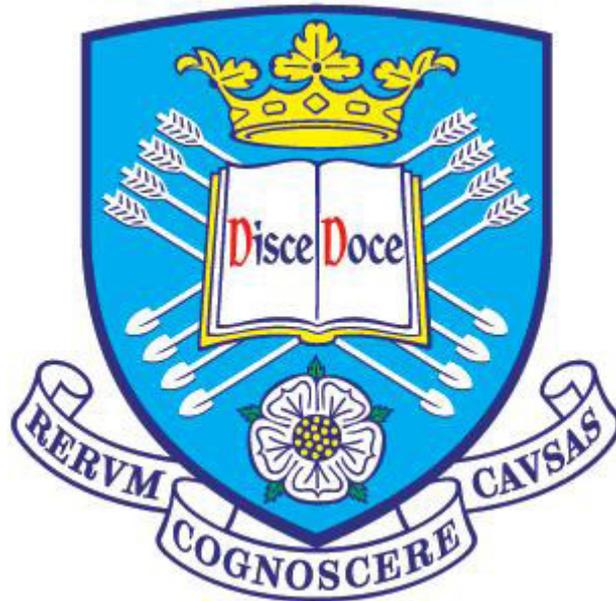


# Visual Speech in Technology-Enhanced Learning



Priya Dey

Department of Computer Science

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

August 2012

## Abstract

This thesis investigates the use of synthetic talking heads, with lip, tongue and face movements synchronized with synthesized or natural speech, in technology-enhanced learning. This work applies talking heads in a speech tutoring application for teaching English as a second language. Previous studies have shown that speech perception is aided by visual information, but more research is needed to determine the effectiveness of visualization of articulators in pronunciation training. This thesis explores whether or not visual speech technology can give an improvement in learning pronunciation.

This thesis investigates techniques for audiovisual speech synthesis, using both viseme-based and data-driven approaches to implement multiple talking heads. Intelligibility studies found the audiovisual heads to be more intelligible than audio alone, and the data-driven head was found to be more intelligible than the viseme-driven implementation.

The talking heads are applied in a pronunciation-training application, which is evaluated by second-language learners to investigate the benefit of visual speech in technology-enhanced learning. User trials explored the efficacy of the software in demonstrating the /b/-/p/ contrast in English. The results indicate that learners showed an improvement in listening and pronunciation after using the software, while the benefit of visualization compared to auditory training alone varied between individuals. User evaluations found that the talking heads were perceived to be helpful in learning pronunciation, and the positive feedback on the tutoring system suggests that the use of talking heads in technology-enhanced learning could be useful in addition to traditional methods.

## Acknowledgements

I would like to thank my supervisor, Steve Maddock for his guidance; Rod Nicolson and Olivier Pascalis for discussions on research plans; Roger Moore and Phil Green for their advice in panel meetings; Pierre Badin, Frederic Elisei, Gerard Bailly and Christophe Savariaux for their work on data acquisition and articulatory modelling in a collaboration at the Département Parole et Cognition, GIPSA-Lab, Grenoble; Oscar Martinez Lazalde for help in visual speech work; James Carmichael for collaboration on user tests; Richard Simpson and the tutors and students from the English Language Teaching Centre for their cooperation and insights into second language learning; Anne Turner for arranging volunteer sessions and helping with data analysis; Tim Harvey for help with data analysis; Joy Stackhouse and the Human Communication Sciences Department for sharing their knowledge in the areas of speech difficulties and intervention; Sarah Creer for advice on speech synthesis, Robin Hofe for providing information on tongue modelling; Rebecca Palmer for information on speech therapy software; Nicolas Martin from the Sheffield School of Clinical Dentistry; and all the volunteers who participated in experiments.

This research has been funded by the Economic and Social Research Council and the Engineering and Physical Sciences Research Council. An International Travel Grant for a research visit to Grenoble was funded by the Royal Academy of Engineering. I am grateful to Loquendo for their support with the use of their software.

## Contents

1	Introduction .....	12
1.1	Contributions .....	14
1.2	Thesis structure.....	15
2	Facial Animation, Visual Speech and Speech Tutoring .....	17
2.1	Speech Production .....	17
2.2	Facial Animation.....	19
2.3	Visual Speech Animation .....	22
2.3.1	Viseme-driven approaches .....	23
2.3.2	Data-driven approaches.....	25
2.3.3	Expression in Speech Animation .....	27
2.4	Talking Heads in Speech Tutoring .....	29
2.4.1	Existing Speech Tutoring Software .....	32
2.5	Summary .....	36
3	Talking Heads in Speech Tutoring.....	38
3.1	Teaching English as a second language for adults.....	38
3.2	Design of Talking Head.....	40
3.3	Graphical User Interface for Speech Tutoring Application .....	41
4	Development of Viseme-Driven Talking Head .....	44
4.1	Implementation of Viseme-Driven Talking Head (THVN) .....	44
4.2	Internal articulator models.....	45
4.2.1	MRI Tongue Contours.....	47
4.3	Text-To-Speech Synthesis.....	51
4.3.1	Festival Speech Synthesis System .....	51
4.3.2	Microsoft Speech Application Programming Interface .....	51
4.3.3	HTS Hidden Markov Model-based Speech Synthesis .....	51
4.3.4	Loquendo TTS .....	52
4.4	Text to Visual Speech .....	52
4.5	Interpolation .....	53
4.6	Principal Component Analysis.....	55
4.7	Coarticulation .....	57
4.7.1	Tuning visual speech model parameters by observation of video.....	58
4.8	Synchronisation between audio and video.....	60
4.9	Expression modelling .....	61
4.9.1	Photo-based viseme-driven talking head (THVP) .....	61
5	Development of Data-Driven Talking Head .....	63
5.1	Data collection.....	63
5.1.1	2D MRI.....	63

5.1.2	Video Capture on Three Cameras .....	65
5.1.3	EMA corpus .....	66
5.1.4	3D MRI corpus .....	68
5.2	Internal Articulatory Modelling .....	68
5.3	External Articulatory Modelling .....	69
6	Evaluation of Quality of Visual Speech .....	73
6.1	Evaluation Approaches .....	73
6.2	Intelligibility Test 1 (THVN): Modified Rhyme Test .....	74
6.2.1	Results of Intelligibility Experiment 1 .....	75
6.3	Subjective test for naturalness (THVN) .....	79
6.4	Subjective Evaluations of Naturalness of Talking Heads THVN and THD .....	81
6.4.1	Online Survey 1: Pilot Evaluation of Naturalness .....	81
6.4.2	Online Survey 2: Further Evaluation of Naturalness .....	84
6.5	Intelligibility Test 2: Modified Rhyme Test on 2 Talking Heads (THVP and THD) .....	86
6.5.1	Results of Intelligibility Test 2 .....	86
6.6	Conclusions from Intelligibility and Naturalness Evaluations .....	92
7	Evaluation of Speech Tutoring Application .....	94
7.1	Choice of Case Study .....	94
7.2	Experiment Design .....	99
7.3	Tutoring Study 1: Evaluation of viseme-driven non-photo-based head .....	100
7.3.1	Study 1 Listening Results .....	102
7.3.2	Study 1 Speaking Results (/b/ and /p/) .....	104
7.3.3	Study 1 Speaking (/b/ scores) .....	105
7.3.4	Speaking (/p/ scores) .....	105
7.3.5	Speaking and Listening Tests combined .....	106
7.3.6	Study 1 User Feedback .....	107
7.4	Tutoring Study 2: Evaluation of viseme-driven photo-based head (THVP) .....	109
7.4.1	Study 2 Experimental Design .....	110
7.4.2	Study 2 Listening Test Results (/b/ and /p/) .....	111
7.4.3	Study 2 Listening Test /b/ sounds .....	112
7.4.4	Study 2 Listening Test /p/ sounds .....	112
7.4.5	Study 2 Speaking (/b/ and /p/ combined) .....	112
7.4.6	Study 2 Speaking (/b/) .....	113
7.4.7	Study 2 Speaking (/p/) .....	113
7.5	Tutoring Study 3: Evaluation of data-driven head (THD) in a tutoring system ..	115
7.5.1	Study 3 Participant Information .....	117
7.5.2	Study 3 Results of Listening Test (/b/ and /p/) .....	117
7.5.3	Study 3 Lowest Listening Scores .....	119
7.5.4	Study 3 Listening (/b/ sounds) .....	120

7.5.5 Study 3 Listening (/p/ sounds).....	121
7.5.6 Study 3 Speaking Results (/b/ and /p/).....	123
7.5.7 Study 3 Speaking Test /b/ sounds .....	126
7.5.8 Study 3 Speaking Test (/p/).....	126
7.5.9 Study 3 User Feedback.....	128
7.5.10 Study 3: Summary of Results.....	131
8 Conclusions and Future Work.....	134
8.1 Future Work .....	139
Appendix A: International Phonetic Alphabet.....	144
Appendix B: Speech Tutoring Application Screenshots .....	145
Appendix C: Words for Video Corpus.....	150
C.1: Phrases for tutoring application.....	150
C.2: MOCHA -TIMIT subset.....	151
Appendix D: Stimulus Words of MRT .....	152
D.1: MRT words used in Intelligibility Test 1: Audio Alone.....	153
D.2: MRT words used in Intelligibility Test 1: Synthetic Talking Head (THVN) .....	154
D.3: MRT words used in Intelligibility Test 1: Natural Video .....	155
D.4: MRT words used in Intelligibility Test 2: Audio Alone.....	156
D.5: MRT words used in Intelligibility Test 2: Viseme-Driven Synthetic Talking Head (THVP).....	156
D.6: MRT words used in Intelligibility Test 2: Data-Driven Synthetic Talking Head (THD) .....	157
D.7: MRT words used in Intelligibility Test 2: Natural Video .....	157
Appendix E: Tutoring Study Stimuli.....	158
E.1: Study 1 Listening Pre/Post Test Stimuli .....	158
E.2: Study 1 Pre/Post Test Speaking Words:.....	158
E.3: Study 1 Pre/Post Test Speaking Phrases .....	159
E.4: Study 1 User Questionnaire .....	160
E.5: Study 2 and 3 Pre/Post Test Listening Words.....	162
E.6: Study 2 and 3 Pre/Post Test Speaking Words .....	162
E.7: Study 2 and 3 Pre/Post Test Speaking Phrases .....	163
E.8: Study 2: User Feedback after using Pronunciation Software Version I (internal and external visualization) .....	164
E.9: Study 2 and Study 3: User Feedback after using Pronunciation Software Version X (external visualization).....	166
E.10: Study 3: User Feedback after using Pronunciation Software Version A (audio alone).....	168
References.....	171

## Figures

Figure 2.1: Anatomy of vocal tract, reproduced under Creative Commons Licence (Flemming 2012).....	18
Figure 2.2: Cohen-Massaro Dominance Functions for the word “stew”, and corresponding lip protusion values over time, reproduced with permission (Massaro 1998).....	25
Figure 3.1: Screenshot of Speech Tutoring Application used in Tutoring Study 1 .....	42
Figure 3.2: Diagram of Speech Tutoring Application used in Tutoring Study 1.....	43
Figure 4.1: Diagram of Speech Tutoring Application .....	45
Figure 4.2: Facegen models for visemes (with names used by Facegen) .....	45
Figure 4.3: Tongue Visemes and Teeth Model (Lazalde et al. 2008) .....	46
Figure 4.4: Mid-sagittal MRI scan for articulation of vowel /e/.....	47
Figure 4.5: Mid-sagittal MRI contours for vowel /e/.....	48
Figure 4.6: Alignment of head meshes with mid-sagittal MRI contours for vowel /e/ .....	48
Figure 4.7: Comparing original tongue mesh with mid-sagittal MRI contour.....	49
Figure 4.8: Modelling tongue mesh based on mid-sagittal MRI contour .....	49
Figure 4.9: Viseme-driven speech synchronised animation.....	53
Figure 4.10: Spline Interpolation, adapted from (Dunlop 2009) .....	54
Figure 4.11: Mesh for Talking Head (THVN).....	57
Figure 4.12: Cohen-Massaro Dominance Functions for the word “stew”, and corresponding lip protusion parameter values, reproduced with permission (Massaro 1998)	58
Figure 4.13: Animation and Video frames for “stew” .....	59
Figure 4.14: Trajectory for sentence “Hello, how are you?” .....	61
Figure 4.12: Facegen meshes for facial expressions.....	61
Figure 4.13: Photos used to create photo-based viseme-driven head.....	62
Figure 4.14: Photo-based viseme-driven head (THVP) .....	62
Figure 4.15: Internal view of photo-based viseme-driven head (THVP) .....	62
Figure 5.1: MRI scan for articulation of vowel /e/ .....	64
Figure 5.2: Video capture with small set of 40 facial markers.....	66
Figure 5.3: Video capture with full set of 168 facial markers.....	66
Figure 5.4: EMA recording .....	67
Figure 5.5: MRI contour tracing .....	69
Figure 5.6: MRI contours .....	69
Figure 5.7: Facial geometric mesh.....	71
Figure 5.8: Articulatory Model .....	71
Figure 5.9: Data-driven talking head.....	72
Figure 6.1: Intelligibility scores. The error bars denote the standard deviation.....	76
Figure 6.2: Confusion Matrix for Synthetic Talking Head.....	78

Figure 6.3: Confusion Matrix for Natural Head.....	78
Figure 6.4: Difference between Synthetic Head and Natural Head .....	79
Figure 6.5: Naturalness Ratings .....	80
Figure 6.6: Naturalness Ratings in Online Survey 1 .....	83
Figure 6.7: Naturalness Ratings in Online Survey 2 .....	84
Figure 6.8: Intelligibility scores. The error bars denote the standard deviation. ....	88
Figure 6.9: Confusion Matrix for audio alone .....	89
Figure 6.10: Confusion Matrix for Viseme-driven Talking Head.....	89
Figure 6.11: Confusion Matrix for Data-driven Talking Head .....	90
Figure 6.12: Confusion Matrix for Natural video.....	90
Figure 6.13: Difference between Viseme-driven Head and Natural Head .....	91
Figure 6.14: Difference between data-driven Head and Natural Head .....	91
Figure 7.1: Animation frames of data-driven head for the word “back” .....	98
Figure 7.2: Animation frames of data-driven head for the word “pack” .....	99
Figure 7.3: Screenshot of Speech Tutoring Application used in Tutoring Study 1 .....	102
Figure 7.4: Study 1 Listening Scores (/b/ and /p/).....	103
Figure 7.5: Study 1 Speaking Scores (/b/ and /p/ combined).....	105
Figure 7.6: Study 1 Speaking Scores (/b/) .....	105
Figure 7.7: Study 1 Speaking Scores (/p/) .....	106
Figure 7.8: Study 1 Speaking and Listening Scores .....	107
Figure 7.9: Screenshot of Speech Tutoring Application used in Tutoring Study 2 .....	109
Figure 7.10: Screenshot of Speech Tutoring Application used in Tutoring Study 2 .....	110
Figure 7.11: Screenshot of Speech Tutoring Application used in Tutoring Study 3 .....	115
Figure 7.12: Study 3 Listening Test /b/ and /p/ scores.....	119
Figure 7.13: Study 3 Listening Test /b/ scores.....	121
Figure 7.14: Study 3 Listening Test /p/ scores.....	122
Figure 7.15: Study 3 Speaking Test (/b/ and /p/ scores).....	124
Figure 7.16: Study 3 Speaking Test Scores (/p/ sounds).....	128
Figure A.1: International Phonetic Alphabet Chart, reproduced with permission (International Phonetic Association 2005).....	144
Figure B.1: Introduction.....	145
Figure B.2: Listening Practice: Sounds .....	145
Figure B.3: Listening Practice: Words.....	146
Figure B.4: Listening Practice: Phrases .....	146
Figure B.5: Listening Test: Sounds.....	147
Figure B.6: Listening Test: Words.....	147
Figure B.7: Speaking Practice: Sounds .....	148
Figure B.8: Speaking Practice: Words .....	148
Figure B.9: Speaking Practice: Phrases .....	149

Figure B.10: End of Lesson ..... 149

## Tables

Table 4.1: Reclassified Visemes .....	50
Table 4.2: Example words used in tuning visual speech .....	59
Table 6.1: Intelligibility Test 1: Mean % words correctly identified .....	76
Table 6.2: Intelligibility Test 1: Visual contribution to intelligibility .....	76
Table 6.3: Results of Online Naturalness Survey 1 .....	83
Table 6.4: Results of Online Naturalness Survey 2 .....	85
Table 6.5: Intelligibility Test 2: Mean % words correctly identified .....	87
Table 6.6: Intelligibility Test 2: Visual contribution to intelligibility .....	87
Table 7.1: Study 1 Listening Scores (/b/ and /p/) .....	102
Table 7.2: Study 1 Listening Scores (/b/) .....	103
Table 7.3: Study 1 Listening Scores (/p/) .....	104
Table 7.4: Study 1 Speaking Scores (/b/ and /p/) .....	104
Table 7.5: Study 1 Speaking and Listening Scores (/b/ and /p/) .....	106
Table 7.6: Study 1 User Feedback .....	108
Table 7.7: Study 2 Listening Scores (/b/ and /p/) .....	112
Table 7.8: Study 2 Listening Scores (/b/) .....	112
Table 7.9: Study 2 Listening Scores (/p/) .....	112
Table 7.10: Study 2 Speaking Scores (/b/ and /p/) .....	112
Table 7.11: Study 2 Speaking Scores (/b/) .....	113
Table 7.12: Study 2 Speaking Scores (/p/) .....	113
Table 7.13: Study 2 User Feedback .....	114
Table 7.14: Study 3 Participant Information .....	117
Table 7.15: Study 3 Listening Scores (/b/ and /p/) .....	118
Table 7.16: Study 3 Mean Listening Scores (/b/ and /p/) .....	118
Table 7.17: Study 3 Lowest Listening Scores (/b/ and /p/) .....	119
Table 7.18: Study 3 Mean Listening Scores (/b/) .....	120
Table 7.19: Study 3 Listening Scores (/b/) .....	120
Table 7.20: Study 3 Listening Scores (/p/) .....	121
Table 7.21: Study 3 Mean Listening Scores (/p/) .....	122
Table 7.22: Study 3 Mean Speaking Scores (/b/ and /p/) .....	123
Table 7.23: Study 3 Speaking Scores (/b/ and /p/) .....	124
Table 7.24: Study 3 Lowest-scoring users' Speaking Scores (/b/ and /p/) .....	125
Table 7.25: Study 3 Mean Speaking Scores (/b/) .....	126
Table 7.26: Study 3 Speaking Scores (/b/) .....	126
Table 7.27: Study 3 Mean Speaking Scores (/p/) .....	127
Table 7.28: Study 3 Speaking Scores (/p/) .....	127
Table 7.29: Study 3 User Feedback .....	129

Table 7.30: Study 3 User Feedback (combined groups)..... 130

# 1 Introduction

The importance of human faces in communication has led to considerable interest in computer facial animation and visual speech synthesis. Human speech perception is aided by visual information, such as the movement of the lips, which aids intelligibility (Sumby et al. 1954; Summerfield 1987). Speech comprehension is also enhanced by facial expressions, which convey meaning and thus support communication (Massaro 1998).

This thesis is concerned with *visual speech synthesis*, computer-generated facial animation synchronized with acoustic speech, and *talking heads*, physiological models with audio-visual speech. This study also encompasses animated *agents* which interact with a user to emulate face-to-face communication with a human assistant. Talking heads can augment the intelligibility of speech, and when combined with animated agents can convey emotions and offer more natural interaction, and these advantages can make them valuable in technology-enhanced learning applications. Animated agents are employed in many software applications, while talking heads with accurate visible articulation are attracting increasing interest for use in pronunciation training (Hazan 2008). For example, the talking head known as “Baldi” is an existing tutor for speech production (Massaro et al. 2008; Massaro 2012), and state-of-the-art technological advances are being utilised in the “ARTUR” articulation tutor (Engwall 2008). Improved modelling of internal articulatory organs has been an important recent development in talking heads for pronunciation training (Badin et al. 2008).

This project applies talking head technology in a speech tutoring application: teaching English as a second language for adults. The aim was to create a pronunciation assistant, to complement traditional methods and to assist the work of a human language tutor. The studies investigate the benefits of visual speech technology in language learning.

The main research question addressed by this thesis is whether or not visual speech technology gives an improvement in learning pronunciation. From this arise subsidiary issues, concerning how the quality of the technology affects the

improvement, and whether it makes a difference to learning. This study seeks to determine the benefit of visual speech in technology-enhanced learning, and the most effective implementations of facial animation for modelling visual speech.

Two approaches were chosen for the development of talking heads: viseme-driven and data-driven speech animation. Previous studies of talking heads in speech tutoring have used viseme-driven techniques (Massaro et al. 2008), which require a smaller amount of data to create key poses for articulators. A disadvantage of data-driven techniques is that they require a large corpus of captured data in order to produce realistic results, and for internal visualization, a corpus of internal articulatory data is required, but the benefit of a data-driven approach based on a real speaker's data is that it can create a more accurate model of articulator movement. Therefore, after the acquisition of a suitable corpus of data, a data-driven head was also created, and the resulting talking heads were compared in intelligibility tests.

The following experiment conditions are used within the thesis:

- Audio alone - this is used in the intelligibility tests (Chapter 5) and the speech tutoring trials (Chapters 6), in comparison with the talking heads that have been implemented.
- Viseme-driven, non-photo-based talking head (THVN) - this head was created using a synthetic mesh in Facegen (Singular Inversions 2008). This head is used in the intelligibility and naturalness tests (Chapter 5) and the first speech tutoring trial (Chapter 6).
- Viseme-driven, non-photo-based talking head, including expression - this head was created using a synthetic mesh in Facegen (Singular Inversions 2008). Facial expressions including eye and head movements were added. This head is used in the web-based naturalness test (Chapter 5).
- Viseme-driven, photo-based talking head (THVP) - this head was created using photographs of the real speaker. This head is used in the second intelligibility test (Chapter 5) and the second speech tutoring trial (Chapters 6).

- Data-driven talking head (THD) – this head was created at GIPSA-Lab, Grenoble, using a corpus of data from a real speaker. This head is used in the second intelligibility test (Chapter 5), the web-based naturalness test (Chapter 5), and the final speech tutoring trial (Chapter 6).
- Real video - video recordings of the real speaker. This is used in the intelligibility and naturalness tests (Chapter 5) in comparison with the talking heads that have been implemented.

Each talking head was applied in a pronunciation training system, and user studies investigated the use of the software in demonstrating the /b-/p/ contrast in English. Few previous studies have explored this contrast, which was chosen as a case study after consultation with English language tutors who revealed that one of their largest groups of students was native Arabic speakers, and the most common difficulty for this group was /b-/p/, because this contrast did not exist in their native language. A difficulty with using /b/ and /p/ as a case study for pronunciation training is that the difference between /b/ and /p/ is produced by voicing, which is difficult to show in a talking head, and although /b/ and /p/ do have some visible differences, as shown by (Lazalde 2010), this difference may not always be salient enough to aid discrimination. The experiments presented in this thesis are widening the range of phonemic contrasts which have been studied, with a less visually salient contrast than those in previous research. The user trials determined the impact of the software in learning perception and pronunciation, and its effectiveness as a teaching tool was evaluated.

## 1.1 Contributions

The contributions of this thesis are as follows:

- A software application using a talking head for teaching pronunciation, which is the first of its kind for teaching British English. This software was used in experiments to investigate the use of talking heads in speech tutoring. This work was shortlisted in the UK ICT Pioneers Competition 2011 (EPSRC 2011).

- A novel corpus, comprising MRI, EMA and video data, which is the first of its kind for a British English female speaker. The corpus was acquired at the "Département Parole et Cognition", GIPSA-Lab, Grenoble, during a research visit funded by an International Travel Grant by the Royal Academy of Engineering. This corpus was used to create a data-driven head in a collaborative project at GIPSA-Lab. The MRI data was used to improve the articulatory modelling of the viseme-driven head (Knight 2011).
- New studies evaluating the visual speech in two different talking heads; one viseme-driven and one data-driven. Intelligibility tests showed that the audiovisual heads were more intelligible than audio alone (Dey et al. 2010a). The data-driven head was found to be more intelligible than the viseme-driven head.
- Original studies evaluating the use of talking heads in learning pronunciation of British English as a second language. The results indicate that learners showed an improvement in listening and pronunciation after using the software, while the benefit of visualization compared to auditory training alone varied between individuals. User evaluations found that the talking heads were perceived to be helpful in learning pronunciation (Dey et al. 2010b).

## 1.2 Thesis structure

The following chapters review the techniques used in producing talking heads, and applications of visual speech. Chapter 2 begins with an overview of the field of facial animation. It reviews existing literature and describes techniques for producing talking heads and visual speech. It then reviews applications of talking head technology, including second language learning, and the teaching of pronunciation. Chapter 3 introduces how this project applies talking head technology in a speech tutoring application. Chapters 4 and 5 describe the data capture and implementation of two different talking heads. Chapter 6 describes the evaluation of the quality of visual speech of the talking heads. Chapter 7 discusses the evaluation of the speech tutoring application, and details the studies

in teaching English as a second language for adults. Finally, conclusions and future work are discussed in Chapter 8.

## 2 Facial Animation, Visual Speech and Speech Tutoring

Visual Speech unites the fields of speech synthesis and computer facial animation. While research in computer graphics involves modelling and animation of the face (Parke et al. 1996; Magnenat-Thalmann et al. 2004), research in speech synthesis focuses on production of acoustically realistic speech. Visual Speech brings the two together in the synthesis of graphically and acoustically realistic speech, complete with synchronised lip, tongue and jaw movements and the modelling of expression (Massaro 2012). The connection between audio and visual research is exemplified by the recent conference Interspeech 2008 which had two special sections, one on talking heads (Engwall 2008; Fagel et al. 2008) and one on visible speech synthesis (Theobald et al. 2008). The modelling of expression also incorporates research in Artificial Intelligence; thus conferences on animated agents also include work on talking heads (Martin et al. 2007).

The following sections discuss the physical mechanisms of speech production, followed by the main techniques for producing facial animation, visual speech, and expression in speech animation.

### 2.1 Speech Production

The process of speech production has three phases: *respiration*, in which the lungs force air through the vocal tract and out through the oral and nasal cavities (Figure 2.1); *phonation* - the vibration of the vocal cords; and *articulation* – the shaping of the upper vocal tract to generate speech sounds.

Speech sounds can be categorised according to the articulatory positions required to produce the sound. A standard representation for transcribing all possible speech sounds has been established by the International Phonetic Association (IPA) (Figure A.1) (International Phonetic Association 2005). Consonants are defined by the place of articulation, its manner and phonation. For example, the place of articulation of the consonant /p/ is *bilabial*, i.e. it is produced by

constriction of airflow between the lips; its manner is *plosive*, i.e. the sound occurs after a blockage of air flow is released; and its phonation is that it is *voiced*. A voiced consonant is produced when the vocal cords are vibrating, whereas an unvoiced consonant is one in which the vocal cords are not vibrating. A table of all consonants, with their place of articulation, manner and phonation is shown in the IPA chart in Figure A.1. Vowels are defined by the location of the tongue within the oral cavity, and the rounding of the lips is also a distinguishing factor. The tongue position can range from the *front* of the mouth, e.g. /i/, to the *back*, e.g. /u/, and the tongue height can range from *close*, e.g. /i/, to *open*, e.g. /a/. A diagram of the positions of all the vowels is shown in Figure A.1.

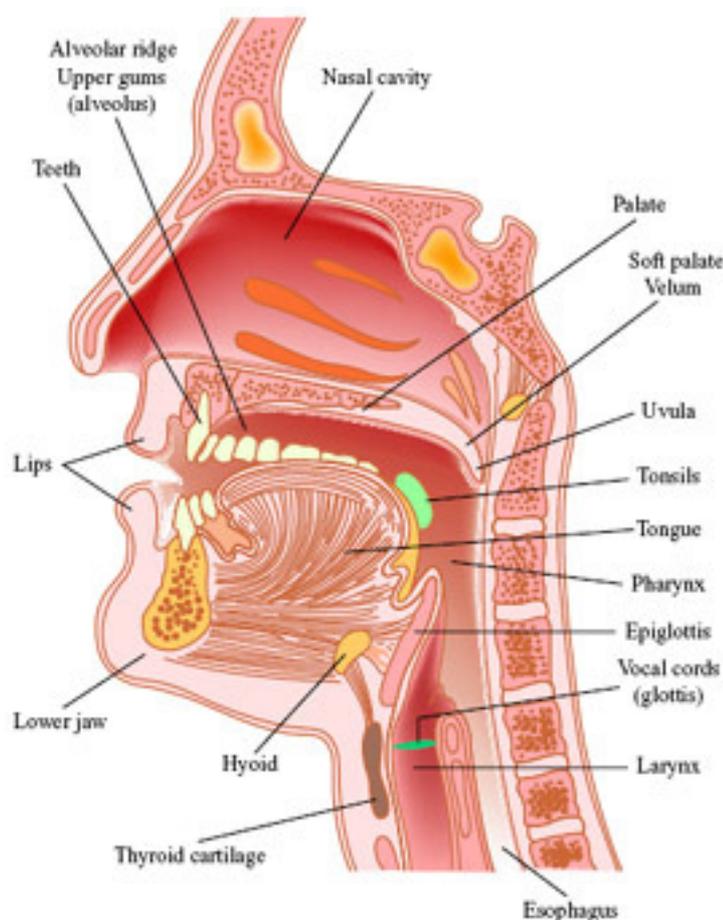


Figure 2.1: Anatomy of vocal tract, reproduced under Creative Commons Licence (Flemming 2012)

## 2.2 Facial Animation

Facial animation involves controlling a face model using geometric manipulations or image manipulations, although many approaches combine several techniques. The techniques can be classified as those involving manipulations of 2D images or 3D models. The 2D approaches include image morphing, which allows transitional images to be generated between a pair of target images (Ezzat et al. 1997). Another image-manipulation technique for 2D animation is video-rewrite, where video footage of an actor is segmented corresponding to phonetic units, which are then concatenated to create animations of a speaker (Bregler et al. 1997). 2D methods have achieved more videorealistic results than 3D methods (Liu et al. 2008), and some 2D animations are perceptually indistinguishable from real video, yet still not as intelligible for lip-reading as real recordings (Geiger et al. 2003). 2D methods can be unrealistic for head movements, and the viewpoint is limited to that of the target images, whereas 3D geometric methods are viewpoint-independent.

3D techniques involving geometric manipulation include interpolation, parameterization and muscle models. Interpolation-based techniques involve modelling portions of the face mesh to approximate expressions, and then blending these different morph targets. This method is fast and offers high fidelity of expressions, but involves intensive manual labour, and is specific to each character. Parameterization was used by Parke, whose facial mesh of 3D points was controlled by a set of conformation and expression parameters (Parke et al. 1996). An advantage over interpolation is that parameters can easily be combined for a wider range of facial expressions. Physically-based muscle models simulate the physical and anatomical characteristics of bones, tissues, and skin (Terzopoulos et al. 1990; Lee et al. 1995). Such methods can be very powerful for creating realism, but the complexity of facial structures make them computationally expensive, and difficult to create. This complexity can be avoided by using mesh deformation to simulate muscle action (Magnenat-Thalmann et al. 1988). Pseudo muscle-based systems often use Ekman and Friesen's Facial Action Coding System (Ekman et al. 1978), which defines 64 basic facial Action Units to represent facial movements caused by muscles.

Waters proposed a vector muscle model in which a muscle was modelled by a linear deformation field affecting the surrounding skin (Waters 1987).

Another technique for 3D animation uses motion capture to map recorded movement onto a character (Zhao et al. 2010). Feature points on an actor's face are recorded, often using reflectance markers placed on the actor, which are tracked by cameras. Ma et al (Ma et al. 2008) used a real-time 3D scanning system to record training data of the high-resolution geometry and appearance of an actor performing a small set of predetermined facial expressions. A set of motion capture markers was placed on the face to track large scale deformations. The large scale deformations were mapped to the finer scale deformations in the form of deformation-driven polynomial displacement maps, encoding variations in medium-scale and fine-scale displacements. For synthesis, the polynomial displacement maps were driven by new motion capture data from a sparse set of motion capture markers. The technique produced accurate reconstructions over most of the face, but the tracking of the contours of mouth and eyes was insufficient for reconstruction of the detailed motions near the edges of the lips and eyelids.

A key issue in motion capture is the accuracy of tracked data, which may include noise from vibration. The motion data is filtered before it is transformed to drive a computer model of a character (Deng et al. 2007). A difficulty which is especially pertinent in facial animation is that feature points are not always visible, for example, the corners of the lips may be occluded during speech. This is a problem for optical tracking systems, which require tracked points to be within the line of sight. Magnetic tracking systems do not have the line-of-sight problem, but are usually less accurate than optical systems, although non-invasive electromagnetic systems are now available which are designed for tracking speech-related facial movements (Northern Digital Inc. 2012).

Markerless vision-based approaches, such as Active Appearance Models (Tresadern et al. 2010), use estimation algorithms to track occluded points. These can be limited in resolution (Poppe 2007), but this issue has decreased as camera technology has improved, and recently markerless facial performance capture has achieved realistic pore-level geometric details, while addressing tracking

problems caused by very fast motion or occlusions (Beeler et al. 2011). Beeler et al. identify anchor frames in a sequence, which are similar to a manually chosen reference frame. Due to the similarity, the image tracker can compute the flow from the reference to each anchor independently and with high accuracy. For example, in a sequence of lip movements where the upper lip becomes occluded by the lower lip, this method is able to track the upper lip backwards from a later anchor frame to the occluded frame, automatically restoring tracking after the occlusion. A high-quality 3D reconstruction technique gave the mesh visually realistic pore-level geometric detail. Another approach achieving photorealistic detail combined optical flow and photogrammetry to reconstruct 3D motion from images (Borshukov et al. 2005). Five synchronized cameras captured the actor's performance, and optical flow was used to track each pixel's motion over time in each camera view. This data was combined with a scanned model of a neutral expression of the actor and photogrammetric reconstruction of the camera positions. Manual correction of optical flow errors was required. The Digital Emily Project (Alexander et al. 2010) used video-based motion analysis in a manually guided process. The animated character's 3D pose was manually set on several example frames, and a model-based optical-flow algorithm calculated the required character pose in the intermediate frames. An advantage over marker-based techniques is that this process is based on video of the entire face, which provides more information about the motion of the eyes and mouth than could be recorded by motion-capture markers. However, this process requires a large amount of manual work to adjust any misaligned poses and remove artifacts.

The problem of following a feature point across its location changes between poses, and mapping it to a corresponding vertex on a target model is a feature correspondence problem (Parke et al. 2008). Retargeting, or cross-mapping, of the geometry involves adaptation of the recorded source data to a target character, which need not have a direct resemblance to the recorded actor (Pighin et al. 2006).

The most popular method of producing facial animation is currently interpolation, because it quickly produces basic facial animations. Many other systems use parameterization, and researchers has extended Parke's parameterized approach to

include more features and functionality; for example Massaro's Baldi talking head was based on Parke's model (Massaro et al. 2005). The popularity of facial animation has led to standardisation of the parameters, in the MPEG-4 standard, which specifies feature points and facial animation parameters (Algirdas 2002) (Ostermann 2002).

## 2.3 Visual Speech Animation

Visual speech animation is the synthesis of realistic facial animations synchronized with acoustic speech. There are various techniques for producing automatic visual speech animation, which are introduced below and discussed in more detail in the subsequent sections.

Parke and Waters classify the approaches to automatic speech synchronisation by the form of input, which may be text-based, pre-recorded acoustic speech, or a combination of inputs (Parke et al. 1996). Waters and Levergood's text-driven approach extracted from input text a timed sequence of phonetic units and control parameters to drive the speech output (Waters et al. 1994). Lewis and Parke's speech-driven method adapted a common speech synthesis method, Linear Prediction Analysis, to provide simple phonetic recognition from recorded speech (Lewis et al. 1987). The recognized phonetic units were then associated with mouth positions to provide keyframes for animation using a parametric model of the human face.

An alternative classification is by Deng and Noh, who classify the approaches to visual speech by the method of output production, which may be viseme-driven or data-driven (Deng et al. 2007). The term *viseme* is defined by Fisher as a unit of speech in the visual domain (Fisher 1968). In viseme-driven speech animation, each key pose is associated with a viseme, i.e. the position of the lips, jaw and tongue when producing a particular sound. Examples of systems that use this approach are given in Section 2.3.1. Data-driven approaches do not require pre-designed key shapes, but use a pre-recorded facial motion database for synthesis using machine learning or concatenation of sample data. Examples of systems that use this approach are given in Section 2.3.2.

A key challenge in visual speech animation is that there is great variation in the realisation of visemes during the production of natural speech. This phenomenon is termed coarticulation, which is the influence of surrounding visemes upon the current viseme (Hardcastle et al. 1999). To account for coarticulation, current systems either explicitly take into account context when blending viseme keyframes, or use a longer unit such as the diphone, which starts at the centre of one phone and ends at the centre of the next, so transitions between phones are preserved.

### 2.3.1 Viseme-driven approaches

Viseme-driven approaches require the creation of key mouth shapes for each phonetic realisation, and then smoothing functions or coarticulation rules are used to synthesize new speech animations. There are difficulties in defining visemes, as there is asynchrony between the acoustic and the visual modalities of speech, where the onset of movement does not always correspond to the onset of the acoustic realisation of a phone. Also, allophones which sound similar can often appear different visually. There is no consensus as to which phones are grouped to form each viseme, and how many visemes to use. Attempts have been made to use machine learning approaches to identify visemes objectively (Hilder et al. 2010), but these have yet to yield a generic set of visemes. However, most viseme-based approaches assume a many-to-one relationship between phones and visemes, and use an approximate set of mouth shapes; for example, Tekalp and Ostermann used 14 visemes (Tekalp et al. 2000).

The most common approach to modelling of coarticulation is by Cohen and Massaro (Cohen et al. 1993), based on Lofqvist's gestural theory of speech production (Löfqvist 1990), using dominance and blending functions. Each dominance function represents the influence over time that a viseme has on a speech utterance. Typically the influence will be greatest at the centre of the viseme and will degrade with distance from the viseme centre. Dominance functions are blended together to generate a speech trajectory, in the same way as spline basis functions are blended together to generate a curve (Figure 2.2). The shape of each dominance function is different according to which viseme it

represents, and what aspect of the face is being controlled (for example, lip width or jaw rotation). This approach to computer-generated speech animation is used in the Baldi talking head (Massaro 2012).

The Cohen-Massaro coarticulation model was extended by Cosi et al by the addition of a temporal resistance function and a shape function for more general cases, such as fast or slow speaking rates (Cosi et al. 2002). This approach is used in the talking head LUCIA (Cosi et al. 2008). Le Goff and Benoit proposed a method to automatically extract the parameters for the dominance function from data measured from a real speaker (Le Goff et al. 1996). King and Parent extended the Cohen-Massaro model by using a curve to replace a single viseme target. Each facial model parameter has a curve that animates that parameter over time. A parameter curve is created by blending the viseme curves of the utterance (King et al. 2005). Bevacqua and Pelachaud proposed additional qualifier parameters to simulate expressivity in lip movements. Visemes for each emotion were derived from recorded speech motion data, and two qualifiers were added to modulate the expressivity of a lip movement; the *tension* qualifier was used to set the intensity of muscular strain, and the *articulation* qualifier controlled the degree to which a lip shape met its target apex (Bevacqua et al. 2004; Deng et al. 2007).

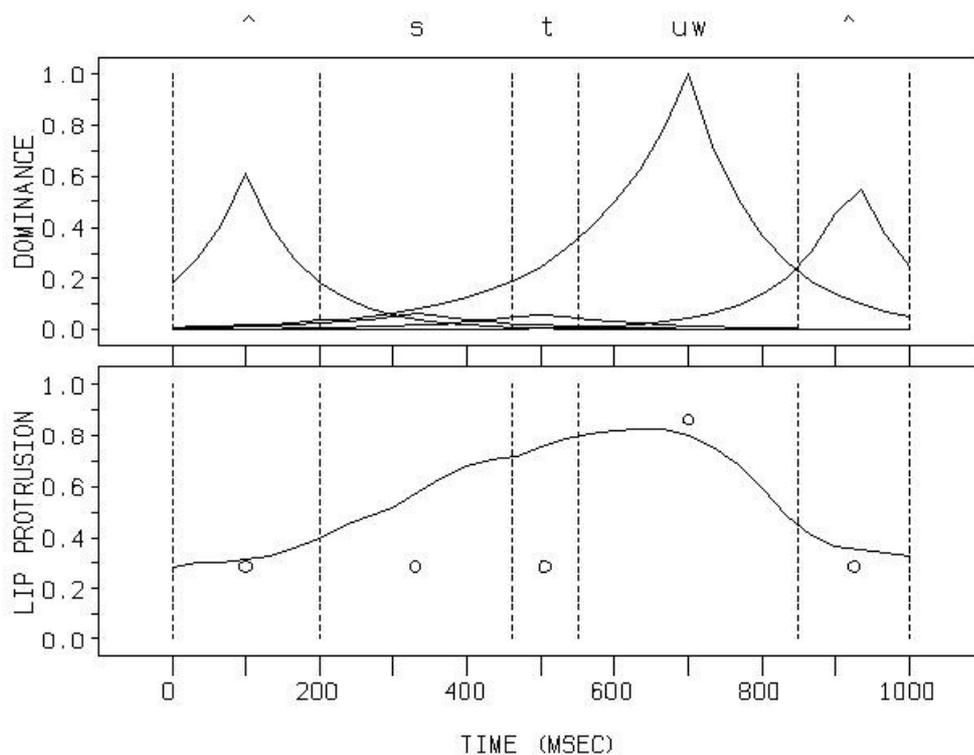


Figure 2.2: Cohen-Massaro Dominance Functions for the word “stew”, and corresponding lip protrusion values over time, reproduced with permission (Massaro 1998)

An alternative to dominance functions is a coarticulation model based on the optimization of a system with constraints (Edge 2004; Edge et al. 2004). The system generates trajectories which pass through appropriate visemes, applying a technique similar to the spacetime constraints method used for articulated body animation (Witkin et al. 1988). This approach was extended by Lalalde (Lalalde 2010), using motion-captured speech mapped to a 3D synthetic face model to derive facial animation parameters and create data for the coarticulated visemes in the constraint-based approach (Lalalde et al. 2010).

### 2.3.2 Data-driven approaches

Data-driven approaches use video or motion capture data captured from a real speaker, which is then used to drive a synthesized talking head. Feature points can be tracked using markers on the face or a markerless tracking system. A 3D surface mesh can be built by reconstructing the facial geometry from feature

points. A model of the speech movements can be built, for example, using Principal Component Analysis to parameterize the data (Bailly et al. 2009). The recorded motion capture sequences can be parameterized, and can be used to drive the talking head directly, to create animations corresponding to the original utterances, but this performance-driven animation approach does not scale to large volumes of speech. Data-driven approaches extract data from the recorded corpus, using concatenative or machine-learning approaches, to synthesize novel sequences.

