Comparison of human and machine-based lip-reading

Sarah Hilder, Richard Harvey and Barry-John Theobald

The Speech Group, CMP
s.hilder@uea.ac.uk, {rwh,bjt}@cmp.uea.ac.uk

Abstract

We investigate the performance of a machine-based lip-reading system using both shape-only parameters and full shape and appearance parameters. Furthermore, we contrast the performance of a machine-based lip-reading system with human lip-reading ability. We find that the automated system outperforms human lip-readers. Curiously however, for relatively simple tasks there is little improvement in recognition accuracy when adding full appearance features to the machine-based system, whereas for human lip-readers we observe significant improvements in performance. Finally, we measure the effect of 'speaker training' on human lip-reading ability and we find even very limited training is sufficient to improve performance.

Index Terms: automated lip-reading, speechreading

1 Introduction

Speech generally is multi-modal in nature [1]. The human speech perception system fuses both acoustic and visual cues to decode speech produced by a talker — illustrated by, for example, the well known McGurk Effect [2]. These visual cues usually include the position and movement of the visible articulators (lips, teeth and tongue), and other cues not directly related to the production of speech (facial expression, head pose, body gestures, and so on). As listening conditions degrade, speech perception increasingly relies on these visual cues to assist with decoding spoken words [3, 4]. Indeed, profoundly deaf people rely solely on the information afforded by visual speech to understand spoken words.

It is also true that automatic speech recognition (ASR) systems benefit from the inclusion of visual features, especially as the acoustic signal degrades [5, 6]. Most work on automatic recognition of speech focuses purely on speech acoustics, or on audio-visual speech, but rarely on pure lip-reading alone (visual-only recognition). There are several problems associated with building a pure automated lip-reading system. Specifically, there is no well defined unit of visual speech and (as yet) there are no speaker-independent features upon which a system can be trained [7].

In this paper we investigate the performance of a speaker-dependent automatic lip-reading system. In particular, we evaluate the system using both shape-only features and the full appearance of the speaker. Furthermore, we compare the performance of the automated system and human lip-reading performance. Finally, we measure the effect of training on human lip-reading ability and compare learning effects with machine-based lip-reading.

2 Active Appearance Models

In this work, active appearance models (AAMs) [8] are used to encode visual speech for automated lip-reading. An advantage of AAMs is that they provide a description of both the shape and the appearance variation of the face. The shape of an AAM is defined as: $\mathbf{s} = \{x_0, y_0, \ldots, x_n, y_n\}^T$ — the concatenated vertex locations that define a triangulated mesh. A compact model that allows a linear variation in the shape is given by:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i, \tag{1}$$

where the coefficients p_i are the shape parameters. The model is usually computed by first hand-labelling a set of images with the vertices that define the mesh, then normalising the meshes for translation, scale and rotation, and finally applying principal component analysis (PCA) [8]. The base shape $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the m largest eigenvalues.

The appearance of an AAM is defined over the pixels within the base mesh $A(\mathbf{x}); \mathbf{x} = (x, y)^{\mathrm{T}} \in \mathbf{s}_0$. AAMs allow linear appearance variation. This means the appearance $A(\mathbf{x})$ can be expressed as a base appearance $A_0(\mathbf{x})$ plus a linear combination of l appearance images $A_i(\mathbf{x})$:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients λ_i are the appearance parameters. As with the shape, the base appearance $A_0(\mathbf{x})$ and appearance images $A_i(\mathbf{x})$ are usually computed by applying PCA to the (shape normalised) training images.

AAMs are usually trained from a few tens of hand-labelled images, and can then be used to automatically annotate entire video sequences. There are a wealth of algorithms for performing this fit. We use the *inverse compositional project-out* algorithm [9]. Hidden Markov models (HMMs) are used as the basis for all automated lip-reading systems described here, in particular the HMM toolkit (HTK) [10], is used.

3 Comparing visual features for lip-reading

The aim of this first experiment is to compare the performance of an automated lip-reading system trained using shape-only features with a system trained using full shape and appearance features. Because we will also compare machine-based lip-reading performance with human lip-reading ability, the task is deliberately made relatively simple to prevent human lip-reading scores from flooring at 0% correct [11].

