EG UK Theory and Practice of Computer Graphics (2010), pp. 1–4
John Collomosse, Ian Grimstead (Editors)

Evaluation of A Viseme-Driven Talking Head

166

Abstract

This paper introduces a three-dimensional virtual head for use in speech tutoring applications. The system achieves audiovisual speech synthesis using viseme-driven animation and a coarticulation model, to automatically generate speech from text. The talking head was evaluated using a modified rhyme test for intelligibility. The audiovisual speech animation was found to give higher intelligibility of isolated words than acoustic speech alone.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Three-Dimensional Graphics and Realism]: Animation

1. Introduction

A talking head has been developed with the goal of achieving realistic speech animation, and is being applied in a pronunciation training system. The aim is to create a pronunciation assistant to complement traditional methods and to assist the work of a human language tutor. Visual speech can be valuable in speech tutoring applications because vision benefits human speech perception, for three reasons as suggested by Summerfield [Sum87]: It helps speaker localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation [PNLM04]. This study aims to elucidate the benefits of visual speech in language learning.

Evaluation of the quality of the talking head's visual speech was carried out to ensure that it was suitable for the task of demonstrating pronunciation in a tutoring system. The talking head was evaluated using a similar approach to that of the LIPS2008 Visual Speech Synthesis Challenge [TFBE08], using subjective quality assessment in terms of intelligibility and naturalness. The intelligibility of the lip animation was evaluated in a word identification test. This was compared with audio speech alone, to determine the benefit of the visual modality, and compared with a real speaker to evaluate the realism of the talking head.

2. Visual Speech Synthesis

Approaches to visual speech can be viseme-driven or data-driven. In viseme-driven speech animation, each key pose is associated with a viseme, i.e. the position of the lips, jaw and tongue when producing a particular sound [LP87]. Data-driven approaches do not require pre-designed key shapes, but use a pre-recorded facial motion database for synthesis using machine learning or concatenation of sample data [DN07]. A key challenge in visual speech animation is that there is great variation in the realisation of visemes during the production of natural speech; this is termed coarticulation, which is the influence of surrounding visemes upon the current viseme. Current systems either explicitly take into account context when blending keyframes, or use a longer unit such as the diphone, which starts at the centre of one phone and ends at the centre of the next, so transitions between phones are preserved.

This talking head was implemented using viseme-driven speech animation. Visually-similar key poses were grouped into 15 visemes, and meshes were created using Facegen modelling software [Sin08] (Fig 1). Visemes for tongue positions were adapted from Lazalde's tongue models [LMM08] (Fig 2). The head was integrated into a GUI for a speech tutoring application, developed using the QT framework [QT09]. The talking head demonstrates how to pronounce sounds at phoneme, word, and sentence level, displaying the appropriate mouth movements, and displays a transverse cross-section though the head, showing the movement of internal parts such as the tongue during speech. Loquendo TTS [Loq08] was used to generate acoustic speech from text, and output phonetic labels and durations. This information was used to create an animation sequence by mapping each phoneme label to the corresponding viseme (Fig 3). Coarticulation was implemented based on Cohen and Massaro's model, using a dominance function to represent

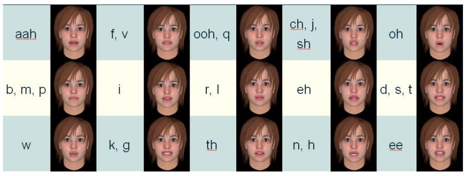
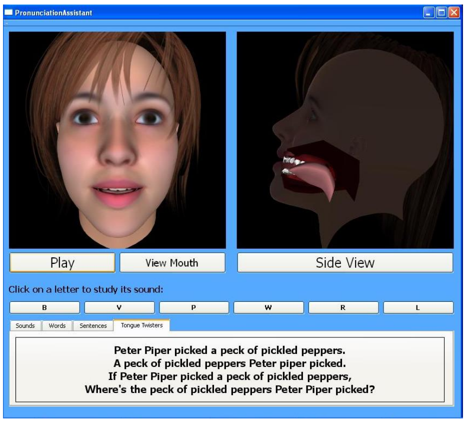


Figure 1: *Visemes*



Peter Piper picked a peck of pickled peppers.
A peck of pickled peppers Peter piper picked.
If Peter Piper picked a peck of pickled peppers,
Where's the peck of pickled peppers Peter Piper picked?