#### 2.3.2.1 Concatenative speech animation

Data-driven, concatenative speech synthesis uses basis units that include context (e.g. diphones, triphones etc.) extracted from a pre-recorded corpus, instead of visemes. As the basis units already incorporate the variation of each viseme according to context and to some degree the dynamics of each viseme, no model of coarticulation is required. Speech is generated by selecting appropriate units from a database and blending the units together. This is similar to concatenative techniques in audio speech synthesis (where, for example, diphone units are concatenated at the spectrally-stable centre points of phones, and the waveforms are transformed so auditory discontinuities are minimised). For auditory speech synthesis a cost function is designed to maintain smoothness in the acoustic signal across concatenation boundaries. The best matching unit is considered the candidate that requires the least modification to form the join. For visual speech synthesis a cost function is designed to ensure a fluent and natural transition between adjacent visual speech gestures (Theobald 2007). The disadvantage to these models is that a large amount of captured data is required to produce natural results. For example, Huang et al used a corpus of 300 sentences, each lasting 5-10 seconds, recorded at 60 frames per second, to give 90000 video frames and a database of 9580 triphones (Huang et al. 2002). Longer units minimise the number of concatenations and preserve coarticulation effects to produce more natural results, but these require larger corpus sizes to ensure good coverage of transitions between the units.

#### 2.3.2.2 Speech animation using Machine Learning

Speech animations can be generated directly from audio by using hidden Markov models or neural nets to transform audio parameters into a stream of control parameters for a facial model. This method can handle voice context, rhythm, tempo, emotion and dynamics without complex approximation algorithms and the training database needs no phonetic units or visemes; the only data needed is the voice and the animation parameters. An example of this approach is the Johnnie Talker system (Takács et al. 2007). Another approach using machine learning techniques uses Gaussian processes to model audio and visual parameters in a shared space (Deena et al. 2010). The limitation is that phonetic labels, which are time-consuming to produce, are needed for both the training and test data.

Two-dimensional shape and appearance models have been used to create near-video-realistic synthetic talking faces (Theobald et al. 2004). Englebienne et al. used an Active Appearance Model to extract features from video frames, and a Hidden Markov Model to align phoneme labels to the audio stream of video sequences, and this information was used to label the corresponding video frames (Englebienne et al. 2007). Their model, trained on these labelled video frames, was able to generate new video-realistic sequences from unseen phoneme sequences. In a web-based test where 33 volunteers compared 12 pairs of video sequences, many of the sequences generated were indistinguishable from real video sequences. The limitations of this approach are that the dataset was from only a single speaker reading aloud, and applying these methods to different contexts, such as conversational speech, would be likely to result in very different results. Also, with no explicit model of coarticulation, effects such as anticipatory coarticulation were not captured as well as they could be if a model of phonetic context was added.

#### 2.3.3 Expression in Speech Animation

A complete facial animation system would include all facial expressions. A fully expressive automated character animation needs to display emotions, head movements such as nodding, and eye movements such as blinking, and eye gaze. Queiroz et al proposed a model for the automatic generation of expressive gaze by

examining eye behaviour in different affective states (Queiroz et al. 2008). Ekman (Ekman 1989) defined facial expressions of emotion (affect displays), and the following non-emotional expressions: Emblems, e.g. nodding in agreement; Manipulators, e.g. blinking; Signals, e.g. raised eyebrows signalling a question; Punctuators, movements at pauses, e.g. smiling; and Regulators, e.g. turning head to listener. These expression overlays were included in Ekman's procedure for complete speech animation. This entailed computing lip shapes by applying rules that transform phonetic units to action units, then computing all action units for emotion and expression. This was followed by spline interpolation between phonetic units, and finally the generation of the complete facial expression image.

Pelachaud's work in facial animation considers the link between intonation and expression. Intonation is the melodic feature of an utterance, and is linked to the attitude of the speaker and conveys emotional signals (Pelachaud 1991). It has three components: type of utterance, attitude and emotion, and affects the pitch, loudness, tempo and pauses within speech. In Pelachaud's system, each emotion corresponds to a particular set of values of these intonational parameters. Emotion affects lip shapes during speech, influencing the muscle tension and the degree of hyper-articulation or hypo-articulation, so these parameters must be modelled in order to give a talking head expressivity (Bevacqua et al. 2004).

Cao et al. used a machine learning approach to model expressive visual behaviour during speech (Cao et al. 2005). A database of high fidelity speech-related facial motions with variations across multiple emotions was recorded with an optical system, tracking markers on the face with 8 cameras. A training set of speech related motions was used to derive a generative model of expressive facial motion. The input of the system is a spoken utterance and a set of emotional tags which can be specified by a user or extracted from the speech signal using a classifier. The output is realistic facial animation that is synchronised to the input audio and conveys the specified emotions. A limitation of this approach is its dependency on the quality of the motion and speech data, and building the database requires laborious manual processing.

## 2.4 Talking Heads in Speech Tutoring

Applications of talking heads in speech tutoring include pronunciation training for speech-impaired children; speech reading for deaf children, and teaching English as a second language for adults. Traditionally dyslexic children are explicitly trained to make articulatory gestures to form words (Wise et al. 1999). Dyslexic children have been found to have difficulty in sensing the position of the tongue, teeth and lips whilst making a specific sound, and in identifying the appropriate picture depicting these positions (Montgomery 1981). Montgomery suggested that dyslexic children would benefit from training in awareness of articulation processes with single phonemes. Therefore a talking head with correct tongue movements for individual phonemes could be beneficial in tongue training.

Visual speech can be valuable in speech tutoring applications because vision benefits human speech perception, for three reasons as suggested by Summerfield (Summerfield 1987): It helps speaker localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation (Potamianos et al. 2004). Potamianos et al. state that human speech perception is bimodal in nature: Humans combine audio and visual information in deciding what has been spoken. The visual modality benefit to speech intelligibility in noise has been quantified by Sumbly and Pollack, who found that the visual contribution to intelligibility ranged from 77% to 81% as the Signal-to Noise ratio increased from -30 dB to -6 dB, for a test using a 32-word vocabulary (Sumbly et al. 1954). Benoît and Le Goff found that visual speech adds intelligibility to the auditory information when the acoustics are degraded (Benoît et al. 1998). Visual speech is of particular importance to the hearing impaired, for whom mouth movement has an important role in communication (Marschark et al. 1998).

There are several theories behind why visual feedback can be beneficial in speech tutoring. Bullock offers theoretical evidence for why spectrograms (visualization of waveforms of acoustic speech) can be useful in speech therapy for children (Bullock 2011a). Established theory is that as children learn phonology, their expressive representations of phonemes (articulatory movements, auditory

feedback) connect with their receptive representations (auditory features of speech). For some children there is a mismatch between their expressive and receptive representations of phonemes. Therefore, another form of feedback should be given so that the learner does not have to rely on their own auditory perception.

Another theoretical framework is the Directions Into Velocities of Articulators (DIVA) computational model (Guenther 2003). It proposes that a child learning to speak is informed by two neural systems. The first is a feed-forward control system, which consists of the speech sound map (motor plan), the cerebellum (coordination and smoothing of movement), and articulatory velocity and position maps (motor cortex). These systems inform the speaker what to do in order to produce a given sound. The second is the feedback system, which consists of orosensory (how the sound feels) and auditory information. Speech production involves a “target” that the motor system aims to achieve in order to produce a particular speech sound. Targets in the DIVA model take the form of convex regions in a planning space consisting of auditory and orosensory dimensions (for example, vocal tract constrictions), which are learned by infants during the babbling phase, when they recognize a specific vocal tract configuration as producing a speech sound. For example, the region for /p/ does not vary over the dimension of lip aperture, because all bilabial stops have full closure of the lips, so it is learned that lip aperture is an important dimension for producing the bilabial stop /p/. Since convex region learning relies on language-specific recognition of phonemes by the infant, the shapes of the resulting regions will vary from language to language. At first a child must rely more on the feedback system than the feedforward system because the movements are not yet mapped. As the child practices speech, it stores more information in the feedforward system about the movements and their associated auditory consequences, which in turn will inform future sound productions. If a child has difficulty perceiving the correctness of a production, its feedback system cannot adequately inform the feedforward system. Bullock argues that an additional feedback component, such as a spectrogram, can help students make judgments about their productions and make adjustments to the feed-forward system. At first, the learner must rely more

on visual feedback and learn to ignore the auditory productions. When sufficient correct productions have been practised and the speech sound and articulatory movement maps have been adjusted, the learner can then start to rely on their auditory feedback system (Bullock 2011c).

These theories are concerned with children learning a first language, but some of the implications can also be extended to second language learning. Most current approaches to second language acquisition can be divided into two groups: nativist models and empiricist models (MacWhinney 1997). Nativists view second language acquisition as repeating the course of first language acquisition, which is aided by a universal language instinct. Empiricist approaches argue that in childhood there is a critical period for language learning, after which the learner can no longer rely on the universal language instinct to facilitate second language learning. The empiricist approach emphasizes the role of input in both first and second language learning and the role of transfer and generalization in second language learning. The *Competition Model* of MacWhinney and Bates (MacWhinney 1992) takes an empiricist approach which views both first and second language learning as data-driven processes that rely not on universals of linguistic structure, but on universals of cognitive structure. This model attributes development to learning and transfer. It assumes that all mental processing uses a common, interconnected set of cognitive structures. Thus a second language learner will initially experience a large amount of transfer from their first language to the second language. Inappropriate interference effects in phonology are eliminated by unlearning some of this direct transfer. Some new sounds are learned that do not mirror first-language sounds. In other cases, newly acquired words need to be unlinked from sounds influenced by first-language segments. Learning at this stage progresses by correct registration of the phonology of the second-language target lexical items. The learner must be encouraged to perceive the mismatch between their output forms and the correct input forms. MacWhinney suggests that this could be achieved through a process in which learners attempt to match their own productions to computer-controlled digitized speech (MacWhinney 1992).

For second-language learners a common difficulty is the conceptual distinction between sounds which are allophones in their native language (for example, /p/ and /b/ are allophones in Arabic, and /r/- /l/ are allophones in Japanese.) The learner needs to work on hearing and producing the contrast in the second language, so practice is essential to activate new concepts relevant to the new language. Listen-and-repeat practice can be helpful to actively learn new concepts at a subconscious level (Fraser 2006). Verbal explanations of pronunciation can be difficult for students with limited vocabulary to understand, and a benefit of visual representations is that they are language-independent. Another benefit of using speech tutoring software, which applies to adult second language learners as well as children learning a first language, is that it enables independent drill to take place. When students use visualization to modify articulatory movements, if they can recognize the salient visual features of the targets, they are actively involved in their learning and this increases their understanding.

#### 2.4.1 Existing Speech Tutoring Software

Applications using talking head technology for second language learning have been developed, but are not in common usage, and the software used by language schools generally uses video recordings rather than synthesized talking heads. For example, the English Language Teaching Centre at the University of Sheffield currently uses the Sky Pronunciation Suite software in teaching (SKY). This uses video and audio recorded from a real human rather than a synthesised talking head. It includes training in the Phonemic Alphabet, in which the system plays a recording, and the user has to click on the correct symbol, and the system gives feedback as a score. The system plays video recordings of a real mouth, which shows the correct mouth movements for each item. The system can record the user's voice, and they can play it back while playing a model speaker's voice simultaneously, so the user can hear how they deviate from the correct pronunciation. They can also visually compare the waveforms. Such spectrogram data, although it can be difficult to read, has been shown to have a beneficial effect when used by speech therapists (Bullock 2011b). This may be because a student who cannot determine when they are producing an incorrect sound is unlikely to change their productions without any external feedback, and the

spectrogram provides an alternative mode of feedback so that the learner does not have to rely on their own auditory perception.

The usefulness of visual feedback depends on how easily it can be interpreted. Visual displays of the acoustic signal may be impractical as a therapy aid due to the abstractness and complexity of the display. This is where talking heads can provide an advantage. Displaying animated models of the articulators can give the learner visual information which is easy to understand. For example, the Speak As You See (SAYS) pronunciation software uses 3D animations to show exactly where to place the tongue in relation to the palate and teeth, in order to produce the correct sounds (Learning Technologies International 2012). This software is used in conjunction with a mirror so that the learner can view their own movements simultaneously. The developers of SAYS state that in an initial test, some students improved their ability to differentiate their pronunciation of /r/ and /l/ sounds after using the software for as little as 15 minutes.

Other existing applications of talking heads for teaching English as a second language include Cohen and Massaro's Baldi (Massaro 2004). Baldi uses terminal analogue synthesis (Klatt 1987), which mimics the final speech product rather than the physiological mechanisms that produce it. The software is based on Parke's talking head, with additional and modified control parameters, texture mapping, and the addition of a tongue. Cohen and Massaro developed a new visual speech synthesis coarticulatory control strategy, using dominance and blending functions (Massaro 1998). Baldi also has controls for paralinguistic information so he can display facial expressions and gestures, and affect in the face, so he can show anger, happiness and sadness (Massaro et al. 2005). The system offers text-to-visible speech synthesis and alignment with natural speech. Baldi can be displayed in various configurations, for example, the skin can be made transparent so that the tongue and inside of the mouth can be viewed, and the head can be rotated to be seen from the back or side (Cosi 2002). Baldi has more recently been augmented with a body, to extend communication through gesture (Massaro et al. 2005).

Baldi has been used in multiple languages (Cosi 2002), (Ouni 2003), (Massaro et al. 2005), (Ouni 2005). For example, Baldi was used to train eleven Japanese

speakers to identify and produce American English /r/ and /l/, using two methods; instruction illustrating the internal articulatory processes of the oral cavity, and instruction providing only the normal view of the tutor's face (Massaro 2004). The perception and production of words by the Japanese trainees improved after training. However, this study did not indicate whether the display of internal articulators gave an advantage over displaying only the external face. In a more recent study, sixteen native English speakers were trained on pairs of similar speech segments in Arabic (/k/ and /q/) and in Mandarin (/i/ and /y/) (Massaro et al. 2008). Participants were trained with auditory speech versus both auditory and visual speech, and with a frontal view versus an inside view of the vocal tract. The participants showed improvements after training with the talking head, compared to the control groups, but the differences were not significant (Massaro et al. 2008). It is possible that the sample sizes were too small, and a larger study could yield more significant results. A tentative finding of this study was that the outside of the face seemed to be more easily processed than a sagittal viewing illustrating the tongue, palate, and velum. This study did not conclusively show that the visualisation of internal articulatory movements was effective in pronunciation training, which supports the view that further research is needed.

Baldi has been used to teach vocabulary to deaf children, and also for language learning with autistic children (Massaro 2012). The speech tutor for deaf children uses Baldi's internal productions. By making the skin transparent or by showing a sagittal view, Baldi can illustrate pronunciation of sounds that are not normally visible. In a trial of the system for the presentation of the internal visemes to deaf children, a significant improvement in learning was found in post-test speech production compared to pre-test. The lack of a control group raises the possibility that some of this learning was independent of the training. However, follow-up tests six weeks later showed that the subjects' productions had deteriorated since the post-test productions, without continued use of the Baldi tutoring system, which suggests that the training system had been a factor in the post-test improvement in speech production (Massaro et al. 2004).

The ARTUR (ARtication TUtoR) project aimed to provide computer assisted pronunciation training for hearing- or speech impaired children and second

language learners. ARTUR is a virtual speech tutor which uses three-dimensional animations of the face and internal parts of the mouth to give feedback on the difference between the user's deviation and correct pronunciation. The facial data, such as jaw position and mouth opening, was extracted from video images of the face. Experiments showed that the augmented reality side-view did not help subjects perform better overall than with the front view only, but it was beneficial for the perception of some articulatory features, such as palatal plosives (Wik et al. 2008) . The automatic mispronunciation detection could make errors, so a Wizard-of-Oz test was used in which a human selected the appropriate feedback and audiovisual instructions based on the user's pronunciation. The subjects' change in articulation during the practice session was monitored with an ultrasound probe. The ultrasound measurements suggested that an improvement was made by following articulatory instructions given by the computer-animated teacher (Engwall 2008).

Other applications in speech therapy include the work of Kroger, who conducted a pilot study using a visual articulatory model as a visual stimulation technique in therapy of articulation disorders and apraxia of speech. The visual recognition of sounds and syllables over the course of therapy was evaluated, and a significant increase in recognition rate was found (Kröger et al. 2005). Grauwinkel and Fagel's talking head with three-dimensional animation of internal articulator dynamics was investigated for use in speech therapy for children with interdental lisps (Grauwinkel et al. 2007). The results showed that most of the children were able to visually identify correct and wrong productions of the talking head. The evaluations showed that the lesson improved the sibilant production of two of the three children. In a subsequent study by the same authors, children's productions of words containing the sounds /s/ and /z/ were recorded and evaluated before and after two short learning lessons with an experimenter using the virtual head to explain the correct pronunciation of these sounds. Results showed that several children significantly improved their speech production of the /s/ and /z/ sounds (Fagel et al. 2008).

Badin et al. developed a French-speaking audiovisual talking head that can display all speech articulators. Three-dimensional models of speech articulators

were derived from volume MRI and multiple view video images acquired on one speaker (Badin et al. 2002). Linear component analysis was used to model these articulators as the weighted sum of a small number of basic shapes corresponding to the articulators' degrees of freedom for speech (Badin et al. 2006). Control parameters for animation were derived from points on the articulators of the same speaker tracked by Electro-Magnetic Articulography (Badin et al. 2008). Badin et al. found that tongue reading can take over the audio information when the latter is not sufficient to supplement lip reading (Badin et al. 2010). This finding has important implications for developing speech tutoring applications, as it lends support to the notion that visualization of tongue movements can contribute to speech perception.

## 2.5 Summary

The development of a complete facial animation system combines visual speech animation with expression modelling. The most common techniques for computer facial animation involve geometric manipulation of 3D facial models, using interpolation or parameterization, to produce graphically realistic computer-generated heads. Visual speech synthesis uses either viseme-driven or data-driven approaches to produce graphically and acoustically realistic speech. Viseme-driven speech animation requires a smaller amount of data to create key poses for internal articulators. A disadvantage of data-driven techniques is that they require a large corpus of captured data in order to produce realistic results, and for internal visualization, a corpus of internal articulatory data is required, but the benefit of a data-driven approach based on a real speaker's data is that it can create a more accurate model of articulator movement. Given the necessary visemes, key-frame interpolation is the simplest and fastest technique to produce animation. However, parameterization can more easily combine expressions, so may be more suitable for producing emotionally expressive speech. Expression overlays convey signals such as the display of emotions, head movements and eye movements, to improve the naturalness of non-verbal communication. All three need to be integrated for the production of believable talking heads.

There are several existing systems which use talking heads in teaching pronunciation, for both second learning and speech therapy, but talking heads are currently not commonly used in speech tutoring. Existing software for teaching second languages generally uses video recordings of real humans talking, which can be useful but cannot show the internal movements during articulation. This is an advantage of talking heads, which can provide multiple views of the mouth and vocal tract. Previous research has suggested that use of talking heads can lead to an improvement in speech production, but more research is needed to determine which aspects of talking heads, for example the visualization of internal articulators, are the most effective in training.

### 3 Talking Heads in Speech Tutoring

This chapter addresses how this project applies talking head technology in learning applications. Cole suggests that pedagogical agents, represented by talking heads or human voices, inspire social agency in interactive media, enabling users to interact with the program as they interact with people (Cole et al. 2007). Studies have shown that users learned more and reported greater satisfaction using programs that incorporated virtual humans or human voices (Moreno et al. 2001; Baylor et al. 2005). This provides motivation for the use of talking heads in tutoring systems.

Existing computer-based speech tutoring systems such as Baldi (Massaro 2012) are assistants used alongside a speech therapist, rather than standalone tutors performing the complete role of teaching. “It is important to point out that even the best computer program could never replace the therapist but only assist and facilitate his or her work. Computer-aided speech training is a complement to traditional methods and has a pedagogical value for the therapist who has a good knowledge of articulatory and acoustic phonetics as well as of the computer technique” (Öster 1996). This research therefore aimed to create a pronunciation assistant rather than a complete tutor.

#### 3.1 Teaching English as a second language for adults

The chosen study was teaching English as a second language for adults. The talking head teaches a particular language feature; a suitable test case was identified in consultation with teachers and users to determine what would be beneficial to the students. The University of Sheffield’s English Language Teaching Centre (ELTC) runs courses for non-native speakers to improve their English; full-time courses to help students reach an appropriate level of proficiency in English for entry into university; and summer schools for students preparing to enter university. Tutors and students from the ELTC were consulted to research how technology could be of benefit in the teaching of English as a second language. Six members of staff and two students were interviewed for

their opinions on the use of technology in second language learning. The findings were that technology is not used a lot in teaching of pronunciation, as most tutors demonstrate speech live. The possibilities for technology are with the routine work of repetition and drilling of vocabulary, which cannot be done to a great extent in class because it can become tedious and takes a lot of time. The ELTC currently use software, Sky (SKY), which shows videos of lip movements for teaching pronunciation. The tutors consulted thought that a 3D model with tongue movements and cross-section visualization would be useful; for example, when demonstrating glottal stops, it is difficult to explain how the vocal tract is used, but an image would make it easier. The tutors currently use charts of dissected heads, and they thought that anything that makes the demonstration seem more alive would be a good asset. One of their students concurred that it is helpful in class when the pronunciation tutor demonstrates tongue positions.

The most common native languages of the students on the ELTC courses are Chinese and Saudi Arabian. The most common problems for the Chinese speakers are /r/ - /l/; for the Arabic speakers, /p/-/b/, voiced and unvoiced similar sounds. A tutor suggested that a suitable test case would use short sentences, which can be broken into sounds, so students can practise strong forms, weak forms and elision. This is because the way a word sounds in isolation differs from the way it is pronounced in connected speech. For example, the strong form of a word, where it is stressed, is phonemically different from the weak form of the word, where it is unstressed. Moreover, elision, the omission of sounds often occurs unintentionally in natural connected speech. Therefore, it is important to train using sentences rather than just individual words, in order to teach natural-sounding speech.

Oster defines guidelines for clinical applications of computer-based speech training for children with hearing impairment (Öster 1996). Oster states that a visual speech training aid has a number of important requirements, which include immediate visual feedback of the child's voice and articulation, and contrastive training, i.e. the correct model of the therapist and the deviant production of the child are shown simultaneously and compared with each other. These guidelines had implications for the design of this system; ways of providing immediate

visual feedback were investigated, while contrastive training could be achieved by adding functionality to play back a recording of the user's production for comparison with a model speaker.

### 3.2 Design of Talking Head

In the design of the Talking Head, one issue to consider was whether to have a whole face or just mouth movement. This research explores what style of talking head is sufficient. For example, the system of Massaro et al allows internal articulatory movements to be viewed, and learning was improved, but this did not prove the effectiveness of showing internal articulatory movements for pronunciation training (Massaro et al. 2008). An English language tutor from the ELTC said that cross-section visualization of tongue positions would definitely be useful, because it is useful to have a practical demonstration of the way lip and tongue positions affect speech. Therefore it was decided that internal visualization would be included in the speech tutoring system (Chapter 4).

The level of natural gestures required was also considered, for example, whether users would respond better to a moving talking head or a stationary talking head. Conversational signals could make the face appear less rigid, to increase its believability. However, the naturalness survey in Chapter 6 found that adding eye and head movements did not improve the perceived naturalness; so these movements were removed to make the animation less distracting for the users, so they could focus on the mouth movements. Psychological studies have found that eye-gaze patterns are concentrated on the “core features” of the face: the nose, the mouth, and especially the eyes (Walker-Smith et al. 1977). If the eyes were hidden, it could help a learner to focus on the lips, but being able to see the whole face can aid intelligibility, as other parts of the face can emphasise the movements. It could be useful to over-accentuate important mouth movements by using caricatures rather than realistic models, so that differences between poses are more identifiable (Frowd et al. 2007). Other issues considered were whether to make the face a cartoon-like character or photographic representation. Caricaturizing some features, for example by making the mouth larger, could give

the head a more cartoon-like appearance to match the level of realism of its behaviour, which could make it more acceptable to users (Mori 1970).

The chosen embodiment for the first talking head was based on an average female face from Facegen (Singular Inversions 2008) (Figure 3.2). The facial features of this face are the average of all races, in an effort to be inclusive of all the potential users of this application, as suggested by the language tutors surveyed. A hair model from Facegen was added. A female embodiment was selected as some studies have shown that female agents are more likely to positively influence learning (Baylor 2005). The talking head was named “Tara” (Talking Articulation Assistant).

### 3.3 Graphical User Interface for Speech Tutoring Application

The talking head, Tara (Talking Articulation Assistant), developed as described in Chapter 4, was integrated into a Graphical User Interface for a speech tutoring application, developed using the QT framework (QT 2009). The speech tutoring application demonstrates how to pronounce sounds at phonetic, word and sentence level, displaying the appropriate mouth movements, and displays a transverse cross-section through the head, showing the movement of internal parts such as the tongue during speech (Figure 3.1). A camera controller was added to allow the user to rotate the head, and buttons were added to show a close-up of the talking head’s mouth in the front view.

The speech tutoring application consists of 10 screens, shown in Appendix B:

- 1: Introduction
- 2: Listening Practice: Sounds
- 3: Listening Practice: Words
- 4: Listening Practice: Phrases
- 5: Listening Test: Sounds
- 6: Listening Test: Words
- 7: Speaking Practice: Sounds

8: Speaking Practice: Words

9: Speaking Practice: Phrases

10: End of Lesson

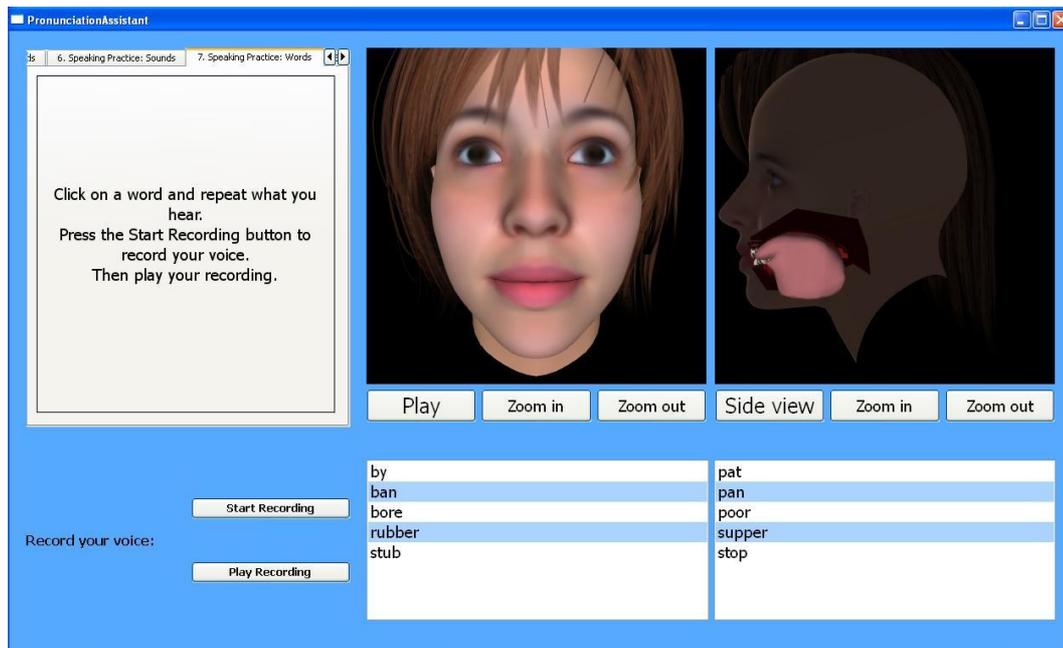


Figure 3.1: Screenshot of Speech Tutoring Application used in Tutoring Study 1

Functionality was added to allow the user to record their voice and play back the recording. Simple speech recognition was implemented using Microsoft 5.1 SAPI and Microsoft English (U.S.) 6.1 Recogniser. To maximize the accuracy of the speech recognition, a forced-alignment approach was used, with a grammar to constraint the recognized vocabulary to the words within the application, and similar variants of these words to allow for variations in speakers' accents e.g. "*bat, pat, but, put*". This allowed the speech recognition to detect when a speaker had said /b/ or /p/, so the talking head could give feedback when it detected the correct sound, for example, saying "*Well done!*" in response to a correct pronunciation. Figure 3.2 shows a diagram of the components of the speech tutoring system.

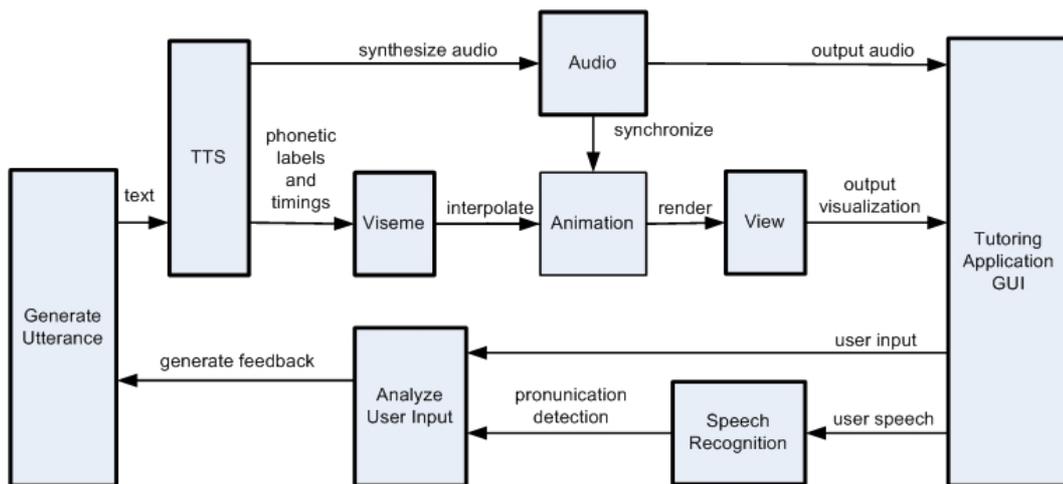


Figure 3.2: Diagram of Speech Tutoring Application used in Tutoring Study 1

## 4 Development of Viseme-Driven Talking Head

This thesis considers two approaches for the development of talking heads: viseme-driven and data-driven speech animation. Viseme-driven techniques require a smaller amount of data to create key poses for articulators. A disadvantage of data-driven techniques is that they require a large corpus of captured data in order to produce realistic results, and for internal visualization, a corpus of internal articulatory data is required, but the benefit of a data-driven approach based on a real speaker's data is that it can create a more accurate model of articulator movement.

This chapter describes the two viseme-driven heads created. The first (THVN) is a generic, non-photorealistic head, while the second viseme-driven head (THVP) was given a more photorealistic appearance using photographs of a real speaker, and the modelling of the internal articulators was improved by the use of MRI data from the same speaker. The following chapter (Chapter 4) describes the acquisition of the corpus of MRI, EMA and video data, in collaboration with GIPSA-Lab, Grenoble, and the creation of a data-driven head (THD).

### 4.1 Implementation of Viseme-Driven Talking Head (THVN)

A viseme-driven, non-photorealistic talking head (THVN) was implemented in C++ on the Windows XP platform, using a Text-To-Speech synthesizer to generate audio to drive the animation (Figure 4.1). Face models were created using Facegen modelling software (Singular Inversions 2008), which provided face meshes for 15 distinct visemes (Figure 4.2), in addition to the neutral face and facial expressions.

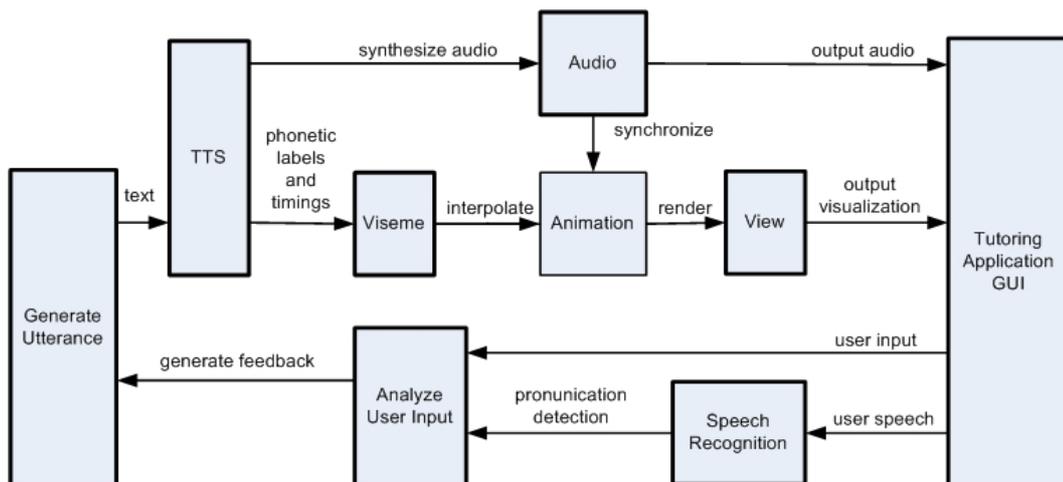


Figure 4.1: Diagram of Speech Tutoring Application

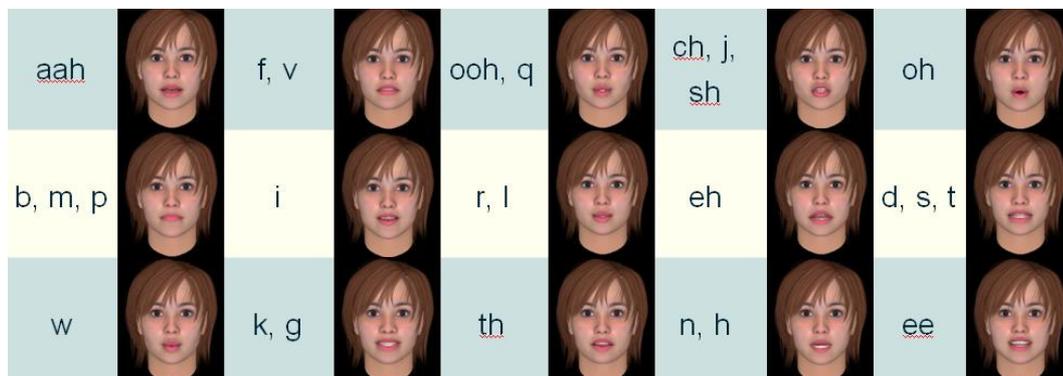


Figure 4.2: Facegen models for visemes (with names used by Facegen)

## 4.2 Internal articulator models

The tongue meshes provided by Facegen were designed to look plausible in the frontal external view, but these tongue shapes were not realistic when viewed mid-sagittally, so would be of limited validity in a tutoring application for demonstrating correct tongue positions during speech. Visemes for the tongue positions were initially adapted from Oscar Martinez Lazalde’s tongue models (Figure 4.3). However, the internal mouth visualization using Lazalde’s teeth and tongue models (Lazalde et al. 2008) required further development to display more accurate articulatory movements.

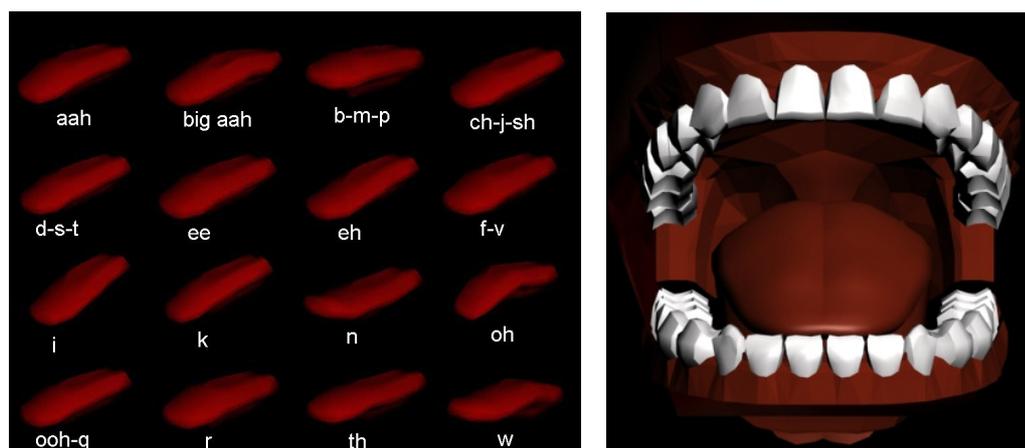


Figure 4.3: Tongue Visemes and Teeth Model (Lazalde et al. 2008)

The Visible Human Project (EPFL 2009) female provides a dataset for tongue anatomy, but more data was needed for modelling tongue dynamics. Badin et al. (Badin et al. 2008) have used MRI, Electro-Magnetic Articulography and video data to produce articulatory models, and Cohen et al. (Cohen et al. 1998) and Engwall (Engwall 2003) have used MRI and electropalatography data in tongue modelling. The Artimate framework uses EMA data for articulatory animation synthesis, to provide animation of the tongue and teeth for a virtual character (Steiner et al. 2012). Existing vocal tract visualization tools include ArtiSynth (Fels et al. 2007) and Vocal Tract Lab (Birkholz et al. 2007). A corpus of articulatory data, MOCHA-TIMIT, has EMA, EPG and laryngograph data of teeth, tongue and velum (Wrench 1999). The “mngu0” corpus consists of MRI data of a single British English male speaker (Steiner et al. 2012).

In order to achieve the most accurate articulatory animation, the speaker used as the source of data for the geometry of the articulatory organs should match the speaker used as the source of the motion capture data. This approach would capture the correct geometric degrees of freedom for modelling articulation, and also respect speaker-specific articulatory strategies (Elisei et al. 2001). For this project, a corpus of articulatory data of a British English female speaker was required, and since none of the existing corpora were suitable, a new corpus was created in collaboration with GIPSA-Lab, as described in Chapter 5.

### 4.2.1 MRI Tongue Contours

The mid-sagittal contours (Figure 4.5) from MRI data (Figure 4.4) captured at GIPSA-Lab as described in Chapter 5 were used to remodel the tongue visemes of the viseme-driven talking head. For each vowel, the corresponding articulation was chosen, while for each consonant, the /e/ context, for example “*epe*”, was chosen because it was the vowel with the most central tongue position (Figure A.1). The vertices of each contour were imported into 3DS Max, and aligned with the head meshes (Figure 4.6). The tongue mesh was remodelled, and deformed manually for each viseme, until its outline was as close a match as possible to the corresponding MRI contour in the mid-sagittal plane (Figures 4.7- 4.8). It now became clear that the visemes had to be reclassified; for example the tongue shape for /m/ was not the same as that for /b/-/p/; therefore a separate viseme was created using the tongue contour for /m/. There was now a total set of 20 visemes (Table 4.1).



Figure 4.4: Mid-sagittal MRI scan for articulation of vowel /e/

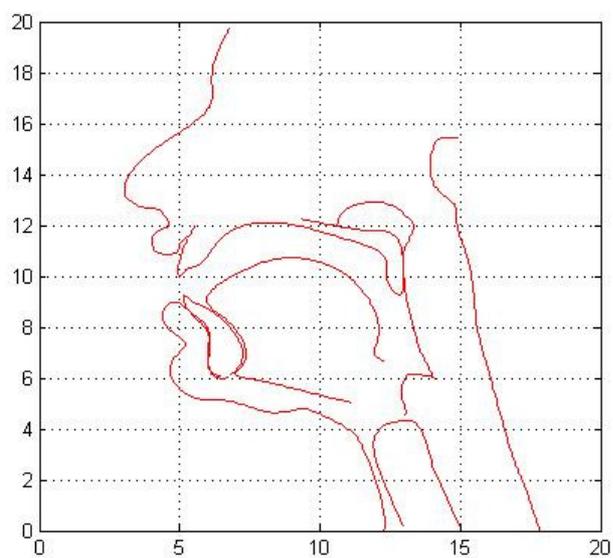


Figure 4.5: Mid-sagittal MRI contours for vowel /e/

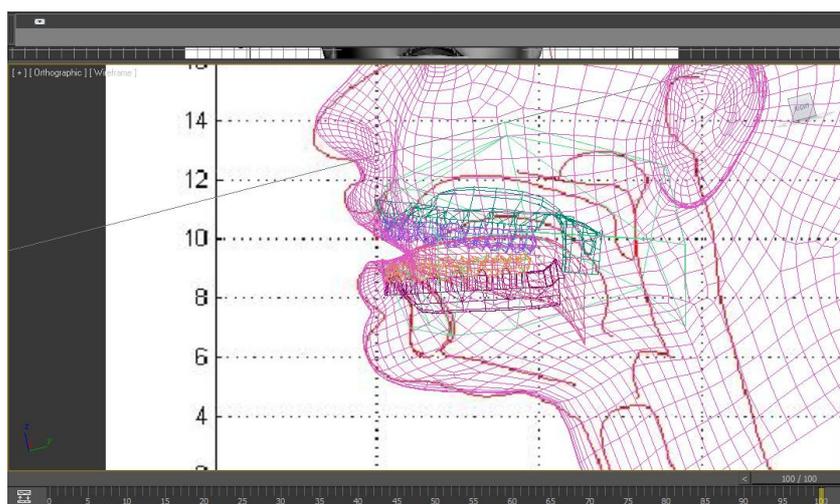


Figure 4.6: Alignment of head meshes with mid-sagittal MRI contours for vowel /e/

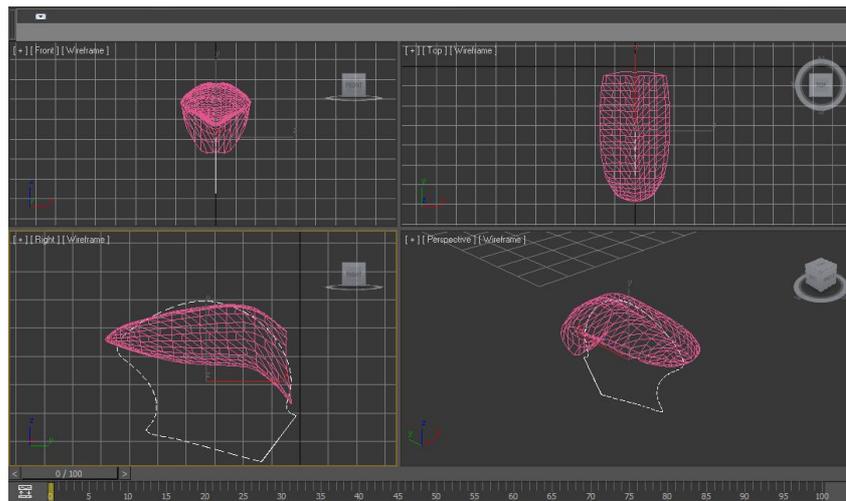


Figure 4.7: Comparing original tongue mesh with mid-sagittal MRI contour

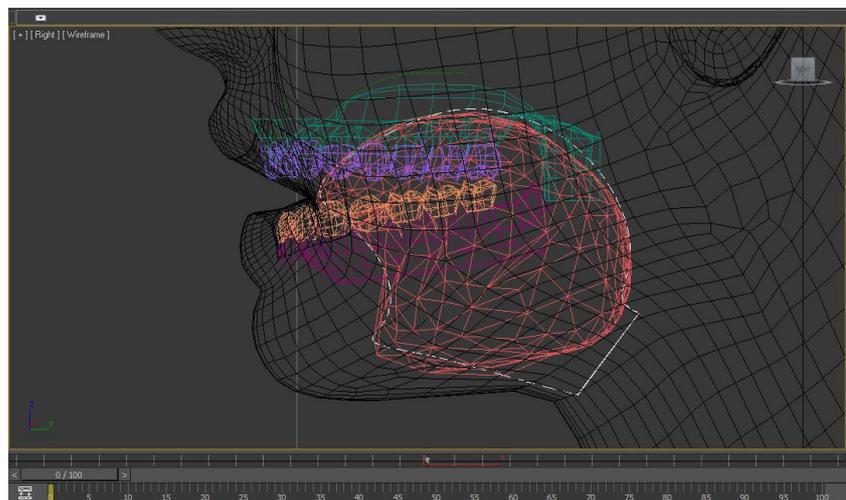


Figure 4.8: Modelling tongue mesh based on mid-sagittal MRI contour

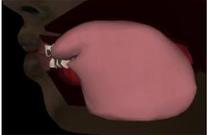
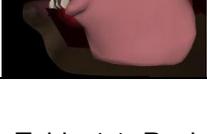
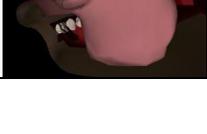
Index	Sound	IPA	Internal Viseme	Index	Sound	IPA	Internal Viseme
1	neutral (jaw closed)	silent		11	n	n	
2	a (in calm)	ɑ:		12	o (in fond)	ɒ	
3	b, p	p		13	oo (in fool)	u:	
4	ch, j, sh	ʃ		14	r	r	
5	d, t	t		15	th	θ	
6	ee (in beep)	i:		16	l	l	
7	e (in less)	e		17	m	m	
8	f, v	f		18	s	s	
9	i (in sit)	ɪ		19	ng	ŋ	
10	k, g	k		20	w	w	

Table 4.1: Reclassified Visemes

### 4.3 Text-To-Speech Synthesis

A text-to-speech (TTS) synthesis system was required to generate the auditory speech and phonetic labels to drive the animation. This was chosen from the existing systems available, which are described in the following sections.

