LIPS2008: Visual Speech Synthesis Challenge

Barry-John Theobald[1]*, Sascha Fagel*[2]*, Gérard Bailly*[3]*, and Frédéric Elisei*[3]

[1]School of Computing Sciences, University of East Anglia, Norwich, UK.
[2]Institute for Speech and Communication, Technical University Berlin, Germany.
[3]GIPSA-lab, Department of Speech and Cognition, University of Grenoble, Grenoble, France.

bjt@cmp.uea.ac.uk, sascha.fagel@tu-berlin.de
gerard.bailly@gipsa-lab.inpg.fr, frederic.elisei@gipsa-lab.inpg.fr

Abstract

In this paper we present an overview of LIPS2008: Visual Speech Synthesis Challenge. The aim of this challenge is to bring together researchers in the field of visual speech synthesis to firstly evaluate their systems within a common framework, and secondly to identify the needs of the wider community in terms of evaluation. In doing so we hope to better understand the differences between the various approaches and to identify the strengths/weaknesses of the competing approaches. In this paper we firstly motivate the need for the challenge, before describing the capture and preparation of the training data, the evaluation framework, and conclude with an outline of possible directions for standardising the evaluation of talking heads.
Index Terms: visual speech synthesis, evaluation

1. Introduction

Visual speech synthesisers have potential for use in a wide range of applications — see [1, 2] for an overview of audio-visual speech synthesis. These applications range from desktop agents on personal computers and characters in computer games, to language translation tools and providing a means for generating and displaying stimuli in speech perception experiments. What has perhaps most prevented widespread use of such systems is that there is no easy way of comparing the overall quality of one system against another allowing next generation synthesisers to built on the strengths of the current state-of-the-art. There are a number of issues: 1) The structure of the data used to train different systems will likely vary. For example, it may differ in the adopted language, the number of sentences, the phonetic makeup of the sentences (balanced for diphone or triphone coverage). 2) A different talker is used during data capture, and some may be better than others in terms of clarity of articulation, speaking style, and so on. 3) The test data to be synthesised and the test methodology differ between various studies: there is no common evaluation scheme. 4) The presentation of the face varies for different systems. Some strive for videorealism, whilst others adopt graphics-based approaches that attempt only to appear human-like. The *uncanny valley* effect [3] suggests the perceived quality of systems with the same underlying speech model can differ significantly. 5) Different evaluation metrics are used to judge the performance. Some systems are evaluated only objectively (e.g. RMS error in geometric features, or articulation parameters), some are also evaluated subjectively (e.g., naturalness, or intelligibility, etc.). 6) Viewing conditions are likely to be different between sites. Some adopt strict viewing conditions with controlled lighting and soundproofed listening conditions, whilst others are less formal.

Recently, researchers in auditory speech synthesis have sought to overcome these issues by running evaluation *en masse* in the form of a competition: a notable example is the *Blizzard Challenge* [4]. These competitions have proven to be popular and provide an excellent framework for unifying and standardising evaluation between research groups. The goal of LIPS 2008 is to adopt this idea for evaluating visual synthesisers and overcome many of the problems associated with comparing systems by providing entrants with the **same training data**. Likewise, all evaluation will be conducted **independently** of any competing research group, using the **same viewers** rating the **same test utterances** using the **same evaluation metrics**. For this purpose a common training corpus is to be captured, labelled phonetically and made available to all challenge entrants.

2. Background

Ultimately, synthesised talking faces require subjective evaluation. Objective measures of performance provide only a *guide* as to the quality of the synthesised output. Comparing synthesised and ground-truth parameters (i.e., measured from a real person speaking) using, for example, correlation, as many systems do, is perhaps an unfair test. A synthesiser cannot be expected to generate a parameter sequence *exactly* as it appears in the ground-truth sequences. Indeed, if a synthesiser were to behave in this manner it would likely be deemed unnatural as the output is entirely predictable and would quickly be perceived as synthetic. Speech produced by a human exhibits natural variation — we never say the same thing in exactly the same way. The question is then are the differences between ground-truth and synthesised sequences perceived as errors, or are they imperceptible and can be attributed to variation observed in natural speech? This cannot be quantified using purely objective measures.

An obvious method for evaluating synthesised visual speech subjectively is to adopt a Turing-type test, where viewers are asked to watch real and synthesised sequences and distinguish those that are real from those that are synthesised [5]. However, the underlying assumption of this form of test is that the goal is to generate *videorealistic* sequences — sequences that are indistinguishable from real video. This is not always the case, and graphics-based systems, e.g. [6], strive only for a realistic model of speech, not a realistic sequence of images. It is important the testing methodology adopted in this challenge is sufficiently general that it can be applied to the full spectrum of systems likely to be entered.

The quality of synthesised visual speech also can be assessed indirectly by having users not make judgements regard-

ing the realism, but instead have users interact with a system [7]. The ease with which they interact with a virtual environment, for example the ability of the system to draw attention using instructions spoken using synthesised audiovisual speech, directing user attention using gestures and gaze, or expressing affect through facial expressions when speaking, can be assessed. The rationale is that receiving output from a system in the form of realistic audiovisual speech is more natural and thus reduces the cognitive load on the user. Cognitive load can be measured in terms of the time required to respond to some instruction, for example typing a number sequence [7] . However, this form of test is again not directly applicable here as the cognitive load will vary between listeners as a function of English speaking ability. Likewise, systems to date tend to focus on the problem of generating *raw*, expressionless speech (i.e., spoken without emotion), so are not concerned with a specific application.

Cosker and colleagues [8] measure the performance of their visual synthesiser using experiments based on the McGurk Effect [10]. It is well known that viewers presented with incongruent audio and visual information perceive neither what was heard or seen. The speech perception system finds the *best fit* for the conflicting audio and visual information. The goal in [8] is to present monosyllabic words in both synthesised and real video conditions, with congruent and incongruent audiovisual stimuli. Viewers were asked to transcribe what they perceived and the strength of the McGurk Effect compared for the real and synthesised stimuli. If the synthesiser generates inaccurate speech gestures, the expected response (predicted from McGurk) will not match the given response — using real video provided a sanity check for the expected responses. The limitation of these experiments is they are concerned only with the short-term aspects of the synthesised speech. Only isolated monosyllabic words are used. This tells us nothing about longer-term dynamics that relate to the overall sense of naturalness. Synthesisers may look realistic and entirely plausible over single words, but longer term coarticulation effects, which are ultimately one of the most difficult aspects of visual speech to capture, may be less well captured.

Most evaluation of visual speech synthesisers consider an *all-in-one* evaluation strategy. The synthesiser generates synthesised video sequences and these are evaluated in some way — i.e. all components of the system are considered simulataneously. A more rigourous evaluation might consider each component of a synthesiser in isolation. For example, the point-light method in [9] separates the underlying speech model from the appearance of the face. This might, for example, help overcome bias with viewer expectations of the way a model is expected to *speak* given the way it looks. The obvious benefit is inadequacies (resp. strengths) of the system are easy to pinpoint. The downside of this form of evaluation is a dense coverage of point-lights scattered about the face is required to accurately capture subtle movements of the lips, jaw and cheeks. This form of evaluation will be a consideration for future challenges.

The evaluation methodology adopted for this challenge is to evaluate systems both in terms of intelligibility [6, 9, 11, 12] and in terms of naturalness [13, 14]. The main difference in the intelligibility tests adopted here is that our evaluation is conducted over sentence-level utterances rather than monosyllabic (VCV) words, and the same test utterances and same viewers will evaluate competing systems. The goal is to identify the strengths and weaknesses of the various synthesis approaches so the field as whole can progress. Automatic speech recognition (ASR), and to an extent auditory speech synthesis, have benefited from a common evaluation paradigm.

3. Data Capture and Preparation

The training data for the corpus is comprised of a single speaker recorded in full-frontal view reciting the phonetically-balanced Messiah sentences, see [15] for a list of the sentences. The sentences were spoken in a neutral speaking style (no expression) and the lighting was adjusted to ensure the visible articulators were clear in the video. They were recorded in three successive batches, representing almost an hour of recording, of which ten minutes contains useful speech. The test data is comprised of 50 semantically unpredictable sentences (SUS) [16] recorded in the same recording conditions as the training data. For evaluating synthesisers in terms of intelligibility, it is important to remove as many non-linguistic clues as possible (e.g., facial expressions, body gestures, and more importantly context) from the stimuli as these all play a role in the speech-reading process in everyday communication. Lip-readers, even experts, use much more than just information from the lips. For example, the missing word in the sentence — "The ******* from a freshly squeezed orange makes a nice drink" — can be guessed without either hearing the audio, or seeing the face.

The video sequences were recorded using an analogue PAL camera rotated to capture in landscape orientation to increase the footprint of the face of the subject in the captured images. The 25Hz interlaced video (two consecutives fields with 720x288 pixels) was post-processed to rotate and de-interlace the images (providing a 50Hz video stream with 576x720 pixels). For easier access to the challenge data, the video will be distributed as individual JPEG images at (90% quality, exported using 'convert' from the ImageMagick toolkit [17]). Example images are shown in Figure 2. The acoustic speech signal was captured using a boom-microphone near the subject, but positioned so as not to obscure the face in the video. The acoustic speech for each utterance is stored in individual RIFF files, with 16-bits/sample and a sampling frequency of 44.1KHz. The total size of the training corpus (zipped video frames by utterance, wav files and segmentation info) is approximately 4.24GB.

Crude phonetic transcriptions of the training corpus were created using HTK in forced-alignment mode, where a person-specific HMM was trained and used to label the data. The crude labels were refined by hand-correcting the transcriptions. This hand-correction was conducted between the three sites with which the co-authors are affiliated.

4. Evaluation

All systems entered into challenge must undergo the same tests and use the same evaluation criteria. The evaluation is designed to measure both the intelligibility and the perceived naturalness of the talking faces. Entrants will generate the test stimuli in a supervised room prior to the evaluation sessions (on the tutorial day of the conference). This is mainly to ensure no handcrafted stimuli are entered into the evaluation process, but also to provide support where required. Entrants will be provided with technical details regarding the format (file type and codec settings, etc.) of the output video they are to provide. This is to ensure both equal technical quality between systems, and the video coding is compatible with the interface used in the tests.

The test sentences that will be provided to the entrants (only upon arrival in the supervised room) will be the acoustic speech and the (hand-corrected) phonetic labels aligned to the audio. The same annotation used in the training corpus will be used in the test corpus. Participants are free to use one or more of the representations of the test utterances (audio, labels, durations,

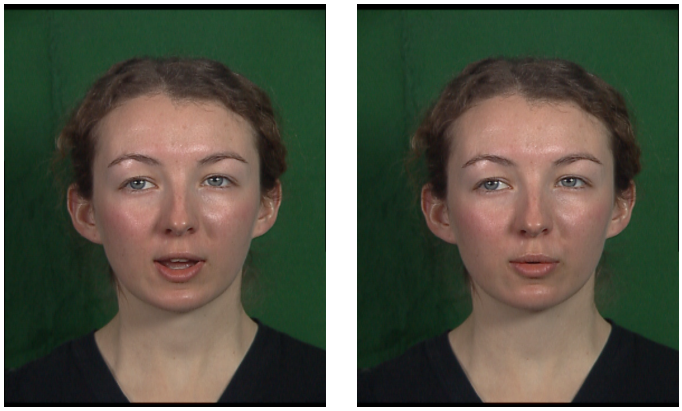

Figure 1: Example images extracted from the training corpus.

text) to generate their synthesised video.

Viewers used to evaluate the systems will be recruited from the conference delegates — the challenge itself will be advertised at the conference to attract participants. The only restrictions on participants are they must firstly have some ability to speak English and they must have normal hearing and normal, or corrected to normal vision. Participants will be asked to rate themselves as native, fluent, proficient, moderate, or poor in English, as the level of English will undoubtedly be a factor in the results. This is especially the case for the intelligibility tests. Noting the level of English of each viewer will allow this to be considered in the analysis.

Systems will be ranked by both their intelligibility and their naturalness scores, so overall there are potentially two *winners*. The intelligibility and naturalness scores will not be combined as different applications may have different demands on the synthesis system. We wish to investigate and highlight the advantages of the various approaches entered into the challenge. The trade-off between intelligibility and naturalness and any minimal requirements regarding both will be discussed by the scientific committee.

4.1. Intelligibility Tests

The acoustic speech waveforms to which the talking faces are to be synchronised will be degraded to a signal-to-noise ratio (SNR) of -10 dB and re-combined with the videos. The reason for supplying degraded audio is that visual-only lip-reading on semantically unpredictable sentences is almost impossible, even for lip-readers [18] and in the case of clean audio the intelligibility is close to optimum either with or without video. For these tests the SNR was adjusted so that the audio alone was barely audible to avoid both ceiling effects. Visual information is known to be especially helpful in environments with degraded audio [19, 20]. Furthermore, combining audible and visible speech reveals information that is imperceptible in both the auditory and visual modality alone [21, 22].

The audio-visual signals will be played back to viewers using a graphical user interface (GUI). After watching each synthesised sentence the viewer will be asked to transcribe orthographically what they believed they heard. Standard speech recognition metrics (measured as a function of insertions, substitutions and deletions) will be calculated from the transcribed listener responses and the original transcriptions, and the performance measured using the SCLITE Scoring Package [23].

To eliminate learning effects and ensure there is no bias resulting from the order that sequences are presented to viewers, the order of presentation will be pseudo-random across systems and viewers. All viewers will see sequences presented in a different order, and the order itself randomised over the systems. In addition, to provide a baseline for performance, original sequences will be included in evaluation. These will provide an upper bound on the expected performance of the synthesisers.

4.2. Naturalness Tests

The naturalness tests will be carried out after the intelligibility test, and will involve playing synthesised video sequences lip-synched to auditory speech (free from degradation) in the same order as in the intelligibility test. After the presentation of each video sequence the viewer will be asked to rate the naturalness of the visual speech gestures along a five point Likert scale [24]. In this instance the clean acoustic speech is required as it forms the basis for judging naturalness — the acoustic signal informs the viewer what the speaker is saying, their task is to determine how likely it is the visible articulator movements produced those sounds. The power of this form of test is viewers are very sensitive to inaccuracies (both static and dynamic) in synthesised facial gestures: the overall lip-shape must be correct, the degree of articulation must be correct, and the auditory and the visual modalities must be synchronised adequately. Quite often a viewer can identify that something is wrong, even if they cannot identify *what* exactly it is that is wrong. These tests will be designed only to gauge the overall sense of how accurately the synthesised visual speech corresponds to the (natural) acoustic speech. Viewers will not be asked to measure specific features such as "How correct are the degrees of articulation?" or "How well are the audio and visual modalities synchronised?" This information will be implicit in the naturalness score.

5. Releasing the Data

The training and test audio-visual corpora will be made available, following acceptance of the licensing conditions, from the LIPS2008 website: `http://www.lips2008.org` (at the time of writing the release of this data is under negotiation). Viewer responses will also be made available via the website, so research teams can later use the same training and test data, and measure the performance of their system against those entered into the original challenge. The tools used during the evaluation will also be available from the website. The size of the training and test corpa are approximately 4.24GB and 0.75GB

respectively. The video frames for each individual utterance are contained in a separate zip archive, and the wav files for all utterances in an additional zip archive.

6. LIPS2008: What next?

LIPS2008 will be somewhat of a pilot study. Firstly our aim is to gauge the level of interest in the research community for developing standardised evaluation metrics and data sets, and secondly to identify the needs and requirements of the many different forms of synthesiser in the literature. To maximise the number of participants in this first in a series of challenges, the condition requiring all systems to use the same training data can be relaxed so pre-trained systems tuned to specific data can also enter. This of course does not meet aims one and four outlined in the introduction. However, we hope to run future challenges, possibly as an annual event (as the Blizzard Challenge has become). In future challenges, it *will* be a requirement that every system uses the training data provided in conjunction with the challenge to overcome effects that might arise due to the appearance of speaker, and so on. We can potentially also broaden the challenge to cover gaze aware talking faces [25] and affective talking faces. Likewise, a more stringent evaluation framework might adopt a modular approach to evaluate individual components of the systems [9].

The scientific committee that will be set up in conjunction with the challenge will discuss the LIPS2008 results and the future directions of this evaluation framework.

7. Acknowledgements

The authors thank Christophe Savariaux and Nick Wilkinson for their assistance with the capture and preparation of the training data, and especially our speaker for agreeing to be recorded reciting the training and test corpora. Barry-John Theobald was supported in part by EPSRC (EP/D049075/1), Sascha Fagel was supported in part by the German Research Council DFG (FA 795/4-1). GIPSA-Lab work was supported in part by Rhône-Alpes Cluster ISLE.

8. References

[1] Bailly, G., Bérar, M., Elisei, F., and Odisio, M., "Audiovisual speech synthesis", International Journal of Speech Technology, 6:331–346, 2003.

[2] Theobald, B., "Audiovisual Speech Synthesis", International Congress on Phonetic Sciences, 285–290, 2007.

[3] Mori, M., "The uncanny valley", Energy, 7(4):33–35, 1970.

[4] Black, A., Bennett, C., Blanchard, B., Kominek, J., Langner, B., Prahallad, K., and Toth, A., "CMU Blizzard 2007: A hybrid acoustic unit selection system from statistically predicted parameters", Blizzard Challenge Workshop, Bonn, Germany, 2007.

[5] Geiger, G., Ezzat, T., and Poggio, T., "Perceptual evaluation of video-realistic speech", Tech Report: CBCL Paper 224/AI Memo 2003-003, MIT, Cambrige, MA, 2003.

[6] Massaro, D., "Perceiving Talking Faces", The MIT Press, 1998.

[7] Pandzic, I., Ostermann, J., and Millen, D., "User evaluation: Synthetic talking faces for interactive services", The Visual Computer, 5:330–340, 1999.

[8] Cosker, D., Marshall, D., Rosin, P., Paddock, S., and Rushton, S., "Towards perceptually realistic talking heads: models, metrics and McGurk", Proceedings of Applied Perception in Graphics and Visualization, 2004.

[9] Bailly, G., Gibert, G., and Odisio, M., "Evaluation of movement generation systems using the point-light technique", IEEE Workshop on Speech Synthesis, 27–30, 2002.

[10] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", Nature, 264:746–748, 1976.

[11] Benoît, C. and Le Goff, B., "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP", Speech Communication, 26:117-129, 1998.

[12] Fagel, S., Bailly, G., and Elisei, F., "Intelligibility of natural and 3D-cloned German speech", Proceedings of Auditory-Visual Speech Processing, 2007.

[13] Theobald, B., Bangham, J.A., Matthews, I., and Cawley, G., "Near-videorealistic Synthetic Talking Faces: Implementation and Evaluation", Speech Communication, 44:127–140, 2004.

[14] Fagel, S., "Auditory-visual integration in the perception of age in speech", International Congress on Phonetic Sciences, 725–728, 2007.

[15] Theobald, B., "Visual speech synthesis using shape and appearance models", Ph.D. dissertation, University of East Anglia, Norwich, UK, 2003.

[16] Benoît, C., Grice, M., and Hazan, V., "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences", Speech Communication, 18(4):381–392, 1996.

[17] `http://www.imagemagick.org`.

[18] Theobald, B., Harvey, R., Cox, S., Owen, G., and Lewis, C., "Lip-reading enhancement for law enforcement", SPIE conference on Optics and Photonics for Counterterrorism and Crime Fighting, 640205.1–640205.9, 2006.

[19] Sumby, W., and Pollack, I., "Visual Contribution to Speech Intelligibility in Noise", Journal of the Acoustical Society of America, 26:212–215, 1954.

[20] Fagel, S., and Madany, K., "Computeranimierte Sprechbewegungen in realen Anwendungen", Verlag der TU Berlin, 2008.

[21] Saldana, H., and Pisoni, D., "Audio-Visual speech perception without speech cues", Proceedings of the International Conference on Spoken Language Processing, 2187–2190, 1996.

[22] Schwartz, J.-L., Berthommier, F., and Savariaux, C., "Audio-visual scene analysis: Evidence for a very-early integration process in audio-visual speech perception", Proceedings of the International Conference on Spoken Language Processing, 1937–1940, 2002.

[23] `http://www.itl.nist.gov/iaui/894.01/tools/`

[24] Likert, R., "A Technique for the Measurement of Attitudes", Archives of Psychology, 140:1–55, 1932.

[25] Raidt, S., Bailly, G., and Elisei, F., "Analyzing and modeling gaze during face-to-face interaction", Proceedings of the International Conference on Intelligent Virtual Agents, 403-404, 2007.