The Influence of Lip Animation on the Perception of Speech in Virtual Environments

J.H. Verwey, E.H. Blake
Collaborative Visual Computing Laboratory
Department of Computer Science
University of Cape Town
Rondebosch, South Africa
{jverwey@cs.uct.ac.za, edwin@cs.uct.ac.za}

Abstract

The addition of facial animation to characters greatly contributes to realism and presence in virtual environments. Even simple animations can make a character seem more lifelike and more believable. The purpose of this study was to determine whether the rudimentary lip animations used in most virtual environments could influence the perception of speech. The results show that lip animation can indeed enhance speech perception if done correctly. Lip movement that does not correlate with the presented speech however resulted in worse performance in the presence of masking noise than when no lip animation was used at all.

1. Introduction

The ability to read a speaker's lips has a significant impact on speech perception [26] . In other forms of media like television and cinema, this visual cue is readily available. For virtual environments lip animations have to be created for every character that will be speaking. Since this can be a time consuming process, most applications provide only very rudimentary lip animations, if at all. While some lip movement certainly contributes to realism and a feeling of presence [6] [10] , it is uncertain whether it can contribute to speech perception. Lip reading in real life provides additional visual information that is integrated with the auditory information in the perception of speech. The perceptual system will however rely more on the visual modality when the auditory cues are weak [9] . Virtual environments are interactive by nature. Sounds can be generated at arbitrary times, which could make it more difficult to hear spoken dialog. In contrast to this, sound tracks for film and television are completely linear and sound engineers have exact control over what the listener will hear. Providing correctly animated lips may therefore be more important in virtual environments where a greater reliance is placed on visual information than in an animated film where the sound track can be edited until the dialog is clear.

Most studies involving lip reading make use of video streams of real faces. It has been shown that video streams with frame rates as low as five frames per second can still contribute to speech perception [14] . Some studies have shown that the artificial reconstruction of lip movement using 3D geometric models can also benefit hearing performance [18] . This benefit may extend to simpler lip animations typically used in virtual environments.

In this study subjects were required to identify spoken words that were accompanied either with simple but correctly constructed lip animation, incorrect animation or no lip animation at all. A noise masker was presented together with the spoken sentence in order to make the task more difficult. We show that correct lip-animation enhances speech perception, but incorrect animation degrades speech perception. The effect is most pronounced when hearing is made difficult by the masking sound. Under these conditions the visual modality is favoured and subjects tend to perceive the visually presented rather than the auditory presented words.

1.1. Visemes for Animation

Auditory speech sounds are classified into units called phonemes. The visual counterpart for a phoneme is called a *viseme* [5] . A viseme represents the shape of the lips when articulating an auditory syllable. Many phonemes however have ambiguous visual representations and map to the same viseme. The Preston Blair phoneme series [2] is a popular set of visemes often used for facial animations in cartoons. In this series only 10 visemes are used to map to all possible phonemes (Figure 1).

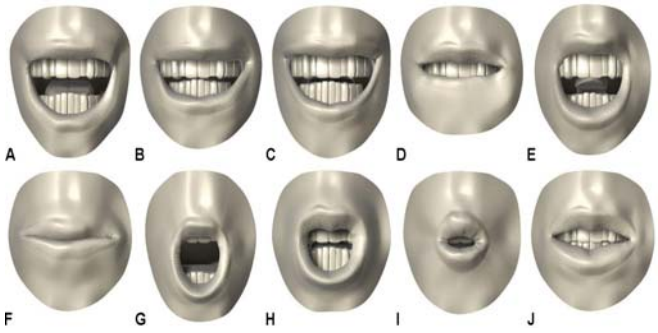

Figure 1. The Preston Blair phoneme series. Each visual representation (viseme) represents one or more auditory phonemes.

Chen *et al* [7] presents an overview of different methods of creating speech-driven facial animations and lip synchronization. Lip animations are constructed by either using a flipbook method, or by using geometry morphing. The flipbook method rapidly displays a list of consecutive

visemes together with the auditory speech to create an impression of lip movement. Since there are a limited number of facial expressions, this method can result in jerky animations when no intermediate frames are drawn for the transition between different visemes. The geometry morphing method requires a 3D model of a face to be constructed. The geometry of the face can be smoothly interpolated between different facial expressions resulting in very smooth animation.

