Visual gesture variability between talkers in continuous visual speech

Helen L Bear helen@uel.ac.uk

1. In visual speech processing we can identify individuals from their unique visual speech signals [1] but in lipreading systems we want to lipread any speaker. Recent work with deep learning implement end-to-end system butin this approach do not develop our knowledge of the visual speech signal which is considered a sequence of gestures (visemes) to represent acoustic utterances. In this work we ask, **how different are speaker-dependent visemes?**

2. We use the Bear speaker dependent viseme algorithm [2] to build 25 sets of visemes.

- a multi-speaker P2V map using **all** speakers' phoneme confusions (MS);
- a speaker-independent P2V map for each speaker using confusions of all **other** speakers in the data (SI);
- a speaker-dependent P2V map for each speaker (SSD).

3. Using 12 speakers of RMAV AV dataset we test as follows; $M_n(p,q)$. M is the visemes of speaker n, p is the training speaker(s), and q denotes the test speaker(s). [3]

MS & SI $M_n(p,q)=M_{(all)}(1,1)$ where p=q for talkers 1 to 12

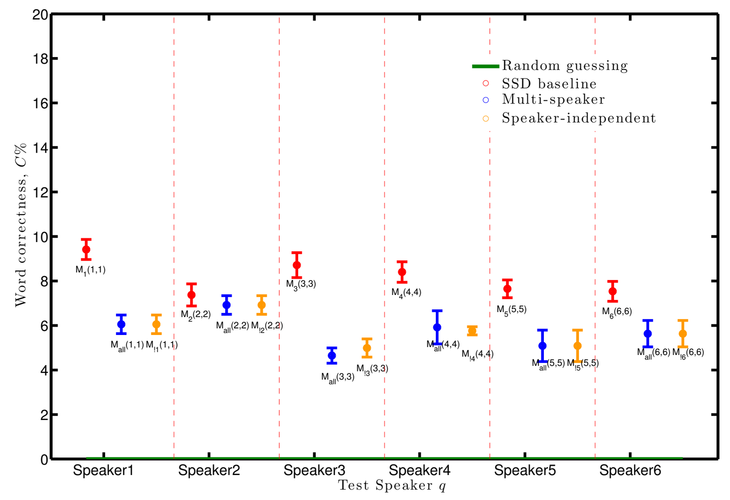

Figure 1: Word classification correctness, $C \pm 1s.e$, of the MS and SI tests for speakers 1-6. Baseline is SSD maps (red)

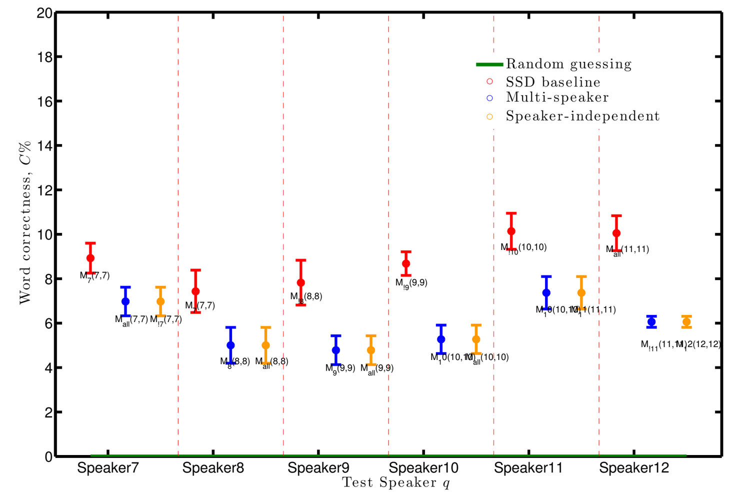

Figure 2: Word classification correctness, $C \pm 1s.e$, of the MS and SI tests for speakers 7-12. Baseline is SSD maps (red)

All speakers bar Speaker 2 are significantly negatively affected by using generalised multi-speaker visemes.
This quantifies lip-reading dependency on speaker identity as dependent on which two speakers are being compared.

DS&D $M_n(p,q)=M_x(x,y)$ where p≠q & n=1:12

Table 1: Different speaker-dependent maps and Data (DS&D) experiments for Speaker one.

Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp2	Sp2	Sp1	$M_2(2,1)$
Sp...	Sp...	Sp1	$M_{..}(...,1)$
Sp12	Sp12	Sp1	$M_{12}(12,1)$

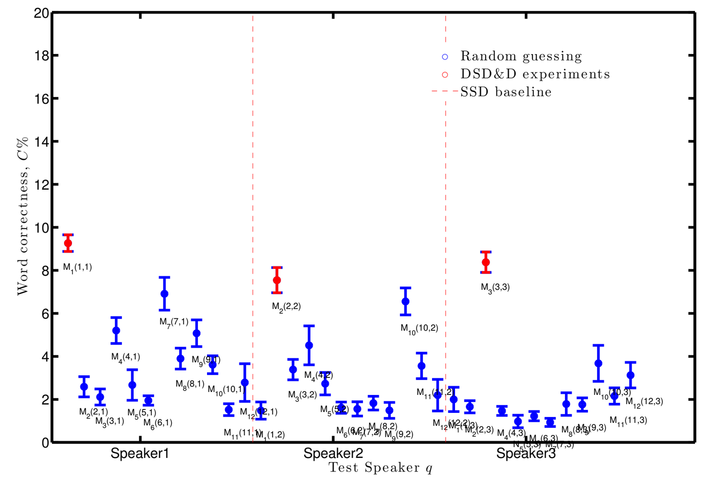

Figure 7: Word classification correctness, $C \pm 1s.e$, of the DSD&D tests for speakers 1-3. Baseline is SSD maps (red)

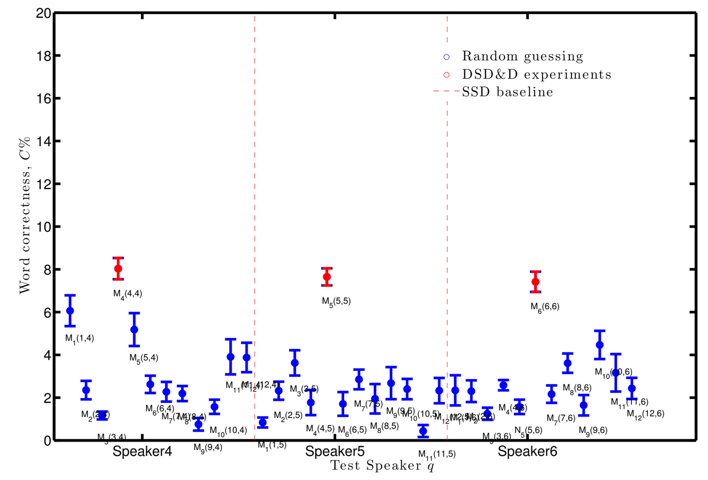

Figure 8: Word classification correctness, $C \pm 1s.e$, of the DSD&D tests for speakers 4-6. Baseline is SSD maps (red)

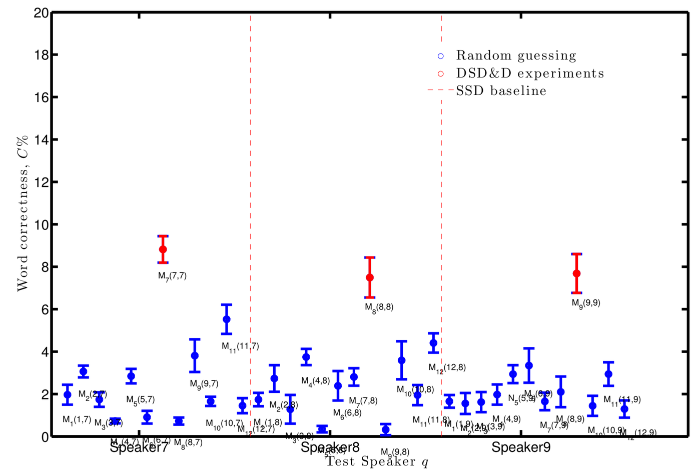

Figure 9: Word classification correctness, $C \pm 1s.e$, of the DSD&D tests for speakers 7-9. Baseline is SSD maps (red)