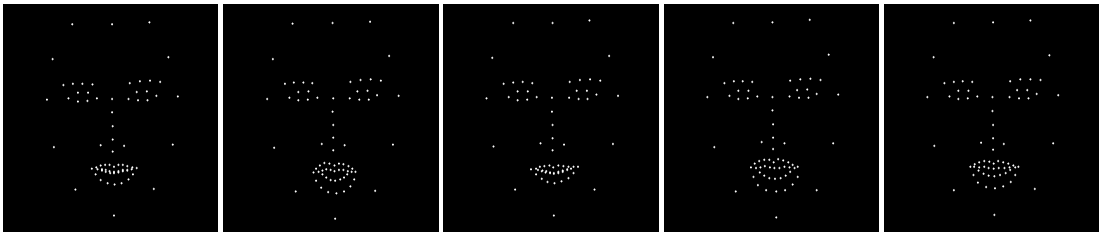

Figure 1: Example frames rendered as point-lights from a video sequence of a talker articulating "B". The speaker was instructed to begin and end with a closed mouth. The first image shows the first frame of the sequence, in the second the mouth is open while the speaker intakes breath, the subsequent frames are during speech.

3.1 Stimuli

Five speakers were filmed in full frontal pose speaking the letters A-F using a tri-chip Thompson Viper FilmStream HD camera. The video sequences were recorded as high definition uncompressed video (1920x1080) at 50Hz progressive scan. The selection of speakers were all male and covered a variety of ethnicities. Each speaker repeated the letters seven times in a neutral speaking style (no emotion), beginning and ending each utterance with a closed mouth. A speaker-specific AAM was trained for each individual and used to encode the videos as shape-only parameters ($\mathbf{p}$) and as shape and appearance ($\{\mathbf{p}; \lambda\}^T$).

To produce shape-only visual stimuli for the human experiments, the mesh vertices located using an AAM were rendered as point-lights, see Figure 1. These sequences exhibit only the shape and kinematic properties of the utterances, and are the visual representation of the AAM shape component. For the full shape and appearance sequences, the original video was used.

3.2 Procedure

Training machine-based lip-reading systems: Word-level HMMs were trained from both shape-only and full shape and appearance features using a leave-one-out cross-validation framework (using repetitions of the letters as folds). The topology of the HMMs was optimised and the best performing topology used in the evaluation. This was a topology of 16 states each with 3 Gaussian mixture components.

Measuring baseline performance of human lip-readers: 17 computing undergraduate students from a UK University volunteered to take part in the experiment. All reported to have normal, or corrected to normal vision. One repetition of each of the letters was selected from each of the five speakers for both the full-appearance and the point-light movies. This provided a set of 60 utterances (six letters $\times$ five speakers $\times$ two stimuli type). Participants were shown, in a randomised order, three repetitions of the 60 test movies (without audio), and they were each asked to circle on an answer sheet the letter they believed was being spoken. This was a closed test where all letters A–F letters were offered as possible answers to all utterances.

3.3 Results

Table 1 presents the results of this experiment as percent correct scores for both the machine-based lip-reading system and the human participants.

For both forms of stimuli we observe the machine-based system out-peforming human lip-readers (although the difference for the full shape and appearance is not statistically significant). Although untrained human viewers achieve only $\approx$43% correct for shape-only, this is significantly above chance ($p < 0.0001$), albeit on a relatively simple (six class) problem. Also note that in our selection of classes half were chosen deliberately to be *difficult* — the letters C, D, and E all end with the same vowel, /i/. (Although this is also the case for B, there is a clear bilabial action in the /b/ consonant, which should be sufficient to discriminate this letter. The differences between C, D, and E are much more subtle). If we reduce the experiment to a four class problem by collecting C, D and E into one group, human shape-only recognition increases to 60.7%, which remains significantly lower than the machines performance of 88.1% ($p < 0.0002$). These results suggest that the poor shape-only recognition is not solely due to the complexity in distinguishing between the *difficult* letters.

Table 1: Human and machine-based lip-reading accuracy for the letters A-F. Human results are the mean over 17 participants, the machine-based results are the mean over cross-validation folds.

	Shape-only	Shape and Appearance
Human	42.9%	71.6%
Machine	74.29%	75.24%

There are two possible causes for the lack of accuracy on the part of the human lip-readers when presented with shape-only stimuli. Firstly, as has been previously noted [12], shape-only lip-reading proficiency depends on the number and positioning of points on the face. It could be that the sparse vertices that define the shape of the AAM provide insufficient information. We consider this unlikely given the performance of the machine-based system trained on the same data. It is more likely that humans just are not used to seeing talking faces presented as (shape-only) point-lights. The world in which we live is inherently appearance-based, and thus the shape and appearance stimuli (as well as presenting more information), are more natural for human viewers.

We note that previous studies, e.g. [5], have reported a significant improvement when training an automated lip-reading system using full shape and appearance features compared with using shape features alone, which is counter to our finding above. However, this could be explained by the complexity of the task. We have considered only a six class problem, whereas in [5] a 26 class problem was considered. This could be verified by increasing the number of classes to include all 26 letters of the alphabet.