Both methods require the different visemes to be synchronized with auditory phonemes as they are spoken. Lip animations can be derived from acoustical speech input by using various computational methods. Lavagetto made use of neural networks for speech-driven facial animation in a multimedia telephone application for the hard of hearing [18] . He showed that the resulting lip animations of a geometric model were useful for enhancing speech perception. Much simpler methods are used for creating animations when using the flipbook method. Software tools like PAMELA [25] extract phonemes from a given text sentences and map them to visemes. The time offset for each viseme can be manually adjusted until the animation looks realistic.

The computational cost involved in creating facial animations directly from the acoustical speech data can be prohibitive for virtual environments that typically spend most processing time on graphics, physics and artificial intelligence computations. The flipbook method is more suitable for these kinds of applications since it uses very few computational resources [7] .

1.2. Accurate Speech Perception

Animations need to be carefully constructed. Phonemes should map to the correct visemes in order for the additional visual information to contribute to speech perception. "The *McGurk* Effect" [21] has shown that different syllables can even be perceived when contradictory visual information is presented together with auditory speech. This effect illustrates how incongruent auditory and visual information can cause a different perception of the auditory stimuli.

For example, when someone hears the auditory syllable /ba/ but sees the visible syllable /ga/ (Viseme (B) followed by viseme (A) in Figure 1) being articulated, it is usually perceived as /da/. The perceived audio-visual syllable has the same visual representation as the presented visual syllable but differs from the presented auditory syllable.

Only some combinations of auditory and visual syllables produce McGurk effects. These studies typically use a limited set of stimuli that usually only consist of single syllables. It has however been shown that the McGurk effect can be obtained for normal words. If the visually and auditory presented words are picked very carefully a completely different English word can be perceived. If for example the auditory word 'mail' were presented together with the visual word 'deal', the word 'nail' would be perceived [9] . It is clear that the visual representation of a spoken sentence can have a significant impact on the perception of the words.

Adverse listening conditions may further aggravate this effect. Interaction in the virtual environment could cause additional sounds to be produced that could drown out spoken dialog. The addition of a masking noise will cause greater reliance on the visual cues. When two sources of information conflict, in this case visual and auditory, the stronger source is usually favoured [9] . Incorrectly constructed lip animations may therefore result in worse hearing performance than when the listener only relies on auditory information.

1.3. Directional Sound

Immersive virtual environments often present a variety of background sounds, music, dialog and effects simultaneously. The human perceptual system has the remarkable ability to pay selective attention to a sound of interest in the midst of other competing sounds. This is often called the "Cocktail-party Effect" [8] [12] . This ability allows listeners to attend to a specific voice while ignoring other voices and background noise. One of the contributing factors in distinguishing sound sources is their physical location [3] . A difference in the location of sound sources greatly enhances the intelligibility of speech in the midst of a masking noise or other competing voices. This is referred to as a *spatial release from masking* or *spatial unmasking* [12] . Directional sound can influence speech perception in real life as well as in virtual environments.

Research in virtual auditory environments has shown that it is possible for sounds to be presented over stereo headphones in such a way that it is perceived as coming from any position in 3D space [1] . Digitized sound data are manipulated to create a stereo sound stream with the separate channels representing the sound that would be perceived at each ear. Slight changes in level, timing and spectrum at each ear will cause virtual sound sources to be perceived at different locations when played over stereo headphones. This is referred to as *sound spatialization* or more commonly, *3D sound*. It has been shown that a release from masking can be obtained in virtual auditory environments where virtual sound sources are spatially separated from one another [11] .

In this study we presented target speech sentences from different locations relative to a masking sound. This allowed us to investigate the influence of the visual lip animation cues at different levels of hearing difficulty.

3. Method

We investigated the problem by performing a large number of trials over an extended period on a few volunteers.

3.1. Subjects

Four paid volunteers were recruited as test subjects for this research. All subjects were between the ages 20 and 30, had self-reported normal hearing and normal or corrected-to-normal vision. Subjects were not informed of the goal of the experiments. The use of four subjects may seem rather

few in the Virtual Reality field. The nature of the phenomena being investigated, however, are such that we do not expect much variation between subjects, provided they have normal hearing. The variation is expected to arise within the experimental subjects and should decrease as the task is learned. The strategy in this sort of research is to choose a few volunteers and then to conduct a very large number of trials with each person. The most rigorous approach would have been to first establish that our subjects did have normal hearing and vision but in practice self-report as well as an initial control of the results for outliers is acceptable. This methodology is consistent with other speech perception studies [4] [16] [24] .