Figure 2: *Speech Tutoring Application*

the influence over time that a viseme has on a speech utterance [CM93]. The dominance functions of each segment were blended together to generate a speech trajectory. The coefficients of the dominance functions were set by observation, comparing the synthesized visual speech against video recordings of a real person saying the same words, until the synthesized speech looked like the recorded speech. For example, in the word "stew", the "u" segment has higher dominance than "s" and "t", and "u" has a low anticipatory rate which causes its domination to extend earlier in time, so the lip protrusion is seen earlier than the vowel is heard. The animation frames were compared against the video frames (Fig 5), and the coefficients were tuned to give the closest match that could be found by observation. The words used for tuning included each sound in initial and final positions (Fig 4). In order to reduce the computation time for the animation, Principal Component Analysis (PCA) was carried out. PCA reduces the dimensionality of the data by transforming it into uncorrelated variables, Principal Components (PCs), which capture the maximal variation in the data [Jol86]. A PCA program [LMM08] was run in Matlab to create PCs for 15 visemes. The talking head system applied dominance functions to the PCs, which were then reconstructed into meshes during the generation of frames for animation. Using PCA reduced the computation time because dominance functions were applied to only a small number of PCs instead of to every vertex of a mesh. Synchronisation between audio and video was achieved by using the audio playback loop to determine which frame to display at each time step.

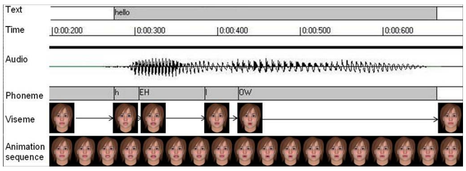
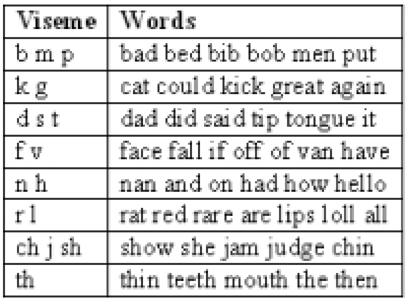


Figure 3: *Viseme-driven speech synchronised animation*



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

Figure 4: *Example words used in tuning visual speech*

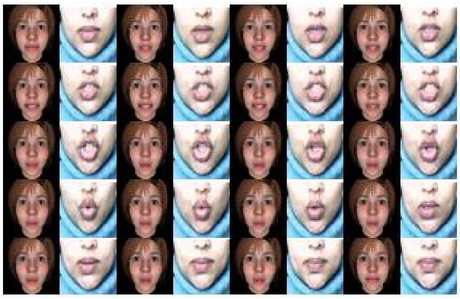


Figure 5: *Animation and Video Frames for "stew"*

3. Evaluation of Visual Speech

The intelligibility of the talking head was evaluated using a Modified Rhyme Test (MRT), an ANSI standard test for statistical intelligibility testing, which has previously been used to evaluate a talking head by Fagel [Fag08]. The MRT used 50 six-word lists of monosyllabic English words, and the words in each list differed only in the initial or final consonant sound; for example, "shop, mop, cop, top, hop, pop" [Mey10].

36 participants with normal hearing and vision were tested individually in an acoustically-isolated booth, with visual images presented on a 15 inch computer screen and acoustic stimuli presented binaurally over headphones. In each trial, participants were shown a six-word list and asked to identify

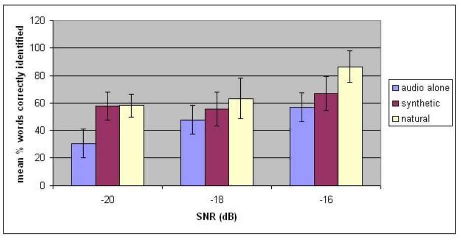


Figure 6: *Intelligibility scores. The error bars denote the standard deviation.*

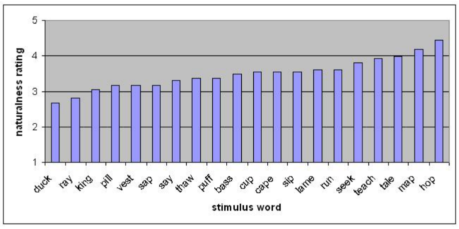


Figure 7: *Naturalness Rating*

which word was spoken. Responses were scored as the number of words identified correctly. 20 words were presented for each of 3 conditions: degraded synthetic audio speech alone; an external view of the talking head with degraded synthetic audio speech, and video of a real person with degraded audio. Different words were used for the 3 different conditions, in order to minimize learning effects. In order to minimize sequence effects, the order of presentation was randomized.