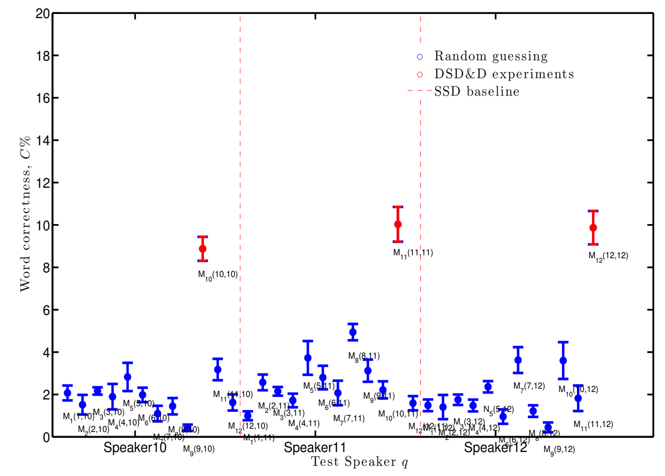

Figure 10: Word classification correctness, $C \pm 1s.e$, of the DSD&D tests for speakers 10-12. Baseline is SSD maps (red)

It is not only a speakers identity but how their gestures are sequenced for lipreading. Similarities between some speakers could adapt to lipread visually-similar speakers.

Table 3: Comparison scores measuring the effect of using speaker-dependent maps for *other* speakers lip-reading.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
Sp01	0	−1	−2	−2	+1	−1	−1	−1	+1	+1	−1	+1
Sp02	+2	0	+1	+1	+2	+2	+1	+1	+2	+2	+1	+2
Sp03	−2	−2	0	−2	+1	−1	−1	−2	−2	−2	−2	+1
Sp04	−2	−1	−1	0	+1	+1	−2	−2	+1	−1	−2	+1
Sp05	−2	−1	+2	−2	0	+1	−1	+2	+1	+2	−1	+2
Sp06	−1	−1	−1	+1	+2	0	+2	−1	−1	+1	+1	+2
Sp07	+1	−1	−1	+1	+1	+1	0	+1	−1	−1	+1	+1
Sp08	−1	−1	+1	−1	−1	−2	−2	0	+1	+2	+1	+1
Sp09	−2	−2	−1	−2	−1	−1	−1	−2	0	−1	−2	+1
Sp10	−2	−2	−1	−1	−1	−2	−2	−2	−2	0	−2	−2
Sp11	−1	+1	−1	+1	+1	−1	+1	−1	−1	+2	0	+2
Sp12	−1	−2	−2	−1	−1	−2	−2	−2	−2	−1	−2	0
Total	−9	−11	−6	−7	**+3**	−5	−8	−9	−3	−4	−8	**+12**

We score effect of sharing visemes, as table 3, M_{12} visemes are optimal for all speaker coverage.

DSD $M_n(p,q)=M_x(y,y)$ where p=q & n=1:12

Table 2: Different Speaker-Dependent maps (DSD) for Speaker one.

Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp2	Sp1	Sp1	$M_2(1,1)$
Sp...	Sp1	Sp1	$M_{..}(1,1)$
Sp12	Sp1	Sp1	$M_{12}(1,1)$

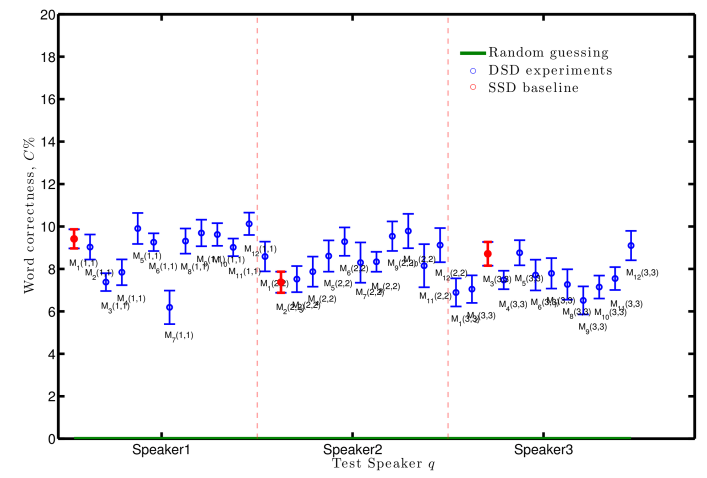

Figure 3: Word classification correctness, $C \pm 1s.e$, of the DSD tests for speakers 1-3. Baseline is SSD maps (red)

