VISUAL SPEECH INTELLIGIBILITY OF DIGITALLY PROCESSED FACIAL IMAGES

N. Michael Brooke and Paul D. Templeton

School of Mathematical Sciences, University of Bath,
Claverton Down, BATH, Avon BA2 7AY, United Kingdom

1. INTRODUCTION

The visual cues conveyed by the movements of a speaker's face can play a role in the perception and recognition of speech which becomes especially important where there is noise or hearing-impairment. Visible speech articulations may be exploited for automatic speech synthesis and recognition in much the same way that acoustical speech signals are. The automatic synthesis of visible speech articulations has already been used to provide optical stimuli for analytical investigations of visual and audio-visual speech perception [7,12]. Studies like these could lead in the longer term to a better understanding and assessment of lip-reading skills and hence to improved rehabilitation of the hearing-impaired. Visual syntheses could also be used to create simultaneous audio and visual speech transmission over band-limited telecommunications networks and to create improved film animations. Visual speech synthesis and its applications have been more fully reviewed elsewhere [1,2]. Automatic visual speech recognition may be used to enhance the performance of conventional acoustic speech recognizers by permitting augmentation of their inputs [e.g. 9,10]. In both synthesis and recognition, a major objective is to estimate the minimum degree of image detail which must be generated or captured in order to convey the essential perceptual cues whilst enabling processing to be as efficient as possible. A number of studies have been carried out using film or video recordings of speakers' faces in an attempt to identify visual cues to the identity of phonetic events. Generally, they have been confined to attempts to find correlations between physical features of the visual stimuli and the perceptual confusions [e.g. 6,8,11]. For reasons which have been given elsewhere, attention has usually focused on the vowels [7]. The availability of digital image-processing techniques and hardware has now made it possible to study more complex and detailed features of visible speech movements of the oral region and in particular, to manipulate complete oral images directly. This paper suggests how image-processing methods may be exploited by describing initial experiments designed to examine how visual vowel intelligibility varies with the information content of oral images whose resolution has been degraded. Whilst these experiments may not themselves necessarily identify the nature of the essential perceptual cues, they may indicate a lower limit beneath which the resolution of the images is insufficient to retain them. Such an estimate would form a first step in estimating the minimum degree of image detail necessary in order to carry out useful visual synthesis and recognition.

2. EXPERIMENTAL PROCEDURES AND RESULTS

Investigation of the variation in visual vowel intelligibility with image resolution was approached: a) by a perceptual vowel identification experiment, and b) by using a self-organising pattern-classification technique to carry out automatic vowel recognition.

2.1. Visual vowel perception using reduced-resolution images
A prototypical experiment was designed and implemented to examine the relationship between visual speech intelligibility and the information content of optical stimuli which consisted of dynamic monochrome displays of the oral region of a speaker's face. A GEMS high-capacity digital framestore controlled by a DEC VAX computer was used to capture, manipulate and display sequences of images of a talker's face. Within the framestore, each image was effectively represented as a two-dimensional array of pixels whose values defined the intensity of the component points of the image. A monochrome television camera was used to record a talker's face and the images were sequentially transferred to the digital framestore at the standard cinematographic rate of 25 images per second. Display of images held in the framestore was carried out by transferring them in sequence to a high-resolution graphics screen or videotape recorder at 25 images per second, thus creating realtime animation. Image manipulation was carried out by selecting and running one from a potentially expansible range of routines, each of which altered the framestore contents in some defined way. Processing options were provided only for altering the resolution of images in the spatial or intensity domains. Raw or image-processed digital image data could also be transferred directly between the framestore and files of computer storage. An integrated package of computer programmes written in Pascal was implemented to manage the handling of the framestore; different combinations of operations could be carried out within a single run by responding to prompts with one from a specified range of options. A full description of the programme package has been given elsewhere [3].