The audio was degraded by adding speech-shaped noise to the acoustic signal [Ass10]. The noise levels were chosen within a range in which the words were barely recognizable; below -20 dB word recognition for audio alone fell to chance levels (16%), while above -16 dB word recognition for natural video became close to optimal. For 16 participants, the SNR was set to -18 dB. For the remaining 16 participants, all words for all three conditions were presented at an SNR of -20 dB, and then repeated at -16 dB.

The naturalness of the talking head was evaluated using subjective quality assessment. After undertaking the intelligibility test at an SNR of -18 dB, 16 participants were presented with the synthetic talking head, for 20 isolated words with no audio degradation. The participants were asked to rate the naturalness of the visual speech along a 5 point Likert scale [Lik32], with a value of 1 for "very unnatural" and a value of 5 for "very natural".

3.1. Results

Visualization improved the intelligibility of the speech at all three SNRs (Fig 6). The word recognition rate was higher for the audiovisual heads than for audio alone, and higher for the natural head than the synthetic head. At the lowest SNR, -20 dB, the synthetic head was significantly more intelligible than audio alone, and the recognition rate for natural video was slightly higher than the synthetic head. At this SNR the improvement in word recognition due to the visualization in the audiovisual head, calculated using a normalized measure [SP54], was 39%; while the improvement due to the natural head was 40%. The visual contribution of the synthetic face relative to the natural face was not invariant as found by [OCIM07], but increased as the SNR decreased. The benefit of visual speech relative to audio alone increased as the SNR decreased, a finding consistent with that of Benoit [BLG98], who found that the poorer the auditory scores the greater the benefit of lip-reading. At a lower SNR the audio alone is less intelligible so listeners rely more on lip movements to decide which word was said. The naturalness scores for the synthetic talking head were, on average across all words and all participants, 3.5 on a scale of 1 to 5 (s.d. 1.0), so the visual speech was rated as moderately natural overall, but for some sounds the animation could be more realistic (Fig 7). The word which scored lowest, "duck", has little external mouth movement compared to "hop", which scored highest, so this may be a factor in the ratings for the animation. The confusion matrix for the synthetic head (Fig 8) compares the visemes presented against the visemes they were perceived as by the participants. For each word presented to participants, the actual animated viseme was plotted against each viseme it was perceived to be. The number of identifications were summed over all participants, for all words spoken by the synthetic talking head, at all SNRs. Each sum was divided by the number of occurences of the animated viseme, to give a percentage of identifications of that viseme. The area of each circle represents the percentage of identifications of that viseme. For example, viseme 6 (r/l) was mistaken for viseme 5 (h/n/ng) as often as it was identified correctly. The two visemes look similar from the outside, and it may be that the tongue movements for (r/l) were less accurately modelled. On the whole, the matrix shows that the correct classifications (on the diagonal) scored the highest, so overall the visemes were identifiable.

For the natural head, the confusion matrix shows that the visemes (h/n/ng) and (g/k) were less well identified than other visemes (Fig 9). This may be because the tongue movements that distinguish these visemes from others were less visible from the external view.

