Towards Perceptually Realistic Talking Heads: Models, Methods and McGurk

Darren Cosker[1], Susan Paddock[2], David Marshall[1] and Paul. L. Rosin[1*]
[1]School of Computer Science, Cardiff University
PO Box 916, Cardiff CF24 3XF, U.K.

Simon Rushton[2†]
[2]School of Psychology, Cardiff University
PO Box 901, Cardiff CF10 3YG, U.K.

Abstract

Motivated by the need for an informative, unbiased and quantitative perceptual method for the development and evaluation of a talking head we are developing, we propose a new test based on the "McGurk Effect". Our approach helps to identify strengths and weaknesses in underlying talking head algorithms, and uses this insight to guide further development. The test also evaluates the realism of talking head *behavior* in comparison to real speaker footage, painting an overall picture of a talking head's performance. By distracting a participant's attention *away* from the true nature of the test, we also obtain an unbiased view on talking head performance - since the participant's *prior* concerning what is synthetic animation and what is real footage is not encouraged to develop.

Our current talking head is a hierarchical 2D image based model, trained from real speaker video footage and continuous speech signals. After training, the talking head may be animated using new continuous speech signals not previously encountered in the training set, and produces realistic lip-synched animations. We apply our McGurk perceptual test to our model and demonstrate how we are able to evaluate and identify some of its strengths and weaknesses. We then suggest how our underlying algorithm may be improved in light of the evaluation.

CR Categories: I.2.6 [Artificial Intelligence]: Learning—Parameter learning; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Video analysis; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Motion; I.5.1 [Pattern Recognition]: Models—Statistical; J.4 [Computer Applications]: Social and Behavioral Sciences—Psychology; I.3.7 [Computing Methodologies]: Three-Dimensional Graphics and Realism—Animation; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Perceptual reasoning; G.3 [Mathematics of Computing]: Probability and Statistics—Time series analysis; G.3 [Mathematics of Computing]: Probability and Statistics—Markov processes

Keywords: Facial animation, McGurk Effect, Perceptual Analysis, Psychological Analysis, Lip-syncing, Learning, Video analysis, Audio analysis, Video synthesis

*e-mail: {D.P.Cosker, Dave.Marshall, Paul.Rosin}@cs.cardiff.ac.uk
†e-mail: {PaddockS, RushtonSK}@cardiff.ac.uk

1 Introduction

Many researchers in computer graphics consider the development of a video-realistic computer generated human, indistinguishable from a real human, as the *holy grail* of computer graphics. This task attracts a great amount of interest from the research community and the movie industry, and unsurprisingly is perhaps one of the most, if not *the* most challenging task in computer animation.

One of the difficulties in achieving this goal is in the animation of the face during speech. Humans are highly sensitive to facial behavior, and are experts in identifying flaws in synthetic facial animation. What is required is a means not only of delivering realistic facial animation but also a means of controlling such animation to deliver lifelike behavior. A quick study of the anatomy of the face will convince a reader that it is a highly complex structure [Parke and Waters 1996]. When we speak we perform a sequence of complex muscle and articulatory actions in both the head and throat, which when combined with airflow from the lungs produces sound. Many researchers have attempted to model and animate the face based on its anatomical structure and behavior during speech [Waters 1987][Kahler et al. 2001], while others have used simpler 3D polygonal meshes [Kalberer and Gool 2002][Reveret et al. 2000]. An alternative method of facial animation delivery is using *Morphable Models* [Ezzat et al. 2002][Theobald et al. 2003] which are trained on real speaker footage, and can produce animation of comparable quality to the video sequence used to train it. These models ignore the underlying anatomy of the face and instead model how the face behaves given examples in the training video. Some of the most realistic facial animations to date have perhaps been seen using these latter models [Ezzat et al. 2002][Geiger et al. 2003].

The animation of a facial model using speech alone has been an ambition of the computer graphics and computer vision communities since the 1970's with the publication of Parke's model [1972]. However, in that time the development of effective and scientific methods of evaluating the realism of such models, and determining the deficiencies of these models - with the aim of addressing these and improving the underlying display and animation algorithms - has been somewhat neglected.

We are currently in the process of developing a 2D image-based talking head based on a hierarchy of sub-facial parts [Cosker et al. 2004][Cosker et al. 2003]. Our model is trained from video footage and continuous speech signals from a real speaker, and after training may be animated using new continuous speech signals not previously encountered in the training set. In order to successfully evaluate our model, and investigate its strengths and weaknesses to direct further development, we have been investigating how we might benefit from employing perceptual analysis and psychological approaches.

In this paper we consider how one perceptual approach we have developed might help the graphics community to achieve more robust and informative perceptual evaluation. We propose an experimental evaluation method based on the *McGurk Effect* [McGurk and MacDonald 1976] and apply it to our talking head model. Using our evaluation we demonstrate how we are able to target strengths and

weaknesses in the underlying algorithms, and draw conclusions as to where development directions lie. As a baseline we perform our evaluations along side real video footage to determine a synthetic animation's overall effectiveness. Concurrently we also consider the optimum display size for presentation of the synthetic animations by evaluating our animations at several different resolutions. We consider our measure as one which compliments existing methods for evaluating the lip-synching and realism of synthetic facial animation, providing a deeper insight into the performance of a talking head from a perceptual perspective, and not solely an analytical one.

