Comparing heterogeneous visual gestures for measuring the diversity of visual speech signals

Helen L Bear[a,*], Richard Harvey[b]

[a]CVSSP, Dept of Electrical Engineering, University of Surrey, Guildford, GU2 7JP, UK
[b]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

Abstract

Visual lip gestures observed whilst lipreading have a few working definitions, the most common two are; 'the visual equivalent of a phoneme' and 'phonemes which are indistinguishable on the lips'. To date there is no formal definition, in part because to date we have not established a two-way relationship or mapping between visemes and phonemes. Some evidence suggests that visual speech is highly dependent upon the speaker. So here, we use a phoneme-clustering method to form new phoneme-to-viseme maps for both individual and multiple speakers. We test these phoneme to viseme maps to examine how similarly speakers talk visually and we use signed rank tests to measure the distance between individuals. We conclude that broadly speaking, speakers have the same repertoire of mouth gestures, where they differ is in the use of the gestures.

Keywords: Visual speech, lipreading, recognition, audio-visual, speech, classification, viseme, phoneme, speaker identity

1. Introduction

Computer lipreading is machine speech recognition from the interpretation of lip motion without auditory support [1, 2]. There are many motivators for wanting a lipreading machine, for example places where audio is severely hampered by noise such as an airplane cockpit, or where placing a microphone close to a source is impossible such as a busy airport or transport hub [3–5]

Conventionally, machine lipreading has been implemented on two-dimensional videos filmed in laboratory conditions [6, 7]. More recently, such datasets have been growing in size to enable deep learning methods to be applied in lipreading systems [8, 9]. Separately there has also been some preliminary work to use depth cameras to capture pose/lip protrusion information [10, 11] or in the RGB colour space for more discriminative appearance features [12]. The challenge with these works are that the results achieved are yet to significantly outperform conventional lipreading systems. The top 1 scores in [8] are less than [13] and to date the best end-to-end system is that of Stafylakis and Tzimiropoulos who achieved an error rate of 11.29% on a 500 word vocabulary [14].

*Corresponding author
Email addresses: dr.bear@tum.de (Helen L Bear), r.w.harvey@uea.ac.uk (Richard Harvey)

In developing lipreading systems we know that speech is a bimodal signal, and we use the the visual channel of information for recognition of visual cues or gestures [15]. The units within this information channel, in sequence form a signal of its own, but it has no formal definition despite a variety of options presented previously [16–20]. Irrespective of the definition in each paper, these units are commonly referred to as 'visemes' and in this paper, we define a viseme as a visual cue (sometimes also referred to as a gesture) that represents a subset of identical phonemes on the lips [21–23]. This means a set of visemes is always smaller than the set of phonemes [24]. These visemes are interesting because they help researchers to answer questions about how best to decipher lip motions when affected by issues such as human lipreading [25], language [26], expression [27], and camera parameters like resolution [28].

Previous work has shown the benefits of deriving speaker-dependent visemes [29, 30] but the cost associated with generating these is significant. Indeed the work by Kricos [29] was limited due to the human subjects required, whereas the data-driven method of Bear [30] could scale if visual speech ground truths for the test speakers were available in advance. The concept of a unique Phoneme-to-Viseme (P2V) mapping for every speaker is daunting, so here we test the versatility and robustness of speaker-dependent visemes by using the algorithm in [31] to derive single-speaker, multi-speaker, and multi-speaker-independent visemes and use these in a controlled experiment to answer the following questions; To what extent are such visemes speaker-independent? What is the similarity between these sets of visemes?

This work is motivated by the many future applications of viseme knowledge. From improving both lipreading and audio-visual speech recognition systems for security and safety, to refereeing sports events and understanding silent films, understanding visual speech gestures has significant future impact on many areas of society.

In our previous work we investigated isolated word recognition from speaker-dependent visemes [21]. Here we extend this to continuous speech. Benchmarked against speaker-dependent results, we experiment with speakers from both the AVLetters2 (AVL2) and Resource Management Audio-Visual (RMAV) datasets. The AVL2 dataset is a dataset of seven utterances per speaker reciting the alphabet. In RMAV the speakers utter continuous speech, sentences from three to six words long for up to 200 sentences each. Our hypothesis is that, with good speaker-specific visemes, we can negate the previous poor performance of speaker independent lipreading. This is because, particularly with continuous speech, information from language and grammar create longer sequences upon which classifiers can discriminate.

The rest of this paper is structured as follows: we discuss the issue of speaker identity in computer lipreading, how this can be a part of the feature extraction method to improve accuracy and how visemes can be generated. We then discuss speaker-independent systems before we introduce the experimental data and methods. We present results on isolated words and continuous speech data. We use the Wilcoxon signed rank [32] to measure the distances between the speaker-dependent P2V maps before drawing conclusions on the observations.

2. Speaker-specific visemes

Speaker appearance, or identity, is known to be important in the recognition of speech from visual-only information (lipreading) [33], more so than in auditory speech. Indeed appearance data improves lipreading classification over shape only models whether one uses Active Appearance Models (AAM) [28] or Discrete Cosine Tranform (DCT) [10] features.

In machine lipreading we have interesting evidence: we can both identify individuals from visual speech information [34, 35] and, with deep learning and big data, we have the potential to generalise over many speakers [8, 36].

One of the difficulties in dealing with visual speech is deciding what the fundamental units for recognition should be. The term *viseme* is loosely defined [19] to mean a visually indistinguishable unit of speech, and a set of visemes is usually defined by grouping together a number of phonemes that have a (supposedly) indistinguishable visual appearance. Several many-to-one mappings from phonemes to visemes have been proposed and investigated [19], [22], or [25]. Bear *et al.* showed in [30] that the best speaker-independent P2V map was devised by Lee [37] when recognising isolated words, but for continuous speech a combination of Disney's vowels [38] and Woodward's [39] consonants were better. From this we inferred that language has a significant effect on the appearance of visemes.

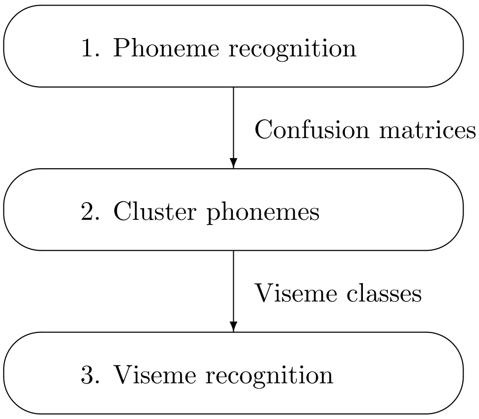

Figure 1: Three step process for recognition from visemes. This figure summarizes the process undertaken by Bear et al. in [31]

The question then arises to what extent such maps are independent of the speaker, and if so, how speaker independence might be examined. In particular, we are interested in the interaction between the data used to train the models and the viseme classes themselves.

More than in auditory speech, in machine lipreading, speaker identity is important for accurate classification [33]. We know a major difficulty in visual speech is the labeling of classifier units so we need to address the questions; to what extent are such maps independent of the speaker? And if so, how might speaker dependent sets of visemes be examined? Alongside of this, it would be useful to understand the interactions between the model training data and the classes. Therefore in this section we will use both the AVL2 dataset [33] and the RMAV dataset [40] to train and test classifiers based upon a series of P2V mappings.

2.1. Speaker-independence

Currently, robust and accurate machine lipreading performances are achieved with speaker-dependent classification models [8], this means the test speaker must be included within the classifier training data. A classification model which is trained without the test speaker performs poorly [33, 41]. Thus speaker independence is the ability to classify a speaker who is not involved in the classifier training [2]. This is a difficult, and as yet, unsolved problem.

One could wonder if, with a large enough dataset with a significant number of speakers, then it could be sufficient to train classifiers which are generalised to cover a whole population including independent speakers. But we still struggle without a dataset of the size needed to test this theory, particularly as we do not know how much is 'enough' data or speakers. Works such as [42] use domain adaptation [43], and [44] use Feature-space Maximum Likelihood Linear Regression (fMLLR) features [45, 46]. These achieve significant improvements on previous speaker independent results but still do not match those of speaker dependent accuracy.

An example of a study into speaker independence in machine lipreading is [33], here the authors also use the AVL2 dataset and they compare single speaker, multi-speaker and speaker independent classification using two types of classifiers (Hidden Markov Models (HMM) & Sieves, sieves are a kind of visual filter [47]). However, this investigation uses word labels for classifiers and we are interested to know if the results could be improved using speaker-dependent visemes.

3. Description of datasets

We use the AVL2 dataset [33], to train and test recognisers based upon the speaker-dependent mappings. This dataset consists of four British-English speakers reciting the alphabet seven times. The full-faces of the speakers are tracked using Linear Predictors [48] and Active Appearance Models [49] are used to extract lip-only combined shape and appearance features. We select AAM features because they are known to outperform other feature methods in machine visual-only lipreading [24]. Figure 2 shows the count of the 29 phonemes that appear in the phoneme transcription of AVL2, allowing for duplicate pronunciations, (with the silence phoneme omitted). The British English BEEP pronunciation dictionary [50] is used throughout these experiments.

Our second data set is continuous speech. Formerly known as LiLIR, the RMAV dataset consists of 20 British English speakers (we use 12, seven male and five female), up to 200 utterances per speaker of the Resource Management (RM) sentences from [51] which totals around 1000 words each. It should be noted the sentences selected for the RMAV speakers are a significantly cut down version of the full RM dataset transcripts. They were selected to maintain as much coverage of all phonemes as possible as shown in Figure 3 [44]. The original videos were recorded in high definition (1920 × 1080) and in a full-frontal position.