4. Conclusion

The talking head shows a gain in intelligibility compared to audio speech alone, and was almost as intelligible as the video of a real speaker. Certain visemes were confused with others, and could be modelled more accurately, but overall the visemes were identifiable. In the subjective naturalness tests, the visual speech was rated to be moderately natural

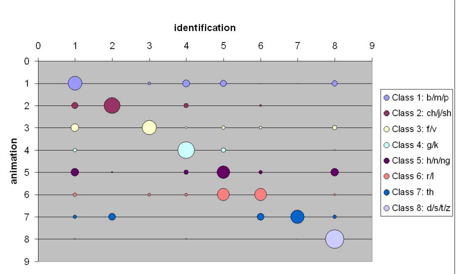


Figure 8: *Confusion Matrix for Synthetic Talking Head*

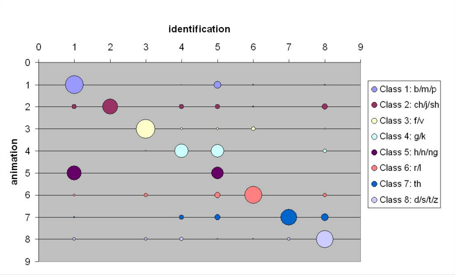


Figure 9: *Confusion Matrix for Natural Head*

overall. Thus the talking head was determined to be sufficiently realistic to be used to demonstrate pronunciation in a tutoring system.

The efficacy as a tutoring system is to be evaluated by user trials involving second language learners of English. The study aims to determine the benefit of visual speech in second language learning, and its effectiveness as a teaching tool for this application. Further studies will include a range of experiments to compare the effects of various aspects of the animation of the talking head; for example, the gain in intelligibility due to the coarticulation model will be investigated, in addition to the effect of more natural head movements and facial expressions.

5. Acknowledgements

This work is sponsored by the Economic and Social Research Council and the Engineering and Physical Sciences Research Council.

References

[Ass10] Assmann P.: Speech Perception Lab, 2010. http://www.utdallas.edu/~assmann/hcs7367. 3

[BLG98] Benoit C., Le Goff B.: Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication 26*, 1-2 (1998), 117–129. 3

[CM93] Cohen M. M., Massaro. D. W.: Modelling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, Thalmann N. M., Thalmann D., (Eds.). Springer-Verlag, 1993, pp. 139–156. 2

[DN07] Deng Z., Noh J.: Computer facial animation: A survey. In *Data-Driven 3D Facial Animation*, Z.Deng, U.Neumann, (Eds.). Springer-Verlag, 2007, pp. 1–28. Chapter 1. 1

[Fag08] Fagel S.: Massy speaks English: Adaptation and evaluation of a talking head. In *Interspeech* (2008). 2

[Jol86] Jolliffe I.: *Principal Component Analysis*. Springer Verlag, 1986. 2

[Lik32] Likert R.: A technique for the measurement of attitudes. *Archives of Psychology 140*, 1 (1932), 55. 3

[LMM08] Lazalde O. M., Maddock S., Meredith M.: A constraint-based approach to visual speech for a Mexican-Spanish talking head. *International Journal of Computer Games Technology*, 3 (2008). 1, 2

[Loq08] Loquendo: Loquendo, 2008. http://www.loquendo.com. 1

[LP87] Lewis J. P., Parke F. I.: Automated lip-synch and speech synthesis for character animation. *SIGCHI Bull. 17* (1987), 143–147. 1

[Mey10] Meyer Sound: Speech intelligibility papers, 2010. http://www.meyersound.com/support/papers/speech/mrt.htm. 2

[OCIM07] Ouni S., Cohen M. M., Ishak H., Massaro D. W.: Visual contribution to speech perception: measuring the intelligibility of animated talking heads. *EURASIP J. Audio Speech Music Process.*, 1 (2007). 3

[PNLM04] Potamianos G., Neti C., Luettin J., Matthews I.: Audio-visual automatic speech recognition: An overview. In *Issues in Visual and Audio-Visual Speech Processing*, Bailly G., Vatikiotis-Bateson E., Perrier P., (Eds.). MIT Press, 2004. 1

[QT09] QT: Qt software, 2009. http://www.qtsoftware.com. 1

[Sin08] Singular Inversions: Facegen, 2008. http://www.facegen.com. 1

[SP54] Sumby W., Pollack I.: Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America 26*, 2 (1954), 212–215. 3

[Sum87] Summerfield A.: Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by Eye: The Psychology of Lip-Reading*, Dodd, B., Campbell R., (Eds.). Lawrence Erlbaum Associates, 1987, pp. 3–51. 1

[TFBE08] Theobald B.-J., Fagel S., Bailly G., Elisei F.: LIPS2008: Visual speech synthesis challenge. In *Interspeech* (2008), pp. 2310–2313. 1