In the next section we describe previous methods used to determine the effectiveness and realism of speech driven facial animation. In Section 3 we then outline our proposed evaluation approach and describe our facial animation system in Section 4. In Sections 5, 6 and 7 we then evaluate our synthetic animations, give results and discuss their implications. We then conclude in Section 8.

2 Perceptual Evaluation of Talking Heads

The quality of synthetic facial animation, produced solely from speech, has been measured using various approaches. These include subjective assessment [Bregler et al. 1997][Cosatto and Graf 2000][Ezzat and Poggio 1998], visual comparison of synthetic versus ground truth animation parameters [Theobald et al. 2003][Cohen et al. 2002], measurement of a test participants ability to perceive audio in noisy environments with the aid of synthetic animation [Cosatto and Graf 2000][Ouni et al. 2003] and through forced choice experimentation [Ezzat et al. 2002][Hack and Taylor 2003].

Subjective evaluation is the most common method and typically entails comments on the animations from the designers and a number of naive test participants. The observations of the participants are then demonstrated using example videos of the animations. Visual comparison of synthetic versus ground truth parameters involves comparing the trajectories of speech synthesized animation parameters, with trajectories of ground truth parameters, typically obtained from a real speaker. The first method provides subjective information on the overall quality of facial animations, but leaves no means of comparison with other systems, or no direct method of determining any strengths or weaknesses inherent in the synthesis algorithm, e.g. are quickly spoken fricatives synthesized adequately? The second method provides more insight into an algorithm's strengths and weaknesses, and a more quantitative measure of a system's overall effectiveness. However, taken on its own it provides no means of communicating the perceptual quality of an animation, i.e. is it video realistic, can it convince a person into thinking it is a real person? [1]

Measurement of the ability of a synthetic talking head to improve the intelligibility of speech in a noisy environment gives a good indication of the quality of an animation when compared to the performance, in the same circumstances, of real speaker footage. This measure, along with comparisons of synthesized trajectories to ground truths, gives a good overall picture of a talking head's lip-synch ability. However we still do not have a good measure of perceived quality if the goal is to produce synthetic animation indistinguishable from real animation.

Perhaps the most thorough method used to date to measure the perceptual quality of talking head animation is using forced choice experiments. In [Geiger et al. 2003] and [Hack and Taylor 2003] a series of experiments are carried out where the user is asked to state whether a displayed animation is real or synthetic. If the animations are indistinguishable from real video then the chance of correctly identifying a synthetic animation is 50/50. The test may be thought of as a kind of *Turing Test* for facial animation. In [Geiger et al. 2003] [Ezzat et al. 2002] the facial animation is produced from phonemes, thus the test determines the quality of lip-synchronization. Animations of the lower part of the face are synthesized and fixed onto real speaker footage to increase the overall impression of realism. In [Hack and Taylor 2003] it is only facial behavior which is measured (with no audio). Thus forced choice is used in each case to measure different animation characteristics.

A drawback of the forced choice approach is that the participants can develop a *prior* or *opinion* about the animations during testing which may influence their decisions. Any artifacts picked up on in the animations e.g. texture flicker or incorrect co-articulation, cause a participant to develop a picture of what is real and what is not. This is due to the participant's knowledge before the test that some animations will be synthetic and some will be real video footage. A further drawback with this method is that good results can be obtained by randomly selecting either "real" or "synthetic" choices for each clip, as you might expect a bored or uninterested participant to behave.

3 A McGurk Test for Perceptual Evaluation

McGurk and MacDonald [1976] noted that auditory syllables such as /ba/ dubbed onto a videotape of talkers articulating different syllables such as /ga/ were perceived as an entirely different syllable, e.g. /da/. During such a test, when a participant closes his or her eyes the illusion created by the integration of both stimuli vanishes, leaving the participant with perception of the auditory signal alone. This raises important questions in audio-visual analysis, such as how do humans integrate and combine auditory and visual stimulus, and why do we combine such information when the auditory signal is by itself sufficient?

MacDonald and McGurk [1978] argue that when information from both visual and auditory sources is available, it is combined and synthesized to produce the "auditory" perception of a best-fit solution. The *McGurk effect* (as the phenomena is now widely known) has been replicated several times [MacDonald et al. 2000][Dekle et al. 1992][Dodd 1977] using varieties of visual and auditory stimuli. An interesting summary of expected misinterpreted audio syllables, given the influence of a differing visual syllable, may be found in [Dodd 1977].

In this paper we propose the McGurk effect as a basis for perceptual evaluation of our talking head animations. The test allows for the evaluation of lip-synching and overall realism. In our tests, participants are shown real and synthesized *McGurk tuples* [2], one video clip at a time, and chosen at random from a database of video clips. To generate the *real* McGurk tuples we re-dub a video of a person speaking a word with audio taken from the same person speaking a different word. To generate *synthetic* McGurk tuples we use a speech input to generate a video sequence using our talking head, and re-dub this video with a different word.

[1] Of course the aim may not to be to produce animation which can fool a person into believing they are watching a real person - in which case the second point may be ignored.

[2] For ease of exposition we refer to a triplet consisting of a visual syllable or word, dubbed with a different audio syllable or word, along with its expected new perceived syllable or word, as a McGurk tuple.

