Automated Lip Synchronisation for Human-Computer Interaction and Special Effect Animation

Christian Martyn Jones and Satnam Singh Dlay
Department of Electrical and Electronic Engineering,
Merz Court, University of Newcastle,
Newcastle upon Tyne, NE1 7RU, United Kingdom
Tel: +44 (0) 191 222 7340, FAX:+44 (0) 191 222 8180
E-mail: c.m.jones@newcastle.ac.uk & s.s.dlay@newcastle.ac.uk

ABSTRACT

The research presents MARTI (Man-machine animation real-time interface) for the realisation of automated special effect animation and human computer interaction. The future developments of the Internet, video communications and multi-media, virtual reality, and animation will rely on the derivation of a natural human-machine interface in order to submerse people, irrespective of technical know-how, into the latest technology, and allow them to interact with computers and one another using their own personality and idiosyncrasies. MARTI introduces novel research in a number of engineering fields to realise the first natural interface and animation system capable of high performance for real-users and real-world applications.

1. INTRODUCTION

Computer game design and special effect filming are major international industries. Each represents the forefront of technology in illusion and deception to allow players escape the confines of the real-world for yet undiscovered virtual realities. Not only is the 'cyber-generation' realising the potential of virtual worlds but also the more reserved fields of architecture, medicine and communications. Architectural design tools enable computer modelling of proposed complexes, permitting historians to re-construct virtual temples and allowing us to visit lost times and lands, and even realise our dreams of the future world within the computer. Virtual characters can assist medical applications and research in the fields of speech therapy, the psychology of speech, and medical linguistics with direct visual feedback of the required speech articulation. Telecommunications and computer interaction is primed to merge with virtual reality where in the future we may live and work solely within a computer generated environment. However, within these virtual worlds we will be deprived of face to face human communication and limited by our own technical capabilities. We must instead devise a simple and effective method of interfacing people into the new technology without loosing our human feelings. Hence the realisation of a natural man-machine interface is paramount for the continued development and success of future virtual applications [1,2]. MARTI provides easy access to the latest technological innovations by providing the 'person in the street' with a friendly life-like interface without loosing the communicative abilities which would prevent 'human-friendliness'.

MARTI is designed as the complete special effects and animation toolbox. The animation industry uses the manipulation of non-human characters to portray the illusion of life, and these manipulations involve the movement of facial expressions to convey emotion and shaping of the mouth for lip synchronisation. Current controls of the facial motion originate from hand puppetry and allow a single performer to produce the required facial image from combinations of manual movements link to multiple arm, hand, and finger joystick analogue controls. The relationship between the human motion and the output face can be highly complex and require exceptional manual dexterity and co-ordination. Hence the industry would benefit from the elimination of complex control systems with the development of automated lip synchronisation. Previous research has considered the extraction of control signals from human facial motion using image recognition systems [3], however the analysis requires that key articulatory points are highlighted. In addition researchers have considered analysis of the acoustic soundtrack [4]. However these systems are limited by the accuracy of their speech recognition and the requirement that they are trained on the single 'puppeteers' voice. Furthermore they do not provide timing and rhythm information to allow accurate synchronisation of the recognised facial images to the original speech. MARTI however overcomes these limitations, allowing automatic lip synchronisation from a single input: speech, and without the normal constraints of head-sets, 'white dot' reflectors and complex puppeteer control hardware. The system does not require pre-training to the performers voice but instead operates in real-time with continuous speech in 'normal', non-laboratory conditions and is invariant to regional accents and dialects, race, age and gender.

2 SPEECH RECOGNITION

Considerable research effort has been directed towards the study of speech recognition over the last four decades. Although significant progress has been made in the field recently, typical vocabulary size remains very

limited and task specific, and recognitive performance still short of the human equivalence. The major contributions to speech recognition have come from artificial neural networks (ANN) and statistical hidden Markov models (HMM), however both have fallen short of the goal. Instead, we are utilising a speech recognition system that benefits from both techniques [5-6].

MARTI uses our latest hybrid connectionist/hidden Markov model speech recognition system (STRUT) [7]. Training and testing of the speech recognition system was completed using the DARPA Acoustic-Phonetic Continuous Speech Corpus (TIMIT). TIMIT employs 60 phonetic symbols in the lexicon and recognition transcriptions, and comprises of 6300 sentences spoken by 630 speakers from 8 major dialect regions of the United States. An extensive test set was employed for performance analysis. This material included 168 speakers, 1344 utterances, and 624 distinct texts.

The objectives of the speech recognition stage have been achieved. We must now consider the relationship between the recognised acoustic elements and the associated facial motion.

3. HUMAN LIP SYNCHRONISATION

The acoustics of speech and the required articulation are obviously related. MARTI addresses speech from a phonetic view-point and recognises elements that are acoustical distinct. These fundamental sounds of speech can then be mapped into facial positions to achieve automated lip synchronisation.

The study of American English combines a theoretical understanding of the articulation of American speech and a practical consideration of conversational speech. A number of American subjects were asked to recite all 60 TIMIT phonemes and example sentences from the test data set whilst their mouth positions were videoed. From these studies we were able to suggest the minimum number of distinct mouth positions, visemes, used in everyday speech, Table 1.

Viseme Groups	
p b m	th dh
bcl pcb	f v
dcl tcl	y iy ih ix
k g ng eng	ey eh
kcl gcl	ae ax ah
d t n nx dx q	axr
em	aa
en	aw
r	ay
hh hv	oy
el l	ow ao
s z	w uw ux uh
sh ch jh zh	er axr

Table 1: Viseme groupings for American speech.

We then use wire-frame parameterized models of the tongue, lips, teeth, and jaw and match these articulation characteristics to the visemes, or control servos on the animatronic system to correspond to the facial images of the human speaker, Figure 1.

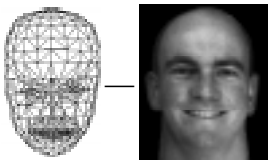

Figure 1: Parameterized modelling of human face.

The speech recognition 'front-end' achieves accuracy in excess of 58 % for speaker independent, continuous speech recognition, without any 'dictionary look-up' and 'grammatical checking' (which limits the user to sensical words and sentences) and provides 60 element phonetic transcriptions with time-signatures, Table 2.

system configuration				error %			perf.
word pen	syntax	feature data	duration model	ins	dels	subs	recog %
5	no	rasta	no	5.76	9.11	29.24	55.89%
6	no	rasta	no	4.18	11.06	28.80	55.96%
5	yes	rasta	no	5.44	9.04	28.91	56.61%
6	yes	rasta	no	4.01	10.76	28.50	56.73%
5	no	plp	yes	5.52	9.12	28.37	56.99%
6	no	plp	yes	4.21	10.77	27.98	57.05%
5	yes	plp	yes	5.07	8.94	27.82	58.18%
6	yes	plp	yes	3.92	10.52	27.40	58.16%

Table 2: Speech recognition performance of STRUT.

The 'recognised' transcription, Figure 2, provides phonetic information together with the start and end times in frames (where each frame represents 10ms). Of the 58 % correctly identified phones, 47 % are recognised to start and end on the correct time frames. In addition, 92 % of the recognised phones have the correct time alignment to within ±2 frames (where the average phone duration is 8 frames), Table 3. Thus we can accurately time-align our recognised phonetic transcription with the original speech sound-track to achieve synchronisation for the lip motion.

1 LS (0,12) dh (13,15) ih (16,20) w (21,27) er (28,34) axr (35,42) aa (43,50) dh (51,55) ax (56,61) f (62,74) aa (75,88) hv (89,93) aa (94,109) z (110,114) uw (115,123) z (124,128) n (129,134) y (135,140) axr (141,147) bcl (148,152) b (153,154) ay (155,175) TS (176,190)

Figure 2: Example of recognition with time data (utterance: 'There were other farmhouses nearby')

The hybrid system consisted of a three layered MLP comprised of 234 input neurons (feature data), 1000 'hidden' neurons, and 61 output classifications

system configuration	recog perf. %	ins saving in errors	del saving in errors	sub saving in errors	ins saving %	del saving %	sub saving %	viseme saving %	overall perf. %
wp5 nosyntax rasta	**55.89**	352/2821	217/4460	3250/14322	0.72	0.45	6.64	7.80	**63.69**
wp5 syntax rasta	**56.61**	303/2665	215/4426	3315/14158	0.62	0.44	6.77	7.83	**64.44**
wp5 nosyn plp dur	**56.99**	481/2705	205/4466	3352/13893	0.98	0.42	6.84	8.24	**65.23**
wp5 syn plp dur	**58.18**	444/2483	206/4377	3453/13623	0.91	0.42	7.05	8.38	**66.56**

Table 4: Human lip synchronisation performance of MARTI using speaker-independent, continuous American speech.

representing the 61 phonemes. Without constraints in input speech this performance represents the highest level of speech recognition to date.

Phone time alignment	Recognition accuracy
correct frame	46.8 %
±1 frame	83.8 %
±2 frames	91.7 %
±3 frames	94.7 %

Table 3: STRUT recognition timing accuracy.

The speech recognition transcription returned is mapped into visemes to be modelled for graphical output. Unfortunately the recognised transcription contains inaccuracies caused by incorrectly inserted, deleted, and substituted phonemes. However, any phoneme that is incorrectly recognised as another but is in the same viseme group as the other will not affect the visual performance of the system, Figure 3. Furthermore, if either of the neighbouring phones of the inserted or deleted error appear in the same viseme group as the error then the output model will not be affected and the visual performance improves. Although our recognition transcriptions now contains only 26 viseme elements as opposed to the original 60 phonemes it has been shown that these groups accurately describe the 'visual' articulation of American speech [8]. The affect of viseme groupings on insertion, deletion and substitution errors has been analysed numerically and shows an overall performance increase for the speech to lip synchronisation of 8.5% to nearly 67%, Table 4.

4. CARTOON LIP SYNCHRONISATION

The animation industry portrays the illusion of life by the motion of non-human characters, including facial expressions and lip synchronisation. Current controls of the facial motion originate from hand puppetry and allow performers to manipulate the character by multiple analogue joystick controls. The relationship between the human motion and the output face can be complex and requires the puppeteer to be highly skilled. Hence the industry would benefit from the elimination of complex control systems with the development of automated lip synchronisation.

MARTI has been used to provide lip synchronisation of animated characters to vocal soundtracks. The speech recognition 'front-end' remains unaltered. The output may be 2D or 3D computer generated models (Disney's film 'Toy Story'), or animatronic systems (Universal Studios: 'Babe'). Cartoon animation does not require the same degree of visual accuracy compared with the human articulation study shown previously. Faces are more simplistic and expressions and lip motion limited and exaggerated. In fact, if we were to make the characters emotions and articulation more human we would loose the very nature and feel of the animation. For the purpose of this study we will be considering the very distinct animation style of Nick Park and Aardman Animation. In particular we develop the viseme groupings for the character 'Wallace' from the three times Oscar winning 'Wallace and Gromit' and present the overall synchronisation performance of the system.

The viseme groupings and the associated character articulations for 'Wallace' were determined from frame by frame studies of the animation, Table 5.

Viseme Groups
p b m em bcl pcl
g t d n z s jh ch k r sh zh dx q ng en eng nx er axr em kcl gcl dcl tcl
f v g gcl
w ow uw aw ux oy
th dh l el
aa ax ow ao
ix uh aw
ih ey eh ah ae ay iy ax ah ng axr y ax-h ix
hh hv

Table 5: Viseme groupings for 'Wallace'

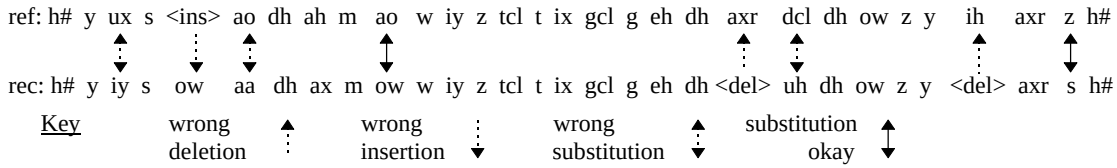

Figure 3: Reference and recognition phonetic transcriptions for the sentence 'You saw them always together those years'

system configuration	recog perf. %	ins saving in errors	del saving in errors	sub saving in errors	ins saving %	del saving %	sub saving %	viseme saving %	overall perf. %
wp6 nosyntax rasta	**55.96**	841/2047	2112/5417	8161/14103	1.72	4.31	16.66	22.69	**78.65**
wp6 syntax rasta	**56.73**	798/1962	2031/5272	8147/13956	1.63	4.14	16.64	22.42	**79.15**
wp5 nosyn plp dur	**56.99**	1355/2705	1735/4466	8443/13893	2.76	3.54	17.24	23.55	**80.54**
wp5 syn plp dur	**58.18**	1186/2483	1670/4377	8486/13623	2.42	3.41	17.33	23.16	**81.34**

Table 6: Cartoon lip synchronisation performance of MARTI for the animation and virtual reality industries.