3.1. Stimuli

Sentences from the Coordinate Response Measure (CRM) corpus [23] were used as auditory stimuli. This corpus has a limited vocabulary with target words consisting of a call sign, a colour and a number. Sentences have the following format:

"Ready (Call sign) go to (Colour) (Number) now."

The call sign can be 'Arrow', 'Baron', 'Charlie', 'Ringo', 'Laker' or 'Tiger'. The possible colours are 'Blue', 'Red', 'White' or 'Green' while the numbers range from one to eight. Subjects were required to identify the correct colour and number combination in a spoken sentence while a masking noise was simultaneously presented. Although the CRM corpus is publicly available for research, all speakers used for the recordings had American accents. Since some subjects might find it difficult to recognize a foreign accent, especially in noisy conditions, it was decided to create a CRM corpus using a native speaker. A native English-speaking female drama student was used as voice talent. Professional sound engineers were employed to record the target stimuli. Each sentence was 2.5 seconds in length on average and was recorded at 48 kHz. The sound files were first edited to make sure every file immediately started with the first word without any delay. The sound files were also trimmed at the end after the last word has been spoken.

Since the call sign was not important for our experiments, only the call sign "Baron" was used. The number 7 was not used in any trials since it is the only two-syllable number and would be easier to recognize. This left four colours and seven numbers in the vocabulary. With 28 possible permutations of colour and number, the chance of a subject guessing both the correct colour and number is 3.6%. In some cases subjects may have been able to recognize only one of the target words, this would clearly be better than recognizing nothing at all. Since this information would be lost when using absolute scoring, it was decided to award a point for answering the correct colour and another point for the correct number. When scored this way the chance of a correct guess is 19.6%.

Since distracting sounds in virtual environments are not limited to speech sources, it was decided not to use speech spectrum or speech shaped noise for these experiments as is common in speech perception studies. White noise of the same length as the longest speech stimulus was generated for the masking stimuli. Ten different masking files were created in this way and were randomly presented during experiments. The root mean square (RMS) energy of a sound file refers to the square root of the mean of the squares of the all the digitized sound sample values. In order to make sure the target-to-noise-ratio was calculated correctly, the RMS energy of the masker and the target sounds were first normalized. All target files were scaled to have data values in the (-1, 1) range. The minimum RMS energy for these files was then calculated and all files were scaled to have the same RMS energy. The masker stimuli were then scaled to have the same RMS as the normalized target stimulus. All stimuli were ramped with a cosine-squared window to remove any clicking at the beginning and end of sentences when presented.

Brungart [5] provides an overview of the creation of spatialized audio over stereo headphones. This involves convolving the impulse responses measured on a KEMAR dummy head with the target stimuli to create a separate set of stereo sound files. KEMAR is a standard audiological research mannequin manufactured by Knowles Electronics. MIT Media Lab measured the impulse responses used in this study [15] . Spatialized sounds were produced by convolving the signals with KEMAR HRTFs for angles 0° and 15° in the horizontal plane. All sounds were created with a zero elevation angle. No further processing was performed on the stereo sound files during presentation.

For visual stimuli 3D models were used to represent the sound producing objects. These models can be seen in Figure 2.

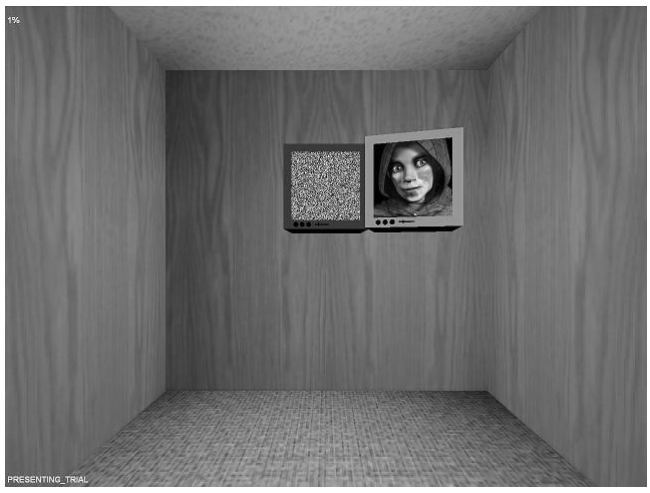

Figure 2. A screen shot of the virtual environment.