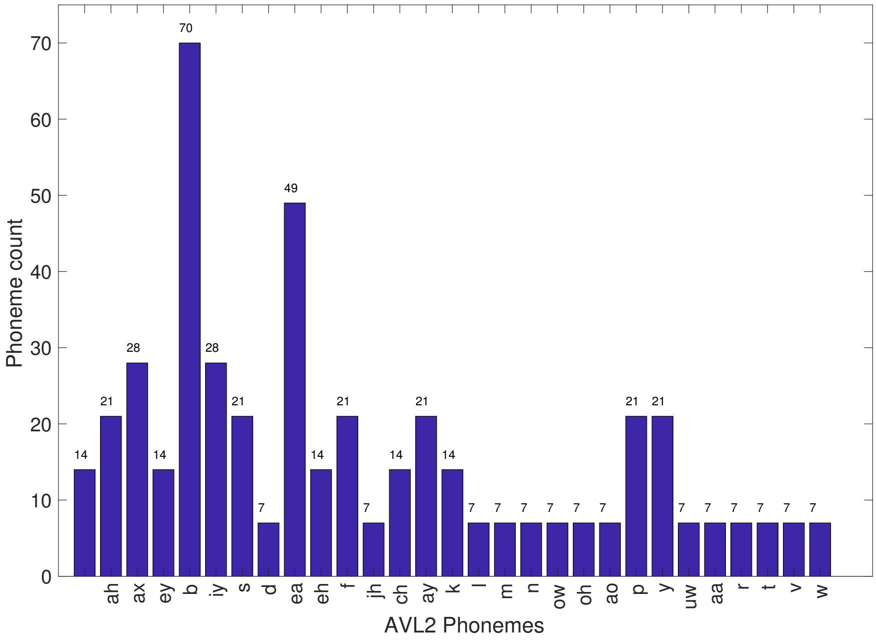

Figure 2: Phoneme histogram of AVLetters-2 dataset

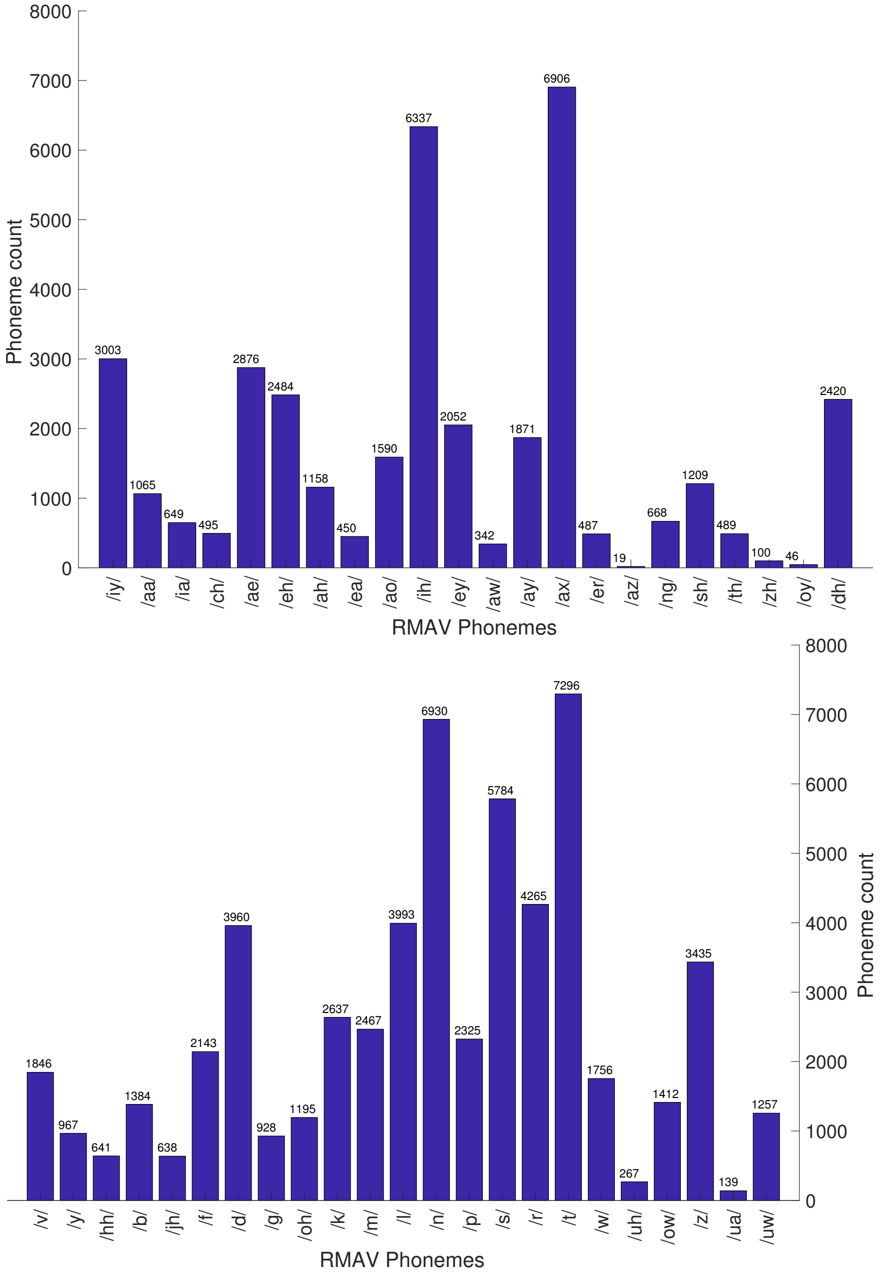

Figure 3: Occurrence frequency of phonemes in the RMAV dataset.

3.1. Linear predictor tracking

Linear predictors have been successfully used to track objects in motion, for example [52]. Here linear Predictors are a person-specific and data-driven facial tracking method [53] used for observing visual changes in the face during speech, linear predictor tracking methods have shown robustness that make it possible to cope with facial feature configurations not present in the training data [48] by treating each feature independently.

A linear predictor is a single point on or near the lips around which support pixels are used to identify the change in position of the central point between video frames. The central points are a set of single landmarks on the outline of speaker lips. In this method both the lip shape (comprised of landmarks) and the pixel information surrounding the linear predictor positions are intrinsically linked, [54].

3.2. Active appearance model features

Individual speaker AAM features [49] of concatenated shape and appearance information have been extracted. The shape features (1) are based solely upon the lip shape and positioning during the speaker speaking e.g. the landmarks in Figure 4 (right) where there are 76 landmarks in the full face (left) and 34 landmarks which are modeling the inner and outer lip contours.

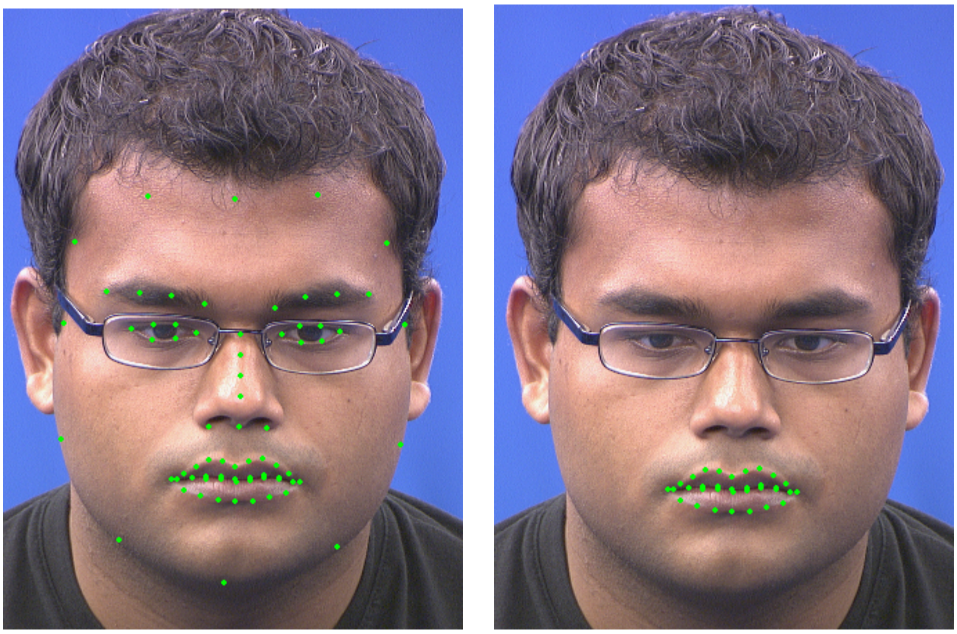

Figure 4: Example Active Appearance Model shape mesh (left), a lips only model is on the right. Landmarks are in green.

The landmark positions can be compactly represented using a linear model of the form:

$$s = s_0 + \sum_{i=1}^{m} s_i p_i \tag{1}$$

where s_0 is the mean shape and s_i are the modes. The appearance features are computed over pixels, the original images having been warped to the mean shape. So $A_0(x)$ is the mean appearance and appearance is described as a sum over modal appearances:

$$A(x) = A_0(x) + \sum_{i=1}^{l} \lambda_i A_i(x) \qquad \forall x \in S_0 \tag{2}$$

Combined features are the concatenation of Shape and Appearance, the AAM parameters of the four AVL2 speakers the twelve RMAV speakers are in Table 1.

Table 1: Number of parameters in shape, appearance and combined shape & appearance AAM features per speaker in AVL2 and RMAV

Speaker	Shape	Appearance	Concatenated
AVL2			
S1	11	27	38
S2	9	19	28
S3	9	17	25
S4	9	17	25
RMAV			
S1	13	46	59
S2	13	47	60
S3	13	43	56
S4	13	47	60
S5	13	45	58
S6	13	47	60
S7	13	37	50
S8	13	46	59
S9	13	45	58
S10	13	45	58
S11	14	72	86
S12	13	45	58

4. Method overview

We used the Bear phoneme clustering approach [31] to produce a series of speaker-dependent P2V maps.

In summary the clustering method is as follows:

1. Perform speaker-dependent phoneme recognition with recognisers that use phoneme labeled classifiers.
2. By aligning the phoneme output of the recogniser with the transcription of the word uttered, a confusion matrix for each speaker is produced detailing which phonemes are confused with which others.
3. Any phonemes which are only correctly recognised as themselves (true positive results) are permitted to be single-phoneme visemes.

4. The remaining phonemes are clustered into groups (visemes) based on the confusions identified in Step 2. Confusion is counted as the sum of both false positives ($FP = N\{p|\hat{p}\}$) and false negatives ($FN = N\{\hat{p}|p\}$), $\forall p \in P$. The clustering algorithm permits phonemes to be grouped into a single viseme, V only if each phoneme has been confused with all the others within V.
5. Consonant and vowel phonemes are not permitted to be mixed within a viseme class. Phonemes can only be grouped once. The result of this process is a P2V map M for each speaker. For further details, see [30].
6. These new speaker-dependent viseme sets are then used as units for visual speech recognition for a speaker.

We present an example to illustrate the results of the phoneme clustering method in Table 3 for the example confusion matrix in Figure 2 [30]. $/v01/$ is a single-phoneme viseme as it only has true positive results. $/v02/$ is a group of $/p1/$, $/p3/$, and $/p7/$ as these all have confusions with each other. Likewise for $/v03/$ which groups $/p2/$ and $/p4/$. Although $/p5/$ was confused with $/p4/$ it was not mixed with $/p2/$ at all so it remains a viseme class of its own, $/v04/$.

Table 2: Example confusion matrix showing confusions between phoneme-labeled classifiers to be used for clustering to create new speaker-dependent visemes from [31]. True positive classifications are shown in red, confusions of either false positives and false negatives are shown in blue. The estimated classes are listed horizontally and the real classes are vertical.

	$/p1/$	$/p2/$	$/p3/$	$/p4/$	$/p5/$	$/p6/$	$/p7/$
$/p1/$	1	0	0	0	0	0	4
$/p2/$	0	0	0	2	0	0	0
$/p3/$	1	0	0	0	0	0	1
$/p4/$	0	2	1	0	2	0	0
$/p5/$	3	0	1	1	1	0	0
$/p6/$	0	0	0	0	0	4	0
$/p7/$	1	0	3	0	0	0	1

Table 3: Example cluster P2V map

Viseme	Phonemes
$/v01/$	$\{/p6/\}$
$/v02/$	$\{/p1/, /p3/, /p7/\}$
$/v03/$	$\{/p2/, /p4/\}$
$/v04/$	$\{/p5/\}$

Our sets of P2V maps are made up of the following:

1. one multi-speaker P2V map using *all* speakers' phoneme confusions (per dataset); and for each speaker;
    2. a speaker-dependent P2V map;
    3. a speaker-independent P2V map using confusions of all *other* speakers in the data.

So we made nine P2V maps for AVL2 (four speaker maps for map types one and three, and one multi-speaker map) and 25 for RMAV (12 speaker maps for map types one and three, and one multi-speaker map). P2V maps were constructed using separate training and test data over cross-validation, seven folds for AVL2 and ten folds for RMAV [55].

With the HTK toolkit [56] we built HMM classifiers with the viseme classes in each P2V map. HMMs were flat-started with `HCompV` and re-estimated 11 times over (`HERest`). We classified using `HVite` and with the output of this we ran `HResults` to obtain scores. The HMMs each had three states each with an associated five-component Gaussian mixture to keep the results comparable to previous work [57].