#### 4.3.1 Festival Speech Synthesis System

Festival provides a general framework for building speech synthesis systems. It offers full text to speech. The system is written in C++ and uses the Edinburgh Speech Tools Library for low level architecture and has a Scheme-based command interpreter for control (CSTR 2008). MBROLA provides voices which can be used with Festival (MBROLA). Festival has been used in existing talking heads (Edge 2004), (Lazalde 2010).

#### 4.3.2 Microsoft Speech Application Programming Interface

The Microsoft Speech API provides a TTS engine for Windows applications. The Microsoft voices available, such as “Microsoft Mary”, sounded too robotic, but the Microsoft Speech SDK 5.1 could be linked with MBROLA (MBROLA) for more natural-sounding voices.

#### 4.3.3 HTS Hidden Markov Model-based Speech Synthesis

Model-based systems are an alternative to concatenative systems such as that used by Festival. An example of a HMM-based Speech Synthesis System is HTS (HTS 2011). HTS can be more consistent but less natural sounding than concatenative synthesis. An HTS sample, compared with a concatenative Festival voice sample, was more intelligible, but sounded monotonic. HTS is easier to manipulate than concatenation; for example, with concatenation, the user does not have much control over prosody and pauses. With concatenation the user can only produce what is pre-recorded in the database, whereas in theory HTS can generate anything, because it uses statistical models rather than a large database. A drawback of HTS is that it is less real-time than concatenation, so it is less suitable for conversation; for example, it could take 2-3 seconds to convert text to speech, which would be a noticeable pause in a real-time interactive application.

#### 4.3.4 Loquendo TTS

Loquendo TTS is a commercially-available concatenative speech synthesis engine (Loquendo 2008). Loquendo TTS 7 was chosen for Text To Speech conversion, because compared to the non-commercial systems available at the time, such as Festival, Loquendo provided more natural-sounding voices, although it still has some artefacts.

Most sentences were automatically generated by Loquendo with the correct pronunciation of segmentals, prosody, intonation and emphasis. There were some difficulties with the synthesis of some words; for example, the isolated word "age" could sound like "aitch", and "dug" could sound like "duck". The word stress sounded unusual in some contexts, e.g. "butter" in the sentence "She said the butter's bitter". However, these issues could usually be overcome by using annotation to control how to pronounce these sounds. A major benefit was the option to use XSAMPA notation to specify exactly which phonetic units would be pronounced. This feature was important when used by the pronunciation-tutoring application to produce instructions on how to pronounce individual sounds.

There were three British voices available, two female and one male. The voice chosen was "Kate", a British female with an English Received Pronunciation (RP) accent, which was selected as the most suitable for the application. The C API of Loquendo TTS was used to integrate the TTS into the application. The callback mechanism of Loquendo TTS provided a means of outputting phoneme labels and durations which could be used directly by the application to produce speech-synchronised animation for the talking head.

### 4.4 Text to Visual Speech

The application takes input as a text file containing the words to be spoken. This text can be annotated with Loquendo markup tags, to control voice parameters such as speed, prosody and pitch. Loquendo TTS generates speech, saves it to a .wav file, and outputs phonetic labels and durations. The outputted phonetic labels are in the XSAMPA format, a machine-readable format for phonetic transcriptions. Each phonetic label is mapped to a Facegen mesh for the

corresponding viseme. When the animation is run, the appropriate viseme mesh is displayed for a particular frame and interpolation is used to create in-between frames. The quality of this interpolation process influences the quality of the resulting animation. A diagram of this process is shown in Figure 4.9.

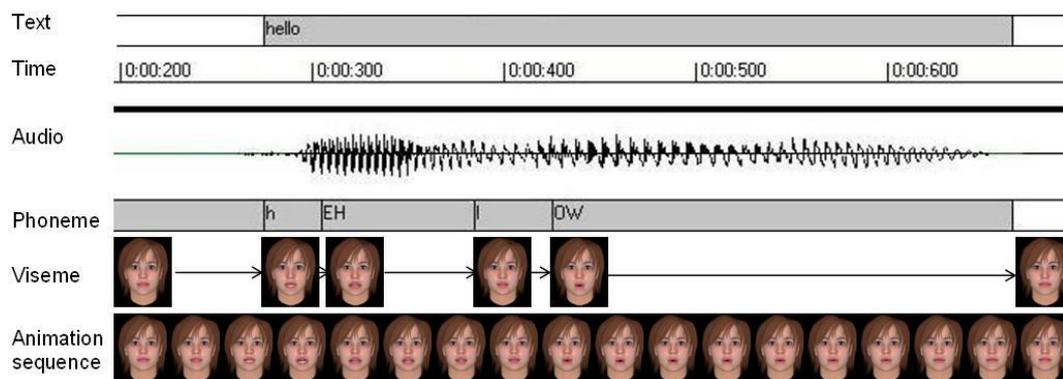


Figure 4.9: Viseme-driven speech synchronised animation

## 4.5 Interpolation

Initially, linear interpolation was used to blend the meshes for a smoother transition between visemes. Linear interpolation is the simplest and fastest method of calculation, but sharp changes of gradient at each keyframe can be visually disturbing. An alternative method which was implemented is Catmull-Rom spline interpolation, which overcomes the gradient change problem and fits the keyframes more smoothly. A Catmull-Rom spline is a cubic curve which passes through all control points (Catmull et al. 1974).

For a keyframe with value  $\mathbf{v}_i$  at time  $t_i$ , where the following keyframe has a value  $\mathbf{v}_{i+1}$  at time  $t_{i+1}$  (Figure 4.10),  $s$  is an interpolation factor in  $[0, 1)$  computed from the keyframe times (Equation 4.1):

$$s = (t - t_i) / (t_{i+1} - t_i)$$

[Equation 4.1]

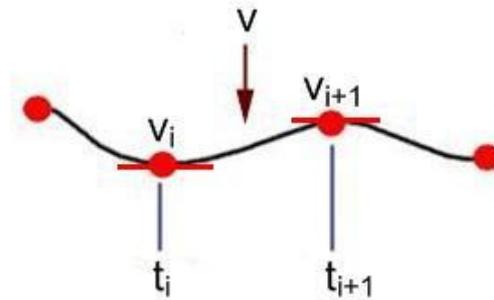


Figure 4.10: Spline Interpolation, adapted from (Dunlop 2009)

Tangents are defined at the end points of the current curve segment:  $\mathbf{T}_i$  at the start point, and  $\mathbf{T}_{i+1}$  at the end point. The interpolation of the curve can be expressed as follows:

$$\mathbf{S} = \begin{bmatrix} | s^3 | \\ | s^2 | \\ | s | \\ | 1 | \end{bmatrix} \quad \mathbf{H} = \begin{bmatrix} | 2 & -2 & 1 & 1 | \\ | -3 & 3 & -2 & -1 | \\ | 0 & 0 & 1 & 0 | \\ | 1 & 0 & 0 & 0 | \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} | \mathbf{v}_i | \\ | \mathbf{v}_{i+1} | \\ | \mathbf{T}_i^0 | \\ | \mathbf{T}_{i+1}^1 | \end{bmatrix}$$

The value  $\mathbf{v}_s$  of the curve at position  $s$  can be calculated using the formula given in Equation 4.2:

$$\mathbf{v}_s = \mathbf{S}^T \mathbf{H} \mathbf{C} \quad \text{[Equation 4.2]}$$

A standard Catmull-Rom spline assumes that the keyframe values are evenly spaced in time, and calculates the tangents  $\mathbf{T}_i^0$  and  $\mathbf{T}_i^1$  as centred finite differences of the adjacent keyframes (Equation 4.3):

$$\mathbf{T}_i = \frac{(\mathbf{v}_{i+1} - \mathbf{v}_{i-1})}{2} \quad \text{[Equation 4.3]}$$

In order to allow for non-uniform spacing of keyframes, additional scaling values are applied to compensate for irregular keyframe timing, and the resultant tangents are given by Equations 4.4 and 4.5, where the scaling factors are given by Equations 4.6 and 4.7 respectively:

$$\mathbf{T}_i^0 = \mathbf{F}_i^- \mathbf{T}_i \quad \text{[Equation 4.4]}$$

$$\mathbf{T}_i^1 = \mathbf{F}_i^+ \mathbf{T}_i \quad \text{[Equation 4.5]}$$

where:

$$\mathbf{F}_i^- = 2 \frac{(t_{i+1} - t_i)}{(t_{i+1} - t_{i-1})} \quad \text{[Equation 4.6]}$$

$$\mathbf{F}_i^+ = 2 \frac{(t_i - t_{i-1})}{(t_{i+1} - t_{i-1})} \quad \text{[Equation 4.7]}$$

$$\mathbf{F}_0^- = \mathbf{F}_0^+ = \mathbf{F}_{N-1}^- = \mathbf{F}_{N-1}^+ = 0 \quad \text{[Equation 4.8]}$$

The tangents of the segments at the extreme ends of the spline are undefined, and are given a value of 0 (Equation 4.8) (Nokia Corporation 2005), (Watt et al. 1992).

## 4.6 Principal Component Analysis

In order to reduce the computation time for the animation, Principal Component Analysis (PCA) was carried out. PCA reduces the dimensionality of the data by transforming it into uncorrelated variables, called principal components, which capture the maximal variation in the data.

Any element,  $v$ , in the original dataset can be represented using Equation 4.9, where  $\mu$  is the mean vector,  $e_i$  is the  $i$ th principal component, and the  $b_i$  are weights uniquely defining  $v$ .

$$v = \mu + \sum_{i=1}^s e_i b_i \quad \text{Equation 4.9}$$

The  $e_i$  principal components can be calculated by finding the eigenvectors of the covariance matrix for the observed dataset, with the corresponding eigenvalues representing the variance,  $s_i$ , accounted for by each component. Components with low eigenvalues represent only small variations in the dataset, and may be culled with little loss of accuracy in the model (Edge 2004).

Code by Lalalde (Lalalde et al. 2008) was used to calculate the PCA and reconstruction functions. The PCA code was run in Matlab, separately for the internal vertices and external vertices, to create Principal Components (PCs) for the final set of 20 visemes. The number of PCs was set to 7 for the internal meshes and 7 for the external meshes, giving a total of 14 PCs, which showed little noticeable loss of accuracy compared to using the maximum 20 PCs. The PC data was loaded into the talking head application, where the dominance functions were applied to the PCs. These PCs were then reconstructed into meshes during the generation of frames for animation. Using PCA reduced the computation time, because the dominance functions were being applied to only a small number of PCs instead of to every vertex of a mesh of 22158 vertices (Figure 4.11). The application of PCA involved a trade-off between minimising the number of PCs to reduce computation time and minimising loss of data, which could cause loss of subtle but salient details in the lip visemes. Separate PCA was run for the external and internal parts, which maintained separation of the Principal Components, and the mesh could be segmented further to apply local PCA to different facial regions, which would simplify each feature space, to give better PCA approximations.

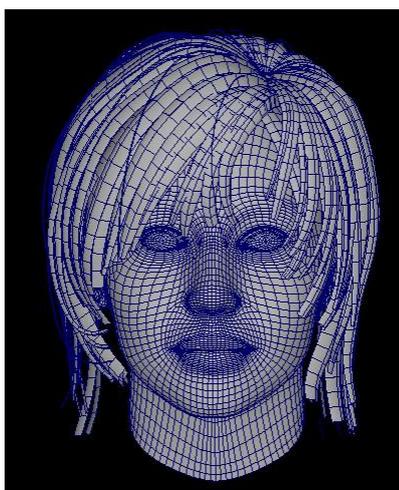


Figure 4.11: Mesh for Talking Head (THVN)

## 4.7 Coarticulation

The Cohen and Massaro model for coarticulation was implemented, using a dominance function to represent the influence over time that a viseme has on a speech utterance (Cohen et al. 1993). Typically the influence is greatest at the centre of the viseme and degrades with distance from the viseme centre (Figure 4.12). The shape of each dominance function is different according to which viseme it represents, and what aspect of the face is being controlled (e.g. lip width, jaw rotation.) Each speech segment has one dominance function for each articulator. Articulatory dominance functions can differ in time offset, duration, and magnitude; different time offsets can capture differences in voicing, while the magnitude can represent the relative importance of a characteristic for a segment (Massaro 1998).

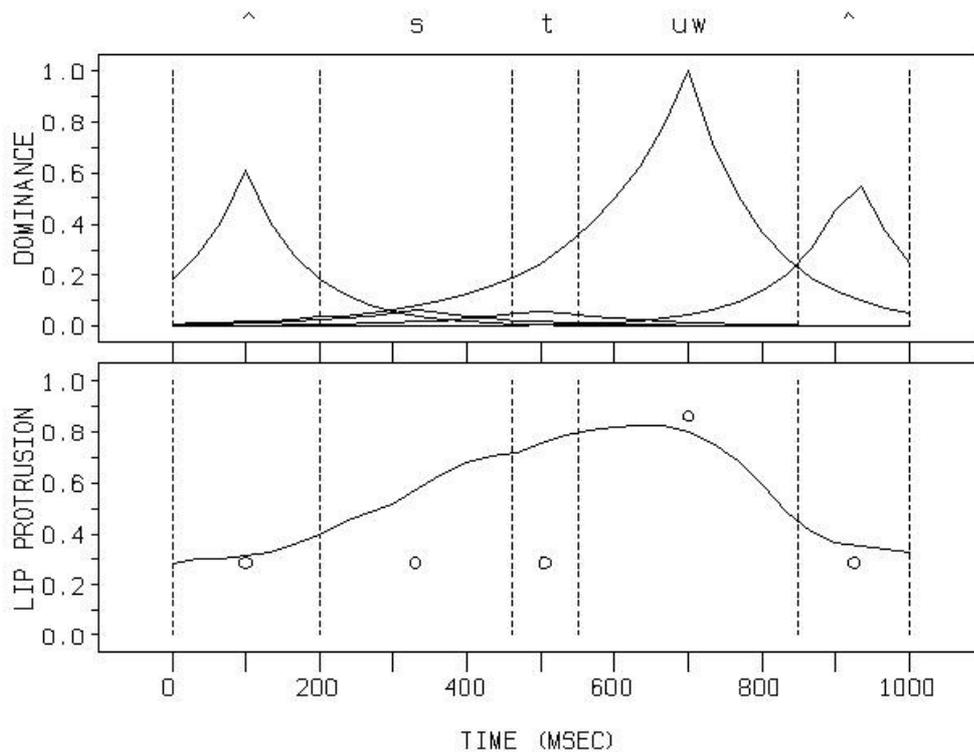


Figure 4.12: Cohen-Massaro Dominance Functions for the word “stew”, and corresponding lip protrusion parameter values, reproduced with permission (Massaro 1998)

#### 4.7.1 Tuning visual speech model parameters by observation of video

The coefficients of the dominance functions were set by observation, comparing the synthesized visual speech in the external frontal view against video recordings of a real person saying the same words, until the synthesized speech looked like the recorded speech. The word lists used for tuning included each sound in initial and final positions, e.g. for the sound /b/, the words “*bad bed bud bib bob ebb*” were used. The dominance functions of each segment were blended together to generate a speech trajectory.

For example, in the word “*stew*”, the /s/ and /t/ segments have low dominance compared to /u/ (Figure 4.12), and the low anticipatory rate of /u/ causes its domination to extend far forward in time. The result is that the lip protrusion extends forward in time from the vowel (Figure 4.13). The animation frames were compared against the video frames (Figure 4.13), and the coefficients were tuned to give the closest match that could be found by observation. Since manually

tuning was a time-consuming process, the number of words that could be used was limited, and not all contexts could be included. The words chosen for tuning included each consonant viseme in initial and final positions (Table 4.2).

Viseme	Words
b m p	bad bed bib bob men put
k g	cat could kick great again
d s t	dad did said tip tongue it
f v	face fall if off of van have
n h	nan and on had how hello
r l	rat red rare are lips loll all
ch j sh	show she jam judge chin
th	thin teeth mouth the then

Table 4.2: Example words used in tuning visual speech

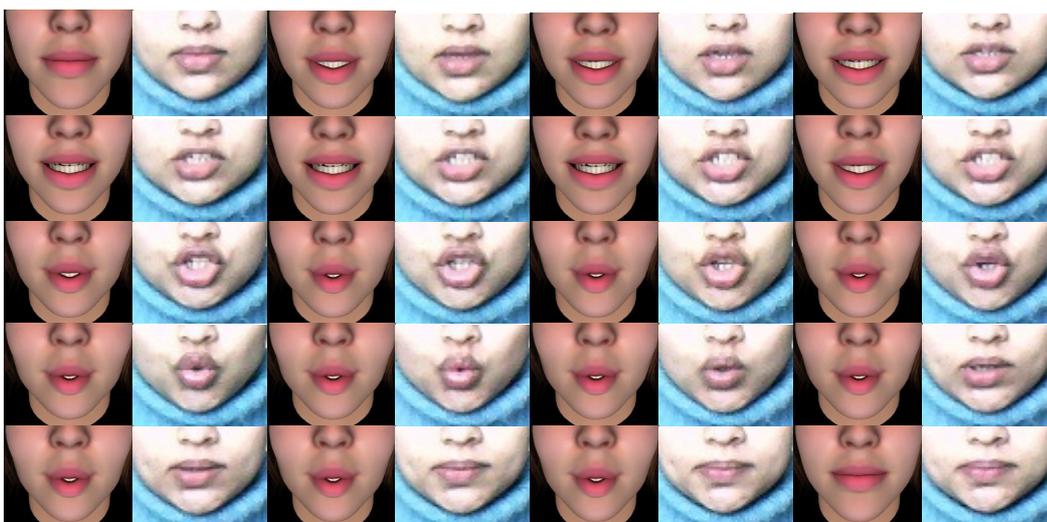


Figure 4.13: Animation and Video frames for "stew"

## 4.8 Synchronisation between audio and video

Synchronisation between audio and video was achieved by using the audio playback loop to determine which frame to display at each time step. The playback offset time is updated by the audio device every 25 milliseconds, so the maximum possible frame rate for playback is 40 FPS. A supersampling technique was employed to smooth the animation. Frames were generated at double the frame rate (80 FPS), averaged in pairs, and the averaged frames were played at 40 FPS.

The frame rate of 40 FPS did not always give smooth animation, so a smoothing filter was applied to the animation. The method used for smoothing the series  $x_t$  was to calculate a weighted moving average by first choosing a set of weighting factors (Equation 4.10):

$$[w_1, w_2, \dots, w_k] \text{ such that } \sum_{n=1}^k w_n = 1$$

Equation 4.10

These weights are used to calculate the smoothed statistics  $s_t$  (Equation 4.11):

$$s_t = \sum_{n=1}^k w_n x_{t+1-n} = w_1 x_t + w_2 x_{t-1} + \dots + w_k x_{t-k+1}$$

Equation 4.11

Two filters were tried: a three-value filter with the weights [1, 2, 1], and a five-value filter with the weights [1, 3, 4, 3, 1]. The five-value filter was found to give smoother animation than the three-value filter, and did not cause perceptible blurring between speech segments.

A graph display window was implemented to display vertex values of animation frames over time. This was used to view the smoothness of the trajectories. Figure 4.14 shows an example for one vertex, at the centre of the bottom lip edge. The graph plots the vertex coordinates against time (in milliseconds), creating three trajectories: Z at the top, X below, and Y at the bottom.

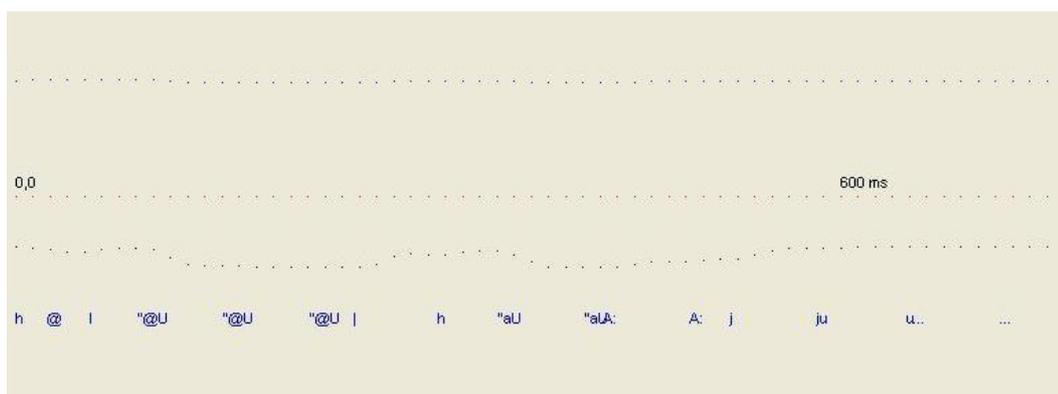


Figure 4.14: Trajectory for sentence “Hello, how are you?”

## 4.9 Expression modelling

Facial expressions, such as blinking, eye movements and smiling, were modelled using Facegen morph targets (Figure 4.12). Head movements, eye movements, and blinking were displayed when the head was idle, with sequences based on Pelachaud’s timings (Pelachaud 1991).



Figure 4.12: Facegen meshes for facial expressions

### 4.9.1 Photo-based viseme-driven talking head (THVP)

A second viseme-driven head (THVP) was created with a more photo-realistic external appearance than the previous viseme-driven head (THVN). Facegen was used with three photographs from the video corpus, taken from three angles (Figure 4.13), with feature points including the mouth corners, nose tip and chin marked on each photograph, to create a head based on the photographs (Figure 4.14). Internal visemes (Figure 4.15) were created as described in Section 4.2.1. Audio recordings of the speaker, taken from the video corpus, replaced the synthetic speech.



Figure 4.13: Photos used to create photo-based viseme-driven head



Figure 4.14: Photo-based viseme-driven head (THVP)



Figure 4.15: Internal view of photo-based viseme-driven head (THVP)

## 5 Development of Data-Driven Talking Head

At the "Département Parole et Cognition", GIPSA-Lab, Grenoble, a collaborative project was undertaken to acquire corpuses of speech using various articulatory measurement methods, and to build a set of articulatory models for a British English talking head. The process involved capturing magnetic-resonance imaging (MRI) scans to create static models of internal articulators during speech, electromagnetic articulography (EMA) recording to capture dynamic motion of the articulators, and 3D video of the face, tracked on multiple cameras, to create a model of the speaker's head and a corpus of audio-visual speech of English speech.

### 5.1 Data collection

The first data capture session involved three hours of 2D MRI recording, to create mid-sagittal scans (Figure 5.1) of the speaker's articulation of a set of vowel and consonant contexts, to be used to create a tongue model in the mid-sagittal plane. Next, in a six-hour session of video capture on two cameras, three hours of video data were recorded, consisting of three views of the speaker's head with facial markers for motion tracking. The video corpus included vowel and consonant sequences, English words and sentences, and a set of facial expressions and neck movements. Four hours were spent recording EMA data, with magnetic coils for motion tracking attached to the speaker's tongue, teeth and nose, with the same corpus content as for video. Finally a 3D MRI corpus was recorded in a two-hour session, creating a set of slices through the whole of the head, for a set of vowel and consonant articulations.

#### 5.1.1 2D MRI

2D MRI data was captured on a 3 Tesla MRI scanner, to be used to create a tongue model in the mid-sagittal plane. A corpus was designed to cover the maximal range of English articulations that the speaker could utter. The corpus consisted of vowels and vowel-consonant-vowel combinations (VCVs). The sounds needed to be sustained for eight seconds of MRI capture, so long vowel

sounds were chosen for the VCVs because they were more sustainable than short vowels. Since the MRI images would not show differences in voicing, only unvoiced consonants were recorded.

12 vowels were chosen, to cover most of the long vowel sounds in the English language. These were combined with the 12 unvoiced consonants of the English language, to create 72 articulations.

- Vowels: i: in beep; ɑ: in calm; ɔ: in cork; u: in fool; ɜ: in burn; ɪ in sit; e in less; æ in apple; ɒ in fond; ʌ in come; U in full; ə in above
- VCVs (6 vowels \* 12 consonants = 72 articulations)  
[ɑ: e i: u: ɔ: ɜ:] \* [p t k f s θ ʃ l r m n ŋ]

E.g. ɑ:pa: ɑ:ta: ɑ:ka: ɑ:fa: ɑ:θɑ: ɑ:sa: ɑ:ʃɑ: ɑ:la: ɑ:ra: ɑ:ma: ɑ:na: ɑ:ŋɑ:

Reference scans were also taken with the incisors in contact, and a dental cast of the speaker's teeth was also scanned in the mid-sagittal plane, to be used to help model the teeth.



Figure 5.1: MRI scan for articulation of vowel /e/

### 5.1.2 Video Capture on Three Cameras

The video capture session used three cameras to provide three different views of the speaker. The subject's face was marked with up to 168 coloured beads glued to the mouth, jaw, nose, cheek, neck and eyebrow areas. First a small subset of the video corpus was recorded with a small number of facial markers (Figure 5.2). Then a larger corpus was recorded with the full set of markers (Figure 5.3). The corpus with the full set of markers would look less natural, but was mandatory to recover a full 3D surface, and would lead to more accurate visual speech synthesis.

The corpus for video consisted of the following:

- MOCHA TIMIT corpus sentences (Wrench 1999)
- VCVs: as for MRI, plus voiced consonants ( $6 \times 25 = 150$ )  
[ɑ: e i: u: ɔ: ɜ:] x [p b t d k g f v θ ð s z ʃ ʒ ʒ l r m n ŋ w j h x]
- Phrases specifically needed for the tutoring application (Appendix C)
- Modified Rhyme Test words (Meyer Sound 2010) (Appendix D)

The dynamic corpus with the small set of 40 markers included the following:

- Small set of vowels: ɑ: in calm, ii: in beep, uu: (u) in fool, ɪ in sit, e in less, ɒ in fond, ə in above
- Small set of VCVs, for the most extreme vowel articulations: [ɑ: i: u:] \* [p b t k f θ s ʃ l r m n]
- Phrases specifically needed for tutoring application (48 phrases)
- MOCHA-TIMIT subset (Appendix C)
- Modified Rhyme Test words (50 tuples) (Appendix D)

The dynamic corpus with the full set of 168 markers included the following:

- 48 phrases for tutoring application (Appendix C)
- MOCHA-TIMIT corpus (460 sentences)
- Modified Rhyme Test words (50 tuples) (Appendix D)

- A full set of 72 VCVs
- A set of facial expressions: neutral, closed smile, open smile, blink, look up, look down, look left, look right
- A set of neck movements: turn left/right, up/down, tilt left/right, forward/backward, shift left/right, lift/release

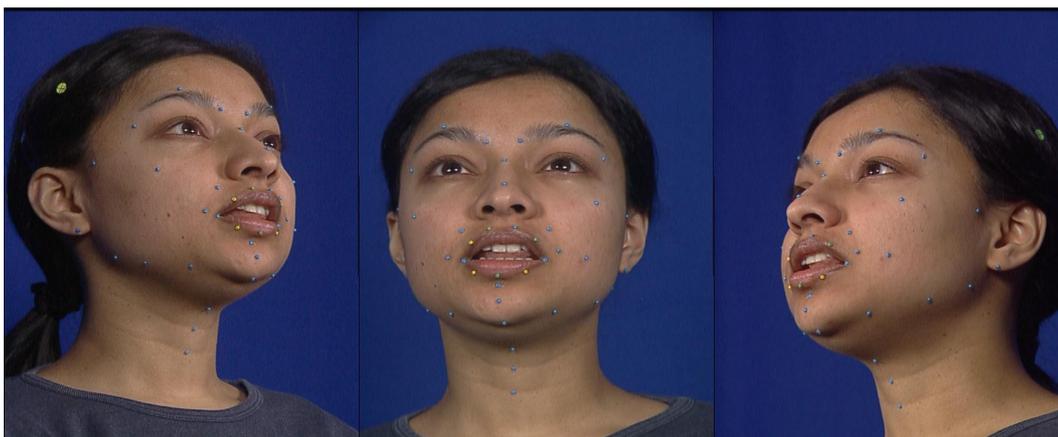


Figure 5.2: Video capture with small set of 40 facial markers

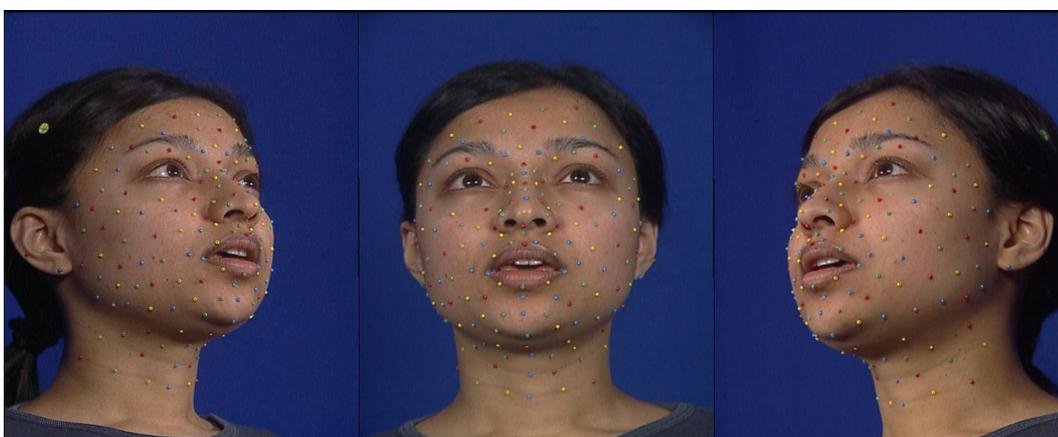


Figure 5.3: Video capture with full set of 168 facial markers

### 5.1.3 EMA corpus

For EMA recording, 6 coils were attached to the speaker's articulators (Figure 5.4):

1. Reference point on bridge of nose, at the point which does not move during speech

2. Upper incisor
3. Lower incisor
4. 0.5 cm from tip of tongue
5. Tongue blade (middle of tongue)
6. Tongue dorsum (back of tongue)

The corpus for EMA was the same as for the video corpus. While the EMA was recorded, video was also recorded with a small set of markers on the facial articulators.



Figure 5.4: EMA recording

#### 5.1.4 3D MRI corpus

For 3D MRI capture, first it was necessary to find machine settings that would give images of sufficient quality within a reasonable recording time. Neutral-pose 3D MRI images taken at various settings were viewed using ImageJ imaging software, to find the setting which produced the best quality images. Eventually a 3D MRI corpus was recorded with articulations sustained for 13.6 seconds. The corpus was a subset of the 2D MRI corpus, with VCVs for the most extreme vowel articulations.

- Vowels: i: (i) in beep, a: (a) in calm, ɔ: in cork, u: in fool, ɜ: in burn, ɪ in sit, e in less, æ in apple, ɒ in fond, ʌ in come, ʊ in full, ə (q) in above
- VCVs: [a: i: u:] \* [p t k f θ s ʃ l r m n ŋ]

## 5.2 Internal Articulatory Modelling

The captured data was processed in order to build models at GIPSA-Lab. The contours of the rigid bony structures involved in the vocal tract (jaw, hard palate, nasal passages, nostrils and sinuses) and the deformable structures (tongue, velum, nasopharyngeal wall) were manually registered in the 2D mid-sagittal MRI images, using a program in Matlab to trace the contours as planar B-spline curves controlled by a limited number of points (Figures 5.5 - 5.6). This MRI data, combined with the EMA data, was used by GIPSA-Lab to make a first data-driven tongue model in the mid-sagittal plane, using a linear modelling approach, involving guided PCA, where *a priori* knowledge was introduced during the linear decomposition (Badin et al. 2006). A first inversion model was created to map the EMA data to articulatory parameters, and then to 2D contours. A limitation of this model was that the back of the tongue was not realistically modelled, due to the EMA coil being too far forward. Further work would be needed to process the 3D MRI data, in order to make a full 3D tongue model, and visualize this in a 3D talking head (Badin et al. 2008).

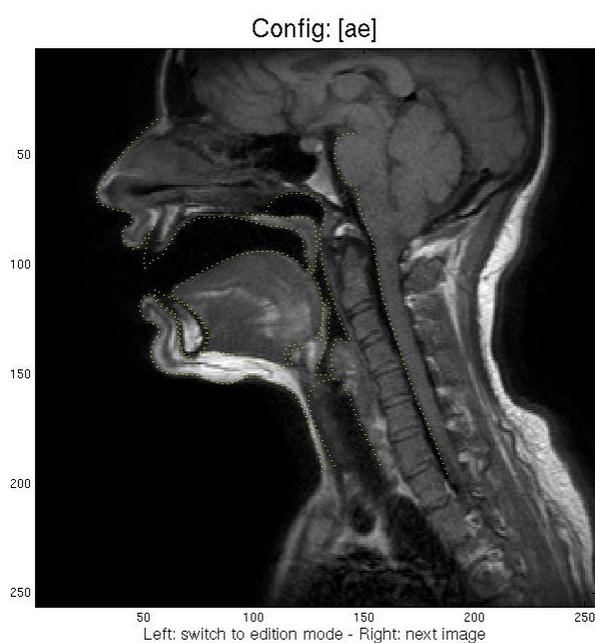


Figure 5.5: MRI contour tracing

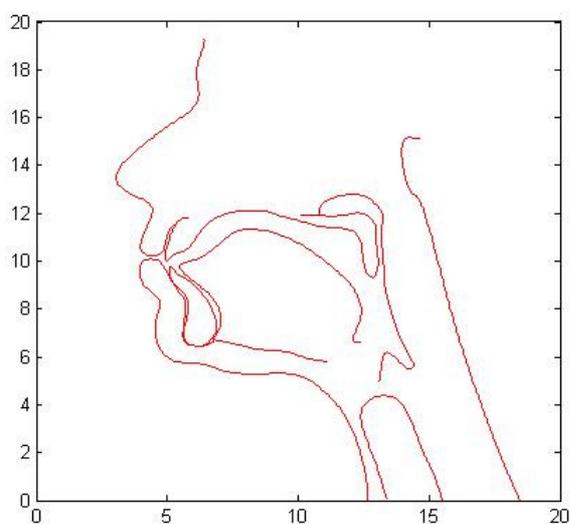


Figure 5.6: MRI contours

### 5.3 External Articulatory Modelling

The audio and video recordings were annotated to identify the vowel and consonant articulations, and the facial marker data was processed in order to extract motion sequences. These were used by GIPSA-Lab to create a data-driven model of the external articulators (Badin et al. 2002) (Figure 5.8). The 3D

movements of facial points were acquired using multicamera photogrammetry (Elisei et al. 2001) (Figure 5.7). A shape model was built using guided PCA where a priori knowledge was introduced during the linear decomposition. This allowed the extraction of six components related to speech movements: The first controlled the opening/closing movement of the jaw; three parameters related to the lips: one controlled a protrusion/spreading movement of both; another controlled the upper lip raising/lowering movement; another controlled the lower lip lowering/raising; the second jaw parameter was associated with a horizontal forward/backward movement of the jaw; and the sixth parameter was related to the vertical movements of the larynx (Bailly et al. 2009).

The shape model of the facial movements was then used to guide a multi-view tracker of the beads using correlation-based techniques. The initial shape model helped to constrain the search space within regions of interest for each vertex of the facial mesh. Automatic tracking of the beads was combined with semiautomatic correction (Bailly et al. 2006). Visemes were selected from the video corpus, and the most salient frames were precisely marked by hand, adding any untracked beads, for example, beads at the sides of the head which could disappear from some of the views. The 3D data was supplemented by lip geometry that was acquired by semi-automatically fitting a parametric lip model to the speaker-specific anatomy and articulation (Bailly et al. 2009).

To achieve a photorealistic appearance, an appearance model was used for computing the colour of each pixel of the face. Selected images of all configurations used for estimating the shape model were warped to a neutral shape, to give shape-free images (Bérar et al. 2003). Linear regression was used to relate the RGB colours of each pixel of the shape-free images to shape parameters. The texture model computes texture maps, which are extracted and blended according to articulatory parameters, so the RGB values of each pixel vary with the values of the articulatory parameters. The texture maps are computed in three steps. First, the shape model is used to track articulations marked by a small subset of beads (Section 5.1.2), for one target image per allophone. Next, shape-free images are extracted by warping the selected images to a neutral shape. The third step is the linear regression of the RGB values of all

visible pixels of the shape-free images, by the values of articulatory parameters obtained in the first step. The speaker-specific shape and appearance models are thus driven by the same articulatory parameters (Bailly et al. 2009). The resulting talking head is a data-driven facial clone of the speaker (Figure 5.9). In the resulting animation frames, some black areas were visible on the lower teeth when the mouth was open, so this could affect the realism of some animations. This was due to these areas appearing and disappearing with the opening and closing of the mouth. To improve the appearance of the inner mouth, precise prediction of the jaw position and tongue position would be required in order to capture changes of appearance due to speech articulation (Bailly et al. 2009).



Figure 5.7: Facial geometric mesh

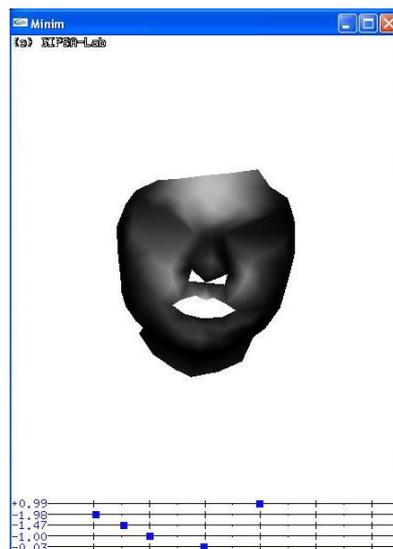


Figure 5.8: Articulatory Model

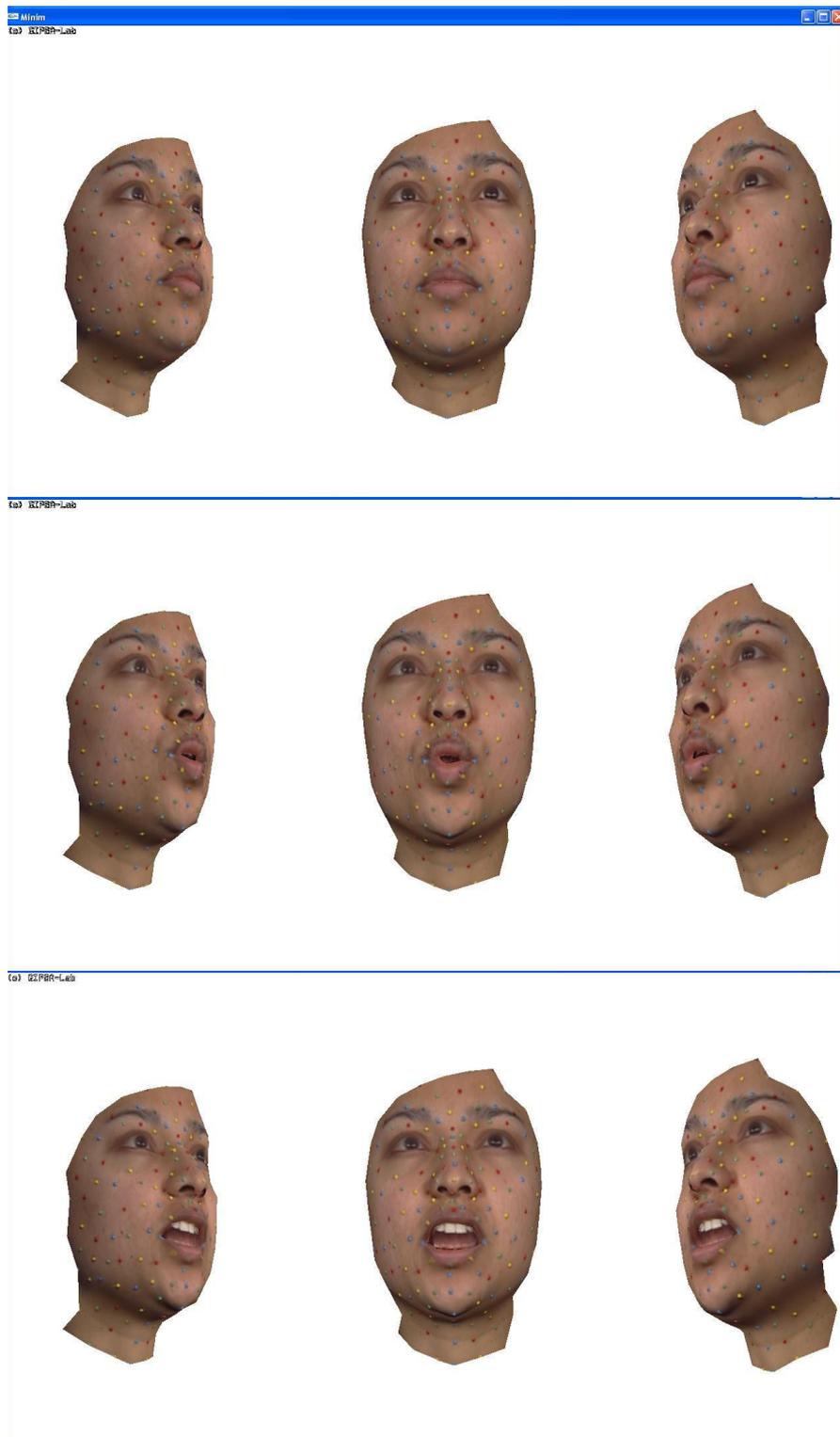


Figure 5.9: Data-driven talking head

## 6 Evaluation of Quality of Visual Speech

The three talking heads described in Chapters 4 and 5 needed to be evaluated to assess the quality of the visual speech, to ensure that it was suitable for the task of pronunciation training. This chapter describes how the visual speech was evaluated using subjective user tests for intelligibility and naturalness (Dey et al. 2010a).