A television screen that displayed a snowy picture, as is common when there is bad reception, represented the masker object. A face representing the target was presented on a separate television in the virtual environment. The snowy television was animated by randomly switching between different noisy images at a constant frame rate.

Illustrations of the Preston Blair phoneme series [2] were used for animating the face of the character. Each animation frame in Figure 3 represents a viseme that corresponds to one or more phonemes. An animation file containing the relative timing offsets of different frames was created with the help of a lip synchronization utility called PAMELA [25] . This tool can determine the correct phonemes to use for any given English sentence. These phonemes were then mapped to appropriate visemes in the Preston Blair series as shown in Table 1.

Phoneme	Example	Viseme
AA	Father = F AA DH ER	J
AE	At = AE T	J
AH	Hut = HH AH T	J
AO	Dog = D AO G	J
AW	Cow = C AW	J
AY	Hide = HH AY D	J
B	Be = B IY	G
CH	Cheese = CH IY Z	C
D	Deed = D IY D	C
DH	Thee = DH IY	C
EH	Ed = EH D	D
ER	Hurt = HH ER T	B
EY	Ate = EY T	D
F	Fee = F IY	F
G	Green = G R IY N	C
HH	He = HH IY	H
IH	It = IH T	J
IY	Eat = IY T	D
JH	Gee = JH IY	H
K	Key = K IY	C
L	Lee = L IY	E
M	Me = M IY	G
N	Knee = N IY	C
NG	Ping = P IY NG	H
OW	Oat = OW T	I
OY	Toy = T OY	I
P	Pea = P IY	G
R	Read = R IY D	C
S	Sea= S IY	C
SH	She = SH IY	H
T	Tea = T IY	H
TH	Theta = TH IY T AH	C
UH	Hood = HH UH D	I
UW	Two = T UW	I
V	Vee = V IY	F
W	We = W IY	A
Y	Yield = Y IY L D	C
Z	Zee = Z IY	C
ZH	Seizure = S IY ZH ER	H

Table 1. Phonemes mapped to visemes.

While Pamela cannot create the correct timing offsets for each frame from the speech file, it does allow the user to adjust the timing offsets until the animation looks correct. For the sentence "Ready Baron, go to blue, one now", Pamela would produce the following phonemes for each word: Ready - R, EH, D, IY, Baron - B, AE, R, AH, N, go - G, O, to - T, UW, blue - B, L, UW, one - W, AH, N, now - N, AW. These phonemes were mapped to the following visemes in Figure 3: Ready - C, D, C, C, Baron - G, J, C, J, C, go – C, I, to – H, I, blue – G, E, I, one – A, J, C, now C, J. Animation files were created in this way for every sentence presented during the experimental trials.

Figure 3. Target speech animation frames.

3.2. Procedure

The experimental software was run on a desktop-based system with a 3000 MHz Intel Pentium processor and 512 MB RAM. The system was also equipped with a GeForce FX5900 graphics card with 128 MB onboard RAM and a Creative Labs Sound Blaster Audigy 2 sound card. A pair of Sennheizer HD 580 circum-aural headphones was used as for the auditory display and a Virtual Research V6 Head-mounted display (HMD) for the visual display. This HMD supports a resolution of 640x480 and can display a 60° field-of-view. The virtual environment application was written in C++ using the Microsoft DirectX API [22] .

All subjects participated in five experimental sessions on five consecutive days. During each experimental session an adaptive method was first used to determine the subject's speech reception threshold (SRT). This refers to the minimum target-to-noise ratio (TNR) at which subjects can reliably perform the task. The transformed up-down method [19] was used to determine the SRT. This method targets the 71% correct response threshold. The method starts with equal target and masking noise levels yielding a target-to-noise ratio of 0 dB. For normal hearing subjects it is very easy to achieve a 100% correct score in this condition.