2.1.1. Stimulus generation. Monochrome frontal images of the oral region of a native British-English speaker enunciating neutrally-stressed /hVd/ utterances (V = non-diphthongal British-English vowels) were recorded. The recordings began and finished with the face in a resting state so that the resultant displays (see below) embodied a short adaptation period. The speaker's head was lightly restrained so that it remained in a constant position. The camera was set level with the oral region and the object distance and focal length of the camera were adjusted so that the oral region filled the frame. The speaker's face was lit by a 1 Kwatt quartz-halogen light placed as close as possible to the camera, so as to produce flat illumination. The aperture of the camera was set to produce optimal contrast. Each recorded utterance was stored within the framestore as a sequence of 64 images corresponding to a duration of about 2.5 seconds. Each image occupied 128 x 128 pixels and each pixel could resolve 256 levels of intensity. The acoustical quality of the utterances was monitored and the image sequence corresponding to one well-formed token of each of the five long vowels (i.e. 'hard', 'heard', 'heed', 'hoard' and 'who'd') was saved on file for further processing. Each of the five saved image sequences was then reloaded into the framestore and image-processed to create seven further image sequences with images of constant size, but reduced spatial and intensity resolutions. The generated image sequences were also stored as files. Table 1 shows the eight display conditions which were thus available for each of the five /hVd/ utterances. A randomised list of 400 test stimuli was generated. These consisted of ten tokens of each of the five utterances in each of the eight display conditions. The stimuli were created by sequentially loading the appropriate stored image sequences into the framestore according to the list and redisplaying them in realtime. The displays were generated at the rate of one every nine seconds. The output of the framestore was simultaneously videorecorded on a Sony U-matics cassette recorder in PAL (625-line) standard. The videorecordings were split into four blocks of 100 stimuli, each block taking about 15 minutes to replay.

Table 1: Display conditions for the visual /hVd/ utterances

Condition	Spatial resolution (number of independent pixels)	Intensity resolution (number of intensity levels)
1	128 x 128	256
2	64 x 64	256
3	32 x 32	256
4	16 x 16	256
5	8 x 8	256
6	128 x 128	16
7	128 x 128	4
8	128 x 128	2

2.1.2. Visual perception experiments and results. The videorecorded stimuli were replayed on a high-quality monochrome television monitor and subjects seated at a comfortable viewing distance were asked to identify which of the five long vowels was being presented in each of the 400 /hVd/ stimuli. No practise stimuli were presented. The four blocks of stimuli were presented with short rest periods (10 minutes) between blocks. Pilot experiments have been carried out with three normal-sighted, normal-hearing native British-English speaking subjects. All were experienced in carrying out visual speech perception experiments. The figure presents the results.

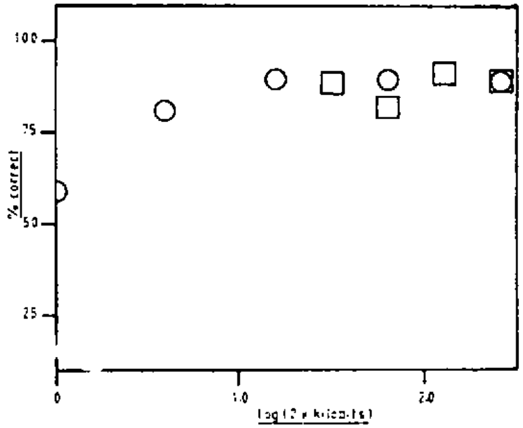

Figure: Results of the visual vowel identification experiment, showing the percentage correct recognition rates over all vowels for all three subjects, plotted against twice the logarithm of the number of kilobits needed to represent each frame of the displayed image sequences. Squares represent varying intensity resolutions and circles varying spatial resolutions.

All three subjects showed broadly similar variations in performance with varying image resolutions; whilst reduction in the number of intensity levels down to two (i.e. pure black and white images) resulted in little deterioration in recognition performance, the degradation of spatial resolution below 32 x 32 pixels resulted in a reduction in performance which became particularly marked when the resolution fell below 16 x 16 pixels. The vowels at the corners of the vowel triangle (/a/, /i/ and /u/) which involved the most extreme horizontal and vertical lip articulations were less frequently misidentified than the two more central vowels, even at the lower resolutions. The central vowels are inherently more difficult to identify [7]. At the lighting level used in the recordings, reduction in the intensity resolution to two levels produced images in which only the oral cavity could be clearly distinguished. Lip separation and mouth width were therefore clearly visible even at this low intensity resolution, as they were with moderate, though not with low, spatial resolutions. At four levels of intensity resolution, some shadowing also appeared under the lower lip when the lip was protruded which may have assisted the discrimination between the rounded and unrounded vowels.