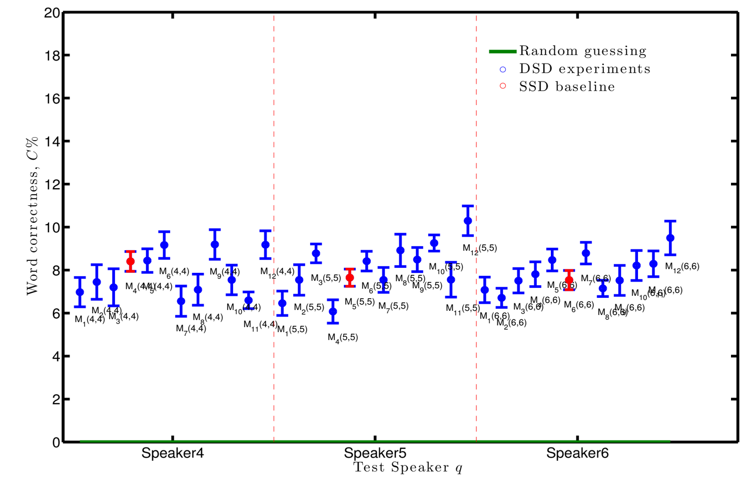

Figure 4: Word classification correctness, $C \pm 1s.e$, of the DSD tests for speakers 4-6. Baseline is SSD maps (red)

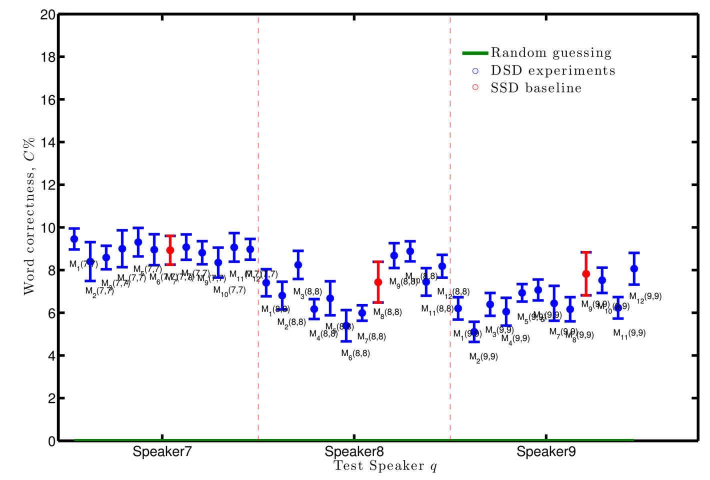

Figure 5: Word classification correctness, $C \pm 1s.e$, of the DSD tests for speakers 7-9. Baseline is SSD maps (red)

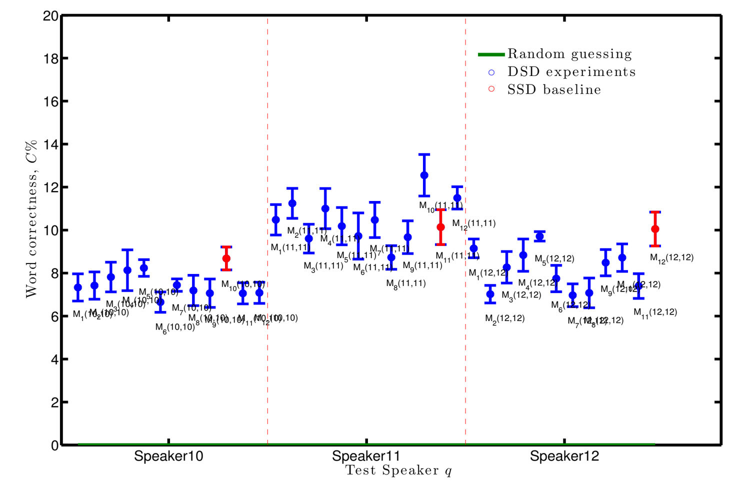

Figure 6: Word classification correctness, $C \pm 1s.e$, of the DSD tests for speakers 10-12. Baseline is SSD maps (red)

Some speakers significantly deteriorate the classification rates when training speakers are not same as the test speakers but others are not significantly affected. This variation is attributed to the speaker identity and language structure.

5. There is risk of over-generalising MS/SI visemes. The lipreading dependency on training speakers by generalising to speakers who are visually similar in viseme usage/tragetory through gestures. Whilst consistent with deep learning, now we should not need such big data volumes to achieve this.

UEL University of East London

References:
1. Speaker identification by lipreading. J. Luettin, N. A. Thacker, and S. W. Beet. ICSLP 1996
2. Phoneme-to-viseme mappings; the good, the bad, and the ugly. H L Bear, RW Harvey, Speech Communication, 2017
3. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. HL Bear, SJ Cox, RW Harvey AVSP 2015