The task is then progressively made more difficult by lowering the volume of the target stimulus whenever the subject scores two correct answers in a row. As soon as the subject gives a single incorrect response the level is adjusted to make it easier again. A reversal happens when

the subject either scores two consecutive trials correct after an incorrect response, or if an incorrect response directly follows two or more correct responses. The process is stopped when the number of reversals reaches a predetermined threshold. During this adaptive procedure the target level is adjusted with varying amounts. Initially big step sizes are used in order to reach the SRT more quickly. The step size is progressively made smaller to obtain a more optimal SRT.

The following values were determined during pilot studies.

- Until 1st reversal, adjust the volume by 5dB.
- Until 3rd reversal, adjust the volume by 3dB.
- Until 7th reversal, adjust the volume by 1dB.
- Until 13th reversal, adjust the volume by 0.5dB.

Once the SRT has been determined for the subject, all experimental trials can be presented at the measured TNR for different conditions. The 71% correct response threshold leaves enough room to show an increase or decrease in performance when the experimental condition is changed.

For the first 3 days subjects had to complete 3 adaptive learning blocks to find an adequate TNR for each subject. During the adaptive trials only audio was presented and the target and masker objects were invisible. The auditory masker was always presented at 0° while the target was presented at 15° to the right. Each of these blocks lasted for about 5-6 minutes. An experimental block of up to 20 minutes followed after this. The average TNR measured in the 3 adaptive blocks was used as the TNR for the experimental block. On the last two days no adaptive blocks were conducted, but two experimental blocks, using the average TNR measured on the third day. Three visual conditions were presented. The face could be correctly animated, incorrectly animated or not animated at all. For incorrectly animated conditions, the animations of a different target sentence were randomly selected. For example, if the auditory stimulus was the sentence “Ready Baron go to blue, one now”, the visual animation for the sentence “Ready Baron go to green, five now” could be presented.

Two spatial conditions were presented. The target object was either presented at 0° or 15° to the right. The masker was always presented directly in front of the listener at 0°. When both the masker and target were presented from the same direction, the target object obscured the masker object as seen in Figure 4.

Figure 4. The co-located condition. The masker object is obscured.

	Target	Masker	Animation
1.	0°	0°	Animated
2.	0°	0°	Non-animated
3	0°	0°	Incorrectly animated
4.	15°	0°	Animated
5.	15°	0°	Non-animated
6.	15°	0°	Incorrectly animated

Table 2. Spatial and visual conditions presented.

The six different conditions presented are summarized in Table 2. The reason for using two spatial positions in this experiment was to investigate the influence of lip-animation at different levels of hearing difficulty. Because of spatial unmasking, the target sentence would be easier to hear when presented at 15° than in the co-located condition.

A slightly transparent input console was superimposed on the display area at the end of each trial. This allowed subjects to provide responses without having to remove the HMD. This can be seen in Figure 5.

Some target words in the CRM corpus are easier to recognize than others [4] . If some words presented in one condition were easier to identify than words presented in another condition, this would create a misleading bias towards one condition. To prevent this, all sentences were presented an equal number of times under all experimental conditions. This ensured that an equal number of easy and difficult sentences were presented for all conditions, removing the bias towards any one condition.

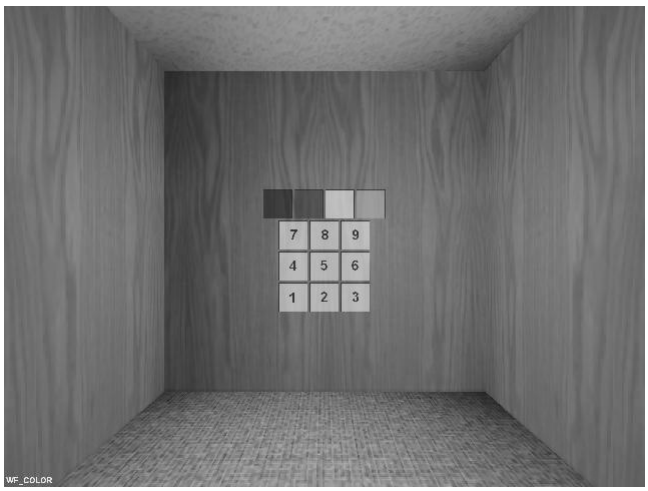

Figure 5. Input console for trial responses.