For each video the participant is simply asked what word they hear while watching. The participants are not informed that some of the clips are real and some are synthetic. The fact that some of the clips *are* generated synthetically should therefore have no influence on the test participant, apart from their perception of the McGurk tuple due to the quality of the animation algorithm. If our lip-synching algorithm produces a poor animation then we expect the audio cue to dominate the visual cue [McGurk and MacDonald 1976], and if good lip-synching is produced from the algorithm we expect a *combined* or *McGurk* response i.e. where the audio and visual data are confused to give a response other than the audio or visual stimuli. We can state that the algorithms lip-synch is effective given either a *combined* or *McGurk* response since the presence of one of these responses depends on correct articulation from the synthetic video, where this video is created from audio alone. Also the illusion of reality produced by the synthetic videos should not be broken given a bad lip-synch animation from the synthesis algorithm, since the participants find, during the course of the experiment, that the audio is *supposed to be* different from the video (i.e. according to the McGurk tuple). A participant's objective knowledge of the illusion (i.e. presence of the McGurk effect) should also not affect his or her perception of the words [McGurk and MacDonald 1976]).

After the test procedure the participants are then asked a sequence of questions to determine whether they noted anything *unnatural* about the videos. This determines the overall effectiveness of the animations in terms of talking head behavior and video-realism of the output. Given sufficiently realistic synthetic video clips, no prior concerning the source of the videos should be developed by the participants, since they are not expecting a mixture of real and synthetic clips. The results of the questioning, along with the success rate of the McGurk effect responses, then form an overall evaluation of the animation algorithm, which may be regarded as both a quantitative measure of lip-synch performance and a Turing test of realism.

To ensure the validity of our McGurk tuples we display real video footage of a speaker as well as synthesized footage. The real and synthesized footage include the same McGurk tuples. This gives us a baseline for our test, i.e. given a McGurk response to a real tuple we might expect a similar response to the synthesized version of the video clip. Given similar responses from a participant to corresponding real and synthetic tuples, for each tuple in the test, we can state that the synthetic lip-synching algorithm is effective. Also, since a McGurk effect response from a participant depends on the correct articulation from the lip-synching algorithm, a non-McGurk response from a participant points to a weakness in the algorithm at co-articulating that specific mouth behavior (i.e. fricatives, consonants, vowels etc). This allows us to analyze our overall results and concentrate on the development of our algorithm in a guided direction, whether that be by providing more training data of certain phrases, or by fine tuning the algorithm to be more sensitive to certain articulatory actions.

4 Talking Head Overview

The talking head used in this study is based on a hierarchical image-based model of the face, projected back into original video footage to increase the illusion of realism. The hierarchical model analyzes independent sub-facial variation from a training video and resynthesizes novel animations for these facial areas given continuous speech. The sub-facial animations are then merged to construct complete facial images for output. Figure 7 shows example synthesized faces before and after registration onto background images. Decomposing the face in a hierarchical fashion has several attractive qualities, such as offering the user of the model a high degree of control over individual parts of the face. Animations are synthesized from the input speech using a Hidden Markov Model(HMM) of co-articulation, which estimates parameters for rendering the different parts of the image-based model from spectral coefficients extracted from the speech. For full details on the system the reader is referred to [Cosker et al. 2004][Cosker et al. 2003]. A thorough description is omitted here since this paper's emphasis is on evaluation and development of talking head models in general, as opposed to our model alone.

5 Experiments

In order to evaluate our talking head using the McGurk perception test we used 20 psychology undergraduate volunteers. This group consisted of 4 males and 16 females, aged between 18 and 29 years (mean age 19.9). All volunteers had normal vision and hearing.

Ten McGurk tuples were used in the experiment. These were chosen through pilot testing from a larger collection of monosyllabic words taken from past research into the McGurk effect [Dekle et al. 1992][Dodd 1977][Easton and Basala 1982]. Table 1 gives the tuples which were chosen. These were used to construct 30 real and 30 synthetic videos, consisting of each tuple, real and synthetic, presented at three different resolutions - 72x75 pixel resolution, 361x289 pixel resolution, and 720x576 pixel resolution. These sizes were chosen as a result of pilot testing, which showed that the 72x75 pixel image produced a McGurk effect roughly 50% of the time in the real video condition, and that the three sizes produced differences between video type conditions (i.e. real or synthetic) and (non-significant) differences between sizes in the proportion of McGurk effects produced. Each video was encoded in the Quicktime movie format. Figure 8 gives an overview of the construction of a synthetic video tuple. The 60 videos (30 real, 30 synthetic) were presented in a random order on a standard PC using a program written in Macromedia Director MX. This program also included two example videos (using the word sets "might-die-night" and "boat-goat-dote", which were not used in the experimental trials) and the option to replay each of the 60 experimental videos before proceeding to the next. Speakers with adjustable volume were plugged in to the PC (adjusted during the example video phase to provide a clear acoustic level) and the experiment took place in a soundproofed laboratory with artificial lighting.