4 The effect of training on human lip-reading ability

The machine-based system tested in Section 3 out-performed human lip-readers, but this system was exposed to a training set containing multiple examples of all five speakers used in testing (note the actual utterances used in testing were **not** included in training). This second experiment was designed to determine whether the (better) performance achieved by the automated system was the result of prior exposure to the speakers. For the previous experiment we cannot simply hold-out a speaker as this requires speaker-independent visual features, which are not yet available [7]. Furthermore, the stimuli used in this experiment was designed to be more challenging to determine if the complexity of the task is attributable to the lack of improvement when adding appearance features to a shape-only lip-reading system.

4.1 Stimuli

A single male speaker recited a subset of 225 monosyllabic words drawn from [13], and 120 bisyllabic/monosyllabic nonsense words in the form of symmetric vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) words. The speaker was recorded using a Sony DSR-PD100AP DV camera, and the video was captured at 25 Hz DV PAL (720x640).

The nonsense VCV and CVC words contain permutations of C = { /p/, /b/, /k/, /g/, /v/, /f/, /ʃ/, /tʃ/, /d/, /t/ } and V = { /i/, /e/, /æ/, /ɪ/, /ʌ/,/ ɔ/ }. The consonants were chosen to contain intra-viseme confusions based on the phoneme-to-viseme mapping in [14].

4.2 Procedure

A subset of 60 monosyllabic real words and 60 mono/ bisyllabic nonsense words (30 each of CVC and VCV words) were chosen randomly. These formed the test data for this experiment and were subsequently removed from the training sequences. 19 computing science undergraduate students volunteered to take part in the experiment. They were first tested to determine their pre-training lip-reading ability. They then underwent a series of training sessions, before being re-tested to determine if the training had any significant effect on lip-reading ability.

Each training session lasted an hour, and participants took part in four sessions over consecutive days (they were advised to take a short break every 20 minutes). Each training clip was presented to viewers in the following conditions: 1) video only, 2) audiovisual, and 3) video only. After the first condition viewers were asked to write the word they had lip-read, next the audiovisual sequence was played to provide feedback, and finally the first condition was presented to again allow the viewer to lip-read the clip, this time having complete knowledge of the word. During training, stimuli were presented on standard individual PCs, so each participant was free to view each condition as many times as required, but the interface used ensured each condition was displayed at least once.

Both the pre-training and post-training tests used the same utterances. However, these were displayed in a different (randomised) order for each test, and no feedback was given after the first test. The tests were conducted as a group exercise, where the stimuli were presented without audio and using a data projector to display to the entire group. The test of monosyllabic (real) words was an open test, where participants were asked to lip-read the entire word without a list of the possible answers. The CVC and VCV test was a closed test, where participants were asked to identify the centre vowel/consonant only, and they were provided a list of possibilities.

An equivalent test was performed for the monosyllabic (real) words using an automated lip-reading system. The training videos were manually segmented by marking the approximate start/stop times of the constituent phonemes in the words according to the visual speech gestures (since we are training a lip-reading system). Phone level HMMs were then trained using visual features extracted from the training utterances. The HMMs were not exposed to any of the test data during the training process, and a topology of three states and five mixture components provided the best results.

4.3 Results

4.3.1 Monosyllabic words

Table 2 shows the effect of training on human lip-reading for real monosyllabic words in our experimental framework. At the word level, this brief training has significantly ($p < 0.02$) improved performance. Although the mean net improvement is only $\approx 4\%$, recall this is an open-test (no prompts) of pure lip-reading ability for isolated monosyllabic words. This is an extremely difficult task as even skilled lip-readers generally require some form of context to aid with understanding.

At the phoneme-level, a significant ($p < 0.04$) improvement in performance of $\approx 3\%$ was recorded with 15 of the 19 viewers seeing some improvement. Taking into account the expected visual confusions of the consonants, a mean improvement of $\approx 2.5\%$ was achieved, which again is significant ($p < 0.007$) — indeed 17 of the 19 participants experience a net benefit in inter-viseme performance.

Table 2: Human lip-reading accuracy for monosyllabic words.

	Monosyllabic Words		
	Word	Phone	Viseme
Pre-train	14.47%	28.40%	32.96%
Post-train	18.42%	31.60%	35.40%

4.3.2 Nonsense words

Table 3 shows the effect of training on human lip-reading ability for nonsense bisyllabic words in our experimental framework. For the case of VCV words, where the task is to identify the central consonant, training resulted in no significant difference when considering phonemes directly. However, taking into account the expected confusions between consonants we see a significant ($p < 0.07$) $\approx 6\%$ improvement in lip-reading accuracy. Thus, post-training viewers were better able to discriminate between classes of consonants (bilabials, labiodentals, dental, etc.), but were no better at discriminating between the intra-viseme phonemes.