To minimize the effect of fatigue, all sessions were kept under one hour and subjects were given a short break between blocks of trials. During pilot testing it was observed that subjects tend to perform better towards the end of a block than at the very beginning. To account for any learning effects within a block, a few warm-up trials were first presented. These trials were not considered for data analysis.

During each experimental block, 28 trials were presented for six different conditions. The last two sessions contained two experimental blocks and no adaptive blocks. A total of seven experimental blocks were conducted over the five days. This resulted in 196 trials per condition. This excludes any adaptive trials since the number of trials presented during each of these blocks naturally varies. To account for learning effects, the first two experimental blocks were not considered for data analysis, leaving 140 usable trials per condition for every subject.

5. Results

Figure 6 shows subject performance for different visual conditions. From left to right the conditions were: correctly animated, non-animated and randomly animated. A spatial release from masking was observed for all three visual conditions. Subjects performed best for correct lip animations and worst when incorrect animations were used. An ANOVA between the co-located and separated conditions showed a significant difference between the two spatial conditions [$F(1,3) = 664.97$, $(p<0.001)$]. Further analysis showed that the difference is significant for all visual conditions. An ANOVA across the different visual conditions also revealed a statistically significant difference between the three visual conditions [$F(2, 6) = 28.2$, $p < 0.001$]. Further comparisons revealed that all visual conditions differ significantly for both spatial conditions.

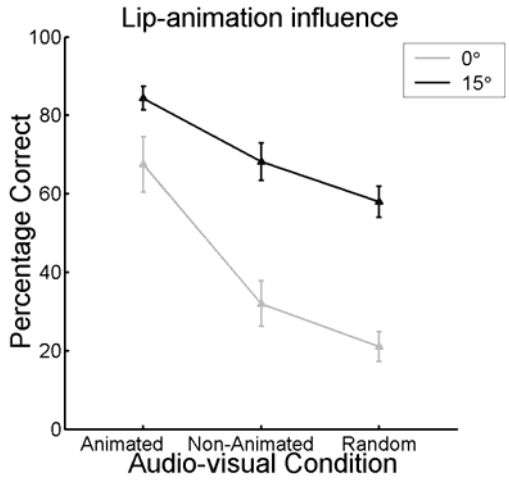

Figure 6. Subject performance under different visual and spatial conditions. 0º means the target and masker were co-located and 15º is the spatially separated condition where the target sound was located to the right. The vertical bars represent the standard deviation.

From Figure 6 we saw that performance for the incorrectly animated condition was worse than the correctly animated and non-animated visual conditions. In this condition animations from different colour and number combinations were used as visual stimuli. The question arises whether this incorrect visual information is merely distracting or whether it created a perceptual bias in favour of the visually presented words.

One could use an alternative scoring to determine how well the subject would have performed if we used the visually presented colour and number as the correct response instead of the auditory. If subjects consistently picked the colours and numbers they saw, one could conclude that subjects relied more strongly on the visual than the auditory cues.

From Figure 7 it is clear that when scoring in this way there is a dramatic difference in the results. On the left the responses are scored according to the auditory presented stimuli. On the right, subject responses are scored against the visually presented stimuli.

For the co-located condition, subjects performed better when using the alternative scoring method. The visual score was significantly higher than the auditory score, which is almost the same as chance (19.6%). This implies that subjects tended to answer according to the visually presented stimuli, that is, the visemes, in the co-located condition. In the separated condition, where spatial unmasking resulted in better hearing conditions, subjects tended to answer according to the auditory presented stimuli, ignoring incongruent visual information. In this condition the visual score was slightly above chance indicating that the incorrect animation still had some impact.

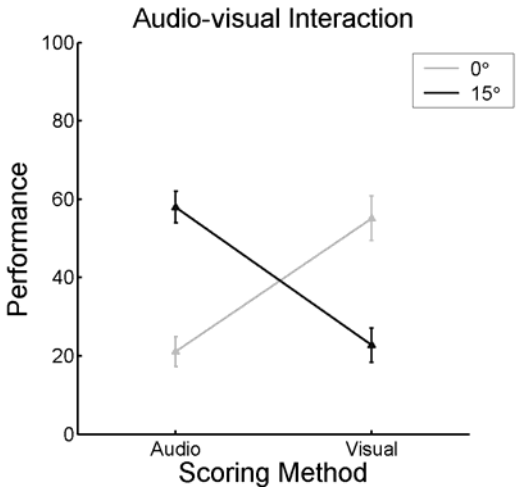

Figure 7. Subject performance when using two different scoring methods for the randomly animated condition. When both the target and the masker were presented at 0° subjects tended to answer according to what they saw rather than what they heard. When the objects were separated by 15°, the auditory cues were stronger and subjects answered according to what they heard while ignoring incorrect visual information.