To measure the performance of AVL2 speakers we noted that a classification network restricts the output to be one of the 26 letters of the alphabet (with the AVL2 dataset). Therefore, a simplified measure of accuracy in this case;

$$\frac{\#\text{ words correct}}{\#\text{ words classified}} \tag{3}$$

For RMAV a bigram word lattice was built with `HBuild` and `HLStats`, and performance is scored as Correctness (4),

$$C = \frac{N - D - S}{N} \tag{4}$$

where N is the total number of labels in the ground truth, D is the number of deletion errors, and S represents the number of substitution errors.

5. Experiment design

The P2V maps formed in these experiments are designated as:

$$M_n(p, q) \tag{5}$$

This means the P2V map is derived from speaker n, but trained using visual speech data from speaker p and tested using visual speech data from speaker q. For example, $M_1(2, 3)$ would designate the result of testing a P2V map constructed from Speaker 1, using data from Speaker 2 to train the viseme models, and testing on Speaker 3's data. Thus we will create (over both datasets); 16 P2V maps where $n = p = q$, two P2V maps where $n \neq p = q$, and 16 P2V maps where $n \neq p \neq q$. A total of 34 P2V maps.

For ease of reading, we provide in Table 4 a glossary of acronyms used to describe our testing methodology.

5.1. Baseline: Same Speaker-Dependent (SSD) maps

For a baseline we select the same speaker-dependent P2V maps as [31]. The baseline tests are: $M_1(1, 1)$, $M_2(2, 2)$, $M_3(3, 3)$ and $M_4(4, 4)$ (the four speakers in AVL2). Tests for RMAV are: $M_1(1, 1)$, $M_2(2, 2)$, $M_3(3, 3)$, $M_4(4, 4)$, $M_5(5, 5)$, $M_6(6, 6)$, $M_7(7, 7)$ and $M_8(8, 8)$, $M_9(9, 9)$, $M_{10}(10, 10)$, $M_{11}(11, 11)$ and $M_{12}(12, 12)$. These tests are Same Speaker-Dependent (SSD) because the same speaker is used to create the map, to train and test the models. Tables 5 depicts how these tests are constructed for AVL2 speakers, the premise is identical for the 12 RMAV speakers.

Table 4: Test method acronyms.

Acronym	Definition
SSD	Single speaker dependent
MS	Multi-speaker
DSD	Different-speaker dependent
DSD&D	Different-speaker dependent and Data
SI	Speaker-independent

Table 5: Same Speaker-Dependent (SSD) experiments for AVL2 speakers. The results from these tests will be used as a baseline.

Same speaker-dependent (SD)			
Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p, q)$
Sp1	Sp1	Sp1	$M_1(1, 1)$
Sp2	Sp2	Sp2	$M_2(2, 2)$
Sp3	Sp3	Sp3	$M_3(3, 3)$
Sp4	Sp4	Sp4	$M_4(4, 4)$

All P2V maps are listed in supplementary materials to this paper. We permit a garbage, $/gar/$, viseme which is a cluster of phonemes in the ground truth which did not appear at all in the output from the phoneme classification (step two of section 4). Every viseme is listed with its associated mutually-confused phonemes e.g. for AVL2 Speaker 1 SSD, M_1, we see $/v01/$ is made up of phonemes {/ʌ/, /iy/, /əʊ/, /uw/}. We know from the clustering method in [31] this means in the phoneme classification, all four phonemes {/ʌ/, /iy/, /əʊ/, /uw/} were confused with the other three in the viseme. We are using the 'strictly-confused' method labeled $B2$ from [30] with split vowel and consonant groupings as these achieved the highest accurate word classification.

5.2. Multi-Speaker (MS) maps

A multi-speaker (MS) P2V map forms the viseme classifier labels in the first set of experiments. This map is constructed using phoneme confusions produced by *all* speakers in each data set. Again, these P2V maps are in the supplementary material.

For the multi-speaker experiment notation, we substitute in the word 'all' in place of a list of all the speakers for ease of reading. Therefore, the AVL2 MS map is tested as follows: $M_{[all]}(1, 1)$, $M_{[all]}(2, 2)$, $M_{[all]}(3, 3)$ and $M_{[all]}(4, 4)$: this is explained in Table 6 and the RMAV MS map is tested as: $M_{[all]}(1, 1)$, $M_{[all]}(2, 2)$, $M_{[all]}(3, 3)$, $M_{[all]}(4, 4)$, $M_{all]}(5, 5)$, $M_{[all]}(6, 6)$, $M_{[all]}(7, 7)$, $M_{[all]}(8, 8)$, $M_{[all]}(9, 9)$, $M_{[all]}(10, 10)$, $M_{[all]}(11, 11)$, $M_{[all]}(12, 12)$.

5.3. Different Speaker-Dependent maps & Data (DSD&D)

The second set of tests within this experiment start to look at using P2V maps with different test speakers. This means the HMM classifiers trained on each single speaker are used to recognise data from alternative speakers.

Within AVL2 this is completed for all four speakers using the P2V maps of the other speakers, and the data from the other speakers. Hence for Speaker 1 we construct

Table 6: Multi-Speaker (MS) experiments for AVL2 speakers.

Multi-Speaker (MS)			
Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp[all]	Sp1	Sp1	$M_{[all]}(1,1)$
Sp[all]	Sp2	Sp2	$M_{[all]}(2,2)$
Sp[all]	Sp3	Sp3	$M_{[all]}(3,3)$
Sp[all]	Sp4	Sp4	$M_{[all]}(4,4)$

Table 7: Different Speaker-Dependent maps and Data (DSD&D) experiments with the four AVL2 speakers.

Different Speaker-Dependent maps & Data (DSD&D)			
Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp2	Sp2	Sp1	$M_2(2,1)$
Sp3	Sp3	Sp1	$M_3(3,1)$
Sp4	Sp4	Sp1	$M_4(4,1)$
Sp1	Sp1	Sp2	$M_1(1,2)$
Sp3	Sp3	Sp2	$M_3(3,2)$
Sp4	Sp4	Sp2	$M_4(4,2)$
Sp1	Sp1	Sp3	$M_1(1,3)$
Sp2	Sp2	Sp3	$M_2(2,3)$
Sp4	Sp4	Sp3	$M_4(4,3)$
Sp1	Sp1	Sp4	$M_1(1,4)$
Sp2	Sp2	Sp4	$M_2(2,4)$
Sp3	Sp3	Sp4	$M_3(3,4)$

$M_2(2,1)$, $M_3(3,1)$ and $M_4(4,1)$ and so on for the other speakers, this is depicted in Table 7.

For the RMAV speakers, we undertake this for all 12 speakers using the maps of the 11 others. In this set of tests we are replicating the format of [21] where $p \neq q$ but we use speaker-dependent visemes to mitigate the effect of speaker independence between training and test data.

5.4. Different Speaker-Dependent maps (DSD)

Now we wish to isolate the effects of the HMM classifier from the effect of using different speaker dependent P2V maps by training the classifiers on single speakers with the labels of the alternative speaker P2V maps. E.g. for AVL2 Speaker 1, the tests are: $M_2(1,1)$, $M_3(1,1)$ and $M_4(1,1)$. (All tests are listed in Table 8).

These are the same P2V maps as in our SSD baseline but trained and tested differently.

5.5. Speaker-Independent maps (SI)

Finally, the last set of tests looks at speaker independence in P2V maps. Here we use maps which are derived using all speakers confusions bar the test speaker. This time we substitute the symbol '$\neg x$' in place of a list of speaker identifying numbers, meaning

Table 8: Different Speaker-Dependent maps (DSD) experiments for AVL2 speakers.

Different Speaker-Dependent maps (DSD)			
Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp2	Sp1	Sp1	$M_2(1,1)$
Sp3	Sp1	Sp1	$M_3(1,1)$
Sp4	Sp1	Sp1	$M_4(1,1)$
Sp1	Sp2	Sp2	$M_1(2,2)$
Sp3	Sp2	Sp2	$M_3(2,2)$
Sp4	Sp2	Sp2	$M_4(2,2)$
Sp1	Sp3	Sp3	$M_1(3,3)$
Sp2	Sp3	Sp3	$M_2(3,3)$
Sp4	Sp3	Sp3	$M_4(3,3)$
Sp1	Sp4	Sp4	$M_1(4,4)$
Sp2	Sp4	Sp4	$M_2(4,4)$
Sp3	Sp4	Sp4	$M_3(4,4)$

'not including speaker x'. The tests for these maps are as follows $M_{\neg 1}(1,1)$, $M_{\neg 2}(2,2)$, $M_{\neg 3}(3,3)$ and $M_{\neg 4}(4,4)$ as shown in Table 9 for AVL2 speakers. Speaker independent P2V maps for all speakers are in this papers supplementary materials.

Table 9: Speaker-Independent (SI) experiments with AVL2 speakers.

Speaker-Independent (SI)			
Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp¬1	Sp1	Sp1	$M_{\neg 1}(1,1)$
Sp¬2	Sp2	Sp2	$M_{\neg 2}(2,2)$
Sp¬3	Sp3	Sp3	$M_{\neg 3}(3,3)$
Sp¬4	Sp4	Sp4	$M_{\neg 4}(4,4)$

6. Measuring the effects of homophenes

Bauman [58] suggests we make 13-15 motions per second during normal speech but are only able to pick up eight or nine. Bauman defines these motions which are so visually similar for distinct words they can only be differentiated with acoustic help as homophenes. For example, in the AVL2 data the words are the letters of the alphabet, The phonetic translation of the word 'B' is '$/b//iy/$' and of 'D' is '$/d//iy/$'. Using $M_2(2,2)$ to translate these into visemes they are identical '$/v08//v01/$'.

Permitting variations in pronunciation, the total number of T tokens (each unique word counts as one token) for each map after each word has been translated to speaker-dependent visemes are listed in Tables 10 and 11. More homophenes means a greater the chance of substitution errors and a reduced correct classification. We calculate the homophene effect, H, as measured in (6). Where T is the number of tokens (unique words) and W is the number of total words available in a single speaker's ground truth

Table 10: Count of homophenes per P2V map

SD Maps		SI Maps	
Map	Tokens T	Map	Tokens
M_1	19	$M_{!1}$	17
M_2	19	$M_{!2}$	18
M_3	24	$M_{!3}$	20
M_4	24	$M_{!4}$	15
$M_{[all]}$	14		

transcriptions.

$$H = 1 - \frac{T}{W} \tag{6}$$

An example of a homophene are the words 'talk' and 'dog'. If one uses Jeffers visemes, both of these words transcribed into visemes become '$/C/$ $/V1/$ $/H/$' meaning that recognition of this sequence of visemes, will represent what acoustically are two very distinct words. Thus distinguishing between 'talk' and 'dog' is impossible, without the use side information such as a word lattice. This is the power of the word network [59, 60].