2.2. Vowel recognition experiments using multi-layer perceptrons (MLPs)
In order to investigate the intrinsic information content of images of the oral region and the degree to which they were capable of being used to extract cues to vowel identification, a multi-layer perceptron (or MLP) model [4] was applied to a prototypical visual vowel recognition experiment. MLP models can be trained to discriminate between a set of events by iteratively computing an optimal mapping from a set of pattern vectors, such as oral images, to the chosen classification space. The mapping can be computed without supplying any additional *a priori* information about the nature of the events.

2.2.1. Generation of input patterns for the MLP models. Fifty tokens of each of the eleven non-diphthongal British-English vowels enunciated by each of three native British-English speakers in a neutrally-stressed /bVb/ context were videorecorded under conditions similar to those described in section 2.1.1. A short-exposure, rotary-shuttered monochrome camera was used to capture clear images. A single frame from the vowel nucleus of each of the recorded tokens was captured and digitised using a framestore with an intensity resolution of 64 levels per pixel. This set of digitised images was stored on computer files. The stored frames were a) standardised to a mid-grey average intensity, b) compressed to 16 x 12 pixels and c) contrast-enhanced by linear expansion of the intensity levels in the central 5/8 of the intensity range to the full intensity range. The spatial resolution was selected, by reference to the visual perception experiments described in section 2.1., as lying somewhere near the lowest useful limit in terms of the retention of useful visible cues. A second set of images, identical except for being quantized to four intensity levels following the processing described above, was generated from the first set of images and stored.

2.2.2. Training and vowel recognition using MLP models. MLP models, with 192 input units (one for each image pixel), 11 output units (one for each of the vowel classes) and 6 'hidden' units in an intermediate layer, were used for the visual vowel recognition experiments. Six intermediate units were found by experiment to give optimal recognition performance; addition of more intermediate units did not improve recognition performance significantly (see below). Experiments were carried out in which MLPs were a) trained and tested on tokens from the same speaker, i.e. speaker-dependent, and b) trained on tokens from three speakers and tested on tokens from any one of the three speakers, i.e. multiple-speaker. In each condition, images quantized to 64 and 4 intensity levels were applied in separate experiments to train and test the MLPs and hence to examine the effects on recognition performance of the intensity resolution of the images. Forty tokens of each vowel supplied by each speaker were used to train the MLPs and ten further tokens were used to test the trained MLPs. Supervised training of the MLPs was carried out by supplying the training tokens together with their vowel class labels. The MLPs were each trained until 99% or more of all training tokens and 95% or more of the tokens of any particular vowel class evoked the correct vowel class label. The recognition experiments were carried out by supplying the trained MLPs with the unlabelled test tokens and recording the vowel classification output by the MLPs. The results are presented as confusion matrices in Tables 2 and 3. The results include ambiguous recognitions, that is, cases where more than one output of the trained MLP was active and exclude unidentified test tokens, where no MLP output was active.

Table 2: Speaker-dependent identification rates (%): 5 MLPs for each of 3 speakers, each trained on 40 tokens of each vowel and tested on 10 tokens of each vowel.
Italic figures are results with 4 levels of intensity quantization.

Stimulus	Response										
	i	a	ɔ	u	ɜ	I	e	æ	ʌ	o	w
i	96 *92*		*1*			2 *5*	1 *1*	*1*			
a		92 *90*			*3*				7 *7*	*3*	
ɔ	*1*		91 *89*	4 *3*						1	2 *6*
u			1 *3*	94 *86*	*1*	1	*1*			1	2 *5*
ɜ		2 *4*			95 *93*		1		1		1 *1*
I	3 *2*				3	87 *93*	7 *3*		*2*		
e	1 *2*					3 *3*	89 *81*	5 *11*	*2*		
æ							7 *12*	92 *86*	1 *3*		
ʌ		6 *6*						1 *1*	90 *89*		--
o	1	1 *3*	*1*	*1*	1 *1*					95 *91*	1 *3*
w		1	2 *5*	2 *7*	2					5 *6*	84 *72*