Training also resulted in a significant improvement ($p < 0.02$) in the recognition of the central vowel in CVC nonsense words, with a mean improvement $\approx 10\%$.

4.3.3 Machine vs. human lip-reading performance

Table 4 shows both the post-training viewer scores and the results achieved by an automated lip-reading system tested on the

Table 3: Human lip-reading accuracy for VCV and CVC words.

	VCV		CVC
	Phone	Viseme	Phone
Pre-train	42.83%	75.00%	55.17%
Post-train	42.17%	81.30%	65.50%

Table 4: Human and machine-based lip-reading accuracy for monosyllabic (real) words encoded using different visual features.

	Monosyllabic Words		
	Word	Phone	Viseme
Human	18.42%	31.60%	35.40%
Shape-only	5%	72.97%	88.11%
Full	3.75%	80.27%	91.6 %

same monosyllabic (real) words as presented to human viewers. The results show that humans perform significantly better than the automated lip-reading system at word level ($p < 0.002$), although this is likely attributable to participants expecting real words as stimuli. Viewers may have exploited prior knowledge of English words in forming their response, i.e. selecting the closest valid word as opposed to what *appears* to have been spoken, whilst the machine-based system was not provided with a list of *allowed* words. This could also explain why the recognition achieved by the automated system improves considerably more than humans from word to phone/ viseme level. Where a number of phones are recognised, humans are able to complete the word, whereas, where the automated system recognises 70-80% of phones, there is no contextual information to complete the word.

At both the phoneme and viseme levels, the automated system performed significantly better than the human participants ($p < 0.002$) and achieved recognition rates of 80.27% and 91.6% (full shape and appearance) compared to 31.6% and 35.4% respectively. The influence of linguistic constraints on the human responses mentioned above may have also led to higher scores at both phone and viseme level, making the machine's recognition performance more profound that it may first appear.

As discussed in Section 3.3, the lack of significant improvement for full shape and appearance compared with shape alone was surprising. We stated that this is likely because for only six classes, albeit three (C, D and E) are very similar, there is sufficient information in the shape to discriminate between the letters. Here on a more difficult task we do indeed find there *is* a significant ($p < 0.02$) improvement over shape alone. We expect that as the complexity of the task increases further still, the more apparent this significance becomes.

5 Discussion

We have compared the performance of human and machine-based lip-reading ability. Relatively simple tasks have been used throughout so the effect of a relatively simple training framework can be measured using non-skilled lip-readers. We also have looked at the contribution of shape and appearance to lip-reading ability and found that for the (6-class) task presented here, appearance is significant for human viewers, but not for the automated system. Generally we find that automatic lip-reading systems outperform human viewers. However, an exception is recognising full, isolated, real words. This may be due to the prior knowledge of English words, which is not provided to the machine-based system. Viewer decision might be based on the validity of a word as opposed to what appears to have been spoken.

6 Acknowledgements

We are grateful to EPSRC for funding (EP/E028047/1) this work.

References

[1] D. Stork and M. Hennecke, Eds., *Speechreading by Humans and Machines: Models, Systems and Applications*, ser. NATO ASI Series F: Computer and Systems Sciences. Berlin: Springer-Verlag, 1996, vol. 150.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[3] N. Erber, "Auditory-visual perception of speech," *Journal of Speech and Hearing Disorders*, vol. 40, pp. 481–492, 1975.

[4] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, March 1954.

[5] I. Matthews, T. Cootes, J. Bangham, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1–16, 2002.

[6] G. Potamianos, C. Neti, G. G., A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[7] S. Cox, Y. Lan, J. Newman, R. Harvey, and B. Theobald, "The variability of lip-reading," in *In Proceedings of the International Conference on Auditory-visual Speech Processing*, 2008, pp. 179–184.

[8] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.

[9] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, November 2004.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," 2002, version 3.2.

[11] B. Theobald, R. Harvey, S. Cox, G. Owen, and C. Lewis, "Lip-reading enhancement for law enforcement," in *SPIE conference on Optics and Photonics for Counterterrorism and Crime Fighting*, G. Owen and C. Lewis, Eds., vol. 6402, September 2006, pp. 640 205–1–640 205–9.

[12] L. Rosenblum and H. Saldaña, "An audiovisual test of kinematic primitives for visual speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, pp. 318–331, 1996.

[13] L. Bernstein and S. Erberhardt, "Johns Hopkins lipreading corpus video-disk set," 1986, Johns Hopkins University, Baltimore, MD.

[14] B. Walden, R. Prosek, and A. Montgomery, "Effects of training on the visual recognition of consonants," *Journal of Speech and Hearing Research*, vol. 20, pp. 130–145, 1977.