### 6.1 Evaluation Approaches

There are currently no standardised evaluation procedures for visual speech. Approaches can be objective, using algorithmic metrics, or subjective, using human participants. A curve of the motion of a point on an animated face can be objectively compared against a curve obtained from motion capture of a real speaker, using a dynamic time warping algorithm for measuring similarity (Salvador et al. 2004).

One subjective method is a Turing test, where viewers are asked whether an animation is real or synthetic, but this test can only be applied to video-realistic talking heads (Theobald et al. 2008). A possible solution is to place markers on both the real face and the synthetic face, and show the viewers the markers only, so that they will compare the motion and not the rendering. Perceptual tests were developed by Cosker (Cosker et al. 2005), who played animations dubbed with an audio word different from that of the video, and asked participants which word they heard. This tested whether the lip-synchronisation was effective enough to confuse the response, due to the “McGurk Effect” (McGurk et al. 1976). This approach can be used for non video-realistic talking heads, but it only uses single words so it cannot test coarticulation effects, although it could be extended to use sentences. Ouni et al. (Ouni et al. 2007) evaluated the intelligibility of the Baldi talking head by using a visual contribution metric (Sumbly et al. 1954) to measure the benefit to intelligibility provided by the synthetic animated face relative to the benefit provided by a natural face. This method quantified that Baldi was 93% as accurate as a real face. Ouni et al also found that for a natural face, the lips alone

were almost as effective as the full face in contributing to intelligibility, but for the synthetic face the lips alone were much less effective for intelligibility.

## 6.2 Intelligibility Test 1 (THVN): Modified Rhyme Test

The intelligibility of the viseme-driven non-photorealistic talking head (THVN) was evaluated subjectively using a Modified Rhyme Test (MRT), an ANSI standard test for statistical intelligibility testing (Meyer Sound 2010). The MRT approach was previously used by Fagel to evaluate the intelligibility of a talking head (Fagel 2008). Fagel found the word recognition rate to be 27% for audio alone, and 50% for audiovisual speech. The MRT used 50 six-word lists of monosyllabic English words, and the words in each list differed only in the initial or final consonant sound, e.g. "*shop, mop, cop, top, hop, pop*" (Appendix D).

32 participants with normal hearing and vision were tested individually in an acoustically-isolated booth, with visual images presented on a 15 inch computer screen and acoustic stimuli presented binaurally over headphones. In each trial, participants were shown a six-word list and asked to identify which word was spoken. Responses were scored as the number of words identified correctly. 20 words were presented for each of 3 conditions:

- degraded synthetic audio speech alone
- an external view of the talking head (THVN) with degraded synthetic audio speech
- video of a real person with degraded audio

Different words were used for the 3 different conditions, in order to minimize learning effects (Appendix D). In order to minimize sequence effects, the order of presentation was randomized. The audio was degraded by adding speech-shaped noise to the acoustic signal. First, the long-term average speech spectrum (LTASS) of the speech waveform was computed, and from this a finite impulse response (FIR) filter was constructed. Gaussian white noise was generated and convolved with the filter. The resulting speech-shaped noise was then added to the original speech (Assmann 2010). Speech shaped-noise has a similar effect to the masking produced by multiple speakers speaking at the same time, and is more

suitable than white noise for speech perception tests, as it simulates real life situations.

The noise levels were chosen within a range in which the words were barely recognizable, after preliminary tests on one listener; below -20 dB word recognition for audio alone fell to chance levels (16%), while above -16 dB word recognition for natural video became close to optimal. For 16 participants, the SNR was set to -18 dB. For the remaining 16 participants, all words for all three conditions were presented at an SNR of -20 dB, and then repeated at -16 dB.

### 6.2.1 Results of Intelligibility Experiment 1

Visualization improved the intelligibility of the speech at all three SNRs (Table 6.1 and Figure 6.1). The word recognition rate was higher for the audiovisual heads than for audio alone, and higher for the natural head than the synthetic head. ANOVA over all conditions shows that the gain in intelligibility due to visualization is highly significant ( $p = 0.01$ ) at each SNR. In post-hoc t-tests, at SNR -16 dB, the natural head was significantly more intelligible than audio alone ( $p = 0.1$ ). Post-hoc T-tests ( $p = 0.1$ ) found no other significant differences between any other conditions at any SNR.

In Table 6.1, Table 6.2 and Figure 6.1 there are two groups of participants, with a different group for SNR -18 dB, which explains the slightly lower intelligibility for the synthetic head at SNR -18 dB, compared to SNR -20 dB.

At the lowest SNR the recognition rate for natural video was only slightly higher than the synthetic head. At SNR -20 dB the improvement in word recognition due to the visualization in the audiovisual synthetic head, calculated using a normalized measure (Sumby et al. 1954), was 39.5%, while the improvement due to the natural head was 39.9% (Table 6.2). The visual contribution of the synthetic face relative to the natural face was not invariant as found by (Ouni et al. 2007), but was higher for the lower SNR of -20 dB, compared to -16 dB. The benefit of visual speech relative to audio alone was higher for the lower SNR (-20 dB, compared to -16 dB), a finding consistent with that of Benoit (Benoît et al. 1998), who found that the poorer the auditory scores the greater the benefit of lip-

reading. At a lower SNR the audio alone is less intelligible so listeners rely more on lip movements to decide which word was said.

Mean % words correctly identified			
audio alone	Synthetic THVN	natural	SNR (dB)
30.3	57.8	58.1	-20
47.8	55.6	63.4	-18
56.9	67.2	86.6	-16

Table 6.1: Intelligibility Test 1: Mean % words correctly identified

Visual contribution to intelligibility (%)		
Synthetic THVN	natural	SNR (dB)
39.5	39.9	-20
15.0	29.9	-18
23.9	68.8	-16

Table 6.2: Intelligibility Test 1: Visual contribution to intelligibility

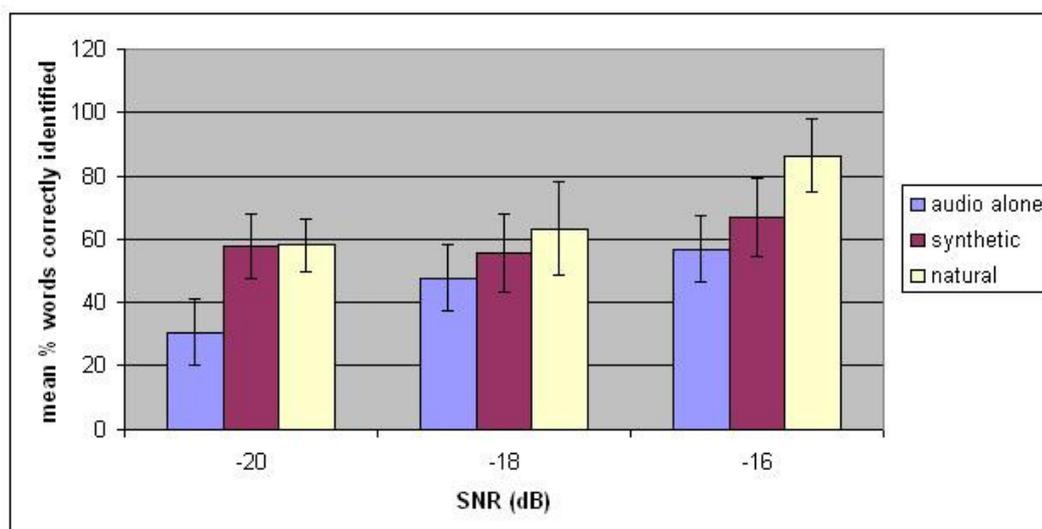


Figure 6.1: Intelligibility scores. The error bars denote the standard deviation.

The confusion matrix for the synthetic head THVN (Figure 6.2) compares the visemes presented against the visemes they were perceived as by the participants. The number of identifications was summed over all participants, for all words spoken by the synthetic talking head, at all SNRs. Each sum was divided by the number of occurrences of the animated viseme, to give a percentage of identifications of that viseme. The area of each circle represents the percentage of identifications of that viseme. For example, viseme 6 (/r/-l/), was mistaken for viseme 5 (/h/-n/-ng/), as often as it was identified correctly. The two visemes look similar from the outside, and the tongue modelling may have been insufficiently accurate to allow discrimination between visemes 5 and 6. Also, in a real speaker there is articulatory movement at the base of the tongue which is visible below the jaw when pronouncing /n/, which was not modelled in the synthetic head. On the whole, the matrix shows that the correct classifications (on the diagonal) scored the highest, so overall the visemes were identifiable.

For the natural head, the confusion matrix shows that the visemes /h/-n/-ng/ and /g/-k/ were less well identified than other visemes (Figure 6.3). This may be because the tongue movements that distinguish these visemes from others were less visible from the external view. This indicates a limitation of the viseme classification: “there are some phones that do not require the use of the visual articulators, and so phonemes such as /k/ or /g/, which are velar consonants articulated at the back of the soft palate, are unlikely to have an associated viseme” (Hilder et al. 2010).

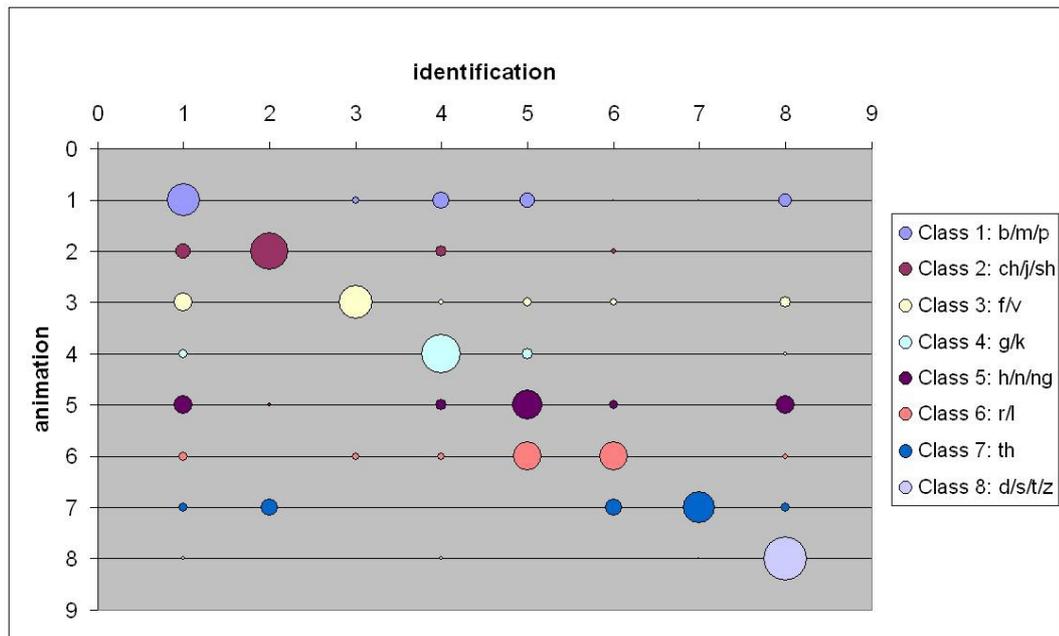


Figure 6.2: Confusion Matrix for Synthetic Talking Head

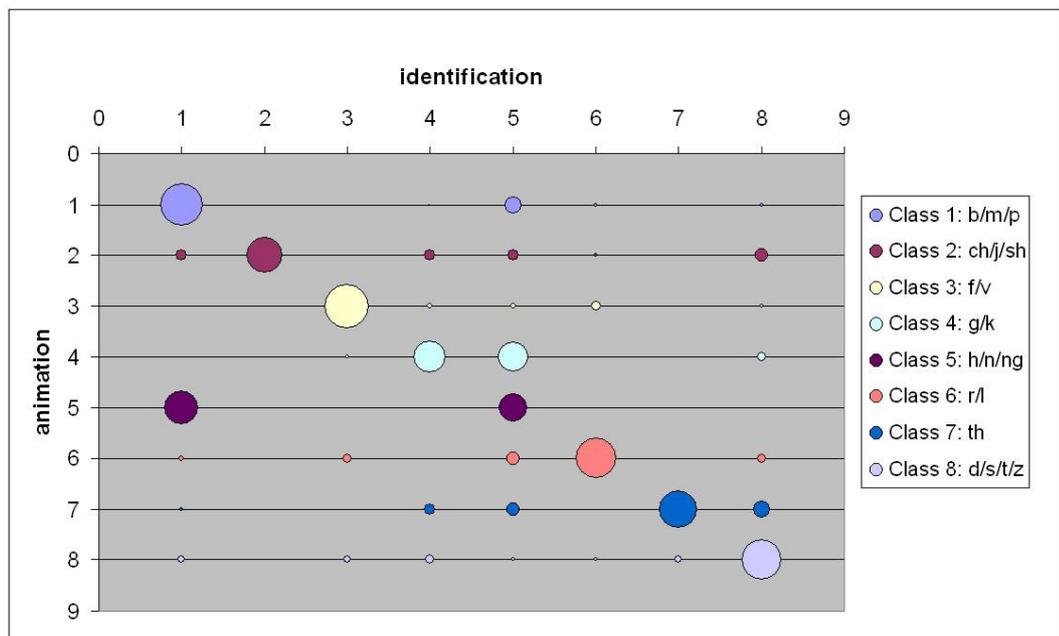


Figure 6.3: Confusion Matrix for Natural Head

Figure 6.4 shows the confusion matrix of the synthetic head minus that of the natural head. This highlights the differences between the two heads and shows the weaknesses of the synthesised model. For example, visemes 5 (/h/-/n/-/ng/) and 6 (/r/-/l/) had high confusions compared with the natural head, and could be more accurately modelled.

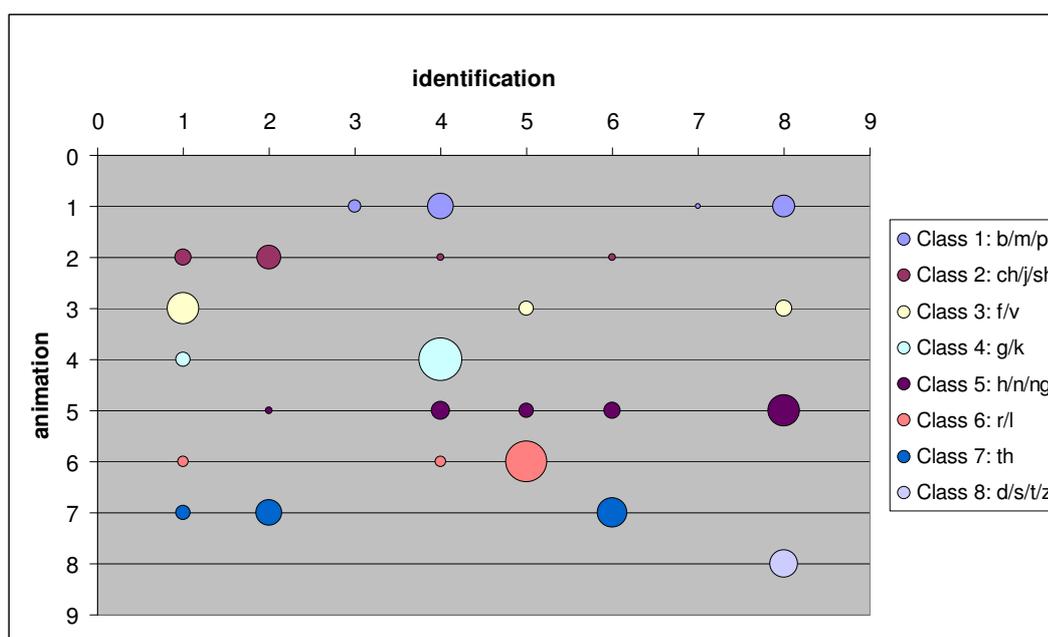


Figure 6.4: Difference between Synthetic Head and Natural Head

### 6.3 Subjective test for naturalness (THVN)

The naturalness of the talking head THVN was evaluated using subjective quality assessment (Theobald et al. 2008). Users were asked to rate the naturalness of the visual speech along a five point Likert scale (Likert 1932). After undertaking the intelligibility test at an SNR of -18 dB, 16 participants were presented with the synthetic talking head, for 20 isolated words with no audio degradation, and were asked to rate the naturalness of the visual speech along a 5 point scale, with 1 for "very unnatural" and 5 for "very natural".

The naturalness scores for the synthetic talking head were, on average across all words and all participants, 3.5 on a scale of 1 to 5 (s.d. 1.0), so the visual speech was rated as "moderately natural" overall, but for some sounds the animation could be more realistic (Figure 6.5). The word which scored lowest, "*duck*", has little external mouth movement compared to "*hop*", which scored highest, so this may be a factor in the ratings for the animation.

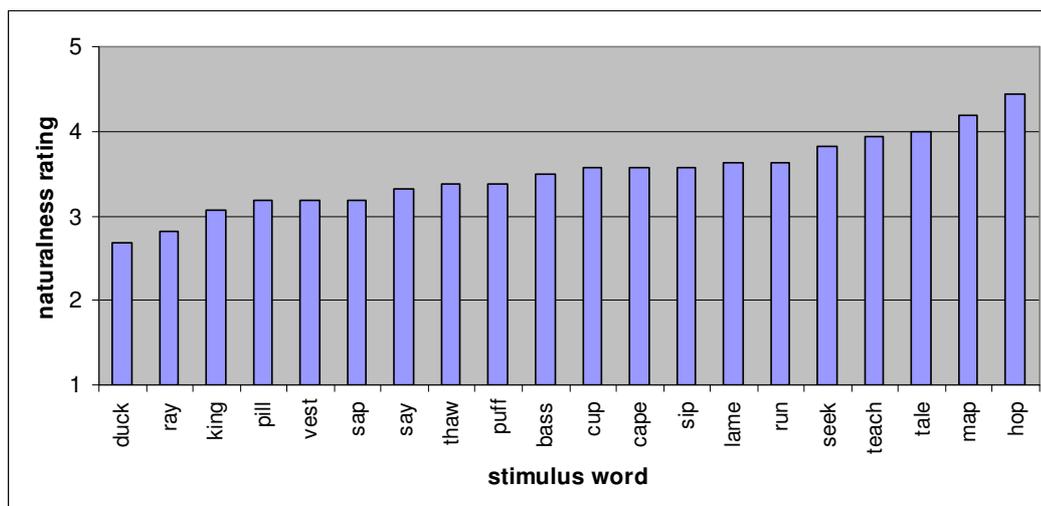


Figure 6.5: Naturalness Ratings

## 6.4 Subjective Evaluations of Naturalness of Talking Heads THVN and THD

Participants were asked to rate the naturalness of talking heads THVN and THD in two online surveys. The first survey was a pilot on a subset of conditions, and the second survey used the full range of conditions. The same 2 sentences from the MOCHA corpus (Wrench 1999), were used for each, chosen to cover a range of sounds:

1. *“Birthday parties have cupcakes and ice-cream.”*
2. *“He will allow a rare lie.”*

The first sentence covered /b/, /p/ and /k/ visemes while the second included /r/ and /l/. The first sentence had 10 syllables, while the second sentence had 7 syllables. These differences may have had effect on the perceived naturalness, so the two sentences were considered separately in the analysis.

### 6.4.1 Online Survey 1: Pilot Evaluation of Naturalness

For the first survey on 6 participants, four conditions were presented for each sentence:

1. The viseme-driven non-photo-based head without eye and head movements (rigid viseme-driven head), with a synthetic voice (THVN)
2. The viseme-driven non-photo-based head with eye and head movements (expressive viseme-driven head) with a synthetic voice (THVN + expression)
3. The data-driven head, with a natural voice (THD)
4. Natural video, with a natural voice

These were the original audiovisual conditions of the talking heads when created; i.e. the viseme-driven heads were presented with a synthetic voice, while the data-driven head and the natural video were presented with the natural voice. The synthetic voice was the “Kate” voice from Loquendo TTS (Section 4.3.4), while the natural voice recordings were taken from the video corpus (Chapter 5). The eye and head movements were created using Facegen morph targets (Section 4.8).

Participants were asked to rate the naturalness of the lip movements on a Likert scale of 1-5, with 5 being the most natural.

A formula was used to convert each Likert score to a percentage, which treated each Likert value as the mid-range value in the possible range of corresponding percentage values (Equation 6.1).

$$\text{likert to percentage} = \frac{\text{likert value}}{\text{likert scale}} - \frac{1}{2 \times \text{likert scale}} \times 100$$

Equation 6.1

Table 6.3 and Figure 6.6 show that the viseme-driven talking head (THVN) was perceived as the least natural, and the eye blinks and head movements did not improve its perceived realism. The data-driven head was perceived as more natural than the viseme-driven conditions. The natural video was perceived as the most natural of all conditions.

The second sentence was rated as slightly more natural than the first, in all conditions except natural video. It is possible that the different visemes had an effect, since the first sentence had /b/, /p/ and /k/ visemes and the second sentence had /r/ and /l/ visemes. However, although the MRT intelligibility experiment had found /r/ and /l/ visemes to be less identifiable for the viseme-driven THVN head (Figure 6.4), this did not seem to affect the perceived naturalness here for the second sentence. The first sentence had 10 syllables, while the second sentence had 7 syllables, so the longer sentence length may have had an effect, because users had more time to notice flaws in the modelling so they perceived the longer sentence as less natural. For the expressive talking head (THVN + expression), both sentences were presented with a blink at the start and the end of the sentence, so for the first sentence there was a longer gap between blinks, which may have been perceived as less natural.

Visual condition	Auditory condition	Mean response for each sentence	Mean for the pair of sentences	Likert score to Percentage (%)
rigid viseme-driven head (THVN)	synthetic voice	3.2	3.5	60
		3.8		
expressive viseme-driven head (THVN + expression)	synthetic voice	3	3.4	58
		3.8		
data-driven head (THD)	natural voice	3.7	3.75	65
		3.8		
natural video	natural voice	4.8	4.75	85
		4.7		

Table 6.3: Results of Online Naturalness Survey 1

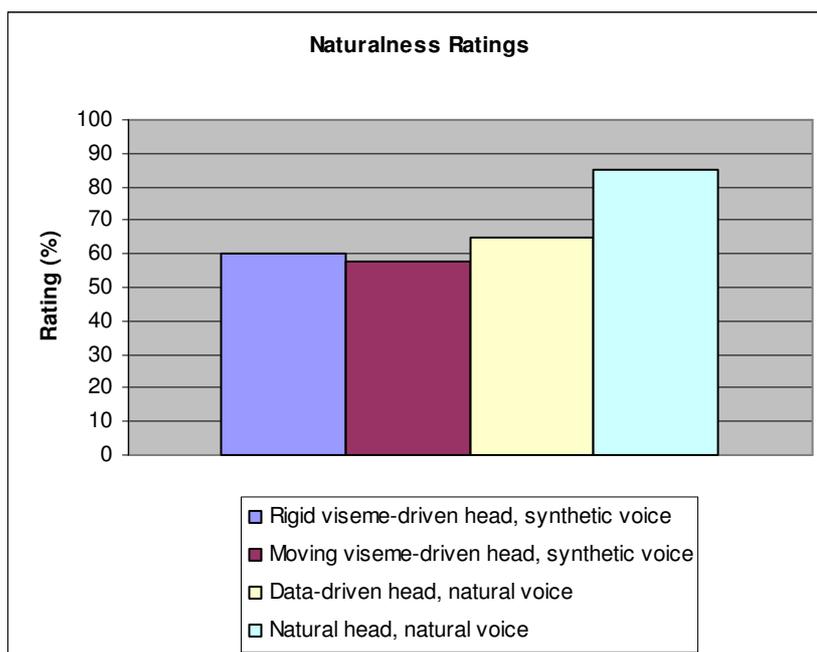


Figure 6.6: Naturalness Ratings in Online Survey 1

### 6.4.2 Online Survey 2: Further Evaluation of Naturalness

For the second survey on 10 participants, eight conditions were presented for each sentence, which included all the audiovisual combinations, i.e. the viseme-driven heads (THVN) were aligned with the natural voice, and the data-driven head (THD) and the natural video were aligned with the synthetic voice. For this survey the Likert scale used was 1-7, to give more precision than a five-point scale.

In this survey, the mean responses showed that the first sentence was rated as more natural than the second in some conditions, such as THVN, but the second sentence was rated as more natural for other conditions, such as THD. This suggests that the differences between the two sentences, such as the number of syllables, did not have much effect on the perceived naturalness. The modal responses showed that the added expressivity slightly improved the naturalness of the viseme-driven head in some conditions, i.e. for the second sentence with the synthetic voice, and the first sentence with the natural voice (Table 6.4 and Figure 6.7). However, the mean results were lower for the expressive head.

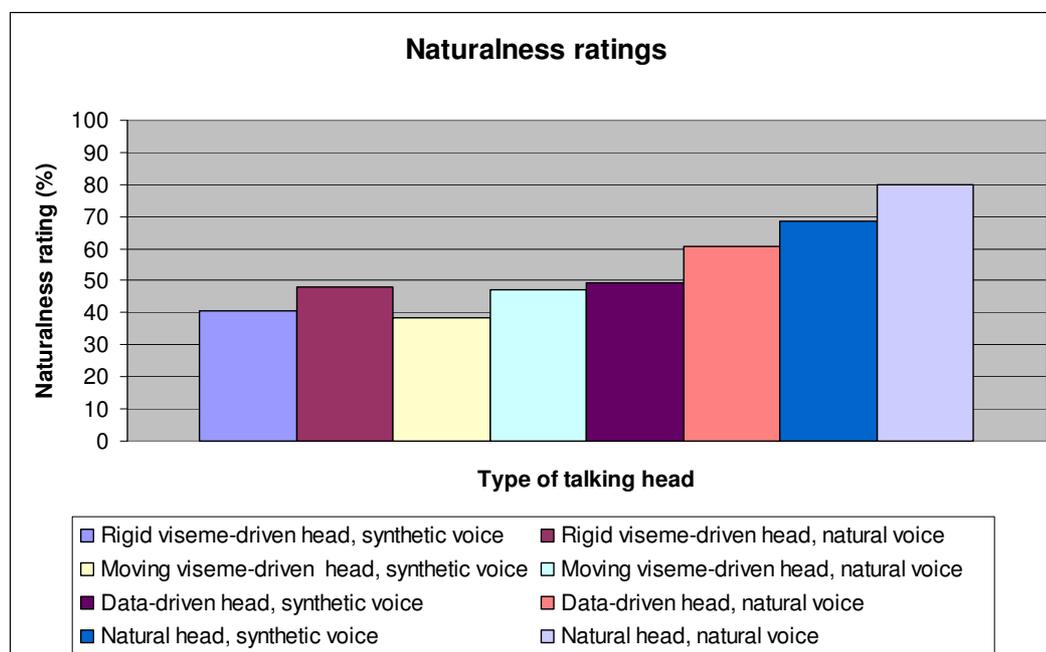


Figure 6.7: Naturalness Ratings in Online Survey 2

Visual condition	Auditory condition	Modal Likert response for each sentence	Mean Likert response for each sentence	Mean Likert response for the pair of sentences	Mean Likert score to Percentage (%)
Rigid viseme-driven head (THVN)	synthetic voice	2, 4	3.8	3.35	40.7
		2	2.9		
Rigid viseme-driven head (THVN)	natural voice	3	4.1	3.85	47.9
		2, 4	3.6		
Expressive viseme-driven head (THVN + expression)	synthetic voice	3	3.2	3.2	38.6
		3	3.2		
Expressive viseme-driven head (THVN + expression)	natural voice	4	4.1	3.8	47.1
		2	3.5		
Data-driven head (THD)	synthetic voice	3	3.8	3.95	49.3
		5	4.1		
Data-driven head (THD)	natural voice	3, 5	4.7	4.75	60.7
		4, 6	4.8		
Natural video	synthetic voice	7	5.7	5.3	68.6
		6	4.9		
Natural video	natural voice	7	6.2	6.1	80
		7	6		

Table 6.4: Results of Online Naturalness Survey 2

The results of both surveys showed that natural speech was always perceived as more realistic than synthetic speech, for each visual condition. The real video with the natural voice was perceived as the most natural of all conditions. Contrary to expectation, the eye blinks and head movements did not improve the perceived realism of the viseme-driven talking head, which indicates that the modelling of the movements could be improved. The data-driven head was always perceived as more natural than the any of the viseme-driven conditions. A limitation of this comparison is that this viseme-driven head THVN was obviously not photorealistic, while the data-driven head THD was based on the same images as

the real video, so this would be likely to influence ratings, regardless of the quality of the speech animation. Consequently a more photo-realistic viseme-driven head (THVP), based on the same speaker, was compared against the data-driven head, as described in the next section.

## 6.5 Intelligibility Test 2: Modified Rhyme Test on 2 Talking Heads (THVP and THD)

The visual speech of the data-driven head (THD), was compared against the photo-based viseme-driven head (THVP), against audio alone, and against real video, in a word identification test. The MRT was used as before, but with 14 words presented for each of 4 conditions:

- degraded natural audio speech alone
- an external view of the viseme-driven talking head (THVP) with degraded natural audio speech
- an external view of the data-driven talking head (THD) with degraded natural audio speech
- video of a real speaker with degraded natural audio

Different words were used for the 4 different conditions, in order to minimize learning effects (Appendix D). In order to minimize sequence effects, the order of presentation was randomized.

The audio was degraded by adding speech-shaped noise to the acoustic signal (Assmann 2010). The noise levels were chosen within a range in which the words were barely recognizable; below -20 dB word recognition for audio alone fell to chance levels (16%), while above -16 dB word recognition for natural video became close to optimal. For 12 participants, all words for all four conditions were presented at an SNR of -20 dB, and then repeated at -16 dB.

### 6.5.1 Results of Intelligibility Test 2

The results of the intelligibility experiment on 12 participants are presented in Table 6.5. The visual contribution to intelligibility is given in Table 6.6.

Mean % words correctly identified				
SNR (dB)	Audio	Viseme-driven (THVP)	Data-driven (THD)	Natural video
-20	25.1	33.5	41.3	63.3
-16	35.9	48.3	52.6	67.4

Table 6.5: Intelligibility Test 2: Mean % words correctly identified

Visual contribution C = head score – audio score / 1 – audio score (%)			
SNR (dB)	Viseme-driven (THVP)	Data-driven (THD)	Natural video
-20	11.2	21.6	51.0
-16	19.3	26.1	49.1

Table 6.6: Intelligibility Test 2: Visual contribution to intelligibility

Visualization improved the intelligibility of the speech at both SNRs (Figure 6.8). The word recognition rate was higher for the audiovisual heads than for audio alone, higher for the data-driven head than the viseme-driven head, and higher for the natural head than the synthetic heads. At each SNR, ANOVA over all conditions shows that the gain in intelligibility from visualization is highly significant at  $p = 0.01$  (SNR -16 dB,  $F(3, 44) = 9.926$ ; SNR -20 dB,  $F(3, 44) = 16.554$ ). ANOVA over the audio condition and the synthetic heads shows that the gain in intelligibility from the synthetic heads, over audio alone, is significant, at both SNRs at  $\alpha=0.05$  (SNR -16 dB,  $F(2, 33) = 4.979$ ; SNR -20 dB,  $F(2, 33) = 3.781$ ). A post-hoc T-test shows that the real video is significantly more intelligible ( $p = 0.05$ ) than audio at SNR – 20 dB. Post-hoc T-tests ( $p = 0.1$ ) found no other significant differences between any other conditions at either SNR.

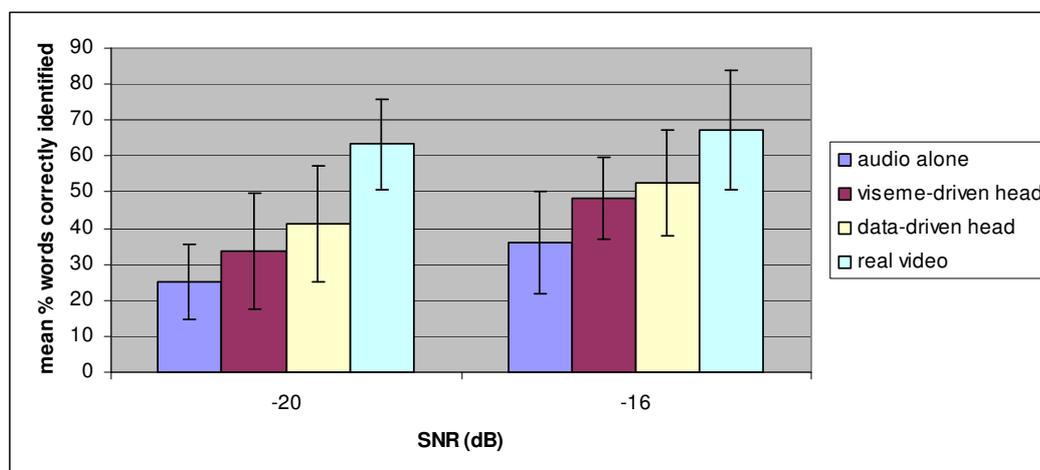


Figure 6.8: Intelligibility scores. The error bars denote the standard deviation.

Each confusion matrix (Figures 6.9 to 6.12) compares the visemes (or phonemes, in the case of audio alone) presented against what they were perceived as by the participants. The number of identifications was summed over all participants, for all words spoken, at both SNRs. Each sum was divided by the number of occurrences of the animated viseme, to give a percentage of identifications of that viseme. The area of each circle represents the percentage of identifications of that viseme. On the whole, for each condition, the matrix shows that the correct classifications (on the diagonal) scored the highest, so overall the visemes were identifiable. There was most confusion for the audio condition (Figure 6.9), followed by the viseme-driven head THVP (Figure 6.10), then the data-driven head THD (Figure 6.11), and then the natural video (Figure 6.12). This shows that the data-driven head THD was more intelligible and more accurately modelled than the viseme-driven head THVP.

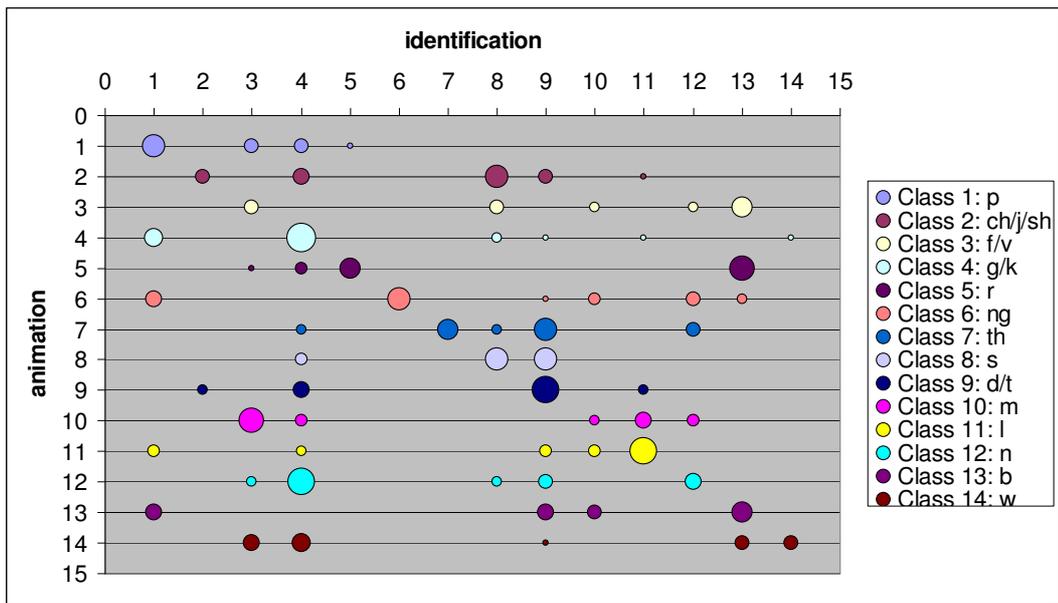


Figure 6.9: Confusion Matrix for audio alone

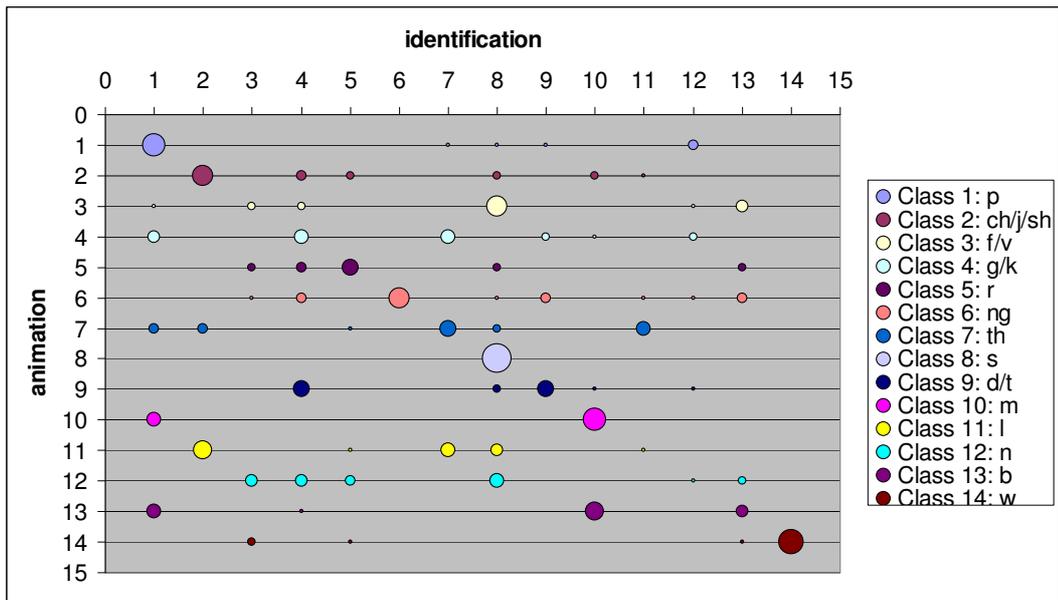


Figure 6.10: Confusion Matrix for Viseme-driven Talking Head

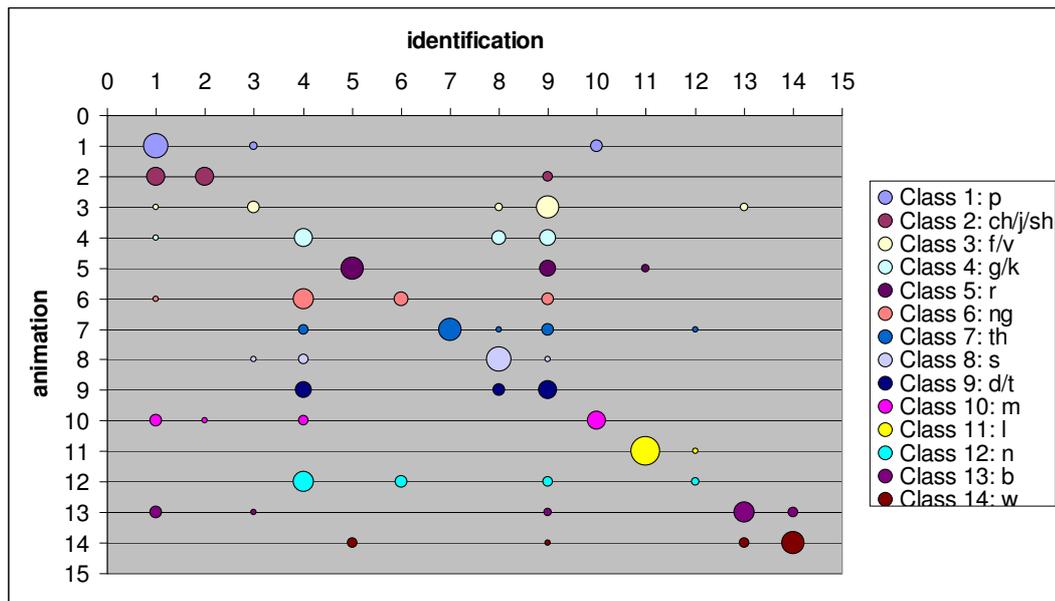


Figure 6.11: Confusion Matrix for Data-driven Talking Head

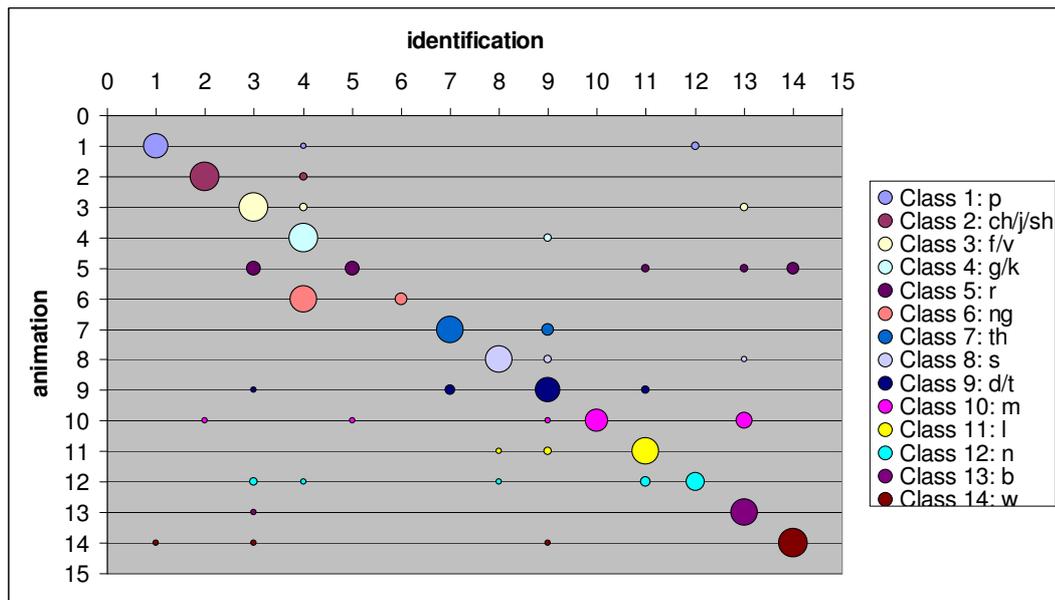


Figure 6.12: Confusion Matrix for Natural video

Figure 6.13 shows the confusion matrix of the viseme-driven head minus that of the natural head. This highlights the differences between the two heads and shows the weaknesses of the synthesised model. For example, visemes 11 (/l/) and 12 (/n/) had high confusions compared with the natural head, and could be more accurately modelled.