Table 11: Homophenes, H in words, phonemes, and visemes for RMAV

Speaker	Word	Phoneme	SD Visemes
Sp01	0.64157	0.64343	0.70131
Sp02	0.72142	0.72309	0.76693
Sp03	0.67934	0.68048	0.73950
Sp04	0.68675	0.68916	0.74337
Sp05	0.48018	0.48385	0.58517
Sp06	0.69547	0.69726	0.74791
Sp07	0.69416	0.69607	0.74556
Sp08	0.69503	0.69752	0.74907
Sp09	0.68153	0.68280	0.73439
Sp10	0.70146	0.70328	0.75243
Sp11	0.70291	0.70499	0.75623
Sp12	0.63651	0.64317	0.70699

7. Analysis of speaker independence in P2V maps

Figure 5 shows the correctness of both the MS viseme set (in blue) and the SI tests (in orange) (Tables 6 and 9) against the SSD baseline (red) for AVL2 speakers. Word correctness, C is plotted on the y-axis. For the MS classifiers, these are all built on the same map M_{all}, trained and tested on the same single speaker so, $p = q$. Therefore the tests are: $M_{all}(1,1)$, $M_{all}(2,2)$, $M_{all}(3,3)$, $M_{all}(4,4)$. To test the SI maps, we plot $M_{!1}(1,1)$, $M_{!2}(2,2)$, $M_{!3}(3,3)$ and $M_{!4}(4,4)$. The SSD baseline is on the left of each speakers section of the figure. Note that guessing would give a correctness of $1/N$, where N is the total number of words in the dataset. For AVL2 this is 26, for RMAV speakers this ranges between 1362 and 1802).

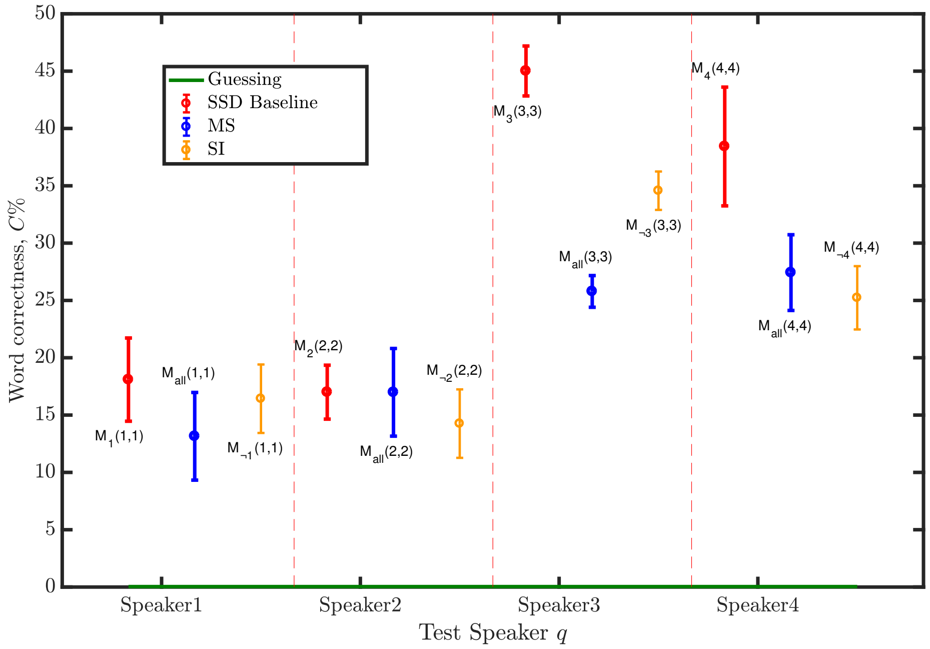

Figure 5: Word correctness, $C \pm 1$s.e., using MS and SI P2V maps AVL2

There is no significant difference on Speaker 2, and while Speaker 3 word classification is reduced, it is not eradicated. It is interesting for Speaker 3, for whom their speaker-dependent classification was the best of all speakers, the SI map ($M_{!3}$) out performs the multi-speaker viseme classes (M_{all}) significantly. This maybe due to Speaker 3 having a unique visual talking style which reduces similarities with Speakers 1, 2 & 4. But more likely, we see the $/iy/$, phoneme is not classified into a viseme in M_3, whereas it is in M_1, M_2 & M_4 and so re-appears in M_{all}. Phoneme $/iy/$ is the most common phoneme in the AVL2 data. This suggests it may be best to avoid high volume of phonemes for deriving visemes as we are exploiting speaker individuality to make better viseme classes.

We have plotted the same MS & SI experiments on RMAV speakers in Figures 6 and 7 (six speakers in each figure). In continuous speech, all but Speaker 2 are significantly negatively affected by using generalized multi-speaker visemes, whether the visemes include the test speakers phoneme confusions or not. This reinforces knowledge of the dependency on speaker identity in machine lipreading but we do see the scale of this effect depends on which two speakers are being compared. For the exception speaker (Speaker 2 in Figure 6) there is only a insignificant decrease in correctness when using MS and SI visemes. Therefore an optimistic view suggests it could be possible with making multi-speaker visemes based upon groupings of visually similar speakers, even better visemes could be created. The challenge remains in knowing which speakers should be grouped together before undertaking P2V map derivation.

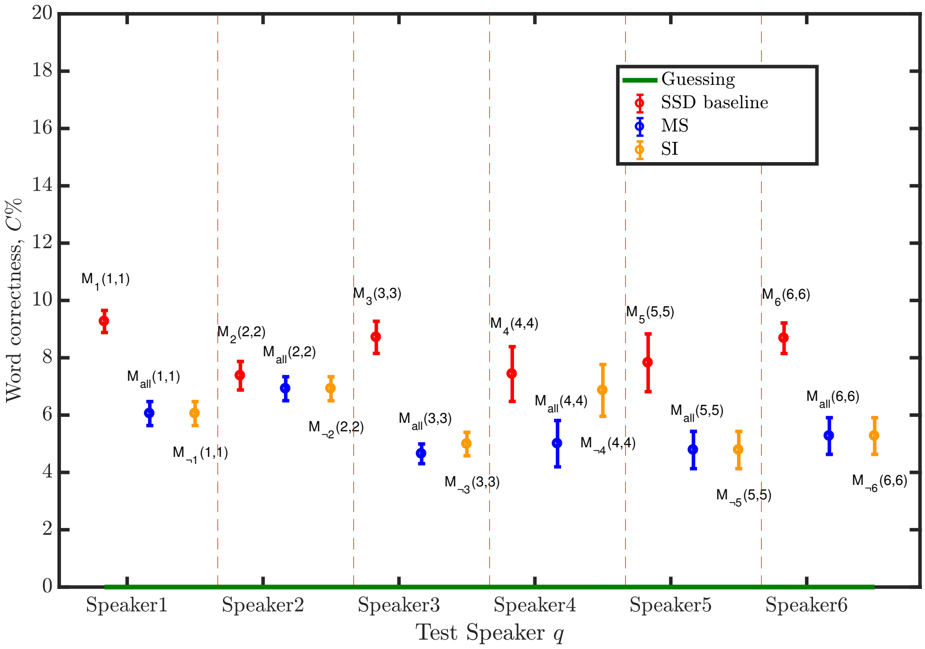

Figure 6: Word correctness, $C \pm 1$s.e., using RMAV speakers 1-6 MS and SI P2V maps

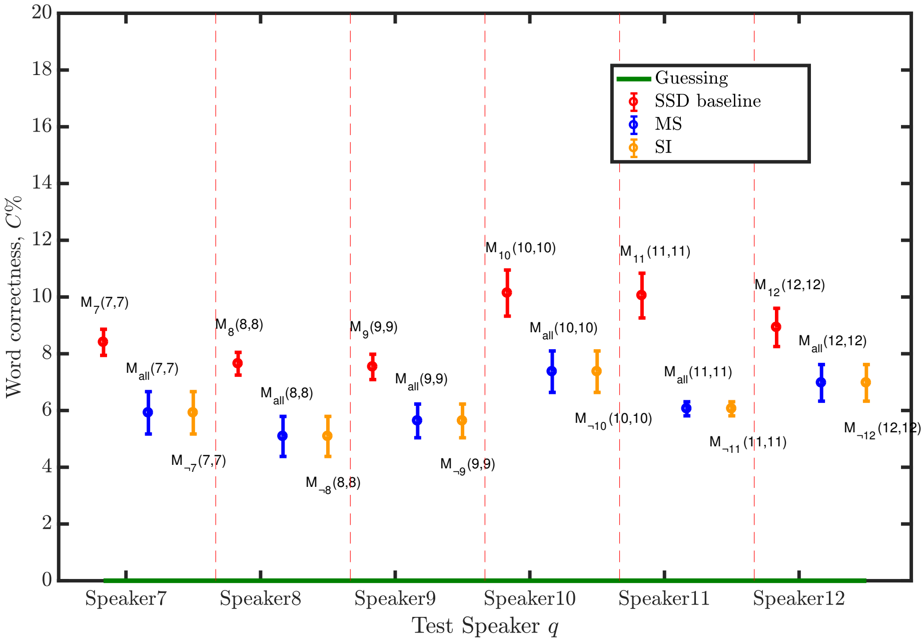

Figure 7: Word correctness, $C \pm 1$s.e. using RMAV speakers 7-12 MS and SI P2V maps

7.1. Different Speaker-Dependent& Data (DS&D) results

Figure 8 shows the word correctness of AVL2 speaker-dependent viseme classes on the y-axis. Again in this figure, the baseline is $n = p = q$ for all M. These are compared to the DSD&D tests: $M_2(2,1)$, $M_3(3,1)$, $M_4(4,1)$ for Speaker 1, $M_1(1,2)$, $M_3(3,2)$, $M_4(4,2)$ for Speaker 2, $M_1(1,3)$, $M_2(2,3)$, $M_4(4,3)$ for Speaker 3 and $M_1(1,4)$, $M_2(2,4)$, $M_3(3,4)$ for Speaker 4 as in Table 7.

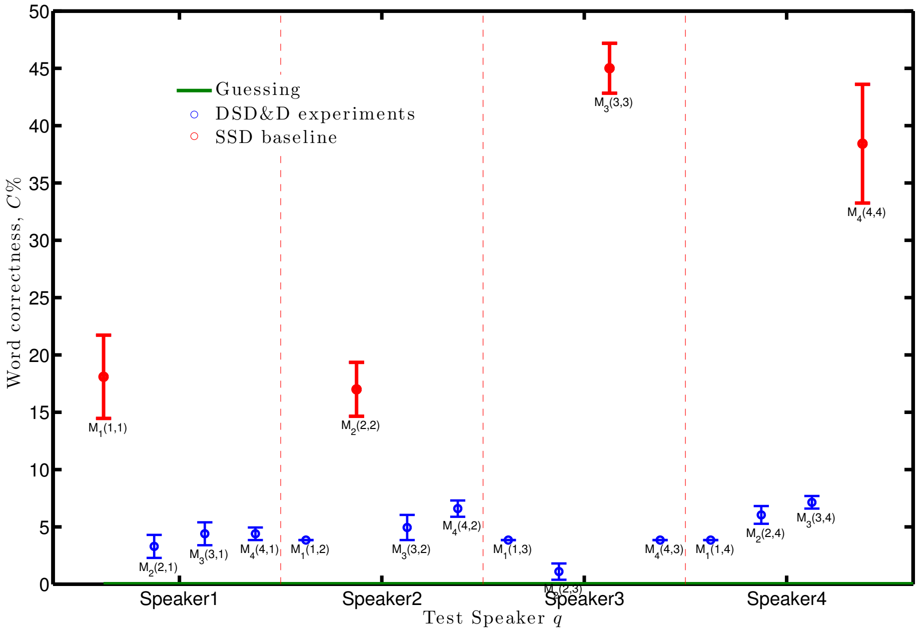

Figure 8: Word correctness, $C \pm 1$s.e., of the AVL2 DSD&D tests Baseline is SSD tests in red.

For isolated word classification, DSD&D HMM classifiers are significantly worse than SSD HMMs, as all results where p is not the same speaker as q are around the equivalent performance of guessing. This correlates with similar tests of independent HMMs in [33]. This gap is attributed to two possible effects, either – the visual units are incorrect, or they are trained on the incorrect speaker. Figures 9, 10, 11, & 12 show the same tests but on the continuous speech data, we have plotted three test speakers per figure.