6. Discussion

Our results show that subjects found it much easier to recognise target words when the noisy television was presented in front and the target television was presented to the right of the masker. Although not the primary objective of this research, this result shows that a spatial release from masking was obtained when the target and masker sounds were presented from different directions. This is consistent with previous findings in the literature [11] .

In this experiment we expected correct lip animation to aid in speech recognition. The results confirmed this and show that correct lip animation significantly contributes to hearing performance. As expected, the incorrect animated condition did result in worse performance than the non-animated condition.

It may be that the interaction between visual and auditory information caused subjects to hear something completely different as is found in experiments involving the McGurk effect. However McGurk experiments are generally very carefully constructed. Only some combinations of strong visual cues with opposing weak auditory cues produce this effect. It is unlikely that the vocabulary of the CRM corpus would result in any McGurk effects when presenting random combinations of auditory and visual stimuli.

In the co-located condition, both the target speech and the masking noise were presented from the same location making it very difficult to hear the target words. The massive increase in performance between the non-animated and correctly animated case in the co-located condition suggests that subjects are able to lip-read very well. It therefore seems likely that the visual cue also had a big influence during the incorrectly animated condition. From Figure 7 we can see that subjects indeed scored higher for the visually presented words than for the auditory ones. This suggests that at least for the co-located condition the visual cue was favoured and subjects answered according to what they saw rather than what they heard.

In the separated condition the target speech and masking noise were spatially removed making it easier to distinguish the target words even though the animation was incorrect. Subjects performed reasonably well and the auditory score was better than the visual score. These results are consistent with findings in the literature that suggests that the stronger cue will usually be favoured when two sources of information conflict [9] .

Overall these results suggest that adding lip animation to characters in virtual environments will significantly increase hearing performance but only if done correctly. What makes these results even more interesting is that the animations used were extremely basic. Other studies suggest that 5 unique frames per second is the bare minimum for visual cues to contribute to speech recognition [14] . Those results were obtained with the use of a video stream and the 5 unique frames did not necessarily include the visemes linked to each phoneme. By constructing the animation in such a way that all visemes are included, a significant increase in hearing performance can still be obtained with minimal effort. Note that these conclusions are only relevant under conditions where it is very difficult to hear. Under normal listening conditions the strong auditory cues will usually be enough to disambiguate any incongruent visual cues. These results do however show that users rely heavily on visual cues under adverse hearing conditions.

7. Conclusion

We expected even rudimentary lip animation to enhance speech perception in virtual environments. A significant improvement in hearing performance for the correctly animated condition over the non-animated condition was demonstrated, confirming the hypothesis. The animations used during this study were extremely basic, consisting of only ten unique frames. We have shown that even such simple animations significantly aid speech perception when correctly synchronized with matching auditory stimuli. The results of subject performance under incorrectly animated conditions show that under adverse hearing conditions, the visual modality is favoured. This implies that when using lip animation, the visemes and phonemes have to match or performance under adverse hearing conditions will be even worse than when no animation was used.

This could have implications for virtual environments with dialog in different languages. Creators of virtual environments do not have exact control over what the user will hear at any given time. Having separate animations for different languages therefore becomes more important for virtual environments than for other forms of media like cartoons or 3D animated films where there is more control over the final audio track.

Although one cannot edit the sound track beforehand, virtual environments do provide additional auditory cues not present in other forms of media. Directional cues can enhance the perception of speech in the midst of competing sounds if the target sound is presented from a different direction than the masking sound. 3D spatialized sound should therefore be used to present speech in virtual environments where the hardware platform supports it.

Acknowledgements

The research was supported by the Innovation Fund of the South African National Research Foundation. The authors would like to thank John Turest-Swartz and his staff from the Contemporary African Music and Arts Archive for providing a recording studio and allowing the use of their offices while conducting our experiments. The assistance in experimental design by Prof. Barbara Shinn-Cunningham, Antje Ihlefeld and Tim Streeter from Boston University is much appreciated.