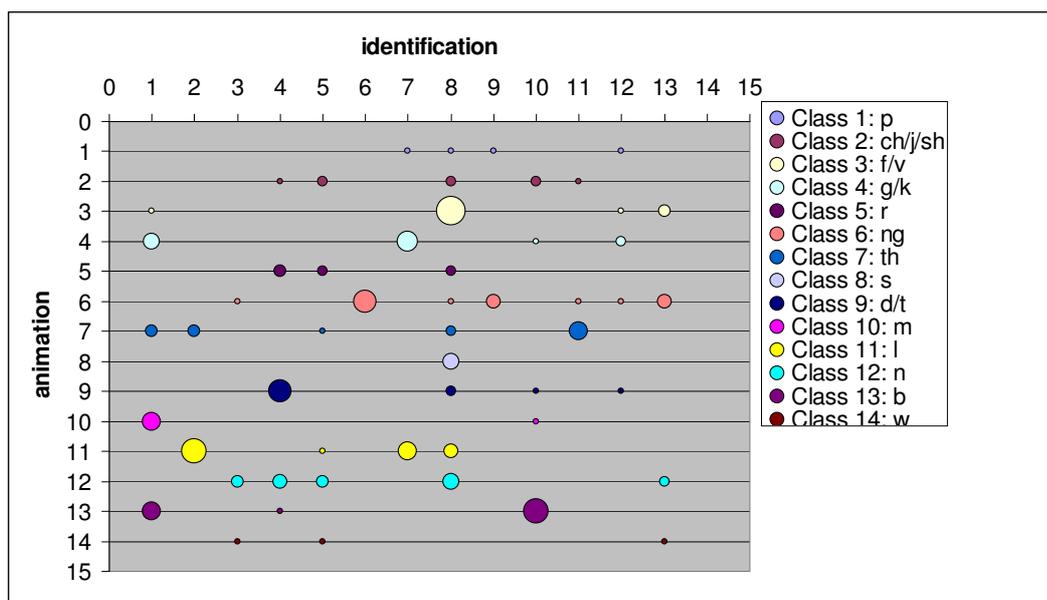


Figure 6.13: Difference between Viseme-driven Head and Natural Head

Figure 6.14 shows the confusion matrix of the data-driven head minus that of the natural head. This shows that visemes 3 (/f/-v/) and 12 (/n/) had high confusions compared with the natural head, and could be more accurately modelled.

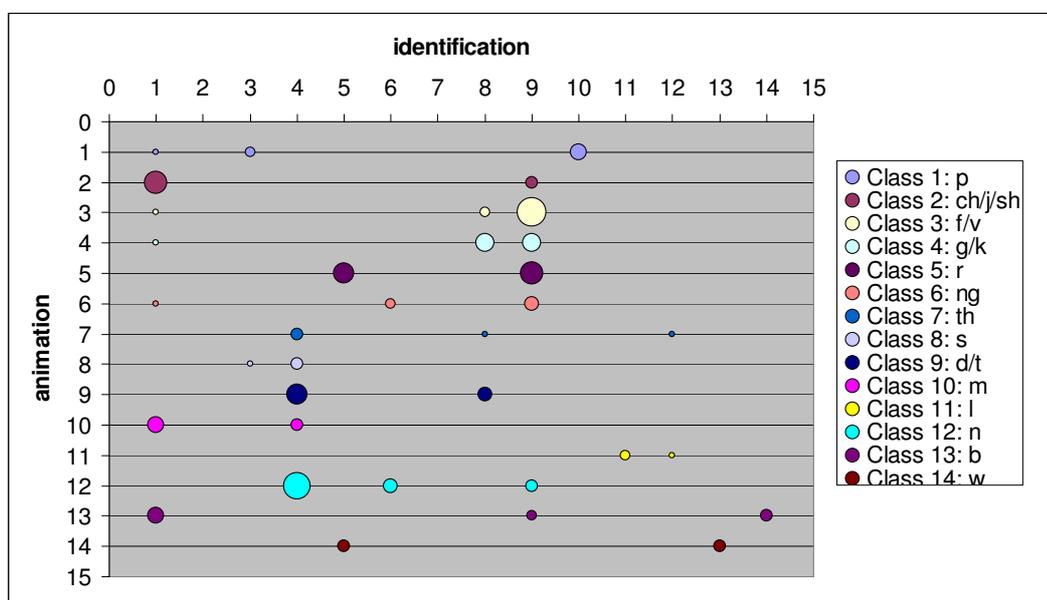


Figure 6.14: Difference between data-driven Head and Natural Head

## 6.6 Conclusions from Intelligibility and Naturalness Evaluations

The naturalness surveys showed that for every visual condition, the talking head with natural speech was always perceived as more natural than the same head with synthetic speech. This is unsurprising, and shows that the quality of the synthesized speech is not yet sufficient for listeners to believe that it is real. Psychological studies into the human reaction to media have found that users are more sensitive to audio quality than video quality, and audio quality has more effect on the user's attention, memory and opinion about what is heard (Reeves et al. 1996). However, these results show that the data-driven head, even with a synthetic voice, was perceived as more natural than the viseme-driven conditions with a natural voice; and the natural visual signal combined with a synthetic voice were always perceived as more natural than any of the synthetic visualizations with a natural voice. So this shows that the perception of audio quality does not completely overrule that of video. Reeves and Nass suggested that poor audio fidelity is more psychologically unfamiliar, as most spoken audio is heard at high-fidelity, whereas we are used to dealing perceptually with low visual fidelity, for example in dim lighting. However, it could be argued that we are also used to dealing with noisy audio conditions; but it is the “robotic” quality of synthesized speech which is unfamiliar and has an eerie effect, reducing our acceptance of the synthetic-voiced talking head. The “uncanny valley” could also explain why adding eye and head movements, which were intended to add more lifelike behaviour to the non-photorealistic head, but were not quite realistic enough, reduced the acceptance of the moving head compared to the rigid head (Mori 1970).

The non-photo-based viseme-driven talking head (THVN) showed a gain in intelligibility compared to audio speech alone, and was almost as intelligible as the video of a real speaker at SNR -20 dB (Figure 6.1) (Dey et al. 2010a). Certain visemes, such as (/r/-/l/) and (/h/-/n/-/ng) were confused with others (Figure 6.5), and could be improved by modelling the internal articulators more accurately, and then reclassifying the visemes into more categories, for example, separating (/r/) from (/l/), which has different tongue movements. Overall the visemes were

identifiable. In the subjective naturalness tests, the visual speech was rated as moderately natural overall. The data-driven head was perceived as more natural than this viseme-driven head.

An intelligibility test comparing the photo-based viseme-driven talking head (THVP) against the data-driven head (THD) confirmed that visualization improved the intelligibility of the speech. The word recognition rate was higher for the audiovisual heads than for audio alone, higher for the data-driven head than the viseme-driven head, and higher for the natural head than the synthetic heads. This shows that the data-driven head was more intelligible and more accurately modelled than the viseme-driven head. For the viseme-driven head, the reclassification of the visemes into more categories, for example separating /r/ and /l/ which had different tongue visemes (Table 4.1), led to viseme (/r/) showing few confusions, while (/l/) had some strong confusions (Figure 6.13), which could be improved by modelling the tongue dynamics more accurately. Viseme (/n/) could be improved by modelling the movement under the jaw more accurately. The data-driven head showed fewer confusions than the viseme-driven head (Figure 6.14), but also showed confusions for the viseme (/n/), which could be improved by modelling the movement under the jaw with more data for this area of the articulators. On the whole, for each talking head, the visemes were identifiable. Thus the talking heads were determined to be sufficiently realistic to be used to demonstrate pronunciation in a tutoring system (Dey et al. 2010b).

## 7 Evaluation of Speech Tutoring Application

The three talking heads (Chapter 4 and 5) were applied in the speech tutoring system (Chapter 3), which was evaluated in user trials. Existing training systems have been evaluated by experiments to assess the performance improvement of users after using the software. The evaluation of this tutoring system's effectiveness in tuition followed a similar approach to that of Massaro (Massaro et al. 2008), with the improvement in speech production assessed by human judgment. The studies aimed to determine whether the system made a difference to learning.

The evaluation of the experiments aimed to determine the effect of using visual speech in learning, and aimed to elucidate the advantages and disadvantages of using this technology. For example, the visualization of the inside of the mouth, showing the tongue movements during speech, is not possible using conventional tutoring. The disadvantages of a computerised system are that its feedback is limited, and it cannot offer as much intervention as a human tutor, so it would not fully replace a human tutor. However, the computer system is intended for use as an assistant, and could help a less experienced person to take the role of the human tutor. A major benefit would be to enable users to use the tutoring system at home in their own time, so students would be able to practise their speech at home with their families, outside of teaching hours.

### 7.1 Choice of Case Study

Key aspects of pronunciation include segmentals – speech sounds (vowels, consonants), and suprasegmentals – rhythm (stress, pausing) and intonation, i.e. prosody (Dabic 2010). For these studies, the focus was on segmentals, and consonants in particular. Although the software does demonstrate the other features indirectly, the featured lesson concentrated on a specific consonant pair, and only the perception and production of these consonant sounds was assessed in the experiments.

The /b-/p/ contrast was chosen as a case study after consultation with tutors from the English Language Teaching Centre at the University of Sheffield. Six members of staff and two students were interviewed for their opinions on the use of technology in second language learning. The tutors revealed that one of their largest groups of students was native Arabic speakers, and the most common difficulty for this group was /b-/p/, because this contrast did not exist in their native language. The /b-/p/ difficulty can also exist in learners from other native languages; for example, Korean (Bauman 2006), Chinese (Swan et al. 2001) and Japanese; “The English /b-/p/ voicing contrast may also lead to confusions as voiceless plosives in Japanese tend to be unaspirated and so English /b/ and Japanese /p/ will be phonetically similar” (Hazan et al. 2005).

A problem for some learners, for example, Korean speakers, is that the lower lip is pressed too close to the top teeth, causing a vibration, which will produce a /v/ (Bauman 2006). Therefore, for those with this difficulty, the internal visualization could be useful, for demonstrating the lip and tooth positions during pronunciation of the /b/ sound.

Lazalde (Lazalde 2010) carried out an objective analysis of the differences between the visemes /b/, /m/ and /p/ for one male and one female speaker, in VCV contexts, speaking at speeds of 100 syllables per minute and 200 syllables per minute. The amount of mouth opening was found to be larger for /b/, followed by /m/ and then by /p/. Physical observation showed that /p/ usually required more lip pressure than /b/ or /m/. Lazalde confirmed that there was a significant difference between the /b/, /p/ and /m/ visemes. Lazalde also found that better results were obtained when synthesizing visual speech using separate visemes rather than when using a single /b-/m-/p/ viseme.

A difficulty with using /b/ and /p/ as a case study for pronunciation training is that the main difference between /b/ and /p/ is produced by voicing, but it is difficult to show voicing in a talking head; it would require the visualization of the glottals. However, it may not be useful to show the movement of the glottals because they cannot be consciously controlled in the same way as a tongue or lip movement.

An existing approach used by clinicians is spectrographic displays which provide feedback of subtle auditory features that are difficult to detect otherwise, such as voice onset time. Spectrograms can visualize the difference between /b/ and /p/, but it can be difficult to learn how to read such data, and since it is an abstract representation, it is not clear how to apply it to one's own productions.

Mahshie used visualization of the subject's own airflow as visual feedback for teaching production of the voicing distinction between /p/ and /b/ (Mahshie 1996). A deaf subject took part in 12 sessions provided over 7 weeks. The training involved coordination of the laryngeal and oral gesture required for production of the voicing distinction for /p/ vs. /b/. It was found that the appropriate production patterns were observed with greater precision when visual feedback was provided than when it was withheld. Evidence for internalization of learning, as reflected in the tests before training at the start of each session, was not observed until later in the training, after several sessions. When the speaker's production was judged for accuracy, while there was an improvement in the production of /p/, the subject's production of /b/ segment voicing actually decreased. Mahshie suggested that this was probably the result of overgeneralizing the production pattern. Therefore this study showed that provision of the feedback resulted in improved performance during some phases of teaching, but not during others.

A study by (Hazan et al. 2005) investigated the effect of audiovisual perceptual training on the perception and production of consonants by 39 Japanese learners of English. This study included /b/-/p/-/v/, but was testing the labiodental contrast of /b/ and /p/ against /v/, so they ignored any difference between /b/ and /p/. Training took part in 10 sessions over 4 weeks. At each session, a perception test was carried out in the form of a minimal-pair identification task. Results found that audiovisual training using natural video gave better perception scores than audio alone, which was better than video alone, although there was no significant difference between any of the conditions. A second study investigated the /r/-/l/ contrast, which was less visually distinctive than the labial/labiodental contrast of /b/ and /p/ vs /v/. This study found that for the perception of the /l/-/r/ contrast, audiovisual training was not more effective than auditory training. In a speech production study of /r/-/l/, native talkers of British English were asked to judge the

speech produced by the learners, using a minimal-pair identification task and a quality rating task. The increase in scores was significantly greater for the learners with natural audiovisual training than for audio alone. The conclusions were that perceptual training resulted in improvements to the pronunciation of the trained consonants in second-language learners, and that audiovisual training was more effective than audio training when the visual contrast was sufficiently salient. Visualization of speech led to a greater improvement in pronunciation, even for contrasts with relatively low visual salience (Hazan, Sennema et al. 2005). The findings of (Hazan et al. 2005) need to be verified with a wider range of phonemic contrasts. The experiments presented in this thesis are widening the range of contrasts studied, with a less visually salient contrast.

For learning the distinction between /b/ and /p/, the usual technique used in existing second-language tutoring involves listening tests, because if a learner cannot perceive the difference, they cannot produce it. Therefore the tutoring application presented in this thesis provides listening practice, comparing /b/ against /p/ sounds. Research has shown that visual speech complements the audio, and the MRT experiments in Chapter 5 have shown that for all these talking heads, the visual signal gives a contribution to intelligibility of the speech. So the visualization of the lips enhances the audio, so would make a talking head in this tutoring application more intelligible than audio alone, and could therefore be more useful than listening practice based on audio alone.

In Chapter 6, the sounds /b/ and /p/ were separated in the confusion matrices, to show how much they were confusable with each other. The confusion matrices show that /b/ was mistaken for a wider range of sounds (/d/-/t/ as well as /m/ and /p/) in the audio alone condition than with the talking heads, which shows that the visualization does make a difference to the perception. The /b/ and /p/ sounds had less confusion for the data-driven head than the viseme-driven head, which shows that the data-driven head was more accurately modelled than the viseme-driven head. The data driven head is modelled on real recordings of the speaker saying /b/ and /p/, so it is more realistic than the viseme-driven heads in showing these sounds. Figure 7.1 shows the animation frames of the data-driven head for the word “back” and Figure 7.2 shows the animation frames for the word “pack”. The

lips appear to be slightly more pursed when saying /p/, in the first two frames of Figure 7.2, than when saying /b/ in the first two frames of Figure 7.1, and the rest of the utterance appears to be pronounced more emphatically for “*pack*” than “*back*”. Thus there are subtle visual differences, and experiments will indicate whether these are salient enough to aid pronunciation.

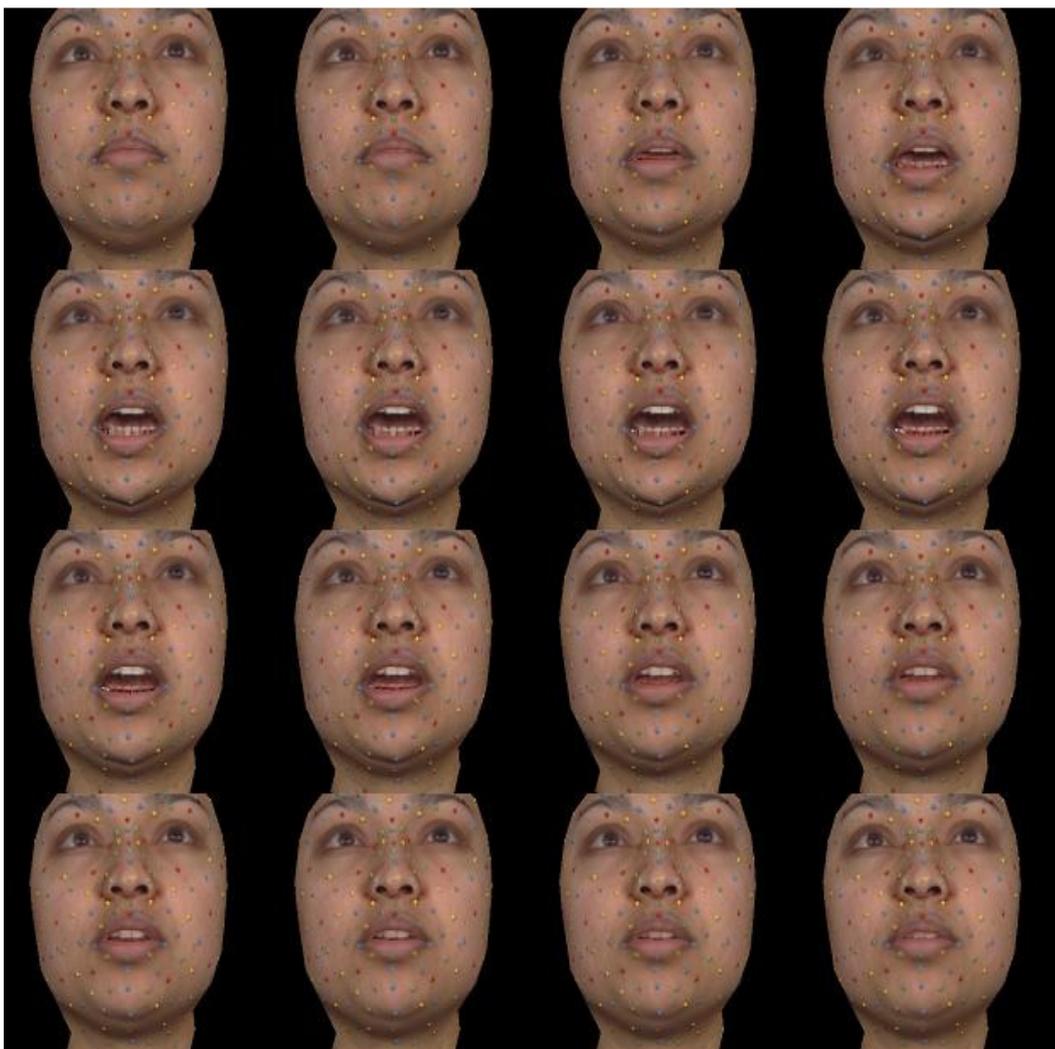


Figure 7.1: Animation frames of data-driven head for the word “*back*”

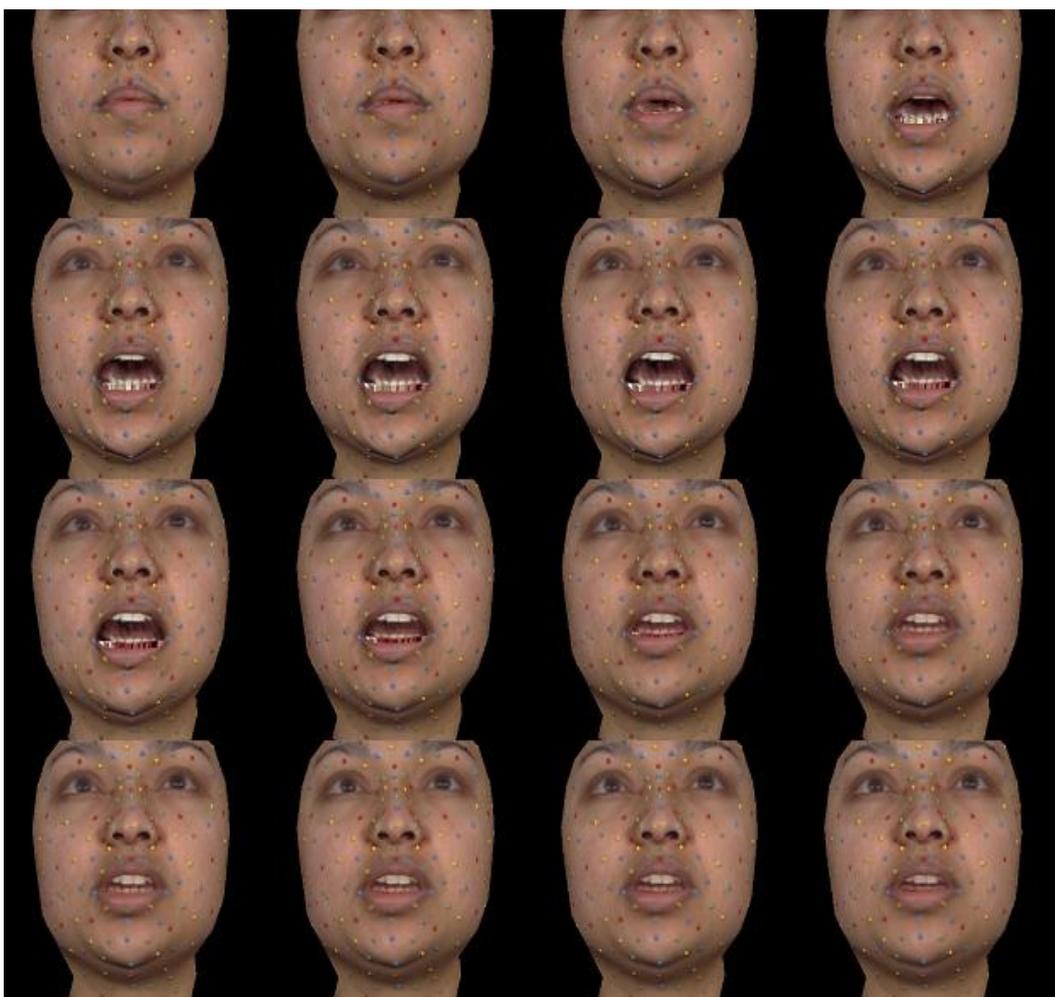


Figure 7.2: Animation frames of data-driven head for the word “pack”

## 7.2 Experiment Design

The studies carried out explored the role of visual speech information on the development of speech perception and production skills in second language learners. It is the presence of voicing and timing differences that allow us to distinguish /b/ from /p/. The visual differences in timing are very subtle and the auditory differences in voicing can be difficult to perceive for some non-native learners. Vision and hearing are complementary and each of these separate channels is more efficient for different verbal information (Ross 1999). Thus putting the audio and visual modalities together may help more than one modality alone.

The experimental protocol took a similar approach to those used by (Mahshie 1996) and (Hazan et al. 2005), with a pre-test before training at the start of each session to give a reflection of internalized learning; a training period using the training software first for listening practice, then for giving feedback on listening and speech production; and finally, a post-training test, to reflect the extent to which the subject was able to produce the pattern following the training period. The pre-test and post-tests recorded speech input (listening ability) with an auditory detection task, to test whether phonological representations were accurate; and speech output (production of speech) with a task involving reading words aloud. Different words were used for the listening and speaking test stimuli (Appendix E), and these differed from those used in the training application, in order to test for generalization of the perception/production patterns to new words. The studies conducted are describes in the following sections. First a pilot study was conducted comparing head THVN against audio alone (Dey et al. 2010b). A small study was conducted using head THVP, to investigate the effect of internal visualization compared with external visualization alone. Finally, a crossover study of 17 participants was conducted comparing the data-driven head THD against audio alone.

### 7.3 Tutoring Study 1: Evaluation of viseme-driven non-photo-based head

A pilot study was conducted using the speech tutoring software with the viseme-driven non-photorealistic talking head (THVN), with no head movements, eye movements or facial expressions, and an earlier implementation of tongue visemes, based on Lazalde's tongue visemes, as described in Chapter 4.7 (Figure 7.3).

The pilot trial was run with five native Arabic speakers, learning English as a second language, who worked through a session lasting one hour with a repeat session one week later. The participants were all from the same English language class, with similar levels of English proficiency.

A pre-test and post-test of pronunciation were carried out, in which the subjects read aloud isolated words and sentences in English and their speech was recorded. A listening pre-test and post-test was also carried out, in which the participants listened to acoustic speech of isolated words and identified which words they heard (Appendix E.1 – E.3).

After the pre-test, the participants were presented with the Pronunciation Assistant software, and asked to work their way through a lesson. Three participants were presented with the complete software, with the talking head in an external frontal view and an internal mid-sagittal view (Figure 7.3), and two were presented with the software with no talking head visualization. The lesson taught the pronunciation of the sounds /b/ and /p/, a contrast which the students found difficult because it did not exist in their native language. The lesson included practice in listening to sounds, words and phrases, and pronunciation practice, in which the software would demonstrate how to pronounce a sound, word or phrase, and then the user would say it aloud, with the option to record their own speech and play it back.

After the session with the Pronunciation Assistant software, the post-test was carried out, in which the listening test and speaking test were repeated. Finally the participants completed a questionnaire about their experience of using the software, which asked the users to rate on a five-point Likert scale how useful they found each feature (Appendix E.4).

The pre- and post pronunciation tests were evaluated by a native English speaker, who was a tutor from the English Language Teaching Centre at the University of Sheffield, and thus experienced in judging pronunciation. The audio recordings were presented to the judge in a random order, and the judge decided whether /b/ or /p/ was heard in the isolated words. For the list of sentences, the judge decided whether each instance of /b/ or /p/ was pronounced correctly. The numbers of correct pronunciations were counted to give an overall score.

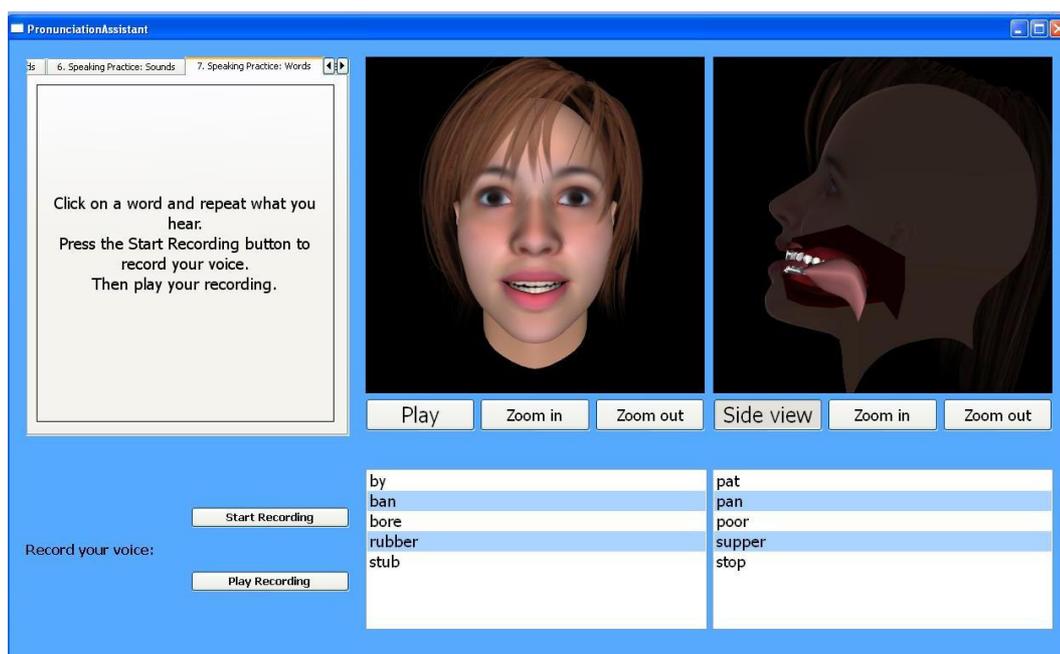


Figure 7.3: Screenshot of Speech Tutoring Application used in Tutoring Study 1

### 7.3.1 Study 1 Listening Results

The listening scores are shown in Table 7.1 and Figure 7.4. Users 1, 3 and 5 tried the audiovisual (talking head) version of the software, while users 2 and 4 had audio alone.

Study 1 Listening Scores (/b/ and /p/) %					
User	Condition	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post
1	audiovisual	75	75	90	90
2	audio	95	90	100	100
3	audiovisual	70	90	90	85
4	audio	60	60	80	60
5	audiovisual	95	85	80	85

Table 7.1: Study 1 Listening Scores (/b/ and /p/)

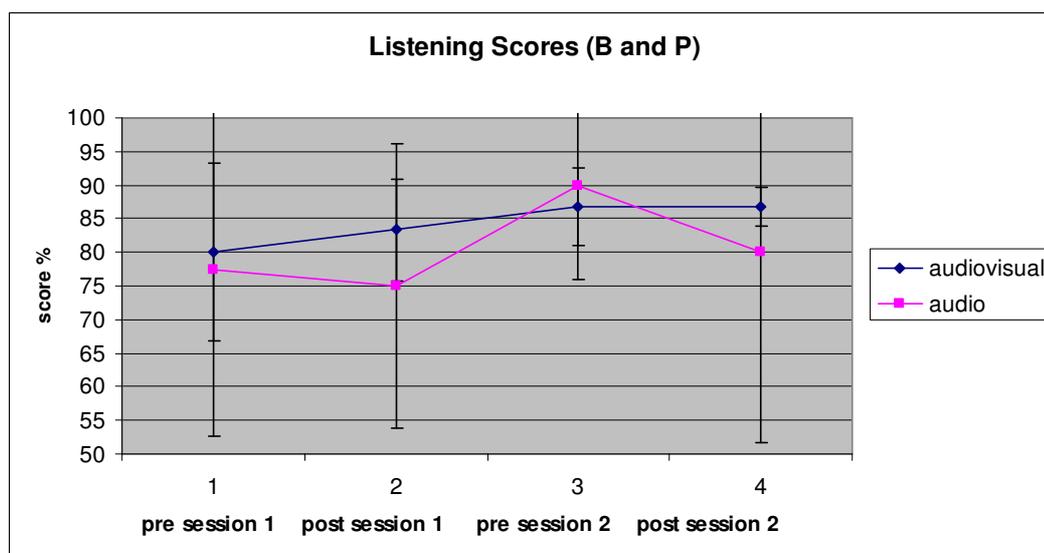


Figure 7.4: Study 1 Listening Scores (/b/ and /p/)

On average, for /b/ and /p/, the talking head gave a higher improvement (mean improvement 6.67%) in listening than audio alone (mean improvement 2.5%). The scores for /b/ and /p/ separately are presented separately in Tables 7.2 and 7.3:

Study 1 Listening (B) %				
User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post
1	60	50	100	80
2	90	80	100	100
3	60	80	80	70
4	60	60	70	40
5	100	80	60	70

Table 7.2: Study 1 Listening Scores (/b/)

For listening to /b/ sounds, there was no improvement on average for either the audio or audiovisual conditions. However, considering each individual, 2 out of 3 participants with the head improved, and 1 out of 2 participants with audio improved.

Study 1 Listening (/p/) %				
User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post
1	90	100	80	100
2	100	100	100	100
3	80	100	100	100
4	60	60	90	80
5	90	90	100	100

Table 7.3: Study 1 Listening Scores (/p/)

For listening to /p/ sounds, there was an improvement on average for both the audio (mean improvement 10%) and audiovisual condition (mean improvement 13.3%), and this improvement was higher for the audiovisually-trained group than audio alone. All participants with the head improved from the start of session 1 to the end of session 2, while 1 out of 2 participants with audio improved.

### 7.3.2 Study 1 Speaking Results (/b/ and /p/)

Table 7.4 and Figure 7.5 show that for speaking (/b/ and /p/), there was an improvement on average for both the audio (mean improvement 0.83%) and audiovisual condition (mean improvement 1.67%), and this improvement was higher for the audiovisually-trained group than audio alone. All participants with the head improved from the start of session 1 to the end of session 2, while 1 out of the 2 participants with audio improved.

Study 1 Speaking Score (/b/ and /p/) %				
User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post
1	88.3	88.3	90	90
2	85	98.3	96.7	95
3	95	100	95	96.7
4	86.7	90	85	78.3
5	85	85	85	86.7

Table 7.4: Study 1 Speaking Scores (/b/ and /p/)

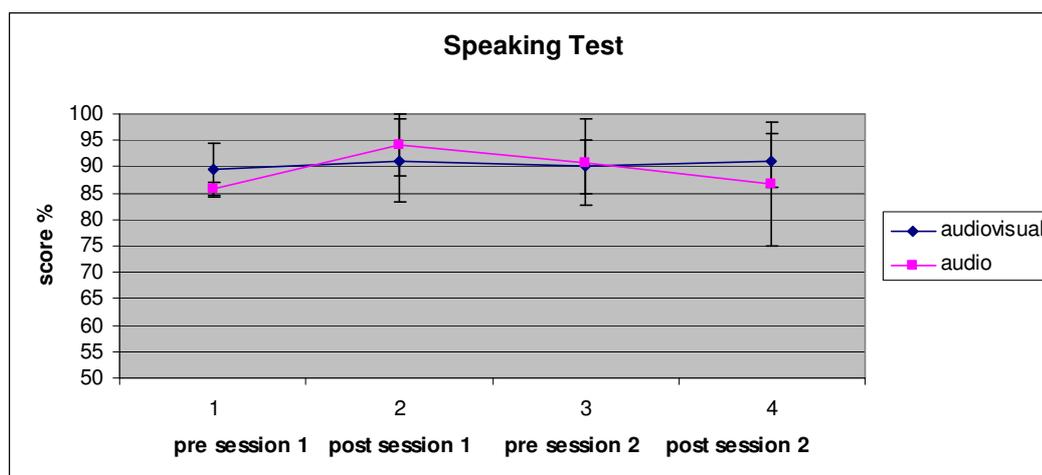


Figure 7.5: Study 1 Speaking Scores (/b/ and /p/ combined)

### 7.3.3 Study 1 Speaking (/b/ scores)

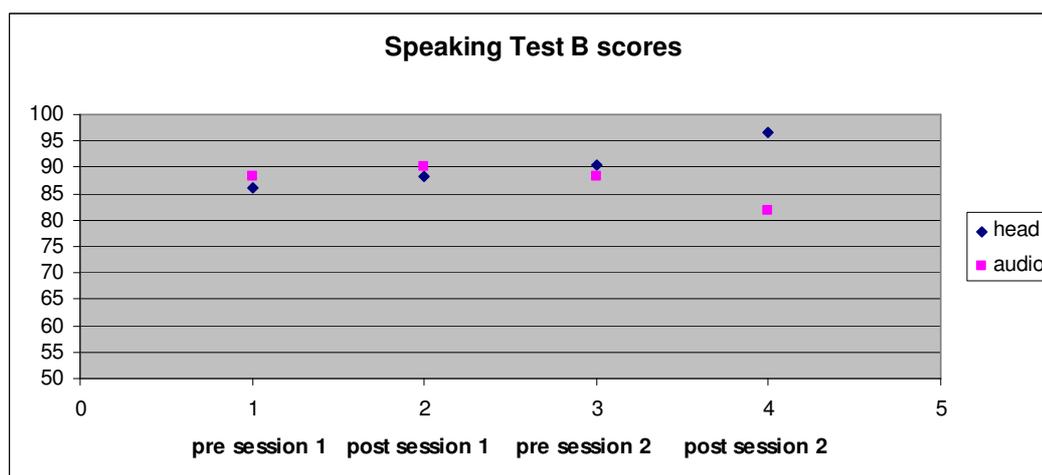


Figure 7.6: Study 1 Speaking Scores (/b/)

Figure 7.6 shows that for speaking /b/ sounds, those who used the viseme-driven head had an improvement in each session, so there was an improvement overall (mean improvement 8.33%). Those who used audio had an immediate improvement in the first session, but not after that, and no overall improvement.

### 7.3.4 Speaking (/p/ scores)

For speaking /p/ sounds, those who used the viseme-driven head showed no improvement (Figure 7.7). Those who used audio had an immediate improvement in the first session, but not after that. The audio-alone condition gave an improvement overall (mean improvement 8.33%).

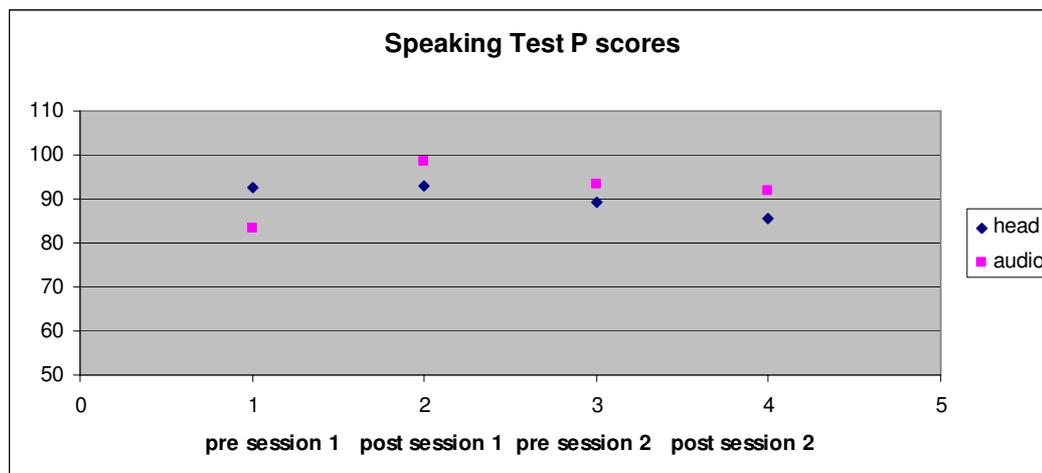


Figure 7.7: Study 1 Speaking Scores (/p/)

### 7.3.5 Speaking and Listening Tests combined

The Speaking and Listening Tests were combined by taking the mean of each corresponding score for listening and speaking (Table 7.5 and Figure 7.8). For speaking and listening combined (/b/ and /p/), there was an improvement on average for both the audio (mean improvement 1.25%) and audiovisual condition (mean improvement 2.92%), and this improvement was higher for the audiovisually-trained group than audio alone.

Study 1 Speaking and Listening scores combined %				
User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post
1	81.7	81.7	90	90
2	90	94.2	98.3	97.5
3	82.5	95	92.5	90.8
4	73.3	75	82.5	69.2
5	90	85	82.5	85.8

Table 7.5: Study 1 Speaking and Listening Scores (/b/ and /p/)

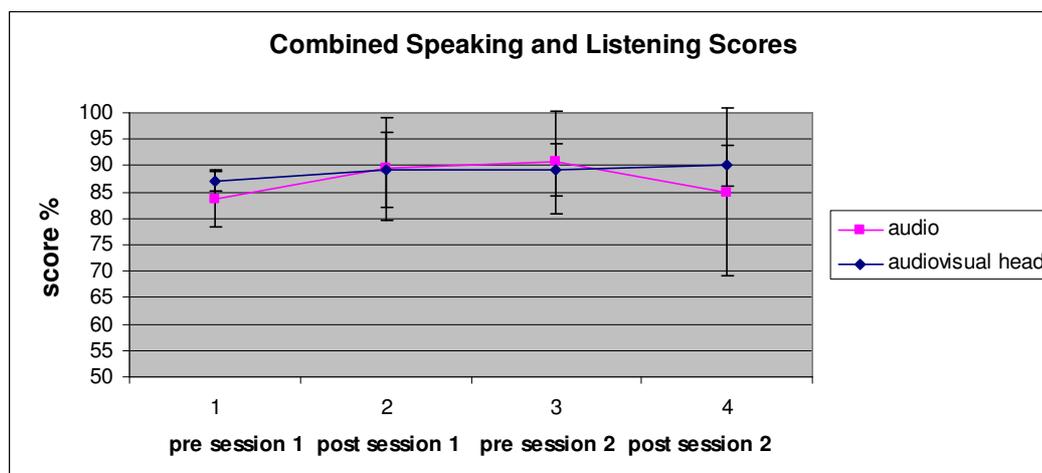


Figure 7.8: Study 1 Speaking and Listening Scores

Generally, there was an improvement in speaking and listening, from the first test (pre-test session 1) to the final test (post-test session 2), for both groups. When the scores /b/ and /p/ sounds were considered separately, there were some differences between them. For speaking, the talking head gave no improvement for /p/ sounds; possibly because /p/ was the harder sound to learn for native Arabic speakers, and this viseme-driven head did not sufficiently convey the difference between /b/ and /p/. However, overall the talking head gave a more consistent improvement than audio alone. The fluctuations in the scores were probably due to individual variations and the small sample sizes. Future tests would require larger groups of participants, and could require longer training times, to show any significant difference in learning.

### 7.3.6 Study 1 User Feedback

The user questionnaire asked the users to rate each feature of the software on a Likert Scale of 1-5 (1= Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree). The responses are shown in Table 7.6.

Study 1 User Feedback: mean response on Likert Scale 1-5		
Question	Audio (2 responses)	Audiovisual (3 responses)
This software was helpful in learning pronunciation	5.0	4.7
I found the external view helpful	NA	4.0
I found the internal view helpful	NA	4.3
I found the listening practice helpful	5.0	5.0
I found the listening test helpful	5.0	5.0
I found the speaking practice helpful	4.5	5.0
I found the recording function helpful	5.0	4.7
The content of the lesson matched my needs.	4.5	4.7
This software is easy to use	5.0	5.0
This software is engaging.	5.0	4.7
The talking head appeared natural	NA	3.3
This software is satisfying to use.	5.0	4.7

Table 7.6: Study 1 User Feedback

The feedback from the questionnaires was generally positive. The students enjoyed using the software, and found the content of the lesson useful. The students who used the talking head agreed that the talking head external view and side view were helpful. Students from both groups liked the practice of pronunciation of words and phrases.

The feedback for the talking head was similar to that for audio; most participants strongly agreed that the software was useful. Each participant was aware of only one version, so they were not comparing the two, and those in the audio-alone group were just as positive about the software. In the talking-head group, when asked if the talking head appeared natural, 66 % agreed and 33% disagreed. Those who had the talking head version unanimously thought that the visualization was useful; 100% strongly agreed that the external view was useful, while for the

internal view, 66% agreed and 33% strongly agreed that the internal view was useful. This indicates that the users thought that the external view was more helpful than the internal view, which may be because the internal view is an unfamiliar view to most, and is not normally used in traditional methods of learning a language, or because the users did not see a difference between the /b/ and /p/ visualizations. This study suggested that the visualization was thought to be useful, but did not test whether it was the internal or external visualization that helped; this was investigated in the next study.

#### 7.4 Tutoring Study 2: Evaluation of viseme-driven photo-based head (THVP)

A user trial evaluated the viseme-driven photo-based head, comparing external and internal visualization (Figure 7.10), against external visualization alone (Figure 7.9). This experiment investigated whether displaying the internal articulators made a difference to learning. While previous studies have investigated whether visualization of articulators gives an improvement in learning (Massaro et al. 2008; Wik et al. 2008), few have investigated the impact of expression in animated characters for speech tutoring (Massaro 2004), and none have proved that talking heads give a significant improvement in learning.

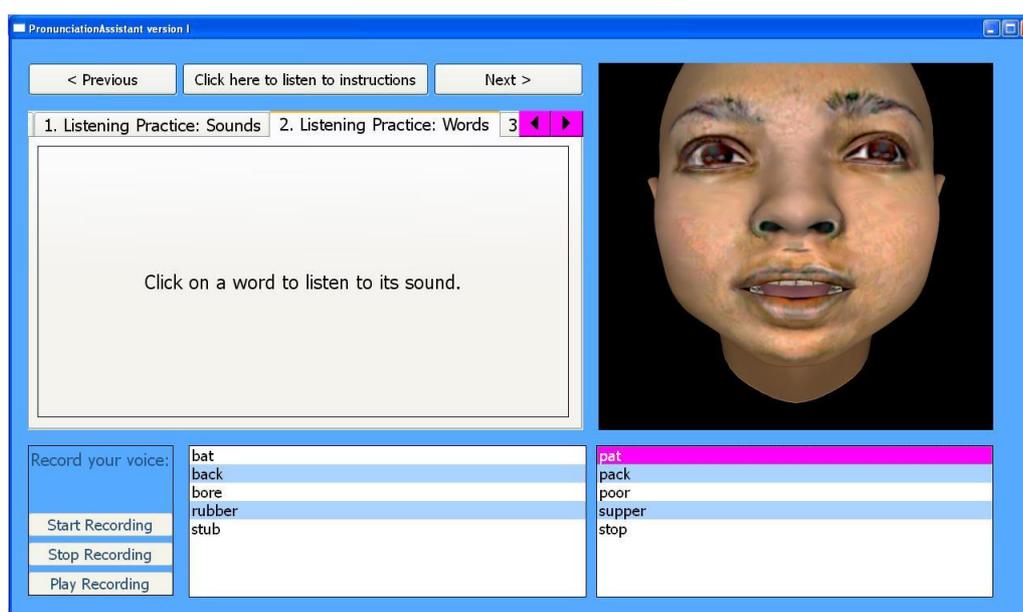


Figure 7.9: Screenshot of Speech Tutoring Application used in Tutoring Study 2

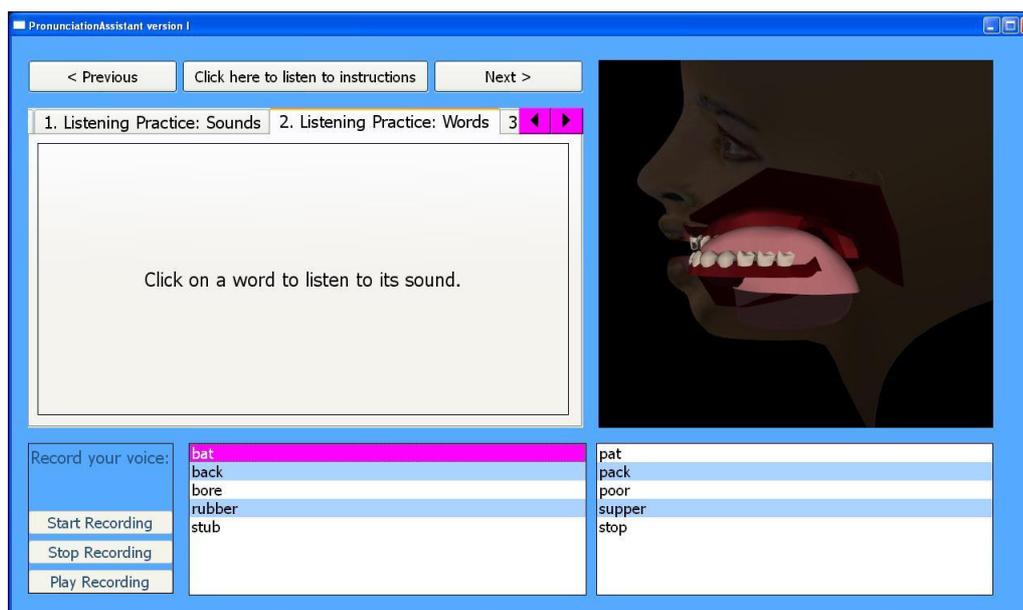


Figure 7.10: Screenshot of Speech Tutoring Application used in Tutoring Study 2

#### 7.4.1 Study 2 Experimental Design

The study was carried out in collaboration with James Carmichael at AlGhurair University in Dubai. The participants were native Arabic speakers, of an intermediate level of English proficiency. The users were divided into two groups, to test the two different conditions. The trials were conducted in sessions lasting one hour over one month.

A listening pre-test and post-test was carried out in the form of a minimum-pair identification task, in which the participants listened to acoustic speech of isolated words and identified which words they heard. A pre-test and post-test of pronunciation was carried out, in which the subjects read aloud isolated words and sentences in English and their speech was recorded (Appendix E.5 – E.7). The words and sentences used in the speaking test were not present in the tutoring application itself, so this experiment would investigate whether the pronunciation training was effective in generalization to new words.

After the pre-test, the participants were presented with the Pronunciation Assistant software, and asked to work their way through a lesson. The lesson teaches the pronunciation of the sounds /b/ and /p/, a contrast which does not exist in the

users' native language. The lesson includes practice in listening to sounds, words and phrases, and pronunciation practice, in which the software demonstrates how to pronounce a sound, word or phrase, and then the user says it aloud, with the option to record their own speech and play it back.

After the session with the Pronunciation Assistant software, the post-test was carried out, in which the listening test and speaking test were repeated. Finally the participants completed a questionnaire about their experience of using the software, which asked the users to rate how useful they found each feature (Appendix E.8 - E.9). For this survey the Likert scale used was 1-7, to give more precision than a 5-point scale.

In order to analyze the speaking test recordings, a native English listener was recruited as a judge, who assessed the speech in the form of a listening test. The reasoning was that if a second-language speaker could be understood by a native English listener, then they were pronouncing the sound correctly. For each word or sentence presented, the judge was asked to determine whether what they heard was /b/, /p/, ambiguous or unintelligible. The judging was carried out as a minimal pair identification task, with carrier sentences that would make sense with either the /b/ or /p/ word; for example, "I would like to put the *bath* here" / "I would like to put the *path* here". The judge did not know what each item was supposed to be, so they were making a decision purely on what they heard, and was unaware of which group each speaker was from, and unaware of whether they were hearing pre- tests or post-tests, so they were listening objectively. The judge's responses were then scored as follows: a correct identification scored 1; an incorrect, ambiguous or unintelligible identification scored 0. The scores were counted to give a score for each participant's pre-test and post-test for each session.

#### 7.4.2 Study 2 Listening Test Results (/b/ and /p/)

User 1, who had the internal visualization, showed no improvement in listening for /b/ or /p/. User 2, who had external visualization only, did show an improvement in listening for /b/ and /p/, separately (Tables 7.8 and 7.9) and combined (Table 7.7).

Study 2 Listening Test %			
User	Session 1 pre-test	Session 2 post-test	Internal Visualization (I) or External only (X)
1	85	80	I
2	60	85	X

Table 7.7: Study 2 Listening Scores (/b/ and /p/)

#### 7.4.3 Study 2 Listening Test /b/ sounds

User	Session 1 pre-test	Session 2 post-test	Internal Visualization (I) or External only (X)
1	80	70	I
2	50	80	X

Table 7.8: Study 2 Listening Scores (/b/)

#### 7.4.4 Study 2 Listening Test /p/ sounds

User	Session 1 pre-test	Session 2 post-test	Internal Visualization (I) or External only (X)
1	90	90	I
2	70	90	X

Table 7.9: Study 2 Listening Scores (/p/)

#### 7.4.5 Study 2 Speaking (/b/ and /p/ combined)

Both users showed an improvement in speaking for /b/ and /p/ combined (Table 7.10). User 2, who had external visualization only, showed a greater improvement.

Study 2 Speaking %			
User	Session 1 pre-test	Session 2 post-test	Internal Visualization (I) or External only (X)
1	52.5	57.5	I
2	65	72.5	X

Table 7.10: Study 2 Speaking Scores (/b/ and /p/)

#### 7.4.6 Study 2 Speaking (/b/)

Both users showed an improvement in speaking for /b/ (Table 7.11).

User	Session 1 pre-test	Session 2 post-test	Internal Visualization (I) or External only (X)
1	40	50	I
2	75	80	X

Table 7.11: Study 2 Speaking Scores (/b/)

#### 7.4.7 Study 2 Speaking (/p/)

User 1, who had the internal visualization, showed no improvement in speaking for /p/. User 2, who had external visualization only, did show an improvement in speaking for /p/ (Table 7.12).

User	Session 1 pre-test	Session 2 post-test	Internal Visualization (I) or External only (X)
1	65	65	I
2	55	65	X

Table 7.12: Study 2 Speaking Scores (/p/)

Overall, User 1, who was trained with both internal and external visualization, showed no improvement in listening, but some improvement in speaking. User 2, who was trained with the external visualization only, improved in both listening and speaking. It may be that for the /p/ and /b/ sounds, the internal visualization was not helpful because the tongue is not used much in producing these sounds, so not much difference was seen between the visualizations. In this experiment, the user with version X saw two views of the external view, whereas the user with version I saw one of each view, so the lips may have been more useful for /b/ and /p/ than the internal view. A limitation of this experiment is that no immediate post-test was carried out after the first session, and no pre-test was carried out at the start of the second session, so in the intervening period between sessions, there could be influences other than the software affecting the users' performance, which were unaccounted for. Furthermore, the numbers of participants was too small for the results to be conclusive. However, some positive feedback was

obtained (Table 7.13). User 1 moderately agreed that the software was useful, and strongly agreed that both the external and internal views were useful. The user liked the talking head and found it to be user-friendly. User 2 strongly agreed that the software was useful, and moderately agreed that the external view was useful. Both users strongly agreed that the software was interesting and satisfying to use.

Study 2 User Ratings on each feature of the software		
	User 1 (Group I)	User 2 (Group X)
The Pronunciation Assistant software was helpful in learning pronunciation.	Moderately agree	Strongly agree
I found the external view of the talking head helpful.	Strongly agree	Moderately agree
I found the internal view of the talking head helpful.	Strongly agree	NA
The talking head looked realistic.	Moderately agree	Strongly agree
The speech animation appeared natural.	Neutral	Slightly agree
I found the listening practice with the talking head helpful.	Strongly agree	Strongly agree
I found the listening test with the talking head helpful.	Strongly agree	Strongly agree
I found the speaking practice with the talking head helpful.	Moderately agree	Strongly agree
I found the recording function in the Pronunciation Assistant software helpful.	Slightly agree	No response
The Pronunciation Assistant software is interesting to use.	Strongly agree	Strongly agree
The Pronunciation Assistant software is satisfying to use.	Strongly agree	Strongly agree
The content of the lesson matched my needs.	Moderately agree	Strongly agree

Table 7.13: Study 2 User Feedback

## 7.5 Tutoring Study 3: Evaluation of data-driven head (THD) in a tutoring system

This experiment investigated whether the data-driven talking head (Figure 7.11), with a photo-realistic appearance and an articulatory model based on a real speaker, was more effective in teaching pronunciation than audio alone.

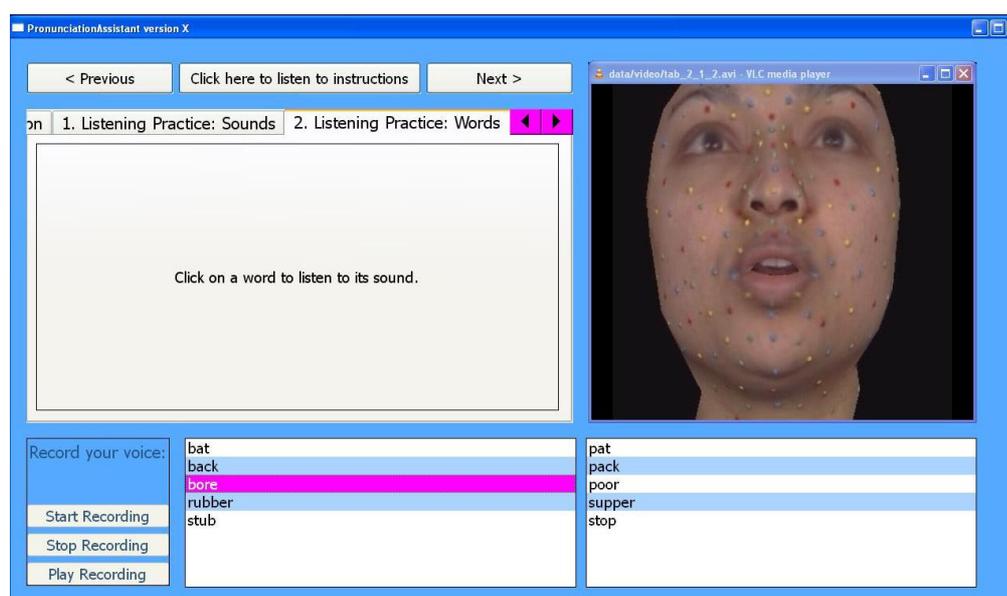


Figure 7.11: Screenshot of Speech Tutoring Application used in Tutoring Study 3

The data-driven talking head was evaluated by second-language learners. Each participant took part in two one-hour sessions, which were carried out within 1 month. The participants were divided into two groups with a crossover design, so that one group, A-AV, were first given the software with audio alone, and then were given the talking head in their second session, while the other group, AV-A, were given the talking head first, and then audio alone in their second session.

In each one-hour session, a participant first took part in a pre-test listening test, where they listened to 20 words and selected which word they heard, and a pre-test speaking test, where their voice was recorded while reading aloud a list of 20 words and 20 sentences (Appendix E.5 – E.7). Then they used the speech tutoring software, which demonstrates the /p/ and /b/ sounds in English. After this training they took part in a post-test listening and speaking test, which took the same

format as the pre-tests. Finally they completed a questionnaire asking their opinions of the software (Appendix E.9 – E.10). The users rated on a 7-point Likert scale how useful they found each feature. In their second session they were asked to rate whether they preferred the version of the software with audio alone or the talking head, and their reasons for this preference.

Participants for the experiment were recruited from the student population of the University of Sheffield. The volunteers were of diverse backgrounds, with a range of native languages including Arabic, Vietnamese, Korean, Chinese and Japanese, and various levels of English proficiency, from moderate to proficient. The results presented are for the 17 participants of the crossover experiment; 8 in group AV-A and 9 in group A-AV.

In order to analyze the speaking test recordings, a native English listener was recruited as a judge, who assessed the speech in the form of a listening test. The reasoning was that if a second-language speaker could be understood by a native English listener, then they were pronouncing the sound correctly. For each word or sentence presented, the judge was asked to determine whether what they heard was b, p, ambiguous or unintelligible. The test was designed to use minimal pairs of words containing /b/ or /p/, and carrier sentences had been created that would be semantically valid with either the /b/ or /p/ word; for example, “I would like to put the *bath* here” / “I would like to put the *path* here”. The judge did not know what each item was supposed to be, so they were making a decision purely on what they heard, and was unaware of which group each speaker was from, and unaware of whether they were hearing pre-tests or post-tests, so they were listening objectively. The judge’s responses were then scored as follows: a correct identification scored 1; an incorrect, ambiguous or unintelligible identification scored 0. Thus points would only be awarded for clear, unambiguous pronunciation. The scores were counted to give a score for each participant’s pre-test and post-test for each session.

The results are presented in the following sections. First, the participant information is given in Table 7.14. Next, the listening results are shown for /b/ and /p/ combined, and then for /b/ and /p/ separately. Then the speaking results are presented similarly. Finally, the subjective user feedback from the questionnaires

is presented. In the tables, the yellow shading indicates the talking head condition, while the unshaded rows are for the audio condition.

### 7.5.1 Study 3 Participant Information

User	Group	Native Language	Level of English (self-rated)	IELTS score
1	AV_A	Arabic	Moderate	5.5
2	AV_A	Arabic	Moderate	5
3	AV_A	Kurdish	fluent	-
4	AV_A	Kannada	fluent	-
5	AV_A	Tamil	fluent	-
6	AV_A	Farsi	Moderate	-
7	AV_A	Chinese	beginner	6.5
8	AV_A	Korean	Moderate	6
9	A_AV	Vietnamese	Proficient	8
10	A_AV	Arabic	Proficient	5.5
11	A_AV	Urdu	Proficient	7
12	A_AV	Malay	Moderate	6.5
13	A_AV	Japanese	beginner	-
14	A_AV	Farsi	Moderate	6.5
15	A_AV	Japanese	Moderate	-
16	A_AV	Turkish	Proficient	-
17	A_AV	Japanese	beginner	-

Table 7.14: Study 3 Participant Information

### 7.5.2 Study 3 Results of Listening Test (/b/ and /p/)

Most users showed an overall improvement from the start of session 1 to the end of session 2 (Figure 7.12). The standard deviation was high, because individual variations caused fluctuations in scores. User 10 in Group AV-A had an unusually low score at the end of the session with the talking head (Table 7.15), which caused the average score for the talking head to decrease (Table 7.16). Another user (User 17) scored 100% in the first listening test, so there was a ceiling effect for his listening results. (This user was still included in the analysis because his speaking scores were lower, so he had the potential for improvement in pronunciation.)

Study 3 Listening (/b/ and /p/) %					
User	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test	Group
1	55	60	75	80	AV-A
2	75	85	80	90	AV-A
3	85	80	95	90	AV-A
4	90	85	90	90	AV-A
5	90	90	90	85	AV-A
6	90	90	95	100	AV-A
7	90	95	90	100	AV-A
8	95	95	90	90	AV-A
9	60	75	85	85	A-AV
10	65	60	75	55	A-AV
11	85	85	90	90	A-AV
12	85	90	80	95	A-AV
13	85	100	100	100	A-AV
14	90	90	95	90	A-AV
15	95	95	95	95	A-AV
16	95	95	95	95	A-AV
17	100	95	100	100	A-AV

Table 7.15: Study 3 Listening Scores (/b/ and /p/)

Study 3 Mean Scores for Listening (/b/ and /p/) %				
Group	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test
AV-A	83.8	85	88.1	90.6
A-AV	84.4	87.2	90.6	89.4

Table 7.16: Study 3 Mean Listening Scores (/b/ and /p/)

Group AV-A, who had the head first, improved after using the talking head, and this group had a greater improvement overall. The audio condition showed a higher average improvement than the talking head, in both groups. (If the outlier, user 10, was removed then all groups and conditions would show an improvement, but the audio condition still would give a greater improvement than the audiovisual condition.) Combining the groups, there was an overall improvement for both audio (mean improvement 2.64%) and audiovisual (mean improvement 0.07%), and the improvement from audio was higher.

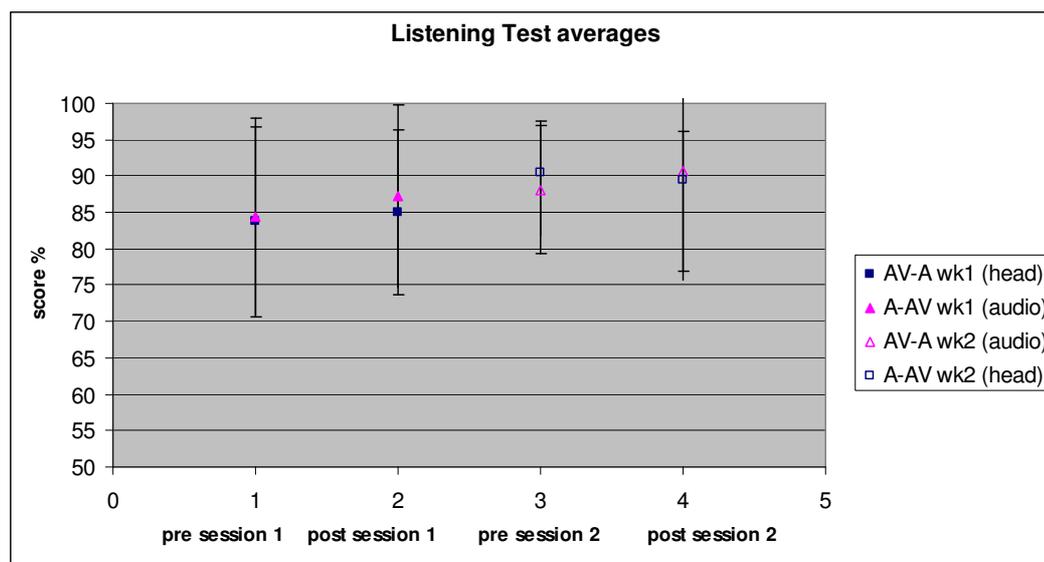


Figure 7.12: Study 3 Listening Test /b/ and /p/ scores

### 7.5.3 Study 3 Lowest Listening Scores

There was a large variation in the scores because the users were of diverse backgrounds, with a range of nationalities and various levels of English proficiency. To separate out those with a clear /b/-/p/ difficulty, the results of those who scored 80% and below are presented in Table 7.17. These users were one Vietnamese and 3 Arabic native speakers, at intermediate levels of English. The results show that individual variations are still high. Two of these users (1 and 2) improved in both conditions, one improved with audio alone, and one showed no improvement in either condition. This shows that some individuals benefit more than others from different ways of learning. User 10 scored lower after both training sessions. Overall for listening, audio alone gave a greater improvement than the talking head (Table 7.17).

User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post	Group
1	55	60	75	80	AV-A
2	75	85	80	90	AV-A
9	60	75	85	85	A-AV
10	65	60	75	55	A-AV

Table 7.17: Study 3 Lowest Listening Scores (/b/ and /p/)

#### 7.5.4 Study 3 Listening (/b/ sounds)

Group AV-A, who had the head first, improved after using the talking head, but showed no improvement in their second session, using audio (Table 7.18). Group A-AV, who had the audio first, improved using audio, but not in their second session, using the talking head. The individual listening scores for /b/ are shown in Table 7.19. Overall, the head gave no improvement for listening to /b/ sounds, while audio alone gave some improvement (Figure 7.13).

Study 3 Mean Scores for Listening (B sounds) %				
Group	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test
AV-A	83.8	85	90	90
A-AV	80	86.7	92.2	88.9

Table 7.18: Study 3 Mean Listening Scores (/b/)

User	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test	Group
1	50	70	100	80	AV_A
2	80	80	70	90	AV_A
3	80	80	100	90	AV_A
4	80	80	90	90	AV_A
5	90	90	90	90	AV_A
6	100	80	90	100	AV_A
7	100	100	90	100	AV_A
8	90	100	90	80	AV_A
9	40	70	90	80	A_AV
10	70	50	90	60	A_AV
11	70	70	80	80	A_AV
12	80	100	80	90	A_AV
13	70	100	100	100	A_AV
14	100	100	100	100	A_AV
15	90	90	90	90	A_AV
16	100	100	100	100	A_AV
17	100	100	100	100	A_AV

Table 7.19: Study 3 Listening Scores (/b/)

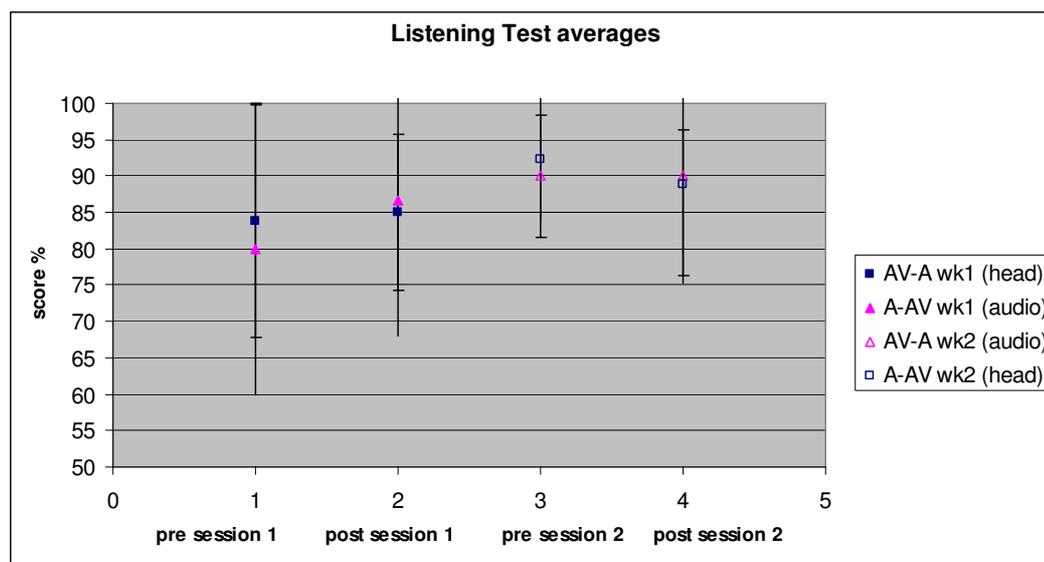


Figure 7.13: Study 3 Listening Test /b/ scores

### 7.5.5 Study 3 Listening (/p/ sounds)

User	Session 1 pre %	Session 1 post	Session 2 pre	Session 2 post	Group
1	60	50	50	80	AV_A
2	70	90	90	90	AV_A
3	90	80	90	90	AV_A
4	100	90	90	90	AV_A
5	80	80	90	80	AV_A
6	80	100	100	100	AV_A
7	80	90	90	100	AV_A
8	100	90	90	100	AV_A
9	80	80	80	90	A_AV
10	60	70	60	50	A_AV
11	100	100	100	100	A_AV
12	90	80	80	100	A_AV
13	100	100	100	100	A_AV
14	80	80	90	80	A_AV
15	100	100	100	100	A_AV
16	90	90	90	90	A_AV
17	100	90	100	100	A_AV

Table 7.20: Study 3 Listening Scores (/p/)

Study 3 Mean Scores for Listening (/p/ sounds) %				
Group	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test
AV-A	82.5	83.8	86.3	91.3
A-AV	88.9	87.8	88.9	90

Table 7.21: Study 3 Mean Listening Scores (/p/)

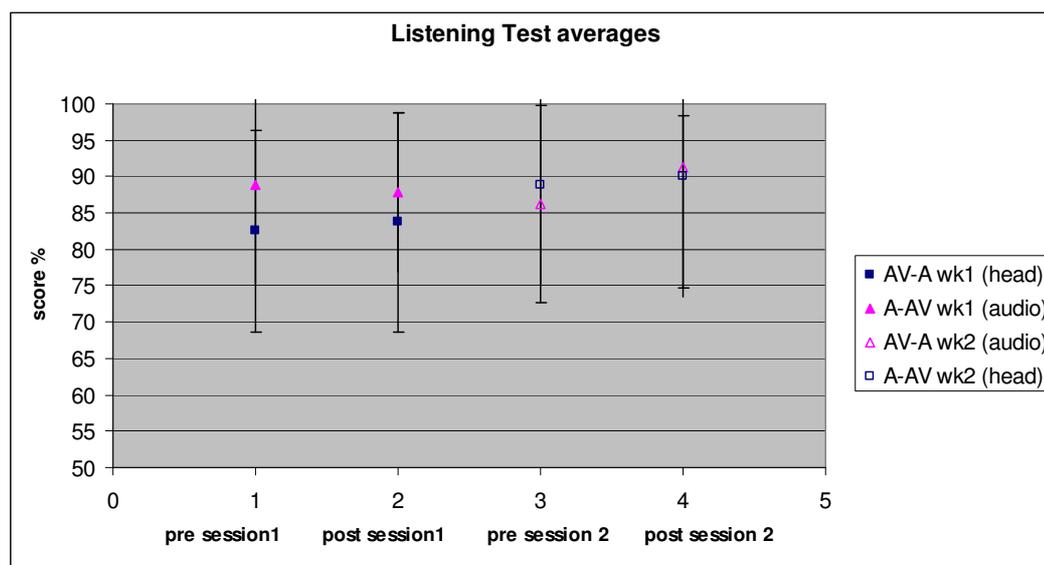


Figure 7.14: Study 3 Listening Test /p/ scores

Both auditory and audiovisual training resulted in an improvement overall for listening to /p/ sounds (Table 7.21 and Figure 7.14). The improvement was slightly higher for auditory training (mean improvement 1.94%) than audiovisual training (mean improvement 1.18%). 4 out of 17 participants improved from audio training, and 5 out of 17 participants improved from audiovisual training. 4 out of 17 participants improved more from audio than audiovisual training, while 4 out of 17 participants improved more from audiovisual than audio training (Table 7.20).

An incubation effect was observed, as the users improved between the two sessions. Incubation is defined as a process of unconscious recombination of thought elements that were stimulated through conscious work at one point in time, resulting in enhanced performance at some later point in time (Seabrook et al. 2003). The results indicate that this effect occurred, as users consolidated the

training into memory and improved their performance by the start of the next session, without any additional practice in the specific task.

An order effect was observed, as those who had auditory training first did not benefit from the audiovisual training in their second session, but those who had the audiovisual training first did benefit from auditory training in their second session. This could be due to “stimulus blocking” (Kamin 1969); training with the impoverished signal first caused the second session with an enhanced signal to make no difference, because the subject had learned how to perform with an impoverished signal, and so they would need to unlearn this before the enhanced signal could be of benefit. This finding suggests that there is value in using the audiovisual head, and that it should be used first, before audio alone.

#### 7.5.6 Study 3 Speaking Results (/b/ and /p/)

Both groups improved in each training session, with a strong within-session improvement each time (Table 7.22 and Figure 7.15), which suggests that the software is beneficial. However, the users deteriorated between sessions, so there was no incubation effect. This suggests an interference effect (Tomlinson et al. 2009), as in the intervening week they reverted to their previous ways, which were incompatible with the new training. The incorrect habits would have to be unlearned before the users could learn new habits, and one week may not have been enough time for this to occur.

Study 3 Mean Scores for Speaking (/b/ and /p/ sounds) %				
Group	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test
AV-A	67.5	72.8	62.8	66.9
A-AV	71.1	79.2	70.8	76.1

Table 7.22: Study 3 Mean Speaking Scores (/b/ and /p/)

User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post	Group
1	50	72.5	47.5	55	AV_A
2	57.5	57.5	40	40	AV_A
3	62.5	62.5	55	65	AV_A
4	62.5	75	62.5	65	AV_A
5	77.5	85	80	80	AV_A
6	65	65	77.5	72.5	AV_A
7	80	85	65	82.5	AV_A
8	85	80	75	75	AV_A
9	75	87.5	82.5	77.5	A_AV
10	52.5	45	52.5	60	A_AV
11	70	87.5	62.5	87.5	A_AV
12	62.5	77.5	80	67.5	A_AV
13	75	75	60	67.5	A_AV
14	80	82.5	72.5	75	A_AV
15	60	87.5	80	87.5	A_AV
16	85	85	87.5	85	A_AV
17	80	85	60	77.5	A_AV

Table 7.23: Study 3 Speaking Scores (/b/ and /p/)

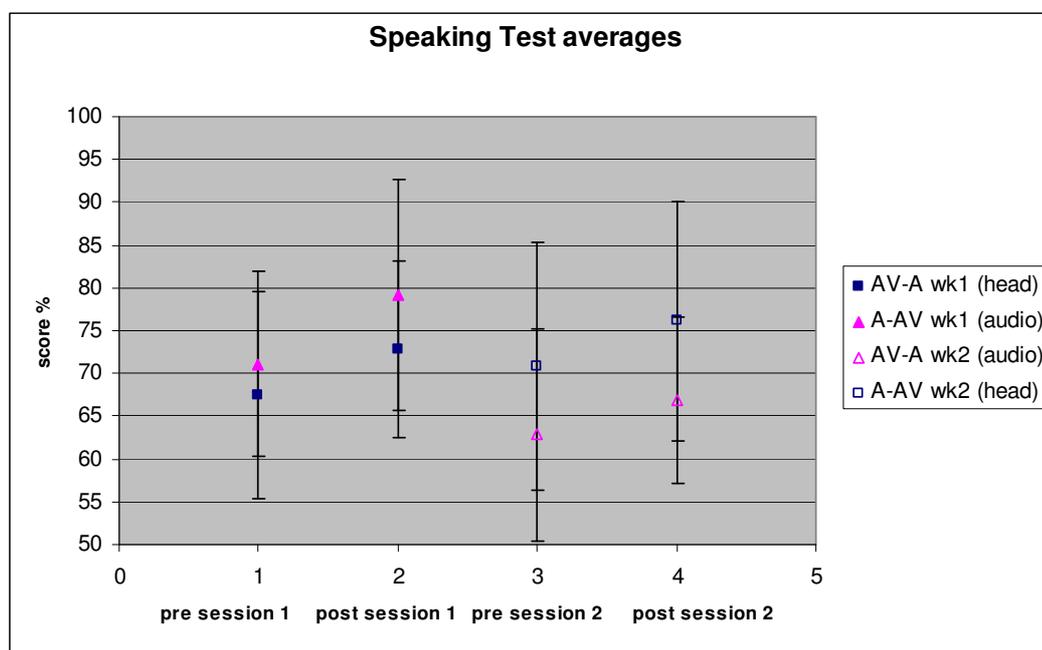


Figure 7.15: Study 3 Speaking Test (/b/ and /p/ scores)

Combining the groups together, the mean improvement was higher from audio (6.06%) than from using the head (5.30%) but the standard deviation is high, showing that there was a lot of variation in the results. 6 out of 17 participants

improved more from audio than from head, while 7 out of 17 participants improved more from the talking head than from audio. This shows that some individuals benefit more than others from the visualization. For the speaking test, the visualization may help with speech production. However, /p/-/b/ is a very subtle visual difference, and is mainly learned by listening to the contrast, so audio alone can be more beneficial in learning to perceive the distinction, which may in turn lead to better speech production.

On average over all users, there was an improvement from the start of the first session to the end of the second session. Ten participants improved using the audio version, and an equal number improved using the head (Table 7.23). Thus the majority did improve from using the software. Most of those who did not improve were those who achieved higher scores to begin with, so had less room for improvement.

In Table 7.24 the results are considered for only the four users who scored lowest overall (users 1, 2, 9 and 10). For speaking /b/ and /p/, for the 4 users who were the lowest scoring users overall, the mean improvement was higher for the head (6.25%) than for audio (3.13). On average for the 4 users who were the lowest scoring users overall, audio was better than the head in the listening test, but the head was better than audio for the speaking test. So this result indicates that for some of those with a definite difficulty in perceiving /p/ vs /b/, the audio alone was more effective when practicing an audio discrimination task, but when practicing speech production, the visualization was more effective than audio alone for teaching how to pronounce the sounds.

User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post	Group
1	50	72.5	47.5	55	AV_A
2	57.5	57.5	40	40	AV_A
9	75	87.5	82.5	77.5	A_AV
10	52.5	45	52.5	60	A_AV

Table 7.24: Study 3 Lowest-scoring users' Speaking Scores (/b/ and /p/)

### 7.5.7 Study 3 Speaking Test /b/ sounds

Both conditions gave an improvement overall for speaking /b/ sounds (Table 7.25). The improvement was higher for audio (4.27%) than audiovisual (0.59%). The individual speaking scores for /b/ are shown in Table 7.26.

Study 3 Mean Scores for Speaking (/b/ sounds) %				
Group	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test
AV-A	65.6	66.3	60	61.9
A-AV	76.1	82.8	75	75.6

Table 7.25: Study 3 Mean Speaking Scores (/b/)

User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post	Group
1	50	65	35	30	AV_A
2	45	30	20	20	AV_A
3	60	70	60	70	AV_A
4	80	75	75	80	AV_A
5	70	80	75	65	AV_A
6	65	65	80	70	AV_A
7	80	85	70	85	AV_A
8	75	60	65	75	AV_A
9	60	80	80	75	A_AV
10	70	50	60	55	A_AV
11	85	90	85	95	A_AV
12	70	85	80	65	A_AV
13	80	75	70	70	A_AV
14	75	85	70	75	A_AV
15	70	100	85	85	A_AV
16	85	85	95	80	A_AV
17	90	95	50	80	A_AV

Table 7.26: Study 3 Speaking Scores (/b/)

### 7.5.8 Study 3 Speaking Test (/p/)

The audiovisual condition showed a higher mean improvement (10.0%) than audio alone (7.85%) in speaking for /p/ sounds (Table 7.27 and Figure 7.16). 11

out of 17 participants improved from audio and 12 out of 17 improved using the head. 6 out of 17 participants improved more using audio, while 8 out of 17 improved more using the head (Table 7.28).

Study 3 Mean Scores for Speaking (/p/ sounds) %				
Group	Session 1 pre-test	Session 1 post-test	Session 2 pre-test	Session 2 post-test
AV-A	69.4	79.4	65.6	71.9
A-AV	66.1	75.6	66.7	76.7

Table 7.27: Study 3 Mean Speaking Scores (/p/)

User	Session 1 pre	Session 1 post	Session 2 pre	Session 2 post	Group
1	50	80	60	80	AV_A
2	70	85	60	60	AV_A
3	65	55	50	60	AV_A
4	45	75	50	50	AV_A
5	85	90	85	95	AV_A
6	65	65	75	75	AV_A
7	80	85	60	80	AV_A
8	95	100	85	75	AV_A
9	90	95	85	80	A_AV
10	35	40	45	65	A_AV
11	55	85	40	80	A_AV
12	55	70	80	70	A_AV
13	70	75	50	65	A_AV
14	85	80	75	75	A_AV
15	50	75	75	90	A_AV
16	85	85	80	90	A_AV
17	70	75	70	75	A_AV

Table 7.28: Study 3 Speaking Scores (/p/)

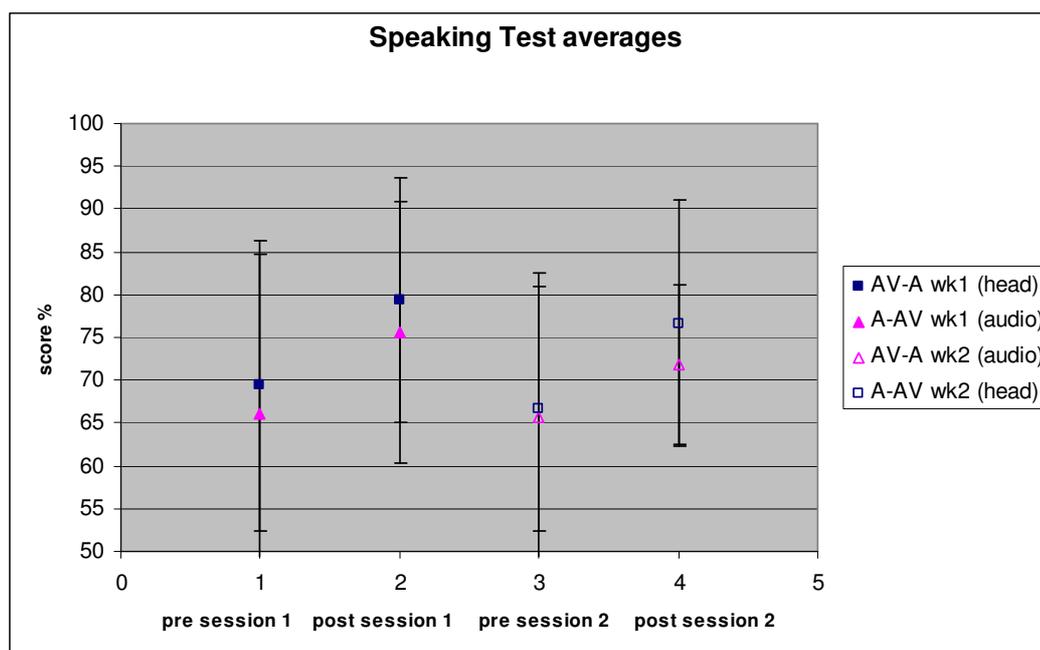


Figure 7.16: Study 3 Speaking Test Scores (/p/ sounds)

### 7.5.9 Study 3 User Feedback

The user questionnaire asked the users to rate each feature of the software on a Likert Scale of 1-7:

- 1= Strongly Disagree
- 2 = Moderately Disagree
- 3 = Slightly Disagree
- 4 = Neutral
- 5= Slightly Agree
- 6 = Moderately Agree
- 7 = Strongly Agree

The responses from the two Groups AV-A and A-AV are presented first separately (Table 7.29) and then combined (Table 7.30).

Study 3 Mean response on Likert scale 1-7				
Question	Group AV-A Audio	Group AV-A Talking head	Group A-AV Audio	Group A-AV Talking head
The Pronunciation Assistant software was helpful in learning pronunciation.	6.5	5.8	6.2	6.3
I found the external view of the talking head helpful.	NA	6.3	NA	6
The talking head looked realistic.	NA	6.4	NA	6.2
The speech animation appeared natural.	NA	6.1	NA	5.8
I found the listening practice with the talking head helpful.	6.4	6	6.3	6.4
I found the listening test with the talking head helpful.	6.3	6	5.4	6.2
I found the speaking practice with the talking head helpful.	6.5	6.4	6.3	6.7
I found the recording function in the Pronunciation Assistant software helpful.	6.4	6.1	6.4	6.4
The Pronunciation Assistant software is interesting to use.	6.1	6.5	6.7	6.2
The Pronunciation Assistant software is satisfying to use.	6.3	6.1	6	5.9
The content of the lesson matched my needs.	5.6	5.6	5.6	5.6
Which version did you prefer? (A rating above 4 shows a preference of the talking head over the audio version)	5.7	NA (their first week)	NA (their first week)	5.4

Table 7.29: Study 3 User Feedback

Combining the two groups, the mean responses of the 17 users are in Table 7.30.

Question	Audio	Talking head	Audio Likert to %	Talking Head Likert to %
The Pronunciation Assistant software was helpful in learning pronunciation.	6.4	6.1	84.3	80
I found the external view of the talking head helpful.	NA	6.1	NA	80
The talking head looked realistic.	NA	6.3	NA	82.9
The speech animation appeared natural.	NA	5.9	NA	77.1
I found the listening practice with the talking head helpful.	6.4	6.2	84.3	81.4
I found the listening test with the talking head helpful.	5.8	6.1	75.7	80
I found the speaking practice with the talking head helpful.	6.4	6.5	84.3	85.7
I found the recording function in the Pronunciation Assistant software helpful.	6.4	6.3	84.3	82.9
The Pronunciation Assistant software is interesting to use.	6.4	6.4	84.3	84.3
The Pronunciation Assistant software is satisfying to use.	6.1	6	80	78.6
The content of the lesson matched my needs.	5.6	5.6	72.9	72.9
Which version did you prefer? (A Likert rating above 4 shows a preference of the talking head over the audio alone.)		5.55		72.1

Table 7.30: Study 3 User Feedback (combined groups)

On average, the users agreed that the software was helpful in learning pronunciation. Group A-AV, who had used the audio version before the head, gave the head a higher rating, whereas Group AV-A gave the head lower ratings, but that was in week 1, before they had tried the audio version.

Most agreed that the external view of the head was useful. All agreed that it looked realistic, and most agreed that the speech animation looked natural. The

ratings for most of the tutoring lesson features were similar for the audio and audiovisual versions. One user found the speech recognition was not very accurate for his voice, and would like this to be improved. He stated that the software was most useful for listening practice. Some users commented that they would like the software to demonstrate a wider range of consonant or vowel sounds, rather than just /b/ and /p/, as not all had a specific /b-/p/ problem; some had more difficulty with another set of sounds, while others wished to improve their pronunciation in general.

Most preferred the talking head, and when asked why, many commented that it was useful to see the lips. One reported that she would prefer the audio version so she was not distracted with visuals, but this was the only comment in favour of the audio alone. One user reported that the head made him feel as if he was practicing with a real tutor. One user preferred audio alone, but showed an improvement in speaking after using the head, and was worse after audio, and showed no improvement for listening. Another user slightly preferred audio, but this user's pronunciation improved more from the head. Conversely, those who did not improve using the head, still rated it as useful. There was no correlation between the users' opinions and their scores. So in conclusion, the talking head was strongly preferred over audio alone by the majority of users.