As expected some speakers significantly deteriorate the classification rates when the speaker used to train the classifier is not the same as the test speaker ($p \neq q$). As an example we look at Speaker 1 on the leftmost side of Figure 9 where we have plotted Word Correctness for the DSD&D tests. Here the test speaker is Speaker 1. The speaker-dependent maps for all 12 speakers have been used to build HMMs classifiers and tested on speaker 1. All C_w for P2V maps significantly reduces except that trained on speaker one. However, in comparison to the AVL2 results, – this reduction in C_w is not as low as guessing. By capturing language in speaker dependent sets of visemes, we are now less dependent on the speaker identity in the training data. This suggestion is supported by the knowledge of how much of conventional lip reading systems accuracies came from the language model.

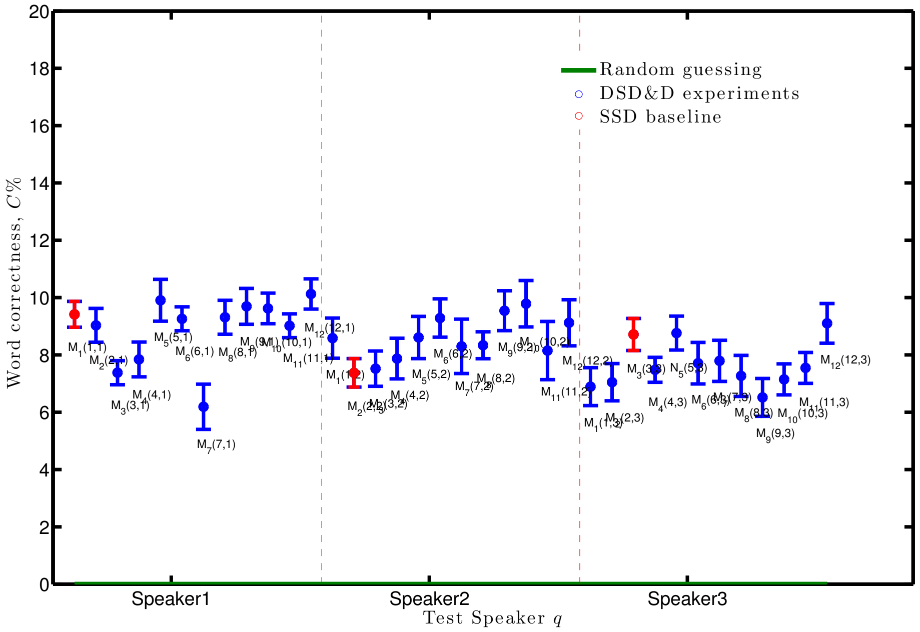

Figure 9: Word correctness, $C \pm 1$s.e., of the RMAV speakers 1-3 DSD&D tests. SSD baseline in red

Looking at Figures 10 to 12 these patterns are consistent. The exception is speaker 2 in Figure 9 where we see that by using the map of speaker 10, M_{10} we do not experience a significant decrease in C_w. Furthermore, if we look at Speaker 10's results in Figure 12, all other P2V maps negatively affect speaker 10's C_w. This suggest that adaptation between speakers may be directional, that is, we could lipread Speaker 2 having trained on Speaker 10, but not vice versa.

7.1.1. Continuous speech gestures, or isolated word gestures?

If we compare these figures to the isolated words results [21], either the extra data in this larger data set or the longer sentences in continuous speech have made a difference. Table 12 lists the differences for all speakers on both datasets and the difference between isolated words and continuous speech is between 3.83% to 37.74%. Furthermore, with isolated words, the performance attained by speaker-independent tests was shown in cases to be worse than guessing. Whilst the poorest P2V maps might be low, they are all significantly better than guessing regardless of the test speakers.

Table 12: Correctness C with AVL2 and RMAV speakers

AVL2	Sp1	Sp2	Sp3	Sp4
	14.06	11.87	42.08	32.75

RMAV	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6	Sp7	Sp8	Sp9	Sp10	Sp11	Sp12
	5.78	4.74	6.49	5.13	5.57	4.92	6.60	5.19	5.64	7.03	7.49	8.04

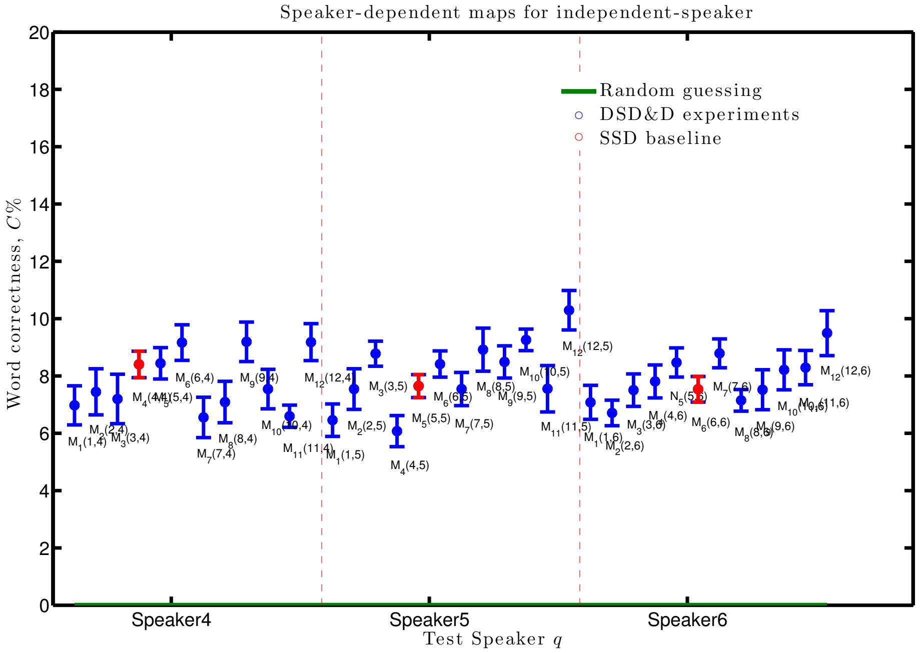

Figure 10: Word correctness, $C \pm 1$s.e., of the RMAV speakers 4-6 DSD&D tests. SSD baseline in red

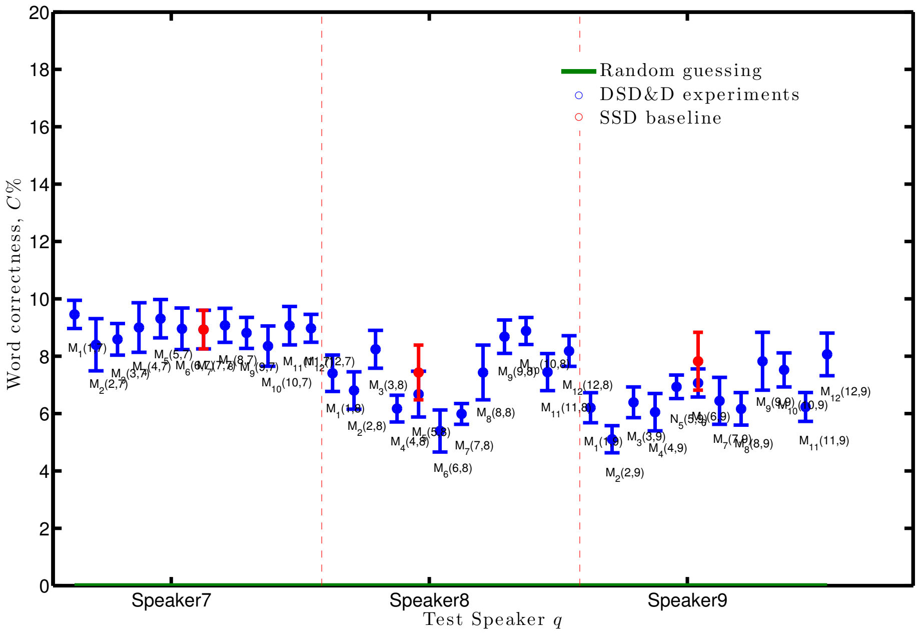

Figure 11: Word correctness, $C \pm 1$s.e., of the RMAV speakers 7-9 DSD&D tests. SSD baseline in red

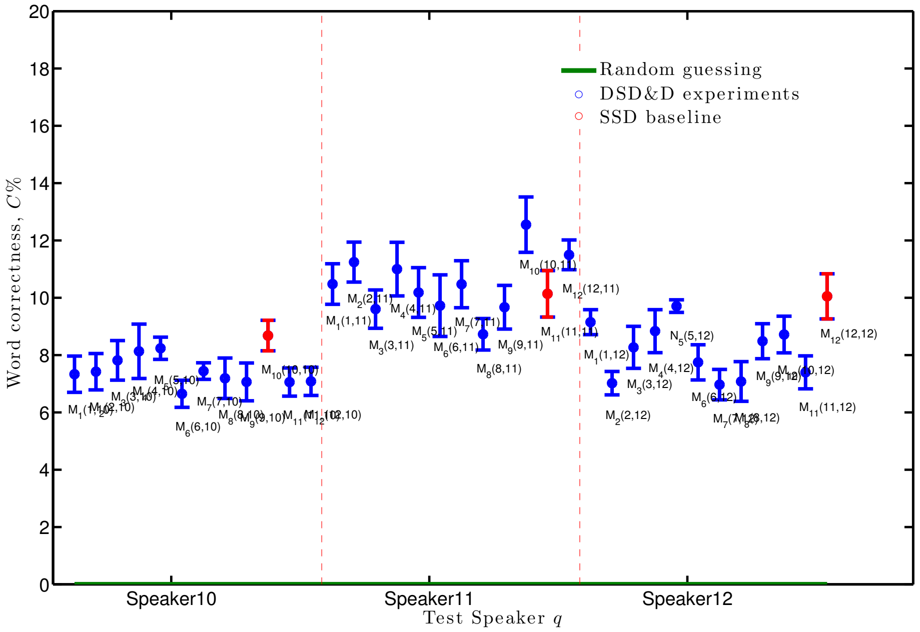

Figure 12: Word correctness, $C \pm 1$s.e., of the RMAV speakers 10-12 DSD&D tests. SSD baseline in red

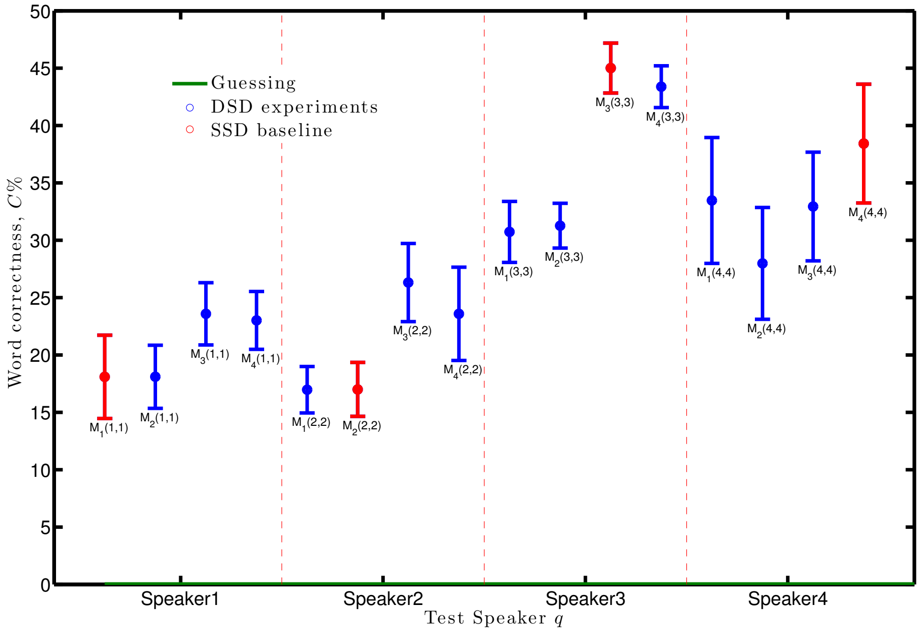

Figure 13: Word correctness, $C \pm 1$s.e., of the AVL2 DSD tests. SSD baseline in red

Figure 13 shows the AVL2 DSD experiments from Table 8. In the DSD tests, the classifier is allowed to be trained on the relevant speaker, so the other tests are: $M_2(1,1)$, $M_3(1,1)$, $M_4(1,1)$ for Speaker 1, $M_1(2,2)$, $M_3(2,2)$, $M_4(2,2)$ for Speaker 2, $M_1(3,3)$, $M_2(3,3)$, $M_4(3,3)$ for Speaker 3 and finally $M_1(4,4)$, $M_2(4,4)$, $M_3(4,4)$ for Speaker 4. Now the word correctness has improved substantially which implies the previous poor performance in Figure 8 was not due to the choice of visemes but rather, the speaker-specific HMMs. The equivalent graphs for the 12 RMAV speakers are in Figures 14, 15, 16 and 17.