Table 1: McGurk Tuples

Dubbed Audio	Source Video	McGurk Response
Bat	Vet	Vat
Bent	Vest	Vent
Bet	Vat	Vet
Boat	Vow	Vote
Fame	Face	Feign
Mail	Deal	Nail
Mat	Dead	Gnat
Met	Gal	Net
Mock	Dock	Knock
Beige	Gaze	Daige

The number of McGurk responses was assessed using an open response paradigm as used by Dodd [1977], requiring each participant to write down the word they had heard after viewing each video. This removes any interpretation bias which may arise if the experimenter transcribes the verbal responses. Participants were

also directed to fix their attention on the mouth of the video during play back, so as to avoid the experience of an audio only stimuli. Participant responses did not have to be real words as the aim was to find exactly what they were hearing, and it could not be guaranteed that all McGurk effects would produce real words.

After viewing all 60 clips each participant was finally asked 3 questions: 1) "Did you notice anything about the videos that you can comment on?", 2) "Could you tell that some of the videos were computer generated?" and 3) "Did you use the replay button at all?". All the videos used in the test, along with the Macromedia Director MX program used to present them, may be found at *http://www.cs.cf.ac.uk/user/D.P.Cosker/McGurk/*.

6 Results

To code the results we used two different interpretation formats: *Any Audio - Any McGurk* and *Expected Audio - Expected McGurk - Other*. Tables 2 and 3 list the words recorded from the test participants which constituted and warranted these formats. In the first format words which were homonyms of the audio, or which sounded very similar and could easily be confused (for example a slightly different vowel sound) were coded as "audio", while all others were coded as "McGurk". In the second format the *Expected Audio* category contained only the audio words and alternative spellings of these. The *Expected McGurk* category included McGurk words, alternative spellings of these, and words with the same consonant sound when presence or absence of voice was ignored (after [Dodd 1977] for example, "fent" was accepted as a response to the word set "bent-vest-vent"). All other responses were placed in the "other" category.

The rationale behind the *Any Audio - Any McGurk* coding method is that it follows in the spirit of the original McGurk observation, i.e. that a different audio word will be heard under misleading visual stimuli. The *Expected Audio - Expected McGurk - Other* category more closely follows word tuples suggested in previous work. Given the variability in different peoples accents and articulatory behavior, it is unrealistic to assume that a fixed McGurk effect response word would apply to every living person. Therefore the *Any Audio - Any McGurk* format is perhaps more reflective of a general McGurk effect.

Table 2: Any Audio - Any McGurk

Tuple	Accepted Audio	Accepted McGurk
Bat-Vet-Vat	Bat	Vat, Fat
Bent-Vest-Vent	Bent,Bint	Vent,Fent,Fint
Bet-Vat-Vet	Bet	Vet,Fet
Boat-Vow-Vote	Boat,Bolt,Bought,But,Boot Booked,Port	Vote,Fault,Foot,Fought,Fot Thought,Faught,Vault,Caught Caugh,Vought
Fame-Face-Feign	Fame	Feign,Fein,Fain,Vain,Vein Fin,Fiend,Feeind,Thin
Mail-Deal-Nail	Main,Male,Meal,Mayo	Nail,Kneel,Neil,Neal
Mat-Dead-Gnat	Mat	Gnat,Nat,Knat
Met-Gat-Net	Met	Net
Mock-Dock-Knock	Mock,Muck	Knock,Nock,Hock
Biege-Gaze-Daige	Beige,Beidge,Beege,Bij Peege	Daige,Deige,Dij,Dage,Dej Age,Stage,Fish,Eige,Eege Vij,Thage,Veige,These, Theign,Vis,Beign

Figure 1 gives the total number of "McGurk" and "Audio" responses, using the *Any Audio - Any McGurk* coding format, given by participants under all conditions (i.e. all video sizes, real and synthetic videos). The results show large variabilities in participant responses, e.g. participants 4, 7, 18, and 19 showed relatively few McGurk perceptions, while participants 6 and 16 tended to favour McGurk responses over audio. Not surprisingly the same results, coded using the *Expected Audio - Expected McGurk - Other* format (Figure 2), show a change in the number of McGurk responses given by participants. This shows the variability which occurs when *strictly* enforcing previously recorded McGurk responses. As previously mentioned, this variability is to be expected across McGurk experiments using clips built from different people. Therefore we are inclined to base the overall interpretation of our results more on evidence from the *Any Audio - Any McGurk* coding format.

Table 3: Expected Audio - Expected McGurk - Other

Tuple	Accepted Audio	Accepted McGurk	Other
Bat-Vet-Vat	Bat	Vat, Fat	
Bent-Vest-Vent	Bent,Bint	Vent,Fent,Fint	
Bet-Vat-Vet	Bet	Vet,Fet	
Boat-Vow-Vote	Boat	Vote	Bolt,Bought,But, Boot,Booked,Port, Fault,Foot,Fought, Fot,Thought,Faught, Vault,Caught,Caugh
Fame-Face-Feign	Fame	Feign,Fein,Fain Vain,Vein	Fin,Fiend,Thin
Main-Deal-Nail	Main,Male	Nail	Meal,Mayo,Kneel, Neil, Neal
Mat-Dead-Gnat	Mat	Gnat,Nat,Knat	
Met-Gal-Net	Met	Net	
Mock-Dock-Knock	Mock	Knock,Nock	Muck,Hock
Beige-Gaze-Daige	Beige, Beege Beidge	Daige,Deige,Dij, Dage,Dej	Bij,Peege,Age Stage,Fish,Eige Eege,Vij,Thage, Veige,These, Theign,Vis,Beign

Figure 3 shows the mean number of "McGurk" responses, for real and synthetic videos, under each video size, and coded using the *Any Audio - Any McGurk* format. It clearly shows that the number of "McGurk" responses increased with video size using real video, and stayed fairly constant using synthetic video. The graph also indicates that more McGurk responses were recorded using the real video clips. Under *t*-test conditions we see that effect of video type (i.e. real or synthetic) was significant ($F(1, 19) = 315.81$, $p < .01$). We also see that the main effect of video size was significant ($F(2, 38) = 75.48$, $p < .01$), with more McGurk responses in the medium and large conditions than the small condition. The interaction between video type and size was also found to be significant ($F(2, 38) = 44.05$, $p < .01$). Figure 4 repeats these observations, showing the mean number of "McGurk", "Audio" and "Other" responses, for real and synthetic videos, under each video size, and coded with the *Expected Audio - Expected McGurk - Other* format.

In terms of directions for further development of our talking head perhaps the most useful perspective on the results is the number of McGurk effects reported for each real tuple versus effects for synthetic tuples. Figure 5 gives the normalized total number of McGurk and Audio responses for each synthetic tuple under large video conditions, while Figure 6 shows the same information for real tuples. Both figures use the *Any Audio - Any McGurk* coding format. We omit the small and medium video results from these figures since our goal is to analyze the McGurk responses under the best viewing conditions (see Figures 3 and 4). The figures confirm the observation that the McGurk effect is stronger in the real videos

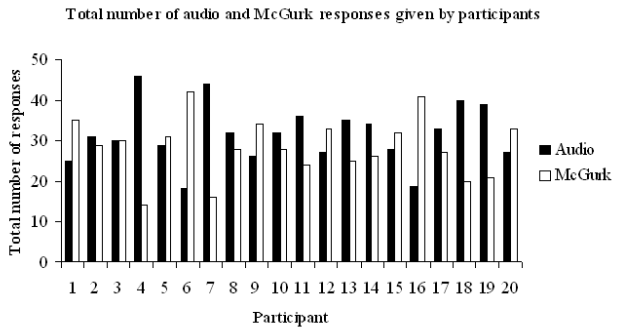

Figure 1: Total Number of McGurk and Audio responses given by participants under all conditions, and using the *Any Audio - Any McGurk* coding format.

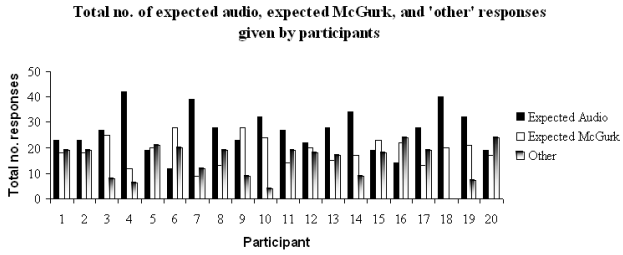

Figure 2: Total Number of McGurk, Audio and Other responses given by participants under all conditions, and using the *Expected Audio - Expected McGurk - Other* coding format.

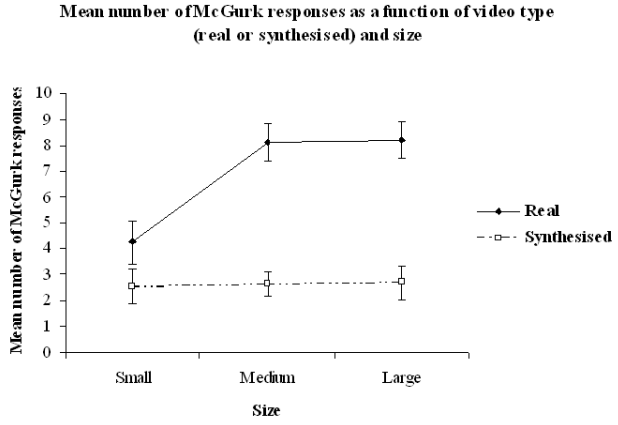

Figure 3: Mean number of McGurk responses given in each condition (i.e. real and synthesized videos and all video sizes), using the *Any Audio - Any McGurk* coding method. Error bars are 2 standard error from the mean.

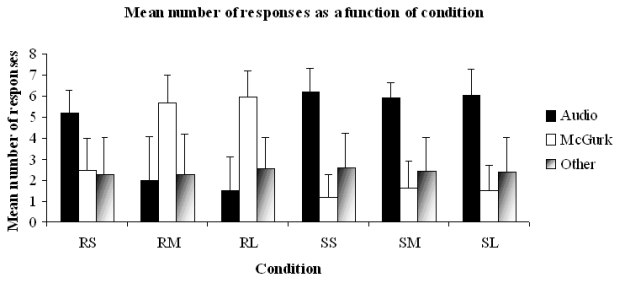

Figure 4: Mean number of responses for real and synthetic videos, under different video size, using the *Expected Audio - Expected McGurk - Other* coding format. *RS, RM and RL* relate to small, medium and large sized real videos respectively, while *SS, SM and SL* relate to small, medium and large sized synthetic videos respectively. Error bars are +1 standard error from the mean.

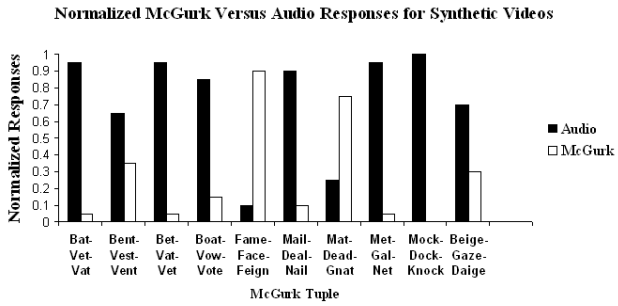

Figure 5: Normalized total number of McGurk and Audio responses for each synthetic video McGurk tuple.

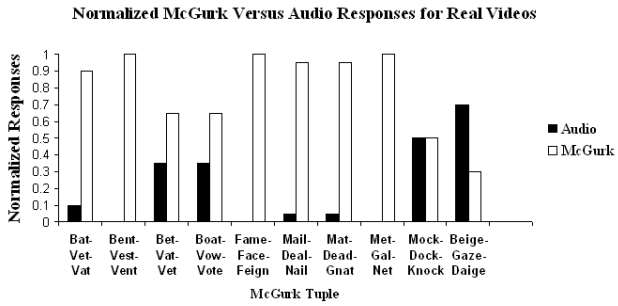

Figure 6: Normalized total number of McGurk and Audio responses for each real video McGurk tuple.

than in the synthetic ones. Using the real videos as a baseline we also notice a bias in the McGurk effect towards certain tuples. This is helpful for future experiments as it allows us to identify weak tuples and remove them from the experiment.

Concerning the end of test questions the following feedback was received: In response to Question 1 most of the participants noticed that the audio did not always match the video, as would be expected given the construction of the McGurk tuples. In response to Question 2 none of the participants noticed that any of the clips were computer generated, although one participant did comment that he thought some of the clips appeared somehow "unnatural". Concerning the use of the replay button (Question 3) most participants chose not to use it, with only a handful opting to use it once, and a single participant using it twice.

In the next section we discuss how the results may be interpreted to identify possible strengths and weakness in our synthetic videos, and how these may be exploited to direct further development of our talking head.

7 Discussion

In terms of overall performance the real video clips produced more McGurk effects from the participants than the synthetic ones. Based on the assumption that production of a McGurk effect implies good lip-synch from the animation algorithm, this points to some lip-synching weakness in our talking head. We consider Figures 5 and 6 in an attempt to identify these weaknesses. The tuple *Beige-Gaze-Daige* performed poorly in the real video and synthetic video trails, with over twice as many "Audio" responses than "McGurk" responses for both video types. This suggests an inherently weak McGurk response for that tuple. The same weakness would also appear to apply to the tuple *Mock-Dock-Knock*, with half the overall responses being "McGurk" and the other half "Audio". These McGurk effect weaknesses are most likely due to the accent of the participant used to create the tuples, and makes the performance of the synthetic versions of these tuples, in comparison to the real tuples, difficult to correctly interpret.

The other real video tuples scored sufficiently high enough to warrant further analysis. The only two synthetic tuples to generate a higher number "McGurk" responses than "Audio" ones were *Mat-Dead-Gnat* and *Fame-Face-Feign*. This suggests satisfactory lip-synch generation for the visemes /D/ and /F/, as well as good articulation throughout the rest of the words (these being *Dead* and *Face* respectively). The high scoring of the real videos of these tuples point to the conclusion that this observation is valid. The content, and poor synthetic video performance, of the remaining synthetic tuples suggests that our animation algorithm currently has difficulty in generating lip-synching for the viseme /V/, as seen in the words *Vet, Vest, Vat and Vow*. To overcome this we believe modification of our training set, to include exaggerated articulation of certain words and consonants, may sensitize our HMM to the presence of these sounds in new talking head input speech signals. Another possibility may be to increase the sampling rate of our speech analysis, in an attempt to better capture short-term speech sounds.

A deeper analysis of the results suggests that our current algorithm is successfully modelling and then synthesizing visemes such as /S/, /A/ and /E/, these being present in the high scoring synthetic tuples. Although further analysis is required in order to confirm this.

Feedback from the questions posed to the participants was particularly encouraging. Out of the 20 participants none stated that they thought any of the clips were computer generated, with only one participant mentioning that "some of the clips seemed somehow unnatural". This points to an overall realistic output, and realistic behavior in our talking head. It also helps support the hypothesis that given no prior, or at best an undeveloped one, a person is less likely to notice a synthetic talking head animation over a real one.

The increase and subsequent plateau in the mean number of McGurk responses to the real video tuples under varying sizes (Figure 3) suggests that intelligibility of the clips degrades between video resolutions of approximately 361x289 pixels and 72x75 pixels, and reaches an optimal level at a resolution between 361x289 pixels and 720x576 pixels - suggesting that an increase in resolution at this point does not contribute to talking head intelligibility. The effect however is less dramatic for the synthetic tuples, and is likely due to the generally low number McGurk responses given in relation to the synthetic clips.

One matter concerning our talking head which we have not yet tested is a participants opinion towards the synthetic videos given longer clips. Given a test of this nature we suspect a participant might become more suspicious of their origin, as reported in experiments using a different talking head in [Ezzat et al. 2002]. In order to perform a test of this nature we are considering using McGurk *sentences*, one example being the audio "My bab pop me poo brive", dubbed onto the video "My gag kok me koo grive", with the expected McGurk effect of perceiving "My dad taught me too drive". Another alternative is to simply create synthetic clips of the talking head speaking a long list of single words (with corresponding real video clips as a baseline).

An alternative to using the McGurk effect as a perceptual test in the manner we have proposed might also be to simply play real and synthetic clips with the *correct* dubbing, and to ask participants what words they hear. Given incorrect dubbing we might then expect McGurk responses from the participants. The attractive aspect of this approach is that the participants are not informed that some clips are real and some synthetic, again reducing their development of a prior.

Although the test described in this study is biased towards determining the effectiveness of artificial lip-synching in near-video realistic talking heads, we envisage that it may also be modified in order to analyze lip-synching in cartoon like animations. One suggested approach to achieving this may be to substitute the "real" video clips with animations generated using motion-capture or key-framing - these approaches being widely used for film and computer game facial animation. In this circumstance the test would then compare the effectiveness of speech driven synthetic facial animation versus animation generated through current industry techniques (based on the assumption that these methods are sufficiently effective enough to generate McGurk effects).

Given a fully developed McGurk talking head test one thing that we do not envisage is that it could be modified to test the synthesis of facial emotion and expression. However, one approach to achieve this kind of testing, from a perceptual point of view, is to play random real and synthetic clips and ask a participant what emotion they are witnessing [Katsyri et al. 2003]. An attractive quality of this type of test is that it reduces prior opinion from the participants. A test of this nature also addresses strengths and weaknesses in the emotion algorithm since it allows the developer to identify which synthetic emotions are confused with real ones (if any), in order to direct the algorithm in rectifying these ambiguities.

One final point which should be made concerns our selection of McGurk tuples, which does not yet include tests for every type of Viseme. We are currently investigating new tuple combinations in order to address this issue.

8 Conclusion

We have described a new perceptual approach for analysis and development of talking heads based on the *McGurk effect*, and have applied this to a 2D talking head we are developing. Our approach addresses weaknesses in current perceptual evaluation techniques, such as forced choice analysis, in that we remove the development of a prior from our participants which can bias trials. Unlike other perceptual evaluation methods our test also gives an insight into the performance of the underlying talking head synthesis algorithm, enabling the identification of its strengths and weaknesses in order to direct further development. By applying the McGurk perception test to our current talking head we have identified several such strengths and weaknesses, and considered ways in which these may be addressed. We have also evaluated how convincingly our talking head *behaves* in comparison with real speaker footage, and found encouraging results. Overall we see our test as one which can compliment existing tests to provide a more rigorous overall evaluation of a talking head.

In order to perceptually evaluate a talking head over long time periods, and to evaluate the synthesis of facial emotion, we have suggested a number of possible solutions. These include using longer McGurk sentences, or long combinations of McGurk words, and comparisons between real and synthetic emotion animations where a user is simply asked "what emotion is being synthesized?".

References

BRAND, M. 1999. Voice puppetry. In *Proc. Computer graphics and interactive techniques*, ACM Press, 21–28.

BREGLER, C., COVELL, M., AND SLANEY, M. 1997. Video rewrite: driving visual speech with audio. In *Proc. of 24th conf. on Computer graphics and interactive techniques*, ACM Press, 353–360.

COHEN, M., MASSARO, D., AND CLARK, R. 2002. Training a talking head. *IEEE Fourth International Conference on Multimodal Interfaces*.

COSATTO, E., AND GRAF, H. P. 2000. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia 2*, 3, 152–163.

COSKER, D., MARSHALL, D., ROSIN, P., AND HICKS, Y. 2003. Video realistic talking heads using hierarchical non-linear speech-appearance models. In *In Proc. of Mirage*, 20 – 27.

COSKER, D., MARSHALL, D., ROSIN, P. L., AND HICKS, Y. 2004. Speech driven facial animation using a hierarchical model. *IEE Vision, Image and Signal Processing (to appear)*.

DEKLE, D., FOWLER, C., AND FUNNELL, M. 1992. Audio-visual integration in perception of real words. *Perception and Psychophysics 51*, 4, 355–362.

DODD, B. 1977. The role of vision in the perception of speech. *Perception 6*, 31–40.

EASTON, R., AND BASALA, M. 1982. Perceptual dominance during lipreading. *Perception and Psychophysics 32*, 6, 562–570.

EZZAT, T., AND POGGIO, T. 1998. Miketalk: A talking facial display based on morphing visemes. *Proc. of Computer Animation Conference*.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable video-orealistic speech animation. In *Proc. of Computer Graphics and Interactive Techniques*, ACM Press, 388–398.

GEIGER, G., EZZAT, T., AND POGGIO, T. 2003. Perceptual evaluation of video-realistic speech. *CBCL Paper 224/AI Memo 2003-003, MIT, Cambridge, MA*.

HACK, C., AND TAYLOR, C. J. 2003. Modelling talking head behavior. *In Proc. of British Machine Vision Conference*.