The number of distinct viseme groups has now been further reduced from the 26 required for accurate human lip synchronisation to only 9. These groups represent the minimum number of discrete lip, teeth and tongue positions for the atypical character animation [9]. The animation performance study adopts the same recognition system and speech database as outlined previously. This time we use the 'Wallace' viseme groups, Table 5, to map the acoustic recognition to the visual output. Use of the 9 'cartoon'-style viseme groups provides an overall performance improvement for human speech to character animation of over 23%, to approaching 81.5%, Table 6 [10].

MARTI has achieved the objective of automated lip synchronisation for computer gaming, virtual reality, and cartoon animation. Animators and puppeteers can devise non-human, and atypical speech patterns and articulations, and use MARTI to display the required facial images for plasticine modelling (stop motion) and cell animation, or for direct control of computer generated characters and animatronic systems.

The performance figures for the human and animation models represent an analytical study, however the speed of normal conversational speech and animation for which MARTI is used, and the effect of smoothing between visemes on the output model, greatly enhances the performance such that users become oblivious to errors. The performance of the system in real-world applications with real-users greatly surpasses the theoretical study.

5. CONCLUSIONS

The research introduces MARTI for the realisation of automated special effect animation and human computer interaction. MARTI has the ability to determine the required lip synchronisation without restricting the 'freedom' of the user by simply performing the analysis on the vocal sound-track.

MARTI achieves a 67% human speech to face synchronisation for any user, including regional dialects and accents, age and gender, and without specific pre-training for that user, and operates with continuous speech within 'normal', non-laboratory conditions. Moreover, MARTI returns in excess of 81% automated lip synchronisation accuracy for cartoon animation such as Nick Parks three time Oscar winning 'Wallace & Gromit' and without the requirement of complex control hardware.

Our research has many real-world applications including natural man-machine interfacing; automated animation; video compression for telecommunications and multi-media; speech training and aids for the handicapped; and player interaction for 'video gaming'.

MARTI is the first natural interface and animation system capable of high performance for real-users and real-world applications.

6. REFERENCES

[1] C.M. Jones, S.S. Dlay, and R.N.G. Naguib, "Berger check prediction for concurrent error detection in the Braun array multiplier", *Microelectronics Journal*, Vol.27 No. 8, pp. 745-755, 1996.

[2] C.M. Jones, S.S. Dlay, and R.N.G. Naguib, *"Berger check prediction for concurrent error detection in the Braun array multiplier"*, Proc. IEEE Int. Conf. on Electronics, Circuits, and Systems (ICECS), Greece, Vol. 1 pp. 81-84, 1996.

[3] S. Morishima and H. Harashima, *"Image synthesis and editing system for a multi-media human interface with speaking head"*, Proc. IEE Int. Conf. on Image Processing and its Applications, pp. 270-273, 1992.

[4] J. Lewis, "Automated lip synchronisation: Background and technique", *Journal of Visualization and Computer Animation*, Vol. 2 Part 4, pp. 118-122, 1991.

[5] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, *"Connectionist probability estimators in HMM speech recognition"*, IEEE Trans. Speech and Audio Processing, Vol. 2 No. 1 Part 2, pp. 161-174, 1994.

[6] H. Bourlard and N. Morgan, "Connectionist Speech Recognition", Kluwer Academic Publ., Massachusetts, 1994.

[7] C.M. Jones, *"Step by step guide to using the speech training and recognition unified tool (STRUT)"*, STRUT package, 1996.

[8] C.M. Jones and S.S. Dlay, *"MARTI: Man-machine animation real-time interface"*, Proc. of IS&T/SPIE Symp on Electronic Imagining (EI), San Jose, California, Vol. 3012, 1997.

[9] C.M. Jones and S.S. Dlay, *"MARTI: Man-machine Animation Real-Time Interface: The Illusion of Life"*, Proc. Int. Conf. on Human-Computer Interaction (HCI), San Francisco, California, 1997.

[10] C.M. Jones and S.S. Dlay, *"Automated Lip Synchronisation for Human-Computer Interaction and Special Effect Animation"*, Proc. IEEE Int. Conf. on Multimedia Computing and Systems (ICMCS), Ottawa, Canada, 1997.