#### 7.5.10 Study 3: Summary of Results

Most users showed an overall improvement from the start of Session 1 to the end of Session 2, although this improvement was not significant given the small number of subjects. The standard deviation was high, because individual variations caused fluctuations in scores. There was no significant difference found using T-tests ( $p = 0.1$ ) between audio alone and the talking head. On average over all 17 users, there was an improvement after using the talking head, and after audio alone, for listening and speaking.

Overall, for the listening test, audio alone gave greater improvement than the talking head. An explanation for this result is that the listening test performance is improved by the practice of listening to the audio contrast. Visualization may not help in this task because the animation may distract users from listening to the

audio contrast. One user, who preferred the audio version, did comment that the animation was distracting. Another who moderately preferred the audio version commented that the motion in the animations was fast, and thought that sequential photos would be better. Long animated sequences may cause learners to have difficulty in remembering the entire process. Constant motion can be disturbing, and learning can improve when there are visual rests, as memory is enhanced when users can stop and think (Reeves et al. 2000). Therefore for the listening test, the audio alone gave higher improvement, because this allowed users to practise listening without the distraction of the animation. Also, since the pre/post tests involved listening to audio alone, without the aid of visualization, it was harder for the audiovisually-trained group, who had less practice of listening without visualization, whereas the audio-trained group had been trained to listen with audio alone, so they performed better in the audio-alone listening post-tests.

For speaking, the audio alone gave a higher improvement than the talking head for /b/ sounds, but the talking head was better than audio for /p/ sounds. Therefore the visualization was more helpful for /p/ than for /b/. This could be because this talking head showed more emphasis for the /p/ sounds; from observation of the data-driven head, the /b/ sounds do not show as much lip movement as the /p/ sound, which had more prominent lip movements. For /b/ sounds, the voicing difference could be heard in the audio, so the audio helped more than the video for /b/. Another explanation is that for Arabic speakers, /b/ is easier to pronounce than /p/ , because the /p/ sound does not exist in their native language (Thelwall et al. 1990), so they may pronounce /p/ as /b/. Also for Japanese speakers the /p/ sound is unaspirated so it can sound like an English /b/ (Hazan et al. 2005). Therefore it could be expected that many of the participants in this experiment would find /b/ easier than /p/. The scores for /b/ and /p/ were in the same range on average, due to the diversity of the participants' native languages and levels of English. However, looking at those with the lowest scores, they scored lower for speaking /p/ than /b/ in the first pre-tests, so they had more room for improvement with speaking /p/ than /b/. Therefore for those with a definite difficulty, the visualization was more beneficial.

On average for the 4 lowest scoring users overall, the head was better than audio for the speaking test, but audio was better than the head in the listening test. So this result indicates that for those with a definite difficulty in perceiving /p/ vs /b/, the audio alone was more effective when practicing an audio discrimination task, but when practicing speech production, the visualization was more effective than audio alone for teaching how to pronounce the sounds. However these findings are limited, and would have to be repeated on larger groups to be able to give significant results.

The talking head was strongly preferred over audio alone by the majority of users, as shown by the final rows of Tables 7.29 and 7.30, where the Likert ratings were all greater than 4, showing a preference of the talking head over the audio alone (Appendix E.10 Question 13). The feedback on all the features of the software was positive, with many users reporting that they thought the talking head was useful, realistic, and interesting to use.

## 8 Conclusions and Future Work

This thesis has explored the application of visual speech in perception and pronunciation training. It has developed and evaluated a new software application featuring a talking head as an aid for pronunciation practice in second language learning, which is the first of its kind for British English. This thesis has provided empirical data which shows that learners liked using talking heads in second language learning, and some learners improved more from using talking heads than from audio alone.

This thesis investigated the development of three talking heads and their deployment in second language learning. Its contribution is to explore a range of techniques, compare different approaches and shed light on the advantages and disadvantages of the different approaches to creating talking heads.

The studies began with a completely generic viseme-driven synthetic head, not based on any person, with a synthetic voice. Parts of the talking head were then replaced with parts specific to one speaker: another viseme-driven head was created with a photo-based face, tongue positions based on MRI images of that speaker, and the real voice of that speaker. A novel corpus was acquired during a research visit to the "Département Parole et Cognition", GIPSA-Lab, Grenoble, comprising MRI, EMA and video data, which is the first of its kind for a British English female speaker. Finally, collaboration with GIPSA-Lab produced a new data-driven head with facial geometry and lip movements modelled on one British English female speaker.

The quality of speech animation was evaluated in Modified Rhyme Tests, which found all the synthetic heads to be more intelligible than audio alone, though less intelligible than real video. The non-photo-based viseme-driven head showed a gain in intelligibility compared to audio speech alone, and was almost as intelligible as the video of a real speaker under similar noise conditions. Certain visemes were confused with others, and could be modelled more accurately, but overall the visemes were identifiable. In a subjective naturalness evaluation survey, the visual speech of the non-photo-based viseme-driven talking head was

rated to be moderately natural. A Modified Rhyme Test found the data driven head to be more intelligible than the photo-based viseme-driven head. A naturalness survey showed that the data driven head was perceived as more natural than the non-photo-based viseme head. These results show that the data driven model is more accurate than viseme-driven. This is because the viseme-driven model is more generic, not entirely modelled on real data, and only an approximation of abstracted parameters, whereas the data driven model is derived from real data from a video corpus of a specific speaker, and captures more subtleties.

The efficacy of the talking heads in a tutoring system was evaluated in three user trials involving second language learners of English. The studies aimed to determine the benefit of visual speech in second language learning, and its effectiveness as a teaching tool for this application.

A pilot trial of the non-photo-based viseme-driven head was run with five native Arabic speakers learning English as a second language. Positive feedback was received from the students, who enjoyed using the software, and found the visualization useful. Generally, there was an improvement in speaking and listening, from the first test to the final test, for both groups. Overall the talking head gave a more consistent improvement in pronunciation than audio alone.

The photo-based viseme-driven head was trialled with two native Arabic speakers learning English as a second language. The user with the internal view showed no improvement, while the user with the external view showed some improvement. Their questionnaire feedback showed that the users thought that both the internal and external views were useful. Previous studies have had similar results, for example, Baldi users reported they preferred the internal visualization to external alone, although no significant difference was found (Massaro et al. 2003). More research will be needed to show whether internal visualization can make a difference in learning pronunciation.

The data-driven head was evaluated in a crossover experiment with 17 second-language speakers, comparing the data-driven head against audio. Generally, there was an improvement in speaking and listening, from the first test to the final test,

for both groups. For the listening test, audio alone gave slightly greater improvement than the head. An explanation for this result is that the listening test performance is improved by the practice of listening to the audio contrast, and visualization may not help in this task because the animation may distract users from listening to the audio contrast. Therefore for this listening test, the audio alone gave a higher improvement, because this allowed users to practice listening without the distraction of visualization. Incubation effects were observed, as the users improved their listening ability between the two sessions. The order in which the training was administered made a difference to the users' listening performance. An interference effect was observed, where those who had auditory training first did not benefit from the audiovisual training in their second session. This suggests that for maximum benefit, the audiovisual training should be administered first.

For the speaking test, some users showed more improvement in the audiovisual condition, while others improved more in the audio condition. The visualization may help with speech production, but although /b/ and /p/ do have some visible differences, this difference may not always be salient enough to aid discrimination: "The distinction between /p/ and /b/ is not likely to be disambiguated by visual cues as visual cues carry little information to the voicing distinction." (Hazan et al. 2002). The main difference between /b/ and /p/ is in voice onset time, and this difference may be mainly learned by listening to the contrast. The speaking results showed that the pronunciation of /p/ improved more after audiovisual training, whereas /b/ improved more from audio training. A similar disparity was also found by (Hazan et al. 2005), who found that /r/ pronunciation improved more for the audiovisual training group than the audio training group, but that /l/ pronunciation did not.

The experimental findings are consistent with previous research which found that talking heads did not improve listening perception over audio alone. (Hazan et al. 2005) found that for the perception of the /l-/r/ contrast, audiovisual training was not more effective than auditory training, and the greatest increase in scores was seen in the audio-trained group.

The experimental results also support evidence that the combining of audio with visualization may be useful to some and not others. Large individual variations were also experienced by (Hazan et al. 2005), who found that some learners improved their pronunciation significantly more than others. “It is well known that individuals vary significantly in terms of their lipreading skills ... and also in their ability to integrate auditory and visual information ... Indeed, it is plausible that, for some learners at least, perceptual training will be more successful when focused on a single modality, with the visual modality acting as a distractor.” (Hazan et al. 2005). Individual learners may use very different learning strategies. Hazan and Kim investigated whether specific auditory or cognitive skills were linked to initial sensitivity to a novel phonetic contrast or to the degree of learning following computer-based phonetic training, and found that rate of learning was not correlated with any of the auditory or cognitive skills tested (Hazan et al. 2010). Therefore it is difficult to predict which learners will benefit from computer-based phonetic training.

These findings verify existing work, and extend it by showing that the findings from the experiments by Hazan et al., which used natural audiovisual stimuli, are also true for a synthetic talking head. These experiments have also extended the range of phonemic contrasts which have been studied. Overall, no significant improvement was found for any of the audio or audiovisual conditions, and in each experiment no significant difference was found between the two conditions. This is not only due to the small sample sizes and large individual variations, but also because over the short training periods, any improvement in pronunciation is likely to be very small. In each experiment, one judge was used for all of each speaker’s responses, to keep the scoring consistent, but multiple judges could be used, taking the average of their judgments. The variation of the experimental results shows that it is difficult to measure pronunciation improvement in such a precise way. The approach usually used by English language tutors is more holistic, assessing the overall quality and intelligibility of the speech. Many teachers do not consider it important to test specific features, because in real-life situations the context allows learner to interpret what they hear, or to be understood even if the sounds are not pronounced correctly. Moreover, testing

oral skills is often difficult to administer given the large number of students to be tested (Bobda 2006). Future studies would be more longitudinal, training students with the software over several months, involving their own tutors and possibly their examination scores from their usual English language classes to assess their long-term improvement.

One benefit of visualization is to make learning more interesting. Animation can bring learning points to life, and relevant imagery increases retention (Shepard 1967). Even if there is very little visual difference between /b/ and /p/, the animated head makes the application more interesting to use, so users are more motivated to concentrate. Fatigue may be reduced if the repetitive practice is made less boring with visual stimuli. Even though no significant improvement was found in listening or pronunciation, the feedback in the questionnaires was positive and the majority of users reported that they found the external and internal visualization useful. These results are consistent with previous studies (Massaro et al. 2003). Baldi was used for /r/-/l/ training on 11 Japanese speakers. Since /r/ and /l/ have different tongue positions it could be expected that internal visualization would be helpful in their studies; however, tests on Baldi did not find a significant improvement, and no difference between internal and external visualization, although users said that they preferred the training with the internal view. While studies to date have not shown a significant difference from using talking heads, this research and previous research have shown that users believe it to be helpful. Since users like the talking heads and enjoy using the software, they may be more motivated to use it for practice, and motivation is an important factor in pronunciation learning. Factors influencing motivation can be negative, e.g. fear of derision (which would not occur using the talking head, as tuition could take place alone with the computer), and positive, e.g. desire to study for the sheer pleasure of learning. A learner who is strongly motivated is more likely to focus on the training, practise more, and succeed in improving their performance (Dabic 2010). Computer-based instruction can provide an increased level of participation through interactivity, which leads to higher levels of cognitive engagement and therefore higher levels of retention (Tobias et al. 2011).

This research has provided a tool for teaching pronunciation, using a computer-generated head to visualize internal articulator motion in a way which cannot be demonstrated by a human tutor alone. Evaluation has shown that the software improved perception and production, even if it was the audio modality which helped more than the visualization, in some cases. Although the experiments have not shown definitively that the visualization improves performance over audio alone, the feedback from users shows that they do think it is helpful. The software can be useful for perception training, giving listening practice and feedback, and by extension of learning perception, speech production can be improved. The application can be used in speaking practice where users listen to sounds and then try to reproduce them, and can listen to their own recordings to compare for themselves their production against the example. Some speech recognition has been incorporated to give feedback on pronunciation, although it is not definitive and does not give specific feedback tailored to the user, so it cannot replace a human tutor, but can be used as an aid to solitary practice. A major benefit is that students can use the software for practice outside of teaching hours, in their own time, at home or anywhere in the world. The self-directed nature of computer-based tuition can also lead to higher retention, as the content is followed at the rate which suits the learner, who can stop and reflect, building internal models, and relating new knowledge to existing knowledge, and repeat the lesson in a way that is not possible in classroom-based tutoring (Kulik et al. 1991; Tobias et al. 2011).

## 8.1 Future Work

A question to consider is why Computer Assisted Language Learning (CALL) has not yet been taken up by many users. A problem is that the technology for providing feedback has not advanced enough to replace the need for human tutors, and there are unsolved issues with the unnaturalness of character embodiment. The realism and believability of appearance, behaviour and interaction need to be improved to increase acceptance of talking head technology.

Further work could improve the realism of the visual speech. The viseme-based head was shown to be less intelligible than the data-driven head, with more

confusion between visemes, so these could be more accurately modelled. For example, a perceptual test based on the McGurk effect (McGurk et al. 1976) could help to identify weaknesses in the synthesis of certain visemes (Cosker et al. 2005). Lalalde has shown that using separate visemes can give better results than a single B-M-P viseme (Lalalde 2010). Utterances with less mouth movement were rated as less natural in the experiment in Section 6.3, so adding extra emphasis could increase the perceived naturalness. More expressive speech could appear to be more natural, because face and head movements may distract attention from the lips, as well as presenting more lifelike behaviour. The perceived intensity of facial expressions can be increased by increasing shape and motion information, and including eye motion (Wallraven et al. 2008). The modelling of non-verbal behaviour could be improved, for example, with more realistic eyelid kinematics (Steptoe et al. 2010). Further experiments could investigate whether a more expressive talking head is perceived as more realistic, and whether this is preferred in the tutoring application. Empirical studies (Reeves et al. 1996) have found that people are inclined to see media as living and animate, and there is a tendency to anthropomorphize objects, for example, inferring a positive attitude and personality from the perception of a smiling expression, so it is likely that more expressive behavior would increase the illusion of humanity, and increase the acceptance of the talking head.

Future work could use the 3D MRI data collected in the corpus to make the first full 3D data-driven articulatory model of a British English female speaker, using similar techniques to those used for a French speaker at GIPSA-Lab (Badin et al. 2008). This would then produce a full 3D internal and external model of a British English speaker to be used in a speech tutoring application. Further experiments would investigate whether the internal view with more accurate articulation modelling would give a greater improvement in learning pronunciation.

Further studies on larger groups of participants could investigate whether a more natural head has a greater effect on learning, and could determine the benefits of using talking heads in learning a language. Further studies could compare the effects of various aspects of the animation of the talking head, such as the impact of more natural facial expressions. Longer training periods would be required on

larger groups of participants to determine whether the use of talking heads can be of benefit in learning pronunciation.

There are many different configurations and variables which could be investigated further. Every speaker has their own articulation strategy, and their own unique facial geometry, and the data-driven head was modelled on only one speaker, so it demonstrates how that person speaks, rather than some “ideal” based on the average articulations of many speakers. It may be possible that this particular data-driven head improved learners’ pronunciation of /p/ because this particular speaker’s facial geometry and articulation strategy showed greater emphasis in the production of /p/ sounds, making this speaker a good candidate for demonstrating that sound. Other speakers could have thinner lips and may not show as much lip protrusion or movement, which could affect the results. The tutoring application could be extended to include a wider range of segmental sounds, and other features of pronunciation, such as prosody. Future studies could experiment with other data-driven and viseme-driven heads, created using different approaches and a range of speakers. The synthetic heads could be compared against natural video, as a baseline for realism. (Hazan et al. 2005) found that when comparing the Baldi talking head against natural video, those trained with the synthetic face showed less improvement than those trained with audio or with natural audiovisual stimuli, but this may be because the quality of the synthetic heads was not yet sufficiently intelligible. The Modified Rhyme Tests in Chapter 5 found the data-driven head to be more intelligible than the viseme-driven head, yet still not as intelligible as natural video. It is expected that more realistic movements would be better for teaching pronunciation, so it is expected that the natural video would give better performance in tutoring than the external view of the data-driven head, which would be better than the external view of the viseme-based head. However, the internal views of the synthetic heads could give an advantage, so these could be used in training in addition to natural video.

A possible future study could experiment with exaggerating and slowing down articulations, which is another possible benefit of synthetic heads, which could make the correct lip and tongue positions easier to see than in natural imaging techniques. Spectrographs show that the sound changes involved in the difference

between /p/ and /b/ last only a few hundred milliseconds, and these rapid changes can be too brief for dyslexic children to recognize (Straub 2001). Tallal found that children with developmental language learning problems are impaired in their ability to process brief, rapidly successive acoustic stimuli, specifically in the tens of millisecond time window. Children were trained to recognize sounds by using modified acoustic speech to amplify and temporally extend the brief, rapidly successive cues, and through adaptive training, the speech was gradually made faster, until it was the normal rate. Trials found that intensive daily training (two hours a day, five days a week, for four weeks) resulted in highly significant improvements in speech discrimination and language processing compared to a control group (Tallal 2001). A similar approach could be extended to visual stimuli, and studies could investigate whether slowed-down, hyper-articulated speech animation could improve learning better than naturally-articulated speech animation.

A future direction for research would be for the tutoring system to give better feedback to the users, and to achieve this, a more accurate speech recognition system would need to be integrated. For example, many hidden Markov model (HMM) recognisers can output a goodness-of-fit probability score indicating how acoustically similar an utterance is to a pre-defined ideal (Green et al. 2003). This would allow the system to rate more precisely how closely a learner's production matches the target the recogniser has been trained to expect, for example, a model British English speaker's production. Visual feedback could be tailored to represent the user's score, which would give them a visual target to aim towards, and an indication of their own improvement, as used in the Ortho-Logo-Paedia articulation program, which provides immediate visual feedback on performance, which can be "right"/"wrong" or graduated (Hatzis et al. 2003). Research in speech therapy suggests that feedback must be both auditory and visual (Palmer 2004). An interesting direction for future research would be visual feedback based on analysis of the users' articulator movements. One approach is to use acoustic-to-articulatory inversion (Ben Youssef 2011) to allow the system to infer which articulator movements the user has made from analysis of the acoustic signal. Alternatively, computer vision techniques can be used, since machine-based lip-

reading systems can now outperform human lip-readers (Hilder et al. 2009). It has been shown that acoustic-to-articulatory inversion can be improved by adding visual features extracted from the speaker's face, as important articulatory information can be extracted using only a few facial measurements (Kjellström 2009). Tongue movements can then be reconstructed from the audio and video information (Kjellstrom et al. 2006). Then the talking head could display the user's own movements, and display what the correct movements should be for comparison, and give appropriate instruction specific to that user. If such a system could give accurate feedback to a user, then this could be of great benefit in practicing speech without the presence of a human tutor and would be a major benefit of using talking heads in addition to traditional methods.



## Appendix B: Speech Tutoring Application Screenshots

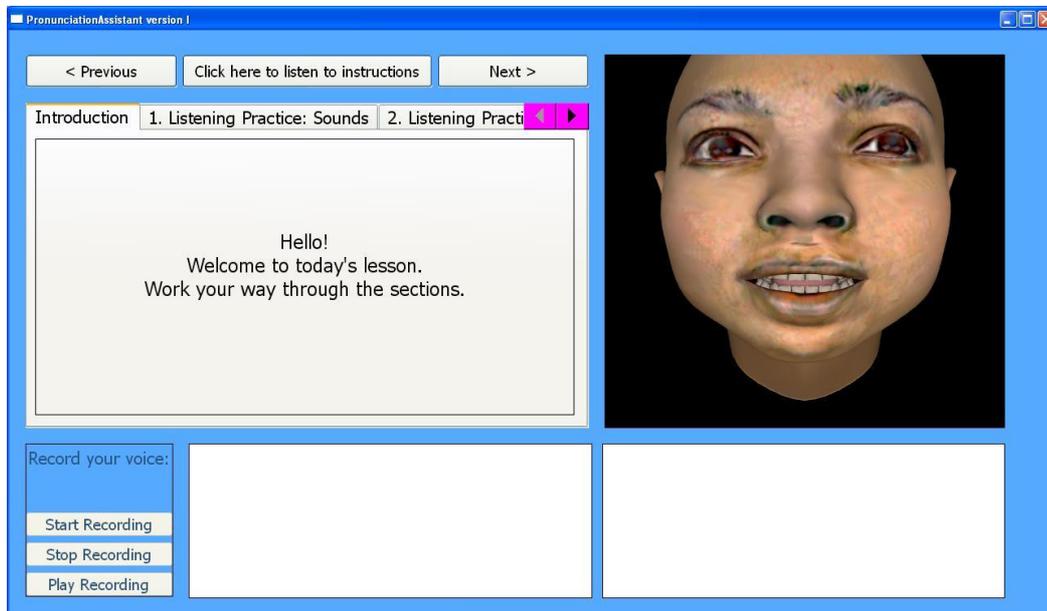


Figure B.1: Introduction

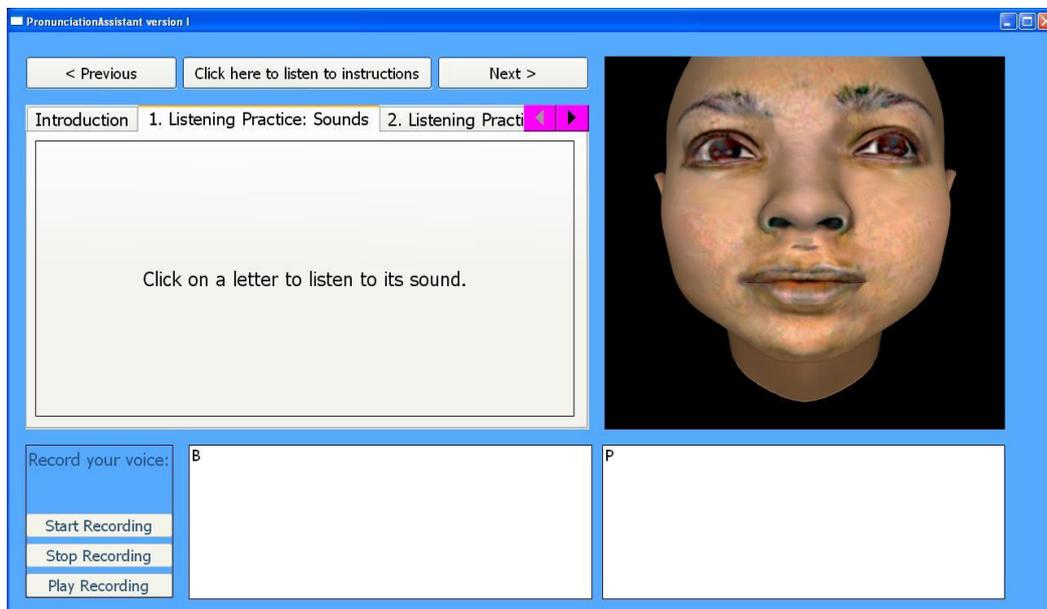


Figure B.2: Listening Practice: Sounds

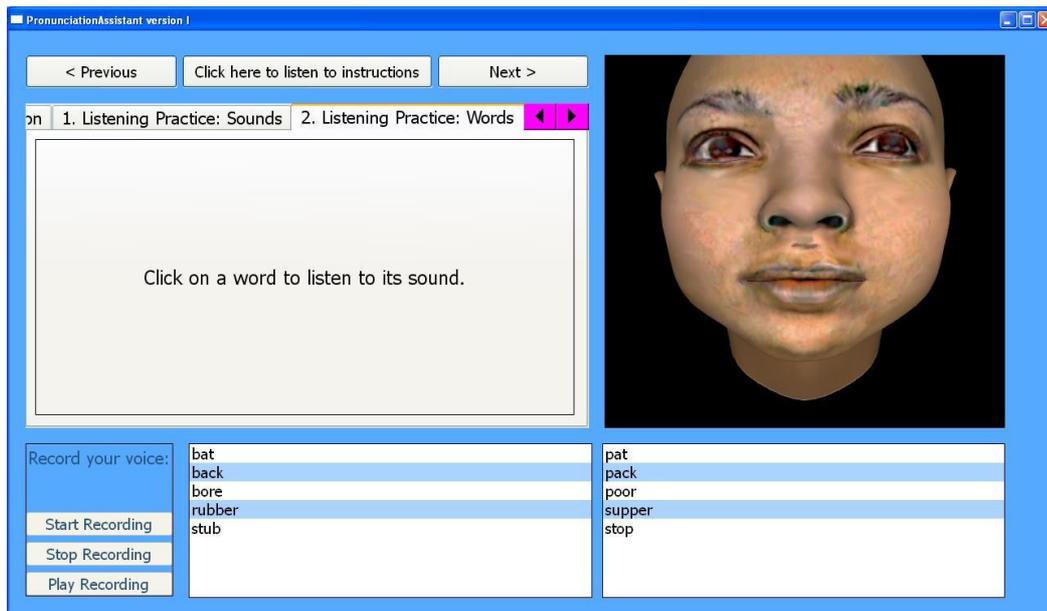


Figure B.3: Listening Practice: Words

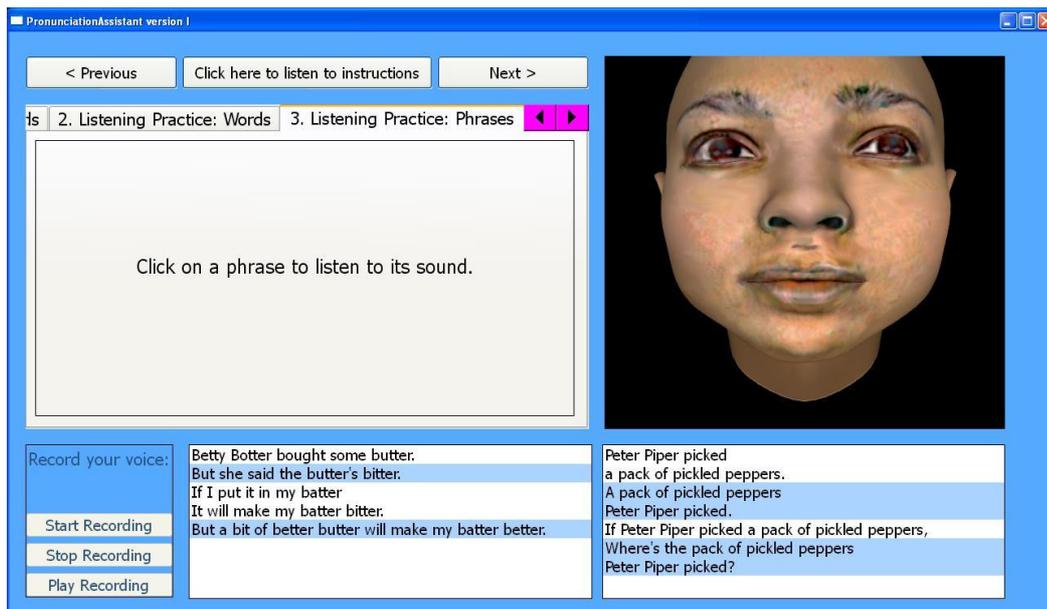


Figure B.4: Listening Practice: Phrases

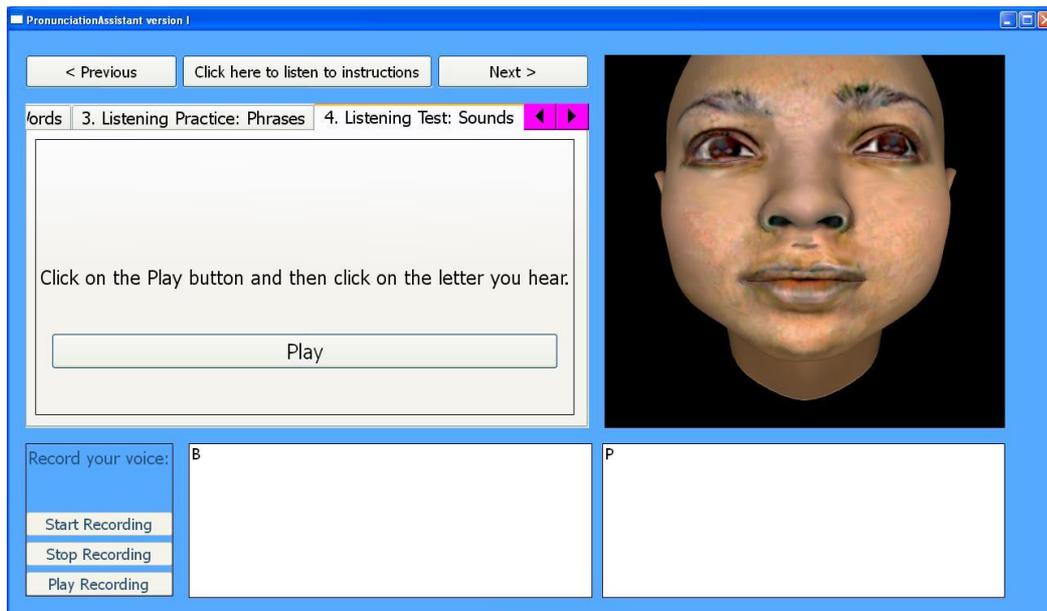


Figure B.5: Listening Test: Sounds

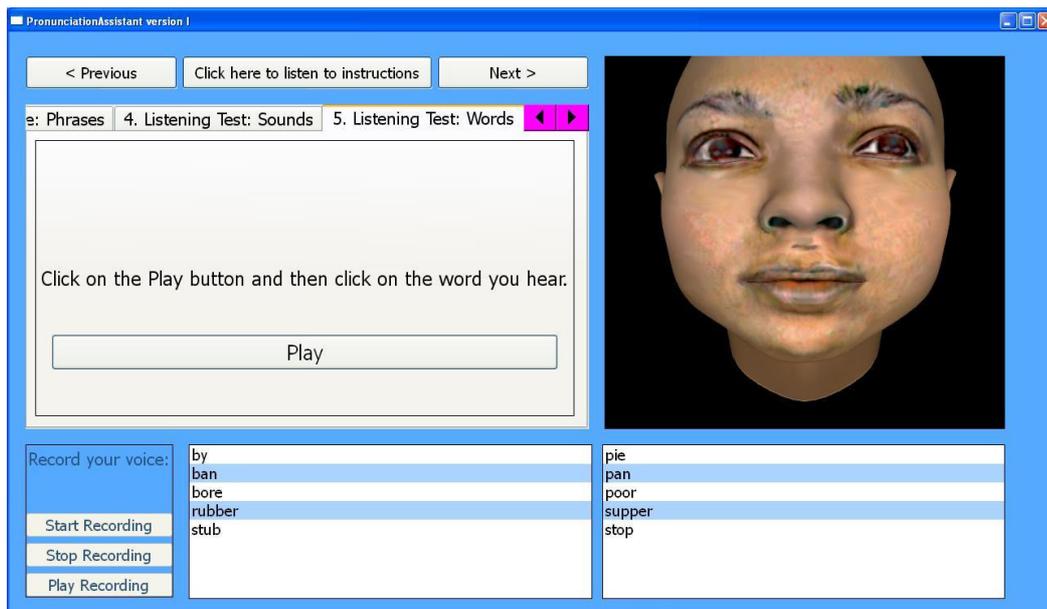


Figure B.6: Listening Test: Words

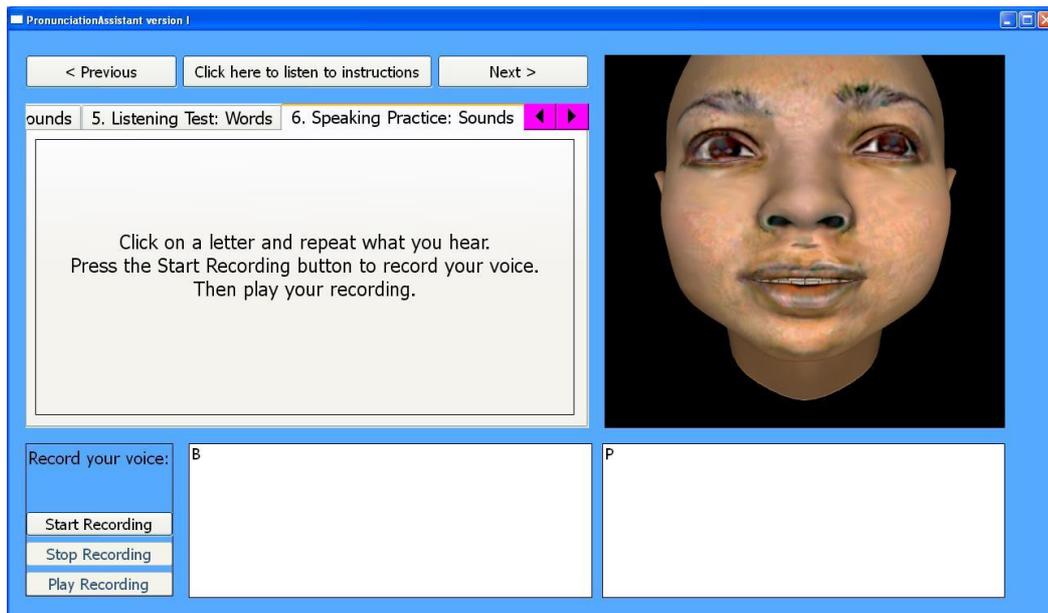


Figure B.7: Speaking Practice: Sounds

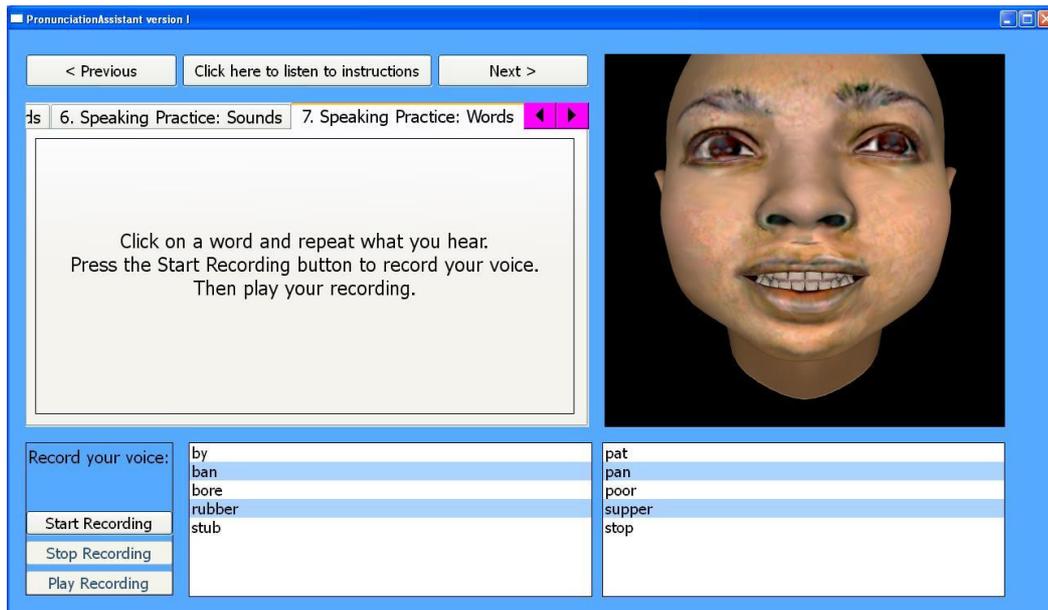


Figure B.8: Speaking Practice: Words

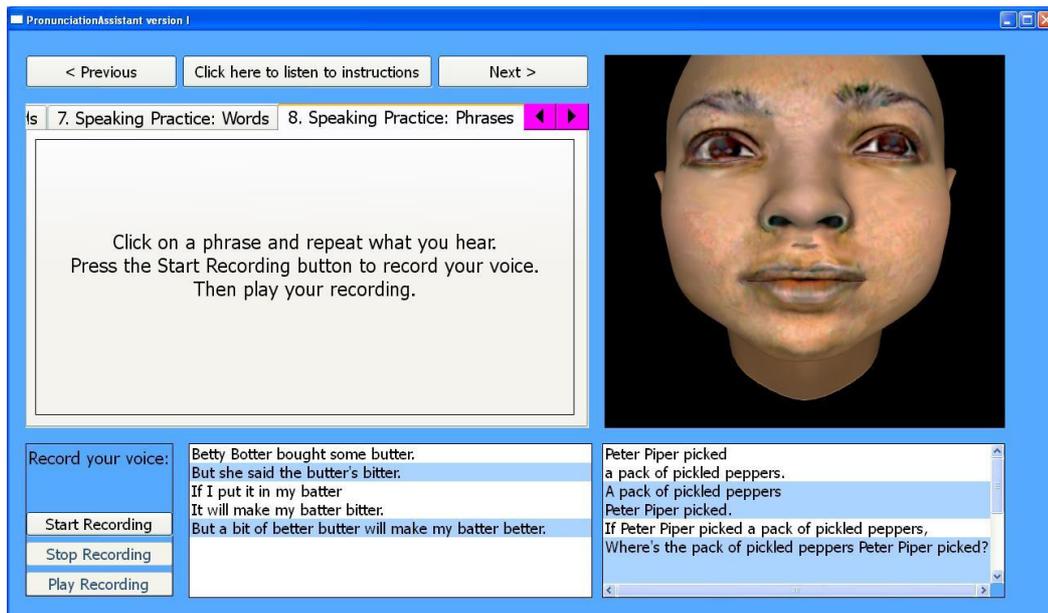


Figure B.9: Speaking Practice: Phrases

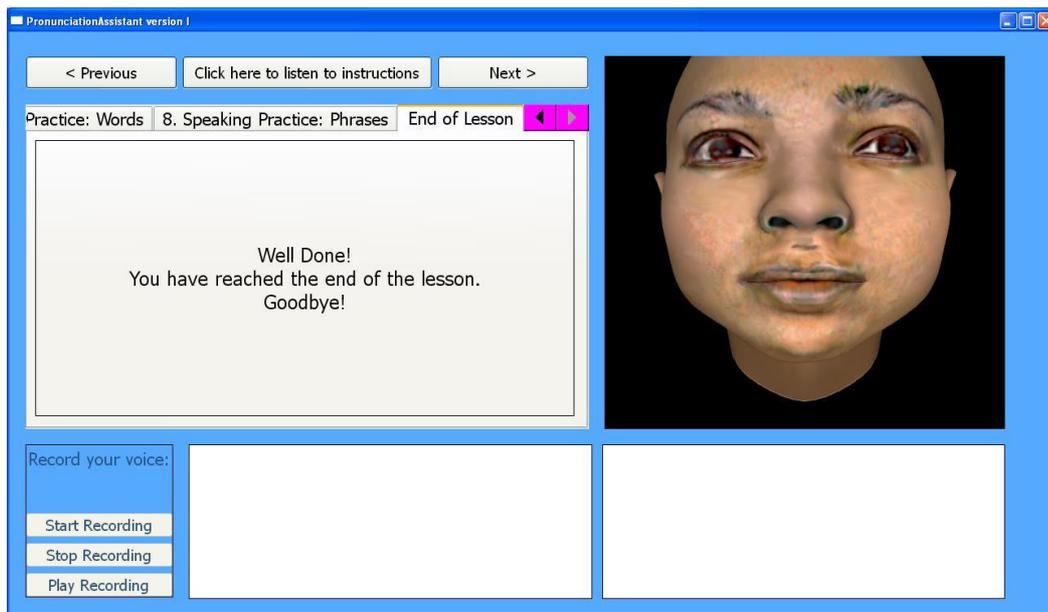


Figure B.10: End of Lesson

## Appendix C: Words for Video Corpus

### C.1: Phrases for tutoring application

All the phrases used in the tutoring application and recorded in the video corpus are listed below.

1. Hello!
2. Welcome to today's lesson.
3. Work your way through the sections.
4. correct
5. incorrect
6. Click on a letter to listen to its sound.
7. b@
8. p@
9. Click on a word to listen to its sound.
10. Bat
11. Back
12. Bore
13. Rubber
14. Stub
15. Pat
16. Pack
17. Poor
18. Supper
19. Stop
20. Click on a phrase to listen to its sound.
21. Betty Botter bought some butter.
22. But she said the butter's bitter.
23. If I put it in my batter it will make my batter bitter.
24. But a bit of better butter will make my batter better.
25. Peter Piper picked a pack of pickled peppers.
26. A pack of pickled peppers Peter piper picked.
27. If Peter Piper picked a pack of pickled peppers,
28. Where's the pack of pickled peppers Peter Piper picked?
29. Click on the Play button and then click on the letter you hear.
30. Click on the Play button and then click on the word you hear.
31. By

32. Ban
33. Pie
34. Pan
35. Click on a letter and repeat what you hear.
36. Press the Start Recording button to record your voice.
37. Then play your recording.
38. Put your lips together.
39. Using your voice, open your lips.
40. Repeat after me...
41. b@, b@, b@:
42. Open your lips with a puff of air.
43. p@, p@, p@
44. Click on a word and repeat what you hear.
45. Click on a phrase and repeat what you hear.
46. Well Done!
47. You have reached the end of the lesson.
48. Goodbye!

## C.2: MOCHA -TIMIT subset

The subset of sentences from the MOCHA-TIMIT corpus (Wrench 1999) recorded in the video corpus is listed below.

011. He will allow a rare lie.
014. Before Thursday's exam, review every formula.
035. Help celebrate your brother's success.
037. Critical equipment needs proper maintenance.
038. Young people participate in athletic activities.
077. Bagpipes and bongos are musical instruments.
109. Birthday parties have cupcakes and ice cream.
158. Our experiment's positive outcome was unexpected.
241. Clear pronunciation is appreciated.
438. The fifth jar contains big, juicy peaches.

## Appendix D: Stimulus Words of MRT

The 300 Stimulus Words of the Modified Rhyme Test (Meyer Sound 2010):

1	went	sent	bent	dent	tent	rent
2	hold	cold	told	fold	sold	gold
3	pat	pad	pan	path	pack	pass
4	lane	lay	late	lake	lace	lame
5	kit	bit	fit	hit	wit	sit
6	must	bust	gust	rust	dust	just
7	teak	team	teal	teach	tear	tease
8	din	dill	dim	dig	dip	did
9	bed	led	fed	red	wed	shed
10	pin	sin	tin	fin	din	win
11	dug	dung	duck	dud	dub	dun
12	sum	sun	sung	sup	sub	sud
13	seep	seen	seethe	seek	seem	seed
14	not	tot	got	pot	hot	lot
15	vest	test	rest	best	west	nest
16	pig	pill	pin	pip	pit	pick
17	back	bath	bad	bass	bat	ban
18	way	may	say	pay	day	gay
19	pig	big	dig	wig	rig	fig
20	pale	pace	page	pane	pay	pave
21	cane	case	cape	cake	came	cave
22	shop	mop	cop	top	hop	pop
23	coil	oil	soil	toil	boil	foil
24	tan	tang	tap	tack	tam	tab
25	fit	fib	fizz	fill	fig	fin
26	same	name	game	tame	came	fame
27	peel	reel	feel	eel	keel	heel
28	hark	dark	mark	bark	park	lark
29	heave	hear	heat	heal	heap	heath
30	cup	cut	cud	cuff	cuss	cud
31	thaw	law	raw	paw	jaw	saw
32	pen	hen	men	then	den	ten
33	puff	puck	pub	pus	pup	pun
34	bean	beach	beat	beak	bead	beam
35	heat	neat	feat	seat	meat	beat
36	dip	sip	hip	tip	lip	rip
37	kill	kin	kit	kick	king	kid
38	hang	sang	bang	rang	fang	gang
39	took	cook	look	hook	shook	book
40	mass	math	map	mat	man	mad
41	ray	raze	rate	rave	rake	race

42	save	same	sale	sane	sake	safe
43	fill	kill	will	hill	till	bill
44	sill	sick	sip	sing	sit	sin
45	bale	gale	sale	tale	pale	male
46	wick	sick	kick	lick	pick	tick
47	peace	peas	peak	peach	peat	peal
48	bun	bus	but	bug	buck	buff
49	sag	sat	sass	sack	sad	sap
50	fun	sun	bun	gun	run	nun

### D.1: MRT words used in Intelligibility Test 1: Audio Alone

Yellow shading indicates the chosen word used in the test.