KAHLER, KOLJA, HABER, JORG, AND SEIDEL. 2001. Geometry-based muscle modeling for facial animation. In *Proc. of Graphics Interface*, 27–36.

KALBERER, G. A., AND GOOL, L. V. 2002. Realistic face animation for speech. *Journal of Visualization and Computer Animation 13*, 97–106.

KATSYRI, J., KLUCHAREV, V., FRYDRYCH, M., AND SAMS, M. 2003. Identification of synthetic and natural emotional facial expressions. In *In Proc. of International Conference on Audio-Visual Speech Processing*, 239–243.

MACDONALD, J., AND MCGURK, H. 1978. Visual influences on speech perception processes. *Perception and Psychophysics 24*, 3, 253–257.

MACDONALD, J., ANDERSON, S., AND BACHMANN, T. 2000. Hearing by eye: How much spatial degradation can be tolerated? *Perception 29*, 10, 1155–1168.

MCGURK, H., AND MACDONALD, J. 1976. Hearing lips and seeing voices. *Nature 264*, 746–748.

OUNI, S., MASSARO, D., COHEN, M., YOUNG, K., AND JESSE, A. 2003. Internationalization of a talking head. *In Proc. of 15th International Congress of Phonetic Sciences, Barcelona, Spain*.

PARKE, F. I., AND WATERS, K. 1996. *Computer Facial Animation*. A. K. Peters.

PARKE, F. 1972. Computer generated animation of faces. In *Proc. of ACM National Conference*.

REVERET, L., BAILLY, G., AND BADIN, P. 2000. Mother: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation.

THEOBALD, B., CAWLEY, G., GLAUERT, J., AND BANGHAM, A. 2003. 2.5d visual speech synthesis using appearance models. In *Proc. of BMVC 2003*, vol. 1, 43–52.

WATERS, K. 1987. A muscle model for animation three-dimensional facial expression. In *Proc. of Computer Graphics and Interactive Techniques*, ACM Press, 17–24.

Figure 7: Background registration of facial images generated by our synthesis algorithm: Background images taken from the training set (Top), and facial images generated from our synthesis algorithm (Middle), are aligned and merged to produce output frames (Bottom).

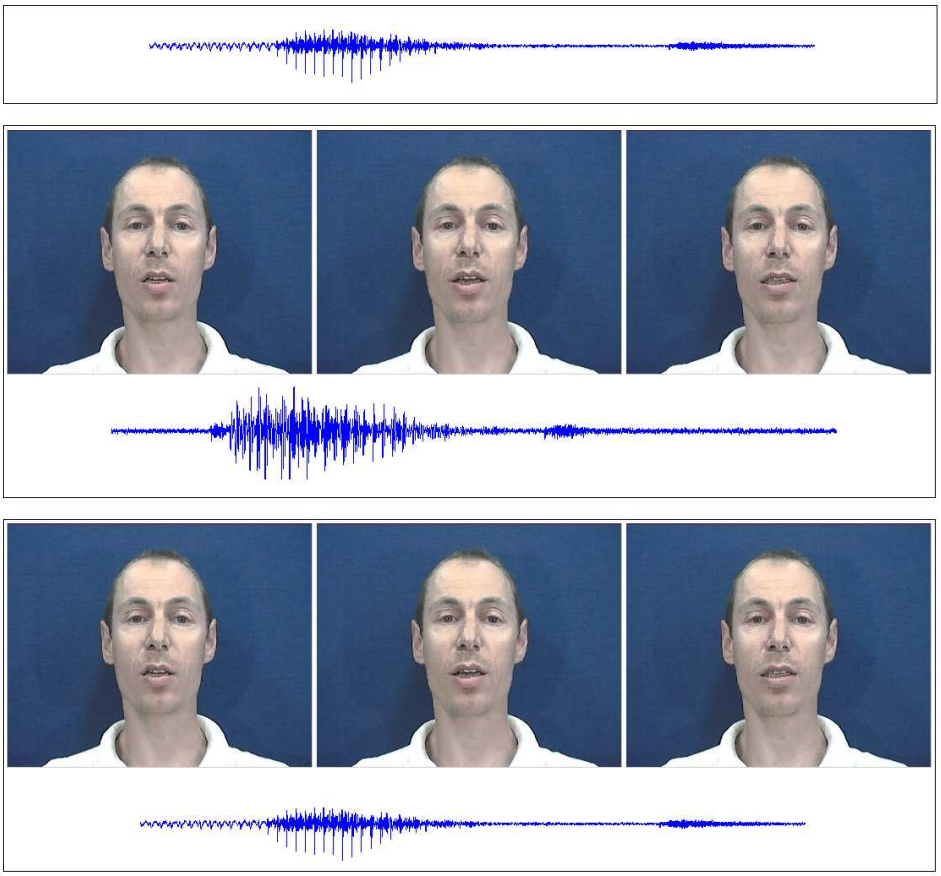

Figure 8: Example preparation of a synthetic McGurk tuple: Audio of the word “Mat” is recorded (Top Box). Video and Audio of the word “Dead” is recorded (Middle Box). Audio for the word “Mat” is dubbed onto video for the word “Dead”, producing the McGurk effect “Gnat” (Bottom Box).