References

[1] R.D. Begault. *3-D Sound for Virtual Reality and Multimedia. Cambridge, MA: Academic Press Professional. 1994.*

[2] P. Blair. Cartoon Animation. *Available online at* http://www.freetoon.com. 2005.

[3] A. Bregman. Auditory scene analysis. *Cambridge, MA: MIT Press.* 1990.

[4] D. Brungart. Evaluation of speech intelligibility with the coordinate response measure. *Journal of the Acoustical Society of America, 109, 2276-2279.* 2001.

[5] D. Brungart. Near-Field Virtual Audio Displays. *Presence: Teleoperators and Virtual Environments. Vol 11, No. 1, 93-106, February 2002.*

[6] D. Burford, E. Blake. Real-Time Facial Animation for Avatars in Collaborative Virtual Environments. *South African Telecommunications Networks and Applications Conference '99, 18-183. 1999.*

[7] T. Chen, R.R Rao. Audio-visual integration in multimodal communication. *Special Issue on Multimedia Signal Processing, 86, 837-852.* 1998.

[8] E.C. Cherry. Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustic Society of America, 25, 975–979.* 1953.

[9] D.J. Dekle, C.A. Fowler, M.G. Funnell. Audiovisual integration in the perception of real words. *Perception & Psychophysics 51, 355-362.* 1992.

[10] S. DiPaola, D. Collins. A Social Metaphor-based 3D Virtual Environment. In *Educators Program From the 30th Annual Conference on Computer Graphics and Interactive Techniques. 1-2.* 2003.

[11] R. Drullman, A.W. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural and three-dimensional auditory presentation. *Journal of the Acoustic Society of America, 107, 2224-2235.* 2000.

[12] M. Ebata. Spatial unmasking and attention related to the cocktail party problem. *Acoustics, Science and Technology, 24, 208-219.* 2003.

[13] R.L. Freyman, K.S. Helfer, D.D. McCall, R.K. Clifton The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustic Society of America, 106, 3578-3588.* 1999.

[14] H.W. Frowein, G.F. Smoorenberg, L. Pyters, D. Schinkel Improved Speech Recognition Through Videotelephony: Experiments with the Hard of Hearing. *IEEE Journal on Selected Areas in Communication, 9, 611-616.* 1991.

[15] Gardner B., Martin K. (1994). HRTF Measurements of a KEMAR Dummy-Head Microphone. *Available online at http://sound.media.mit.edu/KEMAR.html*

[16] K.W. Grant, P.F. Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustic Society of America, 108, 1197-1207.* 2000.

[17] K.P. Green, J.L. Miller. On the role of visual rate information in phonetic perception. *Perception & Psychophysics, 38, 269-276.* 1984.

[18] F Lavagetto. Converting speech into lip movements: A multimedia telephone for hard of hearing people. *IEEE Transactions on Rehabilitation Engineering, 3, 1-14.* 1995.

[19] H. Levitt. Transformed Up-Down Methods in Psychoacoustics. *Journal of the Acoustic Society of America, 49,* 467-476. 1971.

[20] F.J. McGuigan. Experimental Psychology. A Methodological Approach. *Prentice Hall Inc.* 1968.

[21] H. McGurk, J. MacDonald. Hearing lips and seeing voices. *Nature, 264, 746-748.* 1976.

[22] Microsoft Corp. (2004). *MS DirectX Programming Guide. Available at http://www.msdn.microsoft.com/directx/*

[23] T. Moore. Voice communication jamming research. *AGARD Conference Proceedings 331: Aural Communication in Aviation, 2:1–2:6.* 1981.

[24] B.G. Shinn-Cunningham, A. Ihlefeld. Selective and divided attention: Extracting information form simultaneous sound sources. *Proceedings of the 2004 International Conference on Auditory Display, 10, 51-59.* 2004.

[25] M. Strous. PAMELA – Lipsynch utility for Moho. *Available online at* http://www-personal.monash.edu.au/~myless/catnap/pamela/index.html.

[26] W.H Sumby, I. Pollack. Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustic Society of America, 26, 212-215.* 1954.