Test stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	27	peel	reel	feel	eel	keel	heel	p	i
2	35	heat	neat	feat	seat	meat	beat	s	i
3	36	dip	sip	hip	tip	lip	rip	h	i
4	38	hang	sang	bang	rang	fang	gang	b	i
5	39	took	cook	look	hook	shook	book	sh	i
6	42	save	same	sale	sane	sake	safe	l	f
7	43	fill	kill	will	hill	till	bill	w	i
8	46	wick	sick	kick	lick	pick	tick	p	i
9	47	peace	peas	peak	peach	peat	peal	ch	f
10	48	bun	bus	but	bug	buck	buff	n	f
11	27	peel	reel	feel	eel	keel	heel	r	i
12	35	heat	neat	feat	seat	meat	beat	f	i
13	36	dip	sip	hip	tip	lip	rip	l	i
14	38	hang	sang	bang	rang	fang	gang	g	i
15	39	took	cook	look	hook	shook	book	h	i
16	42	save	same	sale	sane	sake	safe	v	f
17	43	fill	kill	will	hill	till	bill	t	i
18	46	wick	sick	kick	lick	pick	tick	k	i
19	47	peace	peas	peak	peach	peat	peal	s	f
20	48	bun	bus	but	bug	buck	buff	s	f

D.2: MRT words used in Intelligibility Test 1: Synthetic Talking Head (THVN)

Test Stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	4	lane	lay	late	lake	lace	<b>lame</b>	m	f
2	7	teak	team	teal	<b>teach</b>	tear	tease	ch	f
3	11	dug	dung	<b>duck</b>	dud	dub	dun	k	f
4	13	seep	seen	seethe	<b>seek</b>	seem	seed	k	f
5	15	<b>vest</b>	test	rest	best	west	nest	v	i
6	16	pig	<b>pill</b>	pin	pip	pit	pick	l	f
7	17	back	bath	bad	<b>bass</b>	bat	ban	s	f
8	18	way	may	<b>say</b>	pay	day	gay	s	i
9	21	cane	case	<b>cape</b>	cake	came	cave	p	f
10	22	shop	mop	cop	top	<b>hop</b>	pop	h	i
11	30	<b>cup</b>	cut	cud	cuff	cuss	cud	p	f
12	31	<b>thaw</b>	law	raw	paw	jaw	saw	th	i
13	33	<b>puff</b>	puck	pub	pus	pup	pun	f	f
14	37	kill	kin	kit	kick	<b>king</b>	kid	ng	f
15	40	mass	math	<b>map</b>	mat	man	mad	p	f
16	41	<b>ray</b>	raze	rate	rave	rake	race	y	f
17	44	sill	sick	<b>sip</b>	sing	sit	sin	p	f
18	45	bale	gale	sale	<b>tale</b>	pale	male	t	i
19	49	sag	sat	sass	sack	sad	<b>sap</b>	p	f
20	50	fun	sun	bun	gun	<b>run</b>	nun	r	i

D.3: MRT words used in Intelligibility Test 1: Natural Video

Test Stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	1	went	sent	bent	dent	tent	rent	b	i
2	2	hold	cold	told	fold	sold	gold	c	i
3	3	pat	pad	pan	path	pack	pass	th	f
4	5	kit	bit	fit	hit	wit	sit	f	i
5	6	must	bust	gust	rust	dust	just	j	i
6	8	din	dill	dim	dig	dip	did	p	f
7	9	bed	led	fed	red	wed	shed	r	i
8	10	pin	sin	tin	fin	din	win	w	i
9	12	sum	sun	sung	sup	sub	sud	n	f
10	14	not	tot	got	pot	hot	lot	p	i
11	19	pig	big	dig	wig	rig	fig	d	i
12	20	pale	pace	page	pane	pay	pave	v	f
13	23	coil	oil	soil	toil	boil	foil	o	i
14	24	tan	tang	tap	tack	tam	tab	b	f
15	25	fit	fib	fizz	fill	fig	fin	z	f
16	26	same	name	game	tame	came	fame	g	i
17	28	hark	dark	mark	bark	park	lark	l	i
18	29	heave	hear	heat	heal	heap	heath	t	f
19	32	pen	hen	men	then	den	ten	th (d)	i
20	34	bean	beach	beat	beak	bead	beam	ch	f

D.4: MRT words used in Intelligibility Test 2: Audio Alone

Test Stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	27	peel	reel	feel	eel	keel	heel	p	i
2	47	peace	peas	peak	peach	peat	peal	ch	f
3	35	heat	neat	feat	seat	meat	beat	f	i
4	46	wick	sick	kick	lick	pick	tick	k	i
5	38	hang	sang	bang	rang	fang	gang	r	i
6	12	sum	sun	sung	sup	sub	sud	ng	f
7	3	pat	pad	pan	path	pack	pass	th	f
8	2	hold	cold	told	fold	sold	gold	s	i
9	39	took	cook	look	hook	shook	book	t	i
10	42	save	same	sale	sane	sake	safe	m	f
11	8	din	dill	dim	dig	dip	did	l	f
12	48	bun	bus	but	bug	buck	buff	n	f
13	28	hark	dark	mark	bark	park	lark	b	i
14	43	fill	kill	will	hill	till	bill	w	i

D.5: MRT words used in Intelligibility Test 2: Viseme-Driven Synthetic Talking Head (THVP)

Test Stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	40	mass	math	map	mat	man	mad	p	f
2	7	teak	team	teal	teach	tear	tease	ch	f
3	33	puff	puck	pub	pus	pup	pun	f	f
4	13	seep	seen	seethe	seek	seem	seed	k	f
5	50	fun	sun	bun	gun	run	nun	r	i
6	37	kill	kin	kit	kick	king	kid	ng	f
7	31	thaw	law	raw	paw	jaw	saw	th	i
8	18	way	may	say	pay	day	gay	s	i
9	4	lane	lay	late	lake	lace	lame	t	f
10	18	way	may	say	pay	day	gay	m	i
11	31	thaw	law	raw	paw	jaw	saw	l	i
12	50	fun	sun	bun	gun	run	nun	n	i
13	45	bale	gale	sale	tale	pale	male	b	i
14	15	vest	test	rest	best	west	nest	w	i

D.6: MRT words used in Intelligibility Test 2: Data-Driven Synthetic Talking Head (THD)

Test Stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	21	cane	case	cape	cake	came	cave	p	f
2	22	shop	mop	cop	top	hop	pop	sh	i
3	30	cup	cut	cud	cuff	cuss	cud	f	f
4	49	sag	sat	sass	sack	sad	sap	k	f
5	36	dip	sip	hip	tip	lip	rip	r	i
6	44	sill	sick	sip	sing	sit	sin	ng	f
7	17	back	bath	bad	bass	bat	ban	th	f
8	41	ray	raze	rate	rave	rake	race	s	f
9	4	lane	lay	late	lake	lace	lame	t	f
10	22	shop	mop	cop	top	hop	pop	m	i
11	16	pig	pill	pin	pip	pit	pick	l	f
12	11	dug	dung	duck	dud	dub	dun	n	f
13	19	pig	big	dig	wig	rig	fig	b	i
14	1	went	sent	bent	dent	tent	rent	w	i

D.7: MRT words used in Intelligibility Test 2: Natural Video

Test Stimulus number	MRT List number							Sound tested	Initial (i) or Final (f)
1	14	not	tot	got	pot	hot	lot	p	i
2	34	bean	beach	beat	beak	bead	beam	ch	f
3	5	kit	bit	fit	hit	wit	sit	f	i
4	26	same	name	game	tame	came	fame	g	i
5	9	bed	led	fed	red	wed	shed	r	i
6	24	tan	tang	tap	tack	tam	tab	ng	f
7	32	pen	hen	men	then	den	ten	th (d)	i
8	23	coil	oil	soil	toil	boil	foil	s	i
9	29	heave	hear	heat	heal	heap	heath	t	f
10	6	must	bust	gust	rust	dust	just	m	i
11	20	pale	pace	page	pane	pay	pave	n	f
12	25	fit	fib	fizz	fill	fig	fin	l	f
13	23	coil	oil	soil	toil	boil	foil	b	i
14	10	pin	sin	tin	fin	din	win	w	i

## Appendix E: Tutoring Study Stimuli

### E.1: Study 1 Listening Pre/Post Test Stimuli

1. path
2. bath
3. best
4. pest
5. bit
6. pit
7. bop
8. pop
9. pun
10. bun
11. tab
12. tap
13. nip
14. nib
15. hob
16. hop
17. pub
18. pup
19. cob
20. cop

### E.2: Study 1 Pre/Post Test Speaking Words:

1. pair
2. bear
3. pond
4. bond
5. sub
6. sup
7. cub
8. cup
9. rip
10. rib
11. beep

12. peep
13. beach
14. peach
15. pin
16. bin
17. pill
18. bill
19. bale
20. pale

### E.3: Study 1 Pre/Post Test Speaking Phrases

1. Bagpipes and bongos are musical instruments.
2. It's absorbed into the bloodstream.
3. The benefits claimed in the report were substantial.
4. The team involved in the project avoided the problem.
5. The blackberries were baked in a pie.
6. Help celebrate your brother's success.
7. Our experiment's positive outcome was unexpected.
8. A batch of biscuits was in the box.
9. Young people participate in athletic activities.
10. Clear pronunciation is appreciated.
11. Birthday parties have balloons and banners.
12. Basic equipment needs proper maintenance.

## E.4: Study 1 User Questionnaire

### **Part 1: Personal Information**

1. In what age group are you?

19 and under

20 - 29

30 - 39

40 - 49

50 - 59

60 +

2. Gender:

Male

Female

3. Please state your level of English:

IELTS Level:

4. Please state your primary or native language.

### **Part 2: General Questions**

1. What have you found most difficult in learning pronunciation?

2. What particular aspects of learning pronunciation do you think software could be useful for?

3. Have you used any other pronunciation software before? If so, please give details.

**Part 3: To be completed after software use**

After using the software, please indicate the extent to which you agree or disagree with the following statements:

SD = Strongly Disagree

D = Disagree

N = Neutral

A = Agree

SA = Strongly Agree

This software was helpful in learning pronunciation  SD  D  N  A  SA

I found the external view helpful  SD  D  N  A  SA

I found the internal view helpful  SD  D  N  A  SA

I found the listening practice helpful  SD  D  N  A  SA

I found the listening test helpful  SD  D  N  A  SA

I found the speaking practice helpful  SD  D  N  A  SA

I found the recording function helpful  SD  D  N  A  SA

The content of the lesson matched my needs  SD  D  N  A  SA

This software is easy to use  SD  D  N  A  SA

This software is engaging  SD  D  N  A  SA

The talking head appeared natural  SD  D  N  A  SA

This software is satisfying to use.  SD  D  N  A  SA

4. What particular aspect(s) of this software did you like?
5. What particular aspect(s) of this software did you dislike?
6. Do you have any suggestions for improvement?

### E.5: Study 2 and 3 Pre/Post Test Listening Words

1. path
2. bath
3. big
4. pig
5. bit
6. pit
7. bark
8. park
9. pun
10. bun
11. tab
12. tap
13. sub
14. sup
15. bale
16. pale
17. pub
18. pus
19. best
20. pop

### E.6: Study 2 and 3 Pre/Post Test Speaking Words

1. bond
2. pull
3. beep
4. peat
5. bang
6. rib
7. peep
8. hob
9. pond
10. pay
11. rip
12. symbol
13. board
14. bull

15. bay
16. hop
17. poured
18. beat
19. pang
20. simple

### E.7: Study 2 and 3 Pre/Post Test Speaking Phrases

1. I would like to put the bath here.
2. I brought a bin.
3. I put the bills on the table.
4. Put this blanket on your back.
5. I put the cub in the basket.
6. Look at how big that bear is!
7. Did you see that pike?
8. Her buns are awful.
9. I will put the patch here.
10. He has too many pets.
11. I would like to put the path here.
12. I brought a pin.
13. I put the pills on the table.
14. I put the cup in the basket.
15. Did you see that bike?
16. Put this blanket on your pack.
17. I will put the batch here.
18. Look at how big that pear is!
19. Her puns are awful.
20. He has too many bets.

## E.8: Study 2: User Feedback after using Pronunciation Software Version I (internal and external visualization)

### 1. Participant ID:

After using the Pronunciation Assistant software, please indicate the extent to which you agree or disagree with the following statements.

### 2. The Pronunciation Assistant software was helpful in learning pronunciation.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

### 3. I found the external view of the talking head (e.g. lip and tongue movements viewed from the front) helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

### 4. I found the internal view of the talking head (e.g. lip and tongue movements viewed from the side) helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

### 5. The talking head looked realistic.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

### 6. The speech animation appeared natural.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

7. I found the listening practice with the talking head helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

8. I found the listening test with the talking head helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

9. I found the speaking practice with the talking head helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

10. I found the recording function in the Pronunciation Assistant software helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

11. The Pronunciation Assistant software is interesting to use.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

12. The Pronunciation Assistant software is satisfying to use.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

13. The content of the lesson matched my needs.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

14. What did you like about this software?

15. What did you dislike about this software?

16. Do you have any comments or suggestions for improvement?

### E.9: Study 2 and Study 3: User Feedback after using Pronunciation Software Version X (external visualization)

1. Participant ID:

After using the Pronunciation Assistant software, please indicate the extent to which you agree or disagree with the following statements.

2. The Pronunciation Assistant software was helpful in learning pronunciation.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

3. I found the external view of the talking head (e.g. lip and tongue movements viewed from the front) helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

4. The talking head looked realistic.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

5. The speech animation appeared natural.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

6. I found the listening practice with the talking head helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

7. I found the listening test with the talking head helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

8. I found the speaking practice with the talking head helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

9. I found the recording function in the Pronunciation Assistant software helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

10. The Pronunciation Assistant software is interesting to use.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

11. The Pronunciation Assistant software is satisfying to use.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

12. The content of the lesson matched my needs.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

13. What did you like about this software?

14. What did you dislike about this software?

15. Do you have any comments or suggestions for improvement?

### E.10: Study 3: User Feedback after using Pronunciation Software Version A (audio alone)

1. Participant ID:

After using the Pronunciation Assistant software, please indicate the extent to which you agree or disagree with the following statements.

2. The Pronunciation Assistant software was helpful in learning pronunciation.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

3. I found the listening practice with the Pronunciation Assistant software helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

4. I found the listening test with the Pronunciation Assistant software helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

5. I found the speaking practice with the Pronunciation Assistant software helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

6. I found the recording function in the Pronunciation Assistant software helpful.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

7. The Pronunciation Assistant software is interesting to use.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

8. The Pronunciation Assistant software is satisfying to use.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

9. The content of the lesson matched my needs.

Strongly disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Strongly agree	Not Applicable
<input type="checkbox"/>							

10. What did you like about this software?

11. What did you dislike about this software?

12. Do you have any comments or suggestions for improvement?

13. If this is your 2nd session, and you have used 2 different versions, which did you prefer?

Strongly prefer audio alone	Moderately prefer audio alone	Slightly prefer audio alone	Neutral	Slightly prefer talking head	Moderately prefer talking head	Strongly prefer talking head	Not Applicable
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

14. If you prefer one version, please comment on why.

## References

- Alexander, O., M. Rogers, W. Lambeth, M. Chiang and P. Debevec (2010). "The Digital Emily Project: Achieving a Photorealistic Digital Actor." IEEE Computer Graphics and Applications.
- Algirdas, P. (2002). MPEG-4 Facial Animation: The Standard, Implementation and Applications, John Wiley and Sons, Inc.
- Assmann, P. (2010). "Speech Perception Lab." from <http://www.utdallas.edu/~assmann/hcs7367/classnotes.html>.
- Badin, P., G. Bailly, L. Reveret, M. Baciú, C. Segebarth and C. Savariaux (2002). "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images." Journal of Phonetics **30**(3): 533-553.
- Badin, P., F. Elisei, G. Bailly and Y. Tarabalka (2008). An Audiovisual Talking Head for Augmented Speech Generation: Models and Animations Based on a Real Speaker's Articulatory Data. Proceedings of the 5th international conference on Articulated Motion and Deformable Objects. Port d'Andratx, Mallorca, Spain, Springer-Verlag.
- Badin, P. and A. Serrurier (2006). Three-dimensional modeling of speech organs: Articulatory data and models. Transactions on Technical Committee of Psychological and Physiological Acoustics, The Acoustical Society of Japan.
- Badin, P., Y. Tarabalka, F. Elisei and G. Bailly (2010). "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding." Speech Communication **52**: 493-503.
- Bailly, G., F. Elisei, P. Badin and C. Savariaux (2006). Degrees of freedom of facial movements in face-to-face conversational speech. International Workshop on Multimodal Corpora, Genoa, Italy.
- Bailly, G., O. Govokhina, F. Elisei and G. Breton (2009). "Lip-Synching Using Speaker-Specific Articulation, Shape and Appearance Models." EURASIP Journal on Audio, Speech, and Music Processing **2009**.
- Bauman, N. (2006). A Catalogue of Errors Made by Korean Learners of English. KOTESOL International Conference
- Baylor, A. L. (2005). Preliminary design guidelines for pedagogical agent interface image. Proceedings of the 10th international conference on Intelligent user interfaces, San Diego, California, USA, ACM.
- Baylor, A. L. and Y. Kim (2005). "Simulating instructional roles through pedagogical agents." International Journal of Artificial Intelligence in Education **15**: 95-115.
- Beeler, T., F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner and M. Gross (2011). High-quality passive facial performance capture using anchor frames. ACM SIGGRAPH 2011 papers. Vancouver, British Columbia, Canada, ACM.
- Ben Youssef, A., Hueber, T., Badin, P. & Bailly, G. (2011). Toward a multi-speaker visual articulatory feedback system. Interspeech, Florence, Italy.
- Benoît, C. and B. Le Goff (1998). "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP." Speech Communication **26**(1-2): 117-129.
- Bérar, M., G. Bailly, M. Chabanas, F. Elisei, M. Odisio and Y. Pahan (2003). Towards a generic talking head. 6th International Seminar on Speech Production, Sydney, Australia.
- Bevacqua, E. and C. Pelachaud (2004). "Expressive audio-visual speech." Computer Animation and Virtual Worlds **15**(3-4): 297-304.

- Birkholz, P. and B. J. Kröger (2007). Simulation of vocal tract growth for articulatory speech synthesis. 16th International Congress of Phonetic Sciences, Saarbrücken, Germany.
- Bobda, A. S. (2006). "Testing Pronunciation." Retrieved 2012, from <http://pedagogie.ac-montpellier.fr/disciplines/anglais/ressources/phonologie/Testing%20pronunciation.htm>.
- Borshukov, G., D. Pionni, O. Larsen, J. P. Lewis and C. Tempelaar-Lietz (2005). Universal capture - image-based facial animation for "The Matrix Reloaded". ACM SIGGRAPH 2005 Courses. Los Angeles, California, ACM.
- Bregler, C., M. Covell and M. Slaney (1997). Video rewrite: visual speech synthesis from video. AVSP.
- Bullock, J. (2011a). Spectrograms For Speech: The Clinical Application Of Spectrographic Displays. Speech and Hearing Sciences, Portland State University. **Master's**.
- Bullock, J. L. (2011b). "History of Spectrograms." 2012, from <http://www.spectrogramsforspeech.com/background/history-of-spectrograms/>.
- Bullock, J. L. (2011c). "Spectrograms for Speech." 2012, from <http://www.spectrogramsforspeech.com/>.
- Cao, Y., W. C. Tien, P. Faloutsos and F. Pighin (2005). "Expressive speech-driven facial animation." ACM Trans. Graph. **24**(4): 1283-1302.
- Catmull, E. and R. Rom (1974). A class of local interpolating splines. New York, Academic Press.
- Cohen, M., J. Beskow and D. Massaro (1998). Recent development in facial animation: an inside view. AVSP.
- Cohen, M. M. and D. W. Massaro (1993). Modelling coarticulation in synthetic visual speech. Models and Techniques in Computer Animation. N. M. Thalmann and D. Thalmann. Tokyo, Springer-Verlag: 139-156.
- Cole, R., B. Wise and S. Van Vuuren (2007). "How Marni teaches children to read." Journal of Educational Technology **47**(1): 14-18.
- Cosi, P. and C. Drioli (2008). LUCIA, a new emotive/expressive Italian talking head. Emotions in the Human Voice. K. Izdebski. San Diego, California, USA, Plural Publishing, Inc. **3**: 153-176.
- Cosi, P. C. (2002). Baldini: Baldi speaks Italian. ICSLP 2002: 7th International Conference on Spoken Language Processing, Denver Colorado.
- Cosi, P. C., E. Magno, G. Perlin and C. Zmarich (2002). Labial coarticulation modelling for realistic facial animation. International Conference on Multimodal Interfaces, Pittsburgh, PA.
- Cosker, D., D. Marshall, P. Rosin, S. Paddock and S. Rushton (2005). "Towards perceptually realistic talking heads: models, metrics and McGurk." ACM Transactions on Applied Perception **2**(3): 270-285.
- CSTR. (2008). "The Festival Speech Synthesis System ", 2008, from <http://www.cstr.ed.ac.uk/projects/festival/>.
- Dabic, S. (2010). To Teach or not to Teach: Pronunciation Challenge in ESL. International Online Language Conference Universal Publishers.
- Deena, S., S. Hou and A. Galata (2010). Visual Speech Synthesis by Modelling Coarticulation Dynamics using a Non-Parametric Switching State-Space Model. ICMI-MLMI'10: Proc. of the 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction, Beijing, China.
- Deng, Z. and J. Noh (2007). Computer Facial Animation: A Survey. Data-Driven 3D Facial Animation. Z.Deng and U.Neumann, Springer Verlag: 1-28.

- Dey, P., S. Maddock and R. Nicolson (2010a). Evaluation of A Viseme-Driven Talking Head. Eighth Theory and Practice of Computer Graphics 2010 Conference, Sheffield, UK, Eurographics Association.
- Dey, P., S. Maddock and R. Nicolson (2010b). A Talking Head for Speech Tutoring. ACM/SSPNET 2nd International Symposium on Facial Analysis and Animation, Edinburgh, UK.
- Dunlop, R. (2009). "Introduction to Catmull-Rom Splines." 2009, from <http://www.mvps.org/directx/articles/catmull/>.
- Edge, J. D. (2004). Techniques for the Synthesis of Visual Speech, University of Sheffield. **PhD**.
- Edge, J. D., M. A. Sanchez Lorenzo and S. Maddock (2004). Animating Speech from Motion Fragments. Technical Report CS-04-02, Department of Computer Science.
- Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. Handbook of Social Psychophysiology. H. Wagner and A. Manstead. Chichester, John Wiley, Ltd.: 143-164.
- Ekman, P. and W. V. Friesen (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, Consulting Psychologists Press.
- Elisei, F., M. Odisio, G. Bailly and P. Badin (2001). Creating and controlling video-realistic talking heads. Auditory-Visual Speech Processing Workshop, AVSP Scheelsminde, Denmark.
- Englebienne, G., T. F. Coote and M. Rattray (2007). A Probabilistic Model for Generating Realistic Lip Movements from Speech. Twenty-First Annual Conference on Neural Information Processing Systems, MIT Press.
- Engwall, O. (2003). "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model." Speech Communication **41**(2-3): 303-329.
- Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes. Interspeech, Brisbane, Australia.
- EPFL. (2009). "Visible Human Project ", 2009, from <http://visiblehuman.epfl.ch/>.
- EPSRC. (2011). "UK ICT Pioneers Competition." from <http://www.epsrc.ac.uk/newsevents/news/2011/Pages/ukictpioneers.aspx>.
- Ezzat, T. and T. Poggio (1997). Videorealistic Talking Faces: A Morphing Approach. Audiovisual Speech Processing Workshop, Rhodes, Greece.
- Fagel, S. (2008). MASSY Speaks English: Adaptation and Evaluation of a Talking Head. Interspeech, Brisbane.
- Fagel, S. and K. Madany (2008). A 3-D Virtual Head as a Tool for Speech Therapy for Children. Interspeech, Brisbane.
- Fels, S. S., J. E. Lloyd, I. Stavness, F. Vogt, A. Hannam and E. Vatikiotis-Bateson (2007). "ArtiSynth: A 3D biomechanical simulation toolkit for modeling anatomical structures " Journal of the Society for Simulation in Healthcare **2**(2): 148.
- Fisher, C. G. (1968). "Confusions among visually perceived consonants." J. Speech Hearing Res. **11**: 796-803.
- Flemming, E. (2012). "24.910 Topics in Linguistic Theory: Laboratory Phonology, Spring 2007. (Massachusetts Institute of Technology: MIT OpenCourseWare), <http://ocw.mit.edu> (Accessed June 26, 2012). License: Creative Commons BY-NC-SA."
- Fraser, H. (2006). "Helping teachers help students with pronunciation: A cognitive approach." Prospect **21**(1): 80-96.
- Frowd, C. D., V. Bruce, D. Ross, A. McIntyre and P. J. B. Hancock (2007). "An application of caricature: how to improve the recognition of facial composites." Visual Cognition **15**(8): 954-984.

- Geiger, G., T. Ezzat and T. Poggio (2003). Perceptual Evaluation of Video-Realistic Speech. CBCL Paper 224/ AI Memo 2003-003, Massachusetts Institute of Technology.
- Grauwinkel, K. and S. Fagel (2007). Visualization of internal articulator dynamics for use in speech therapy for children with Sigmatismus Interdentalis. AVSP-2007.
- Green, P., J. Carmichael, A. Hatzis, P. Enderby, M. Hawley and M. Parker (2003). Automatic Speech Recognition with Sparse Training Data for Dysarthric Speakers. 8th European Conference on Speech Communication Technology (Eurospeech) Geneva, Switzerland.
- Guenther, F. H. (2003). Neural control of speech movements. Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities. A. Meyer and N. Schiller. Berlin, Mouton de Gruyter.
- Hardcastle, W. J. and N. Hewlett (1999). Coarticulation: Theory, Data and Techniques, Cambridge University Press.
- Hatzis, A., P. Green, J. Carmichael, S. Cunningham, R. Palmer, M. Parker and P. O'Neill (2003). An Integrated Toolkit Deploying Speech Technology for Computer Based Speech Training with Application to Dysarthric Speakers. 8th European Conference on Speech Communication Technology (Eurospeech) Geneva.
- Hazan, V. (2008). Talking heads and pronunciation training: a review. Interspeech, Brisbane.
- Hazan, V. and Y. H. Kim (2010). Can we predict who will benefit from computer-based phonetic training? Interspeech 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology, Tokyo, Japan.
- Hazan, V., A. Sennema and A. Faulkner (2002). Audiovisual Perception In L2 Learners. ICSLP.
- Hazan, V., A. Sennema, M. Iba and A. Faulkner (2005). "Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English." Speech Communication.
- Hilder, S., R. Harvey and B.-J. Theobald (2009). Comparison of human and machine-based lip-reading. AVSP, Norwich, UK.
- Hilder, S., B.-J. Theobald and R. Harvey (2010). In Pursuit of Visemes. International Conference on Auditory-Visual Speech Processing.
- HTS. (2011). "HMM-based Speech Synthesis System (HTS)." Retrieved 2012, from <http://hts.sp.nitech.ac.jp/?Home>.
- Huang, F. J., E. Cosatto and H. P. Graf (2002). Triphone based unit selection for concatenative visual speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL.
- International Phonetic Association (2005). IPA chart.
- Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. Punishment and aversive behavior. B. A. Campbell and R. M. Church. New York, Appleton-Century-Crofts: 279–96.
- King, S. A. and R. E. Parent (2005). "Creating speech-synchronized animation." IEEE Trans. Vis. Graph **11**(3): 341–352.
- Kjellström, H., & Engwall, O. (2009). "Audiovisual-to-articulatory inversion." Speech Communication **51**(3): 195-209.
- Kjellstrom, H., O. Engwall and O. Balter (2006). Reconstructing Tongue Movements from Audio and Video. Interspeech, Pittsburgh.
- Klatt, D. (1987). "Review of text-to-speech conversion." Journal of the Acoustical Society of America **82**: 737–793.
- Knight, H. (2011) "Talking with Tara." New Scientist **Volume**, 23 DOI:

- Kröger, B. J., J. Gotto, S. Albert and C. Neuschaefer-Rube (2005). "A visual articulatory model and its application to therapy of speech disorders: a pilot study. ." Speech production and perception: Experimental analyses and models, *ZASPiL* (40): 79-94
- Kulik, C. C. and J. A. Kulik (1991). "Effectiveness of computer based instruction; an updated analysis." Computers in human behaviour 7(1): 75-94.
- Lazalde, O. M. (2010). Analyzing and evaluating the use of visemes in an interpolative synthesizer for visual speech. Department of Computer Science, University of Sheffield. **PhD**.
- Lazalde, O. M., S. Maddock and M. Meredith (2008). "A Constraint-Based approach to Visual Speech for a Mexican-Spanish Talking Head." International Journal of Computer Games Technology 2008(3).
- Lazalde, O. M. M. and S. Maddock (2010). Comparison of different types of visemes using a constraint-based coarticulation model. EG UK Theory and Practice of Computer Graphics, Sheffield, UK, Eurographics Association.
- Le Goff, B. and C. Benoit (1996). A text-to-audiovisual-speech synthesizer for French. International Conference on Spoken Language Processing (ICSLP).
- Learning Technologies International. (2012). "Speak As You See (SAYS)." 2012, from <http://www.learningtechnologiesinternational.com/SAYSdescription.pdf>.
- Lee, Y., D. Terzopoulos and K. Waters (1995). Realistic modelling for facial animation. ACM SIGGRAPH.
- Lewis, J. P. and F. I. Parke (1987). "Automated lip-synch and speech synthesis for character animation." SIGCHI Bull. 17(SI): 143-147.
- Likert, R. (1932). "A Technique for the Measurement of Attitudes." Archives of Psychology 140(1): 55.
- Liu, K. and J. Ostermann (2008). Realistic Facial Animation System for Interactive Services. Interspeech 2008, LIPS 2008: Visual Speech Synthesis Challenge, Brisbane, Australia.
- Löfqvist, A. (1990). Speech as audible gestures. Speech Production and Speech Modeling. W. J. H. a. A. Marchal. Dordrecht, Kluwer Academic Publishers. 55: 289–322.
- Loquendo. (2008). "Loquendo." 2008, from <http://www.loquendo.com/en/>.
- Ma, W.-C., A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung and P. Debevec (2008). "Facial performance synthesis using deformation-driven polynomial displacement maps." ACM Trans. Graph. 27(5): 1-10.
- MacWhinney, B. (1992). Transfer and competition in second language learning. Cognitive processing in bilinguals. R. Harris. Amsterdam, Elsevier: 371-390.
- MacWhinney, B. (1997). Second language acquisition and the Competition Model. Tutorials in bilingualism. J. Kroll and A. D. Groot. Mahwah, NJ, Lawrence Erlbaum.
- Magenat-Thalmann, N., E. Primeau and D. Thalmann (1988). "Abstract muscle action procedures for human face animation." The Visual Computer 3(5): 290-297.
- Magenat-Thalmann, N. and D. Thalmann (2004). Handbook of virtual humans. Chichester Wiley.
- Mahshie, J. J. (1996). Feedback Considerations For Speech Training Systems. ICSLP.
- Marschark, M., D. LePoutre and L. Bement (1998). Mouth movement and signed communication. Hearing by Eye II. R. Campbell and B. Dodd, and Burnham, D. Hove, United Kingdom, Psychology Press Ltd. Publishers: 245–266.
- Martin, J.-C., C. d'Alessandro, C. Jacquemin, B. Katz, A. Max, L. Pointal and A. Rilliard (2007). 3D Audiovisual Rendering and Real-Time Interactive Control of Expressivity in a Talking Head. Intelligent Virtual Agents, Paris, France, Springer.
- Massaro, D. (2012). Animated speech: Research progress and applications. Audiovisual Speech Processing. G. Bailly, P. Perrier and E. Vatiokis-Bateson, Cambridge University Press.

- Massaro, D. W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. Cambridge, MA, MIT Press.
- Massaro, D. W. (2004). Symbiotic value of an embodied agent in language learning. IEEE Proc. of 37th Annual Hawaii Intl. Conference on System Sciences.
- Massaro, D. W., S. Bigler, T. Chen, M. Perlman and S. Ouni (2008). Pronunciation Training: The Role of Eye and Ear. Interspeech, Brisbane.
- Massaro, D. W. and J. Light (2003). Read My Tongue Movements: Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/. Eurospeech (Interspeech), 8th European Conference on Speech Communication and Technology., Geneva, Switzerland.
- Massaro, D. W. and J. Light (2004). "Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss." J Speech Lang Hear Res **47**(2): 304-320.
- Massaro, D. W., S. Ouni, M. M. Cohen and R. Clark (2005). A Multilingual Embodied Conversational Agent. 38th Annual Hawaii International Conference on System Sciences, Los Alimitos, LA.
- MBROLA. "The MBROLA Project homepage." 2008, from <http://tcts.fpms.ac.be/synthesis/mbrola.html>.
- McGurk, H. and J. MacDonald (1976). "Hearing lips and seeing voices." Nature **264**: 746-748.
- Meyer Sound. (2010). "Speech Intelligibility Papers - Glossary." from <http://www.meyersound.com/support/papers/speech/mrt.htm>.
- Montgomery, D. (1981). "Do dyslexics have difficulty accessing articulatory information?" Psychological Research **43**(2): 235-243.
- Moreno, R., R. E. Mayer, H. A. Spires and J. C. Lester (2001). "The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedagogical Agents?" Cognition and Instruction **19**(2): 177-213.
- Mori, M. (1970). "Bukimi no tani [The uncanny valley]." Energy **7**(4): 33-35.
- Nokia Corporation. (2005). "Java Developer's Library." from <http://www.forum.nokia.com/infocenter/>.
- Northern Digital Inc. (2012). "Wave Electromagnetic Articulograph Speech Research Motion Capture System " Retrieved 2012, from <http://www.ndigital.com/lifesciences/products-speechresearch.php/research.php>.
- Öster, A.-M. (1996). Clinical applications of computer-based speech training for children with hearing-impairment. 4th International Conference on Spoken Language Processing, Philadelphia, USA.
- Ostermann, J. (2002). Face Animation in MPEG 4. MPEG-4 Facial Animation: The Standard, Implementation and Applications. I. S. P. R. Forchheimer, John Wiley and Sons.
- Ouni, S., M. M. Cohen, H. Ishak and D. W. Massaro (2007). "Visual contribution to speech perception: measuring the intelligibility of animated talking heads." EURASIP J. Audio Speech Music Process. **2007**(1): 3-3.
- Ouni, S. C. (2005). "Training Baldi to be multilingual: A case study for an Arabic Badr." Speech Communication **45**(2): 115-137.
- Ouni, S. M. (2003). Internationalization of a Talking Head. 15th International Congress of Phonetic Sciences (ICPhS'03), Barcelona, Spain.
- Palmer, R., Enderby, P., and Cunningham, S.P. (2004). "The effect of three practice conditions on the consistency of chronic dysarthric speech." Journal of Medical Speech and Language Pathology **12**: 183-189.
- Parke, F. I. and K. Waters (1996). Computer facial animation, A. K. Peters, Ltd.
- Parke, F. I. and K. Waters (2008). Computer facial animation, A. K. Peters, Ltd.

- Pelachaud, C. (1991). *Communication and Coarticulation In Facial Animation*, University of Pennsylvania.
- Pighin, F. and J. P. Lewis (2006). Facial motion retargeting. *ACM SIGGRAPH 2006 Courses*. Boston, Massachusetts, ACM.
- Poppe, R. (2007). "Vision-based human motion analysis: An overview." *Comput. Vis. Image Underst.* **108**(1-2): 4-18.
- Potamianos, G., C. Neti, J. Luetin and I. Matthews (2004). Audio-Visual Automatic Speech Recognition: An Overview. *Issues in Visual and Audio-Visual Speech Processing*. G. Bailly, E. Vatikiotis-Bateson and P. Perrier, MIT Press.
- QT. (2009). "QT Software." 2009, from <http://www.qtsoftware.com/>.
- Queiroz, R. B., L. M. Barros and S. R. Musse (2008). "Providing expressive gaze to virtual animated characters in interactive applications." *Computers in Entertainment* **6**(3): 1-23.
- Reeves, B. and C. Nass (1996). *The Media Equation: how people treat computers, television, and new media like real people and places*, Cambridge University Press.
- Reeves, B. and C. Nass (2000). "Perceptual user interfaces: Perceptual Bandwidth." *Communications of the ACM* **43**(3): 65-70
- Ross, M. (1999). "Speechreading." 2012, from <http://www.therubins.com/geninfo/speechrd.htm>
- Salvador, S. and P. Chan (2004). *FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space*. KDD Workshop on Mining Temporal and Sequential Data.
- Seabrook, R. and Z. Dienes (2003). *Incubation in Problem Solving as a context Effect*. Proc. 25th Meeting Cognitive Science Society, Lawrence Erlbaum Associates, Mahwah, NJ.
- Shepard, R. N. (1967). "Recognition memory for words, sentences, and pictures." *Journal of Verbal Learning and Verbal Behavior* **6**: 156-163.
- Singular Inversions. (2008). "FaceGen." from <http://facegen.com/>.
- SKY. "SKY software." Retrieved 2009, from <http://www.skysoftwarehouse.com/>.
- Steiner, I. and S. Ouni (2012). *Artimate: an articulatory animation framework for audiovisual speech synthesis*. Workshop on Innovation and Applications in Speech Technology.
- Steiner, I., K. Richmond, I. Marshall and C. D. Gray (2012). "The magnetic resonance imaging subset of the mngu0 articulatory corpus." *Journal of the Acoustical Society of America* **131**(2): JASA Express Letters.
- Stephoe, W., O. Oyekoya and A. Steed (2010). "Eyelid kinematics for virtual characters." *Computer Animation And Virtual Worlds* **21**: 161–171.
- Straub, R. O. (2001). *Faculty Guide for The Scientific American Frontiers Video Collection for Developmental Psychology*, Worth Publishers.
- Sumby, W. H. and I. Pollack (1954). "Visual contribution to speech intelligibility in noise." *Journal of the Acoustical Society of America* **26**(2): 212–215.
- Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. *Hearing by Eye: The Psychology of Lip-Reading*. Dodd and R. B. and Campbell. London, United Kingdom, Lawrence Erlbaum Associates: 3–51.
- Swan, M. and B. Smith (2001). *Learner English: A Teacher's Guide to Interference and Other Problems*, Cambridge University Press.
- Takács, G., A. Tihanyi, G. Feldhoffer, T. Bárdi and O. Balázs (2007). *Synchronization of acoustic speech data for machine learning based audio to visual conversion* 19th International Congress On Acoustics, MADRID.
- Tallal, P. (2001). Experimental Studies of Language Learning Impairments: From Research to Remediation. *Speech and Language Impairments in Children: Causes, Characteristics, Intervention and Outcome*. D. V. M. Bishop and L. B. Leonard, Psychology Press.

- Tekalp, A. M. and J. Ostermann (2000). "Face and 2-D Mesh Animation in MPEG-4." Signal Processing: Image Communications **15**: 387-421.
- Terzopoulos, D. and K. Waters (1990). "Physically-based facial modeling, analysis, and animation." Journal of Visualization and Computer Animation **1**(4): 73-80.
- Thelwall, R. and M. A. Sa'Adeddin (1990). "Arabic." Journal of the International Phonetic Association **20**(2): 37-39
- Theobald, B.-J. (2007). Audiovisual Speech Synthesis. ICPhS Saarbrücken.
- Theobald, B.-J., G. Cawley, A. Bangham, I. Matthews and N. Wilkinson (2008). Comparing text-driven and speech-driven visual speech synthesisers. Interspeech, Brisbane.
- Theobald, B.-J., S. Fagel, G. Bailly and F. Elisei (2008). LIPS2008: Visual Speech Synthesis Challenge. Interspeech, Brisbane.
- Theobald, B. J., J. A. Bangham, I. A. Matthews and G. C. Cawley (2004). "Near-videorealistic synthetic talking faces: implementation and evaluation." Speech Communication **44**(1-4): 127-140.
- Tobias, S. and J. D. Fletcher (2011). Computer Games and Instruction. Charlotte, North Carolina, Information Age Publishing
- Tomlinson, T. D., D. E. Huber, C. A. Rieth and E. J. Davelaar (2009). An interference account of cue-independent forgetting in the no-think paradigm. Proceedings of the National Academy of Sciences 106.
- Tresadern, P., P. Sauer and T. F. Cootes (2010). Additive Update Predictors in Active Appearance Models. British Machine Vision Conference BMVA Press.
- Walker-Smith, G. J., A. G. Gale and J. M. Findlay (1977). "Eye movement strategies involved in face perception." Perception **6**(3): 313 - 326.
- Wallraven, C., M. Breidt, D. W. Cunningham and H. H. Bulthoff (2008). "Evaluating the Perceptual Realism of Animated Facial Expressions." ACM Transactions on Applied Perception **4**(4).
- Waters, K. (1987). A muscle model for animating three-dimensional facial expression. SIGGRAPH
- Waters, K. and T. Levergood (1994). An automatic lip-synchronization algorithm for synthetic faces. Proceedings of the second ACM international conference on Multimedia. San Francisco, California, United States, ACM.
- Watt, A. and M. Watt (1992). Advanced Animation and Rendering Techniques: Theory and Practice, ACM Press.
- Wik, P. and O. Engwall (2008). Can visualization of internal articulators support speech perception? Interspeech Brisbane, Australia.
- Wise, B. W., J. Ring and R. K. Olson (1999). "Training Phonological Awareness with and without Explicit Attention to Articulation." Journal of Experimental Child Psychology **72**: 271-304
- Witkin, A. and M. Kass (1988). "Spacetime Constraints." Computer Graphics **22**(4).
- Wrench, A. (1999). "MOCHA-TIMIT", from <http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html>.
- Zhao, J., W. Lirong, Z. Chao, S. Lijuan and Y. Jia (2010). Pronouncing Rehabilitation of Hearing-impaired Children based on Chinese 3D Visual-speech Database. Fifth International Conference on Frontier of Computer Science and Technology.