The overall results are summarised in Table 4. The good recognition rates for the vowels in the speaker-dependent case suggest that there is sufficient information in the mouth images at the vowel nuclei to provide visual cues to their identity. When the intensity resolution was reduced to four levels, the degradation in performance was relatively small. The multiple-speaker experiments give a slightly poorer recognition performance than the speaker-dependent tests, but nonetheless suggest that the MLPs were able to generalise successfully over the variations in the mouth gestures of three speakers [8]. Reduction in the intensity resolution for the multiple-speaker tests resulted in approximately the same deterioration in performance as in the speaker-dependent case. The vowel confusions were similar in all cases.

Table 3: Multiple speaker identification rates (%): 5 MLPs each trained on 40 tokens of each vowel from each of 3 speakers and tested on 10 tokens of each vowel from each speaker.
Italic figures are results with 4 levels of intensity quantization.

Stimulus	Response										
	i	a	ɔ	u	ɜ	I	e	æ	ʌ	o	w
i	96 *91*					3 *6*	*3*				
a		88 *87*							9 *13*		
ɔ			89 *84*	7 *3*		*1*				1	3 *10*
u	*1*		3 *3*	94 *86*	1	*1*					2 *6*
ɜ		1 *3*		*1*	91 *83*	4 *1*	*1*		1 *1*	1 *1*	*3*
I	6 *9*				1 *3*	87 *78*	5 *5*				
e	1 *3*				1 *3*	5 *3*	81 *69*	11 *13*	1 *3*		
æ	1						9 *12*	85 *86*	2 *2*		
ʌ		9 *12*					*2*	3 *2*	87 *80*		
o		1 *1*	*1*		1 *3*					94 *91*	1 *3*
w			6 *5*	7 *9*	1 *1*					4 *5*	80 *77*

3. DISCUSSION

The preliminary results of the perceptual experiment appear to suggest that oral images with a spatial resolution of 16 x 16 pixels or an intensity resolution of just 2 levels may be adequate to capture many of the essential visual cues to the identification of the vowels tested. However, the experiment was severely constrained: a) only the five long vowels of British English were tested and three of those lay at the corners of the vowel triangle which represent articulatorily extreme positions; b) the vowels were tested using isolated syllable utterances with a fixed phonetic context so that no coarticulatory variations were involved in the productions; c) only three, non-naive subjects have so far been tested; d) the test tokens were produced by a single speaker and different speakers are known to generate different mouth movements [8]. In addition, consonant identification was not tested [7]. It may therefore be inappropriate at present to assert that these resolutions represent a reliable estimate of the lower limit below which images of vowel productions cannot embody the essential visual cues, or to generalise from these results. For example, it is possible that the visually less distinctive vowels may require a higher resolution for recognition. Also, a full analysis would need to take into account the inherent uncertainty in visually identifying particular vowels, even under the best viewing conditions. The preliminary experiments, however, indicate the feasibility of carrying out a more comprehensive perceptual experiment which would overcome at least some of these limitations, for example, by using suitably constructed word or sentence lists, such as the FAAF lists [5]. These embody differing phonetic contexts and can employ continuous utterances, since the test words are usually embedded in carrier sentences.

Table 4: Summary of MLP visual vowel recognition results

Image resolution (intensity levels)	Correct identification rate (%)			
	Speaker-dependent		Multiple speaker	
	Overall	Poorest vowel	Overall	Poorest vowel
64	91	84	88	80
4	87	72	82	69

The results of the pilot experiments with MLP recognizers are not directly comparable with those of the perceptual experiment because: a) a larger set of eleven vowels in a fixed /bVb/ phonetic context was used; and b) single frame images were selected so that no dynamic cues were available, but the frames were taken from the vowel nuclei. Nonetheless, the results indicate that good visual vowel recognition, comparable with that demonstrated by human subjects viewing videorecorded faces uttering /bVb/ syllables [7], can be obtained, even with intensity resolutions as low as four intensity levels and spatial resolutions of only 12 x 16 pixels. This in turn suggests that essential visual cues to vowel identity in a fixed phonetic context can be embodied in single, nuclear images of low resolution. The degradation in performance in changing from 64 to four levels of intensity resolution is also consistent with that observed in the perceptual experiment, but results at other intensity resolutions are not at present available. The visual vowel recognition results obtained from the multiple speaker MLP experiments also indicate that MLPs can successfully generalise, at least to a limited extent, by learning how to map from the image space to the classification space when the training images are provided by more than one speaker. Although the mappings do not reveal the nature of the visual cues to vowel recognition, they do suggest that at least some of the underlying cues common to several speakers may be represented at low spatial and intensity resolutions.

Reducing the spatial and intensity resolutions is only one way of achieving compression of image data, which is a major objective in visual speech synthesis and recognition. The use of MLPs suggests a different and highly efficient method, since the activations of the small number of units in the hidden layer of trained MLPs themselves represent a coded version of an image [4]. A form of training for MLPs involving an identity mapping, which generates an output identical to the input, could therefore be envisaged as a means of teaching MLPs how to code images of talkers' faces. The trained MLPs could then be used either to code very compactly the captured images of talkers in the case of speech recognition. Conversely, they could be used to generate images for visual syntheses from a small set of driving data.

4. ACKNOWLEDGMENTS

This work was carried out a) under a research grant (GR/E09450) from the Science and Engineering Research Council and b) during a Short-term Research Fellowship by NMB at the British Telecom Research Laboratories at Martlesham Heath, to both of whom the authors express their thanks.

5. REFERENCES

[1] N M BROOKE, 'Computer graphics animations of speech production', Computing and the Humanities, **1** (in press)

[2] N M BROOKE, 'Computer graphics synthesis of talking faces', in Proceedings of the ESCA Tutorial and Research Workshop on Speech Synthesis (Autrans, France) (to appear)

[3] N.M. BROOKE, 'Intelligibility of visible speech cues with oral images of reduced information content', BTRL Short-term Fellowship Report, pp.33 (1989)

[4] J L ELMAN & D ZIPSER, 'Learning the hidden structure of speech', Journal of the Acoustical Society of America, **83**, 1615-1626 (1988)

[5] J R FOSTER & M P HAGGARD, 'Four-alternative auditory feature test (FAAF) - linguistic and psychometric properties of the material with normative data in noise', British Journal of Audiology, **21**, 165-174 (1987)

[6] V FROMKIN, 'Lip positions in American English vowels', Language and Speech, **7**, 215-225 (1964)

[7] M McGRATH, A Q SUMMERFIELD & N M BROOKE, 'Roles of lips and teeth in lipreading vowels', Proceedings of the Institute of Acoustics (Autumn Conference, Windermere), **6(4)**, 401-408 (Institute of Acoustics, Edinburgh, 1984)

[8] A A MONTGOMERY & P L JACKSON, 'Physical characteristics of the lips underlying vowel lipreading performance', Journal of the Acoustical Society of America, **73**, 2134-2144 (1983)

[9] E D PETAJAN, 'Automatic lipreading to enhance speech recognition', Proceedings of the Global Telecommunications Conference (Atlanta, Georgia), 265-272 (IEEE Communication Society, 1984)

[10] E D PETAJAN, N M BROOKE, B J BISCHOFF & D A BODOFF, 'Experiments in automatic visual speech recognition', Proceedings of the 7th. Symposium of the Federation of Acoustical Societies of Europe (FASE), 1163-1170 (Institute of Acoustics, Edinburgh, 1988)

[11] G L PLANT, 'Visual identification of Australian vowels and diphthongs', Australian Journal of Audiology, **2(2)**, 83-91 (1980)

[12] Q SUMMERFIELD, A MACLEOD, M McGRATH & M BROOKE, 'Lips, teeth and the benefits of lipreading', in A.W. Young and H.D. Ellis (Eds.), 'Handbook of Research on Face Processing', 223-233 (North-Holland, Amsterdam, 1989)