With continuous speech we can see the effects of unit selection. Using Speaker 1 for example, in Figure 14 the three maps M_3, M_7 and M_{12} all significantly reduce the correctness for Speaker 1. In contrast, for Speaker 2 there are no significantly reducing maps but maps M_1, M_4, M_5, M_6, M_9, and M_{11} all significantly improve the classification of Speaker 2. This suggests that it is not just the speakers' identity which is important for good classification but how it is used. Some individuals may simply be easier to lip read or there are similarities between certain speakers which when learned on one speaker are able to better classify the visual distinctions between phonemes on similar other speakers.

In Figure 16 we see Speaker 7 is particularly robust to visual unit selection for the classifier labels. Conversely Speakers 5 (Figure 15) and 12 (Figure 17) are really affected by the visemes (or phoneme clusters). Its interesting to note this is a variability not previously considered, some speakers may be dependent on good visual classifiers and the mapping back to acoustics utterances, but others not so much.

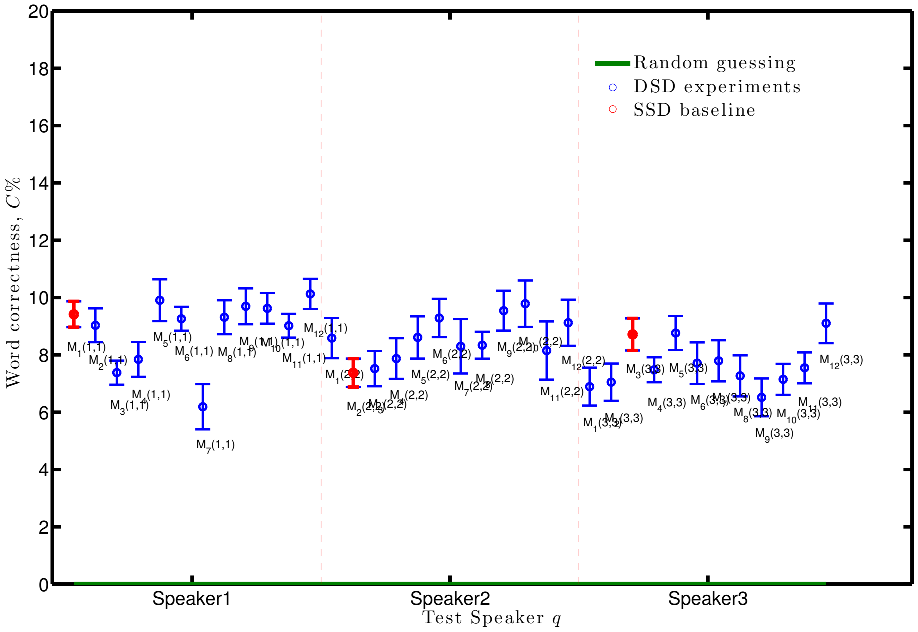

Figure 14: Word correctness, $C \pm 1$s.e., of the RMAV speakers 1-3 DSD tests. SSD baseline in red

Figure 18 shows the mean word correctness of the DSD classifiers per speaker in RMAV. The y-axis shows the % word correctness and the x-axis is a speaker per point. We also plot random guessing and error bars of one standard error over the ten fold mean. Speaker 11 is the best performing speaker irrespective of the P2V selected. All speakers have a similar standard error but a low mean within this bound. This suggests subject to speaker similarity, there is more possibility to improve classification correctness with another speakers visemes (if they include the original speakers visual cues) than to use weaker self-clustered visemes.

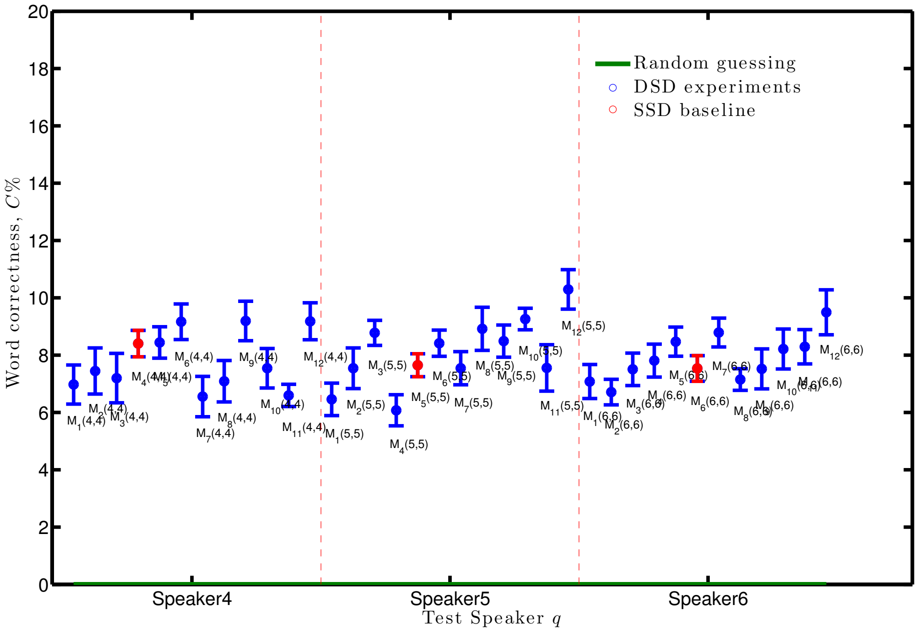

Figure 15: Word correctness, $C \pm 1$s.e., of the RMAV speakers 4-6 DSD tests. SSD baseline in red

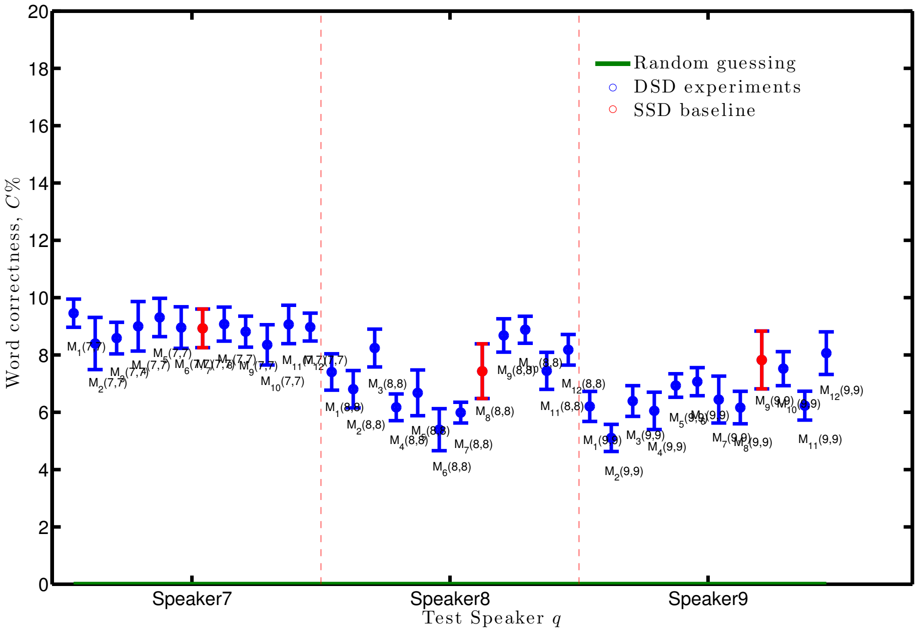

Figure 16: Word correctness, $C \pm 1$s.e., of the RMAV speakers 7-9 DSD tests. SSD baseline in red

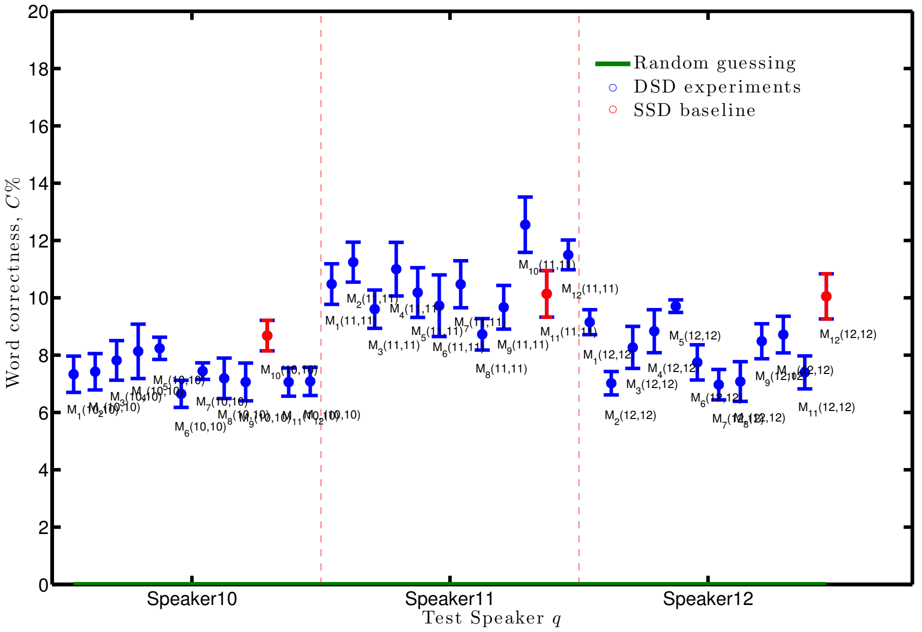

Figure 17: Word correctness, $C \pm 1$s.e., of the RMAV speakers 10-12 DSD tests. SSD baseline in red

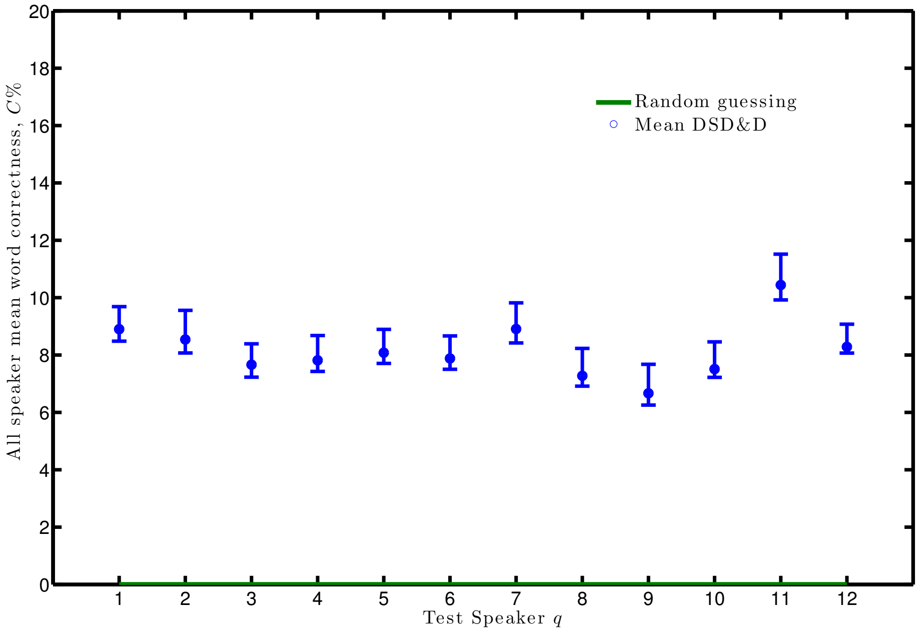

Figure 18: All-speaker mean word correctness, $C \pm 1$standard error of the DSD tests

7.3. Weighting the M_n effect on other speakers

To summarize the performance of DSD versus SSD we use scores. If DSD exceeds SSD by more than one standard error we score +2, or −2 if it is below. The scores ±1 indicate differences within the standard error. The scores are shown in Tables 13 and 14. M_3 scores the highest of the four AVL2 SSD maps, followed by M_4, M_2 and finally M_1 is the most susceptible to speaker identity in AVL2. It seems that the more similar to phoneme classes the visemes are, then the better the classification performance. This is consistent with Table 10, where the larger P2V maps create fewer homophones [61]

Table 13: Weighted ranking scores from comparing the use of speaker-dependent maps for *other* AVL2 speakers

	M_1	M_2	M_3	M_4
Sp01	0	+1	+2	+2
Sp02	−1	0	+2	+1
Sp03	−2	−2	0	−1
Sp04	−1	+1	−1	0
Total	−4	0	**+3**	**+2**

Table 14: Weighted scores from comparing the use of speaker-dependent maps for *other* speaker lipreading in continuous speech (RMAV speakers).

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
Num of visemes	16	14	16	15	18	16	16	14	19	15	15	13
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**

In Table 7 of our supplementary material, we list the AVL2 speaker-dependent P2V maps. The phoneme pairs {/ə/, $/eh/$}, {$/m/$, $/n/$} and {$/ey/$, $/iy/$} are present for three speakers and {/ʌ/, $/iy/$} and {$/l/$, $/m/$} are pairs for two speakers. Of the single-phoneme visemes, {/tʃ/} is presented three times, {$/f/$}, {$/k/$}, {$/w/$} and {$/z/$} twice. We learn from Figure 13 that the selection of incorrect units, whilst detrimental, is not as bad as training on alternative speakers.

Table 14 shows the scores for the 12 RMAV speakers. The speaker dependent map of Speaker 12 (right column) is one of only two (M_{12} and M_5) which make an overall improvement on other speakers classification (they have positive values in the total row at the bottom of Table 14), and crucially, M_{12} only has one speaker (Speaker 10) for whom the visemes in M_{12} do not make an improvement in classification. The one other speaker P2V map which improves over other speakers is M_5. All others show a negative effect, this reinforces the observation that visual speech is dependent upon the individual but we also now have evidence there are exceptions to the rule. Table 14 also lists the number of visemes within each set. All speaker-dependent sets are within the optimal range of 11 to 35 illustrated in [62].

8. Speaker independence between sets of visemes

For isolated word classification the main conclusion of this section is shown by comparing Figures 13 & 5 with Figure 8. The reduction in performance in Figure 8 is when the system classification models are trained on a speaker who is not the test speaker. This raised the question if this this degradation was due to the wrong choice of P2V map or speaker identity mismatch between the training and test data samples. We have concluded that, whilst the wrong unit labels are not conducive for good lipreading classification, is it not the choice of P2V map which causes significant degradation but rather the speaker identity. This regain of performance is irrespective of whether the map is chosen for a different speaker, multi-speaker or independently of the speaker.

This observation is important as it tells us the repertoire of visual units across speakers does not vary significantly. This is comforting since the prospect of classification using a symbol alphabet which varies by speaker is daunting. There are differences between speakers, but not significant ones. However, we have seen some exceptions within the continuous speech speakers whereby the effect of the P2V map selection is more prominent and where sharing HMMs trained on non-test speakers has not been completely detrimental. This gives some hope with similar visual speakers, and with more 'good' training data speaker independence, whether by classifier or viseme selection, might be possible.

To provide an analogy; in acoustic speech we could ask if an accented Norfolk speaker requires a different set of phonemes to a standard British speaker? The answer is no. They are represented by the same set of phonemes; but due to their individuality they use these phonemes in a different way.

Comparing the multi-speaker and SI maps, there are 11-12 visemes per set whereas in the single-speaker-dependent maps we have a range of 12 to 17. It is M_3 with 17 visemes, which out performs all other P2V maps. So we can conclude, there is a high risk of over-generalising a speaker-dependent P2V map when attempting multi-speaker or speaker-independent P2V mappings as we have seen with the RMAV experiments.

Therefore we must consider it is not just the speaker-dependency which varies but also the contribution of each viseme within the set which also contributes to the word classification performance, an idea first shown in [5]. Here we have highlighted some phonemes which are a good subset of potentially independent visemes $\{/ə/, /eh/\}$, $\{/m/, /n/\}$ and $\{/ey/, /iy/\}$, and what these results present, is a combination of certain phoneme groups combined with some speaker-dependent visemes, where the latter provide a lower contribution to the overall classification would improve speaker-independent maps with speaker-dependent visual classifiers.

It is often said in machine lipreading there is high variability between speakers. This should now be clarified to state there is not a high variability of visual cues given a language, but there is high variability in trajectory between visual cues of an individual speakers with the same ground truth. In continuous speech we have seen how not just speaker identity affects the visemes (phoneme clusters) but also how the robustness of each speakers classification varies in response to changes in the viseme sets used. This implies a dependency upon the number of visemes within each set for individuals so this is what we investigate in the next section.

Due to the many-to-one relationship in traditional mappings of phonemes to visemes, any resulting set of visemes will always be smaller than the set of phonemes. We know a benefit of this is more training samples per class which compensates for the limited data in currently available datasets but the disadvantage is generalization between different articulated sounds. To find an optimal set of viseme classes, we need to minimize the generalization to maintain good classification but also to maximize the training data available.

9. Distance measurements between sets of heterogeneous visemes

Our statistical measure is the Wilcoxon signed rank test [32]. Our intent is to move towards a distance measurement between the visual speech information for each speaker. We use a non-parametric method as we can not make assumptions about the distributions of the data, the individual P2V mappings re-distribute the data samples.

The signed rank test a non-parametric method which uses paired samples of values, to rank the population means of each pair-value. The sum of the signed ranks, W, is compared to the significance value. We use $\rho = 0.05$ for a 95% confidence interval to determine significance, p. If $W < \rho$ then $p = 1$ else $p = 0$. The null hypothesis is there is no difference between the paired samples. In our case, this means that the speaker variation (represented in P2V maps) is not significant. In finding speakers who are significantly different, we hope to identify speakers who will be easier to adapt features between due to similarity in lip trajectory during speech.

To compare the distances between the speaker-dependent P2V mappings, we use the Wilcoxon signed rank test which allows non-parametric pairwise comparison of speaker mean word correctness scores. Table 15 is the signed ranks r. Scores are underlined where the respective significance $\rho = 1$. The respective continuous speech comparison is in Table 16. Both tables are presented as a confusion matrix to compare all speakers with all others. The on-diagonal is always $r = 1$ (in Tables 15 & 16), this confirms speakers are identical when paired with themselves.

Table 15: Wilcoxon Signed Rank, r, for the AVL2 speakers

	Sp01	Sp02	Sp03	Sp04
M_1	1.000	0.844	0.016	0.031
M_2	0.844	1.000	0.016	0.016
M_3	0.016	0.016	1.000	0.625
M_4	0.031	0.016	0.625	1.000

In Table 15 we see an immediate split in the speakers. We can group speakers 1 and 2 together, and separately group speaker 3 with speaker 4. The similarity between speaker 1 and 2 ($r = 0.844$) is greater than between speakers 3 and four ($r = 0.625$). It is interesting that with a small dataset and a simple language model, there are clear distinctions between some speakers.

Table 16: Wilcoxon signed rank, r, for the RMAV speakers

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
M_1	1.000	$\underline{0.037}$	0.695	0.160	0.084	$\underline{0.020}$	0.275	0.193	0.193	0.375	0.508	0.275
M_2	$\underline{0.037}$	1.000	0.084	$\underline{0.037}$	1.000	0.922	0.084	1.000	0.625	0.064	$\underline{0.037}$	$\underline{0.020}$
M_3	0.695	0.084	1.000	0.922	0.232	0.160	0.770	0.432	0.492	0.846	0.193	0.322
M_4	0.160	$\underline{0.037}$	0.922	1.000	0.322	0.232	0.492	0.432	0.334	0.922	0.105	0.105
M_5	0.084	1.000	0.232	0.322	1.000	1.000	0.275	1.000	1.000	0.131	$\underline{0.037}$	0.064
M_6	$\underline{0.020}$	0.922	0.160	0.232	1.000	1.000	0.193	1.000	1.000	0.152	0.064	0.064
M_7	0.275	0.084	0.770	0.492	0.275	0.193	1.000	0.275	0.375	0.770	0.375	0.232
M_8	0.193	1.000	0.432	0.432	1.000	1.000	0.275	1.000	0.922	0.232	$\underline{0.025}$	0.160
M_9	0.193	0.625	0.492	0.334	1.000	1.000	0.375	0.922	1.000	0.322	0.084	0.232
M_{10}	0.375	0.064	0.846	0.922	0.131	0.152	0.770	0.232	0.322	1.000	0.322	0.232
M_{11}	0.508	$\underline{0.037}$	0.193	0.105	$\underline{0.037}$	0.064	0.375	$\underline{0.025}$	0.084	0.322	1.000	0.770
M_{12}	0.275	$\underline{0.020}$	0.322	0.105	0.064	0.064	0.232	0.160	0.232	0.232	0.770	1.000

Table 16 is the respective analysis for the RMAV speakers, these results are not clear cut. Four of the RMAV speakers are not significantly different from all others others, these are speakers 3, 7, 9, and 10. The significantly different speaker pairs are:

- M_1, M_2
- M_1, M_6
- M_2, M_4
- M_2, M_{11}
- M_2, M_{12}
- M_5, M_{11}
- M_{11}, M_{11}

This observation reinforces the notion that some individual speakers have unique trajectories between visemes to make up their own visual speech signal, and idea first presented in [5], but here, others speakers (3, 7, 9, and 10) demonstrate a generalized pattern of visual speech units.

We postulate that these four speakers could be more useful for speaker independent systems as generalizing from them is within a small data space. Also, adapting features between the other speakers would be more challenging as they have a greater distance between them. It is also possible that speaker adaptation may be complicated with our observation in section 7.1, that adaption between speakers could be directional. For example, if we look at speakers 1 and 2 from RMAV, we know they are significantly distinct (Table 16) but, if we also reference the effect of the P2V maps of these speakers in Table 14, the visemes of speaker two insignificantly reduces the mean classification of speaker one whereas the visemes of speaker one significantly increases the mean classification of speaker two. This means that for this pair of speakers we prefer the visemes of speaker one. But this is not consistent for all significantly different visual speakers. Speaker pair 1 and 6 demonstrated both speakers classified more accurately with their own speaker-dependent visemes. This shows the complexity at the nub of speaker-independent lipreading systems for recognizing patterns of visual speech units, the units themselves are variable.

10. Conclusions

By comparing Figure 5 with Figure 8 we show a substantial reduction in performance when the system is trained on non-test speakers. The question arises as to whether this degradation is due to the wrong choice of map or the wrong training data for the recognisers. We conclude that it is not the choice of map that causes degradation since we can retrain the HMMs and regain much of the performance. We regain performance irrespective of whether the map is chosen for a different speaker, multi-speaker or independently of the speaker.

The sizes of the MS and SI maps built on continuous speech are fairly consistent, at most only ±2 visemes per set. Whereas the SSD maps have a size range of six. We conclude there is high risk of over-generalizing a MS/SI P2V map. It is not only the speaker-dependency that varies but also the contribution of each viseme within the set which affects the word classification performance, an idea also shown in [5]. This suggests that a combination of certain MS visemes with some SD visemes would improve speaker-independent lipreading. We have shown exceptions where the P2V map choice is significant and where HMMs trained on non-test speakers has not been detrimental. This is evidence that with visually similar speakers, speaker-independent lipreading is probable. Furthermore, with continuous speech, we have shown that speaker dependent P2V maps significantly improve lipreading over isolated words. We attribute this to the co-articulation effects of visual speech on phoneme recognition confusions which in turn influences the speaker-dependent maps with linguistic or context information. This is supported by evidence from conventional lipreading systems which show the strength of language models in lipreading accuracy.

We provide more evidence that speaker independence, even with unique trajectories between visemes for individual speakers, is likely to be achievable. What we need now is more understanding of the influence of language on visual gestures. What is in common, is the language between speakers. What we are seeking is an understanding of how language creates the gestures captured in visual speech features.

We can address lipreading dependency on training speakers by generalizing to those speakers who are visually similar in viseme usage/trajectory through gestures. This is consistent with recent deep learning training methods. However here, we show that we should not need the big data volumes to do this generalization and presented evidence that adaptation between speakers may be directional meaning we can recognise speaker A from speaker B data, but not vice versa.

These are important conclusions because with the widespread adoption of deep learning and big data available, we trade-off data volumes and training time for improved accuracy. We have shown that if we can find a finite number of individuals whose visual speech gestures are similar enough to cover the whole test population, one could train on this much smaller data set for comparable results to lipreading big data.

We have measured the distances/similarity between different speaker-dependent sets of visemes and shown there is minimal significant correlation supporting prior evidence about speaker heterogeneity in visual speech. However, these distances are variable and require further investigation.

Our conclusion that it is the use, or trajectory of visemes, rather than the visemes themselves which vary by speaker suggests that there might be alternative approaches for finding phonemes in the visual speech channel of information. By this we mean that, using the linguistic premise that phonemes are consistent for all speakers, there could be a way of translating between strings of visemes which provide more information, thus are more discriminative for recognizing the phonemes actually spoken. This approach is consistent with deep learning methods which have excellent results when lipreading sentences rather than short units such as in [63].

Acknowledgement

We gratefully acknowledge the assistance of Professor Stephen J Cox, formerly of the University of East Anglia, for his advice and guidance with this work. This work was conducted while Dr Bear was in receipt of a studentship from the UK Engineering and Physical Sciences Research Council (EPSRC).

References

[1] T. Stafylakis, G. Tzimiropoulos, Combining residual networks with LSTMs for lipreading, in: Interspeech, 2017.

[2] H. L. Bear, S. Taylor, Visual speech processing: aligning terminologies for better understanding, in: BMVC Deep Learning for Machine Lip Reading workshop, BMVA, 2017.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, J. Zhou, Audio-visual speech recognition, in: Final Workshop 2000 Report, Vol. 764, 2000.

[4] S. S. Morade, S. Patnaik, Lip reading by using 3D discrete wavelet transform with dmey wavelet, International Journal of Image Processing (IJIP) 8 (5) (2014) 384.

[5] H. L. Bear, G. Owen, R. Harvey, B.-J. Theobald, Some observations on computer lip-reading: moving from the dream to the reality, in: SPIE Security + Defence, International Society for Optics and Photonics, 2014, pp. 92530G–92530G.

[6] N. Harte, E. Gillen, TCD-TIMIT: An audio-visual corpus of continuous speech, IEEE Trans on Multimedia 17 (5) (2015) 603–615.

[7] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, The Journal of the Acoustical Society of America 120 (5) (2006) 2421–2424.

[8] J. S. Chung, A. Zisserman, Lip reading in the wild, in: Asian Conference on Computer Vision, 2016.

[9] K. Thangthai, R. Harvey, Improving computer lipreading via DNN sequence discriminative training techniques, in: Interspeech, 2017.

[10] T. Heidenreich, M. W. Spratling, A three-dimensional approach to visual speech recognition using discrete cosine transforms, arXiv preprint arXiv:1609.01932.

[11] T. Watanabe, K. Katsurada, Y. Kanazawa, Lip reading from multi view facial images using 3D-AAM, in: Asian Conference on Computer Vision, Springer, 2016, pp. 303–316.

[12] A. Rekik, A. Ben-Hamadou, W. Mahdi, An adaptive approach for lip-reading using image and depth data, Multimedia Tools and Applications 75 (14) (2016) 8609–8636.

[13] M. Wand, J. Koutník, J. Schmidhuber, Lipreading with long short-term memory, in: Int Conf Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 6115–6119.

[14] T. Stafylakis, G. Tzimiropoulos, Deep word embeddings for visual speech recognition, in: Int Conf Acoustics Speech and Signal Processing, 2017.

[15] S. Petridis, B. Martinez, M. Pantic, The MAHNOB laughter database, Image and Vision Computing 31 (2) (2013) 186–202.

[16] L. Cappelletta, N. Harte, Viseme definitions comparison for visual-only speech recognition, in: European Signal Processing Conference, 2011, pp. 2109–2113.

[17] S. Hilder, R. W. Harvey, B.-J. Theobald, Comparison of human and machine-based lip-reading, in: AVSP, 2009, pp. 86–89.

[18] T. Chen, R. R. Rao, Audio-visual integration in multimodal communication, Proc of the IEEE (5) (1998) 837–852.

[19] C. G. Fisher, Confusions among visually perceived consonants, Journal of Speech, Language, and Hearing Research 11 (4) (1968) 796–804.

[20] T. J. Hazen, K. Saenko, C.-H. La, J. R. Glass, A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments, in: Int Conf on Multimodal Interfaces, ACM, 2004, pp. 235–242.

[21] H. L. Bear, S. J. Cox, R. Harvey, Speaker independent machine lip reading with speaker dependent viseme classifiers, in: Int Conf on Facial Analysis, Animation and Audio-Visual Speech Processing (FAAVSP), ISCA, 2015, pp. 115–120.

[22] E. B. Nitchie, Lip-Reading, principles and practise: A handbook, Frederick A Stokes Co, New York, 1912.

[23] H. L. Bear, Visual gesture variability between talkers in continuous visual speech, in: BMVC Deep Learning for Machine Lip Reading workshop, BMVA, 2017.

[24] L. Cappelletta, N. Harte, Phoneme-to-viseme mapping for visual speech recognition., in: Int Conf Pattern Recognition Applications and Methods (ICPRAM), 2012, pp. 322–329.

[25] J. Jeffers, M. Barley, Speechreading (lipreading), Thomas Springfield, IL:, 1971.

[26] J. L. Newman, S. J. Cox, Language identification using visual features, IEEE Trans on audio, speech, and language processing 20 (7) (2012) 1936–1947.

[27] A. Metallinou, C. Busso, S. Lee, S. Narayanan, Visual emotion recognition using compact facial representations and viseme information, in: Int Conf on Acoustics Speech and Signal Processing, IEEE, 2010, pp. 2474–2477.

[28] H. L. Bear, R. Harvey, B.-J. Theobald, Y. Lan, Resolution limits on visual speech recognition, in: Int Conf on Image Processing (ICIP), IEEE, 2014, pp. 1371–1375.

[29] P. B. Kricos, S. A. Lesner, Differences in visual intelligibility across talkers, The Volta Review 82 (1982) 219–226.

[30] H. L. Bear, R. Harvey, Phoneme-to-viseme mappings: the good, the bad, and the ugly, Speech Communication 95 (2017) 40 – 67. doi:https://doi.org/10.1016/j.specom.2017.07.001.
URL http://www.sciencedirect.com/science/article/pii/S0167639317300286

[31] H. L. Bear, R. W. Harvey, B.-J. Theobald, Y. Lan, Which phoneme-to-viseme maps best improve visual-only computer lip-reading?, in: Advances in Visual Computing, Springer, 2014, pp. 230–239.

[32] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics bulletin 1 (6) (1945) 80–83.

[33] S. Cox, R. Harvey, Y. Lan, J. Newman, B. Theobald, The challenge of multispeaker lip-reading, in: Int Conf on Auditory-Visual Speech Processing, 2008, pp. 179–184.

[34] J. Luettin, N. A. Thacker, S. W. Beet, Speaker identification by lipreading, in: Proc.Int Conf Spoken Language, 1996.

[35] H. E. Cetingul, Y. Yemez, E. Erzin, A. M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, IEEE Trans on Image Processing 15 (10) (2006) 2879–2891.

[36] M. Wand, J. Schmidhuber, Improving speaker-independent lipreading with domain-adversarial training, in: Interspeech, ISCA, 2017.

[37] S. Lee, D. Yook, Audio-to-visual conversion using hidden markov models, in: Proc. Pacific Rim Int Conf on Artificial Intelligence, Springer, 2002, pp. 563–570.

[38] J. Lander, Read my lips. facial animation techniques, http://www.gamasutra.com/view/feature/131587/read_my_lips_facial_animation_.php. Accessed: 2014-01.

[39] M. F. Woodward, C. G. Barber, Phoneme perception in lipreading, Journal of Speech, Language and Hearing Research 3 (3) (1960) 212.

[40] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, R. Bowden, Improving visual features for lip-reading., in: AVSP, 2010, pp. 7–3.

[41] M. M. Rashid, R. Mustafa, M. Sanaullah, An investigation of the various methods of lip reading systems.

[42] M. Wand, J. Schmidhuber, Improving speaker-independent lipreading with domain-adversarial training, in: Interspeech, 2017.

[43] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Int Conf on Machine Learning, 2015, pp. 1180–1189.

[44] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, R. Bowden, Improving visual features for lip-reading, Int Conf on Audio-Visual Speech Processing (AVSP) 7 (3).

[45] Y. Miao, L. Jiang, H. Zhang, F. Metze, Improvements to speaker adaptive training of deep neural networks, in: Spoken Language Technology Workshop, IEEE, 2014, pp. 165–170.

[46] M. H. Rahmani, F. Almasganj, Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features, in: Int Conf Pattern Recognition and Image Analysis (IPRIA), IEEE, 2017, pp. 195–199.

[47] J. A. Bangham, R. Harvey, P. D. Ling, R. V. Aldridge, Nonlinear scale-space from n-dimