INSIGHTS INTO MACHINE LIP READING

Yuxuan Lan, Richard Harvey and Barry-John Theobald

School of Computing Sciences, University of East Anglia, Norwich, UK

ABSTRACT

Computer lip-reading is one of the great signal processing challenges. Not only is the signal noisy, it is variable. However it is almost unknown to compare the performance with human lip-readers. Partly this is because of the paucity of human lip-readers and partly because most automatic systems only handle data that are trivial and therefore not representative of human speech. Here we generate a multiview dataset using connected words that can be analysed by an automatic system, based on linear predictive trackers and active appearance models, and human lip-readers. The automatic system we devise has a viseme accuracy of $\approx 46\%$ which is comparable to poor professional human lip-readers. However, unlike human lip-readers our system is good at guessing its fallibility.

***Index Terms*—** automated lip-reading, speech recognition, visual speech

1. INTRODUCTION

Automated lip-reading involves extracting features from regions of interest in images containing the mouth of a speaker and then mapping the temporal patterns observed in these features to the underlying spoken words [1]. This is notoriously difficult as speech involves more than just the visible articulators, so there is ambiguity in the process as many sounds are visually indistinguishable. In addition the visible articulators are free to adopt the position of upcoming sounds if there is no immediate requirement on their position for the current sound (e.g. early lip-rounding during /s/ in anticipation of the rounded vowel /u/ for the word "soon") — a phenomenon known as coarticulation. This means that for the same underlying sound, the visible speech articulators can be in different positions and so the visual features can be very different.

Despite increasing research interest over the past decade or so, the performance of automated lip-reading systems[1] falls significantly below the performance of acoustic speech recognisers. This partly is because of the reasons highlighted above, but more fundamentally the choice of visual features is important. It has been shown [2, 3] that data-driven model-based features tend to be more reliable than general image-based features, but still the poor performance prevents any real-world application of these systems. There is a real need for lip-reading systems, e.g. human lip-readers have been used in criminal court cases to transcribe video evidence when the accompanying acoustic channel is unavailable. This is expensive as lip-readers with the required ability are rare. In this paper we are interested in comparing the performance of our automated lip-reading system with the performance of expert human lip-readers.

[1] We consider visual-only recognition rather than the more common audiovisual recognition.

2. DATA CAPTURE

An audiovisual corpus of 12 speakers, 7 male and 5 female, each reciting 200 sentences selected from the Resource Management Corpus [4] was recorded. The database has a vocabulary size of approximately 1000 words, and was recorded in full-frontal view using using a tri-chip Thomson Viper FilmStream high-definition camera. The speakers were instructed to keep their head relatively still, and each was recorded in a single sitting to ensure reasonably constant illumination.

3. AUTOMATED LIP-READING

Fig. 1. The first three modes of variation of a combined shape and appearance model. The modes are shown at +3 standard deviations (top row) and -3 standard deviations (bottom row).

The basis of our lip-reading system [3] is a set of viseme-level Hidden Markov Models (HMMs), which are built and manipulated using HTK [5]. The visual features from which these models are trained are Active Appearance Model (AAM) parameters [6] as these have been shown to be more reliable than traditional image-based features [2]. The modes of variation of the AAM are shown in Figure 1. The AAM

features are transformed using a LDA transform similar to [7], see Figure 2 for an overview. These features are then upsampled to 100 frames per second and augmented with the first and second derivatives — see [3] for a more detailed description. The effect of this is to increase the separation in the features between viseme classes and to reduce the dependency on speaker identity. See Figure 3 for example.

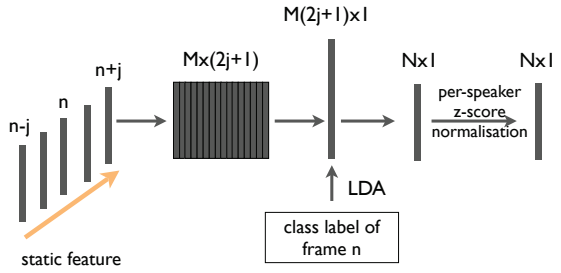

Fig. 2. An overview of the window-based LDA transform.

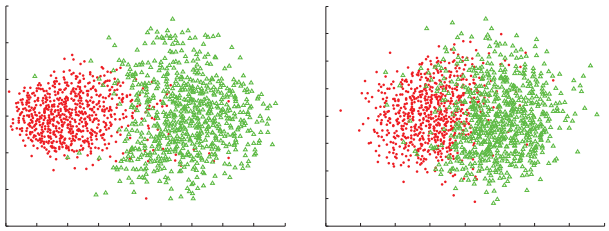

Fig. 3. Sammon projection of (A) MFCCs and (B) LDA transformed AAM features. In both cases the features represent 12 speakers and two visemic classes (/f/ (red dots) and /u/ (green triangles). Notice the good separation between the different viseme classes for both the acoustic and the visual data. The Hi-LDA transform effectively removes the influence of the speaker from AAM features which are highly speaker dependent [8].

To generate viseme-level transcriptions, the acoustic speech from the dataset is force-aligned to generate phone-level transcriptions, and these are converted to viseme transcriptions using a standard phoneme-to-viseme mapping [9]. These viseme-level transcriptions are then used to train and test both the visual-only and audio-only HMMs. Correspondingly, the pronunciation dictionary is also translated from phoneme-level pronunciations to viseme-level pronunciations. 14 HMMs are trained on visual features: one for each viseme and one to model 'visual silence'. A 'short pause' model is tied to the middle state of the silence model. Left-right HMMs with three states and a diagonal covariance Gaussian Mixture Model (GMM) associated with each state are used.

The system is trained to lip-read a single speaker. Single Gaussian HMMs are initialised using flat start training via HTK command `HCompV`, and this is followed by a series of embedded re-estimation (`HERest`). The number of Gaussian mixture components is increased from 1 to 2, 5, and 9. A bigram word language model is constructed from the training data. During recognition, various insertion penalties $p = \{-20, 0, 10\}$ and the grammar scale factors $s = \{0, 1, 5, 15\}$ are tested. Table 1 lists methods and some key parameters used by the automated lip-reading system. The set of parameters that provide the highest accuracy on a validation set are applied to the test set, which results in a viseme recognition accuracy of 45.67% and a word recognition accuracy of 14.08%. We are interested in determining how this accuracy compares with the performance of expert human lip-readers.

stage	methods	key parameters
1.feature extraction	AAM shape and appearance [6]	98% variation
	PCA on shape and appearance [6]	98% variation
	forms hyper vector, see also Figure 2	$n = 2$
	LDA on hyper vector	99% variation
	global and speaker z-score normalisation [3], upsample to 100fps	
2.HMM training	flat start training `HCompV`	3 state, 1 GMM component
	embedded re-estimation `HERest`	4 iterations
	GMM component increment	1 to 2, 5, and 9
	construct language model	2-gram word
3.HMM recognition	Viterbi decoding HVite	$p = 10$, $s = 5$, #GMM component = 5

Table 1. Key methods and parameters used by the automated lip-reading system.

4. HUMAN LIP-READING

Six expert human lip-readers participated in a two-part exercise to transcribe (silent) video sequences from our RM corpus. All are practising professional lip-readers and were paid to produce transcripts of the best possible quality. Each lip-reader worked alone and independently. For the first stage of the test, the lip-readers were provided with 10 videos of a single speaker and the accompanying transcriptions. They were free to use these videos as they wished so they could acquaint themselves with the style of the speaker. The lip-readers were then asked to transcribe 10 test videos of the same speaker for which the transcriptions and the audio were not available. The aim of this test was to measure the raw ability of the lip-readers — no information (e.g. regarding the vocabulary, the domain of discourse, etc.) was provided other than the tran-

scriptions for the ten training videos.

The second stage of the experiment was designed to measure the affect of having knowledge analogous to the training data. The lip-readers were supplied with transcriptions for 1000 sentences from the dataset (but not the videos) and a list of the 971 words that form the vocabulary of the corpus. They were then asked to repeat the task of transcribing the previous ten test videos and revise their transcripts where requried. Four of the six lip-readers completed both stages of the experiment, whistle two stated that their results in stage 2 did not differ from stage 1 and that we should use their stage 1 results for stage 2.

5. RESULTS

For both tasks measuring human performance, we the word recognition accuracy and the viseme recognition accuracy by comparing with the ground-truth transcriptions. The results are presented in Table 2 and Figure 4.

Table 2. Word and viseme accuracy for six individual lip-readers compared with the performance of an automated system. Note: "*" indicates that a participant recommended we used their stage 1 score in stage 2.

lip-reader	stage 1		stage 2	
	word acc%	viseme acc%	word acc%	viseme acc%
#1	30.99	57.01	0	39.7
#2	8.45	34.63	8.45*	34.63*
#3	18.31	56.12	40.85	60.6
#4	0	51.04	7.04	64.78
#5	2.82	30.75	2.82*	30.75*
#6	46.48	73.73	69.01	85.37
computer	-	-	14.08	45.67

There is a noticeable change in accuracy from stage one to stage two for the four lip-readers that conducted both stages of the experiment. It seems that most lip-readers were able to make positive use of the contextual and grammar information provided at the beginning of stage two, and the improvement in accuracy implies perhaps a general language model used in stage one was replaced by a more targeted language model that is specific to the task. If this information is not incorporated with visual cues correctly, it can lead to misunderstanding, for example, see the case of lip-reader #1.

From these results we can make several observations. First, all lip-readers are different. They recognise different words and sentences from the same material, and some lip-readers are more accurate than others. Second, lip-readers apply complicated language modelling during lip-reading. At stage one, lip-readers received very limited information about the subject and the topic that was carried by training material. The transcriptions from lip-readers were largely in the scope of day-to-day language, reflecting a general, broad-

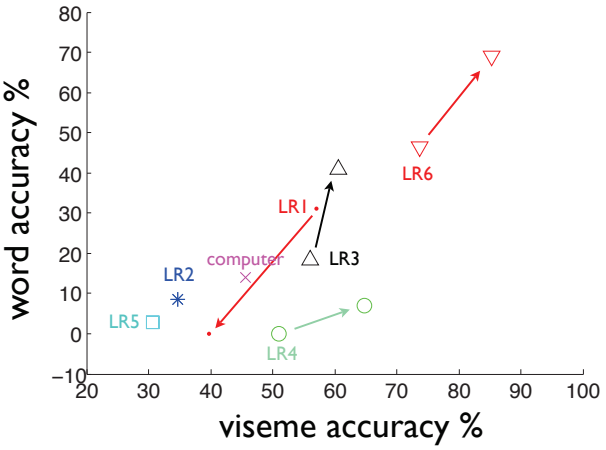

Fig. 4. Word accuracy plotted against viseme accuracy for six expert lip-readers. The colour-coded arrow indicates the direction of change from the first stage to the second stage. The performance of an automated lip-reading system evaluated on the same material (marked as "computer") is also included.

scoped language model that was applied. The exception is lip-reader #6, who seems to have tuned the model towards the scope of the dataset. More information was provided at stage two, which resulted in transcriptions becoming more focussed around the domain of the RM dataset. This would imply a narrow-scoped, targeted model was learned using the provided information. It is worth pointing out the case of lip-reader #3, who had a 4.4% increase in viseme performance, but gained a 22.5% absolute improvement on word accuracy. One explanation is that the person perceived the same visual cues in both stages, but a suitable language model helped to make right decision during recognition. Finally, the performance of automated lip-reading falls within that of the human lip-readers, although it remains a lot worse than an acoustic speech recognition system. Given the importance of context, it is frustrating that we were not able to persuade all human lip-readers to persevere with stage 2 - human are not insightful about their true word accuracy.

6. FUTURE WORK

We previously have extensively investigated the choice of feature for use in automated lip-reading [2, 3], and in this paper we have shown that the performance of a state-of-the-art lip-reading system using reliable features is comparable to a reasonably capable human lip-reader on a similar task. A significant focus of our future work will be to improve the language modelling to bridge the gap in performance between current levels and the performance of the better human lip-readers. In [3] we showed that a key problem is the number of deletions in the recogniser output, see the relative proportion between the leading diagonal and the column labelled *Del* in

Figure 5. These deletions undoubtedly are due to coarticulation effects, and we suspect that human lip-readers make better use of knowledge of the language to help overcome these issues (e.g. the significant improvement in performance for three of the four lip-readers that undertook both parts of the test after being provided with example transcripts and the vocabulary). There is no more visual information provided to the lip-readers between stages one and two, rather they are better able to make use of the limited visual information with increased knowledge of the language. An interesting question is how can this aspect of language modelling be incorporated into an automated lip-reading system.

ground truth \ prediction	ah	eh	f	ao	t	uh	w	k	p	iy	aa	ch	oo	Del	%c	%d
ah	295	11	10	8	14	10	6	16	5	29	6	1	10	240	70.1	36.31
eh	28	330	2	9	11	6	4	9	3	19	3	2	3	189	76.9	30.58
f	2	1	205	2	4	1	9	4	0	1	0	1	0	66	89.1	22.3
ao	12	7	3	70	8	5	5	4	0	4	7	0	6	83	53.4	38.79
t	37	23	30	17	949	8	28	60	8	24	14	11	5	426	78.2	25.98
uh	2	3	3	3	1	30	2	1	0	3	3	3	3	49	52.6	46.23
w	2	1	6	2	7	0	227	17	1	3	0	4	1	132	83.8	32.75
k	23	12	13	16	28	1	25	638	7	18	9	7	13	408	78.8	33.5
p	9	8	10	0	11	2	29	12	241	2	3	1	5	129	72.4	27.92
iy	27	32	7	13	10	4	12	21	6	359	3	7	6	287	70.8	36.15
aa	1	3	0	0	1	2	1	3	1	1	20	0	1	25	58.8	42.37
ch	4	1	6	2	10	1	7	8	1	2	3	56	5	73	52.8	40.78
oo	2	4	1	1	5	1	2	0	0	0	0	0	44	37	73.3	38.14
Ins	62	45	30	22	69	13	55	73	8	50	27	10	24			
Total															75.26	31.78

Fig. 5. Confusion matrix for a speaker-independent lip-reading system trained and tested using LDA transformed AAM features.

7. CONCLUSIONS

Automated lip-reading (visual-only speech recognition) has been receiving increasing interest in recent years, but the performance of state-of-the-art systems still is significantly below the performance of acoustic speech recognisers on the same task. In this paper we have attempted to quantify the performance of automated systems in terms of the performance of expert human lip-readers on the same task. We found that the accuracy of human lip-readers can be improved significantly when limited knowledge of the language is introduced, and that in less than ideal conditions (i.e. strange vocabulary and sentence structure), good human lip-readers are able to exploit language well and improve their accuracy. We conclude that the performance of an automated lip-reading system on a reasonably large vocabulary (≈ 1000 words) and continuous speech is comparable to a reasonable human lip-reader. Note that all human lip-readers that took part in these tests consider themselves expert and all offer their services as expert lip-readers.

A potential criticism on this work is that the automated lip-reading system are provided with informations such as vocabulary and training transcripts that can lead to a sophisticated language model, yet only 2-gram model is applied. In the future, there is a plan to apply a higher order language model, e.g., 3-gram, 4-gram, and utilise a discriminative training scheme.

8. ACKNOWLEDGEMENTS

The authors are grateful to the lip-readers that took part in the evaluation, and to the UK Home Office for funding the work.

9. REFERENCES

[1] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior, "Recent advances in the automatic recognition of audio-visual speech," in *Proceedings of the IEEE*, Sept 2003, vol. 91, pp. 1306–1326.

[2] Y. Lan, R. Harvey, B. Theobald, E-J. Ong, and R. Bowden, "Comparing visual features for lipreading," in *Proc. of International Conference on Auditory-visual Speech Processing*, 2009, pp. 102–106.

[3] Y. Lan, B. Theobald, R. Harvey, E-J. Ong, and R. Bowden, "Improving visual features for lip-reading," in *Proceedings of International Conference on Auditory-Visual Speech Processing*, 2010.

[4] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status.," in *In Proceedings of the DARPA Speech Recognition Workshop.*, 1986.

[5] S. Young, G. Evenmann, D Kershaw, G. Moore, J. Odell, D. Ollason, V Valtchev, and P. Woodland, *The HTK Book (version 3.2.1)*, 2002.

[6] 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.

[7] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. 2004, MIT Press.

[8] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The callenge of multispeaker lip-reading," in *International Conference on Auditory-Visual Speech Processing 2008*, 2008, pp. 179–184.

[9] C. G. Fisher, "Confusions among visually perceived consonantsnants," *Journal of Speech and Hearing Research*, vol. 11, pp. 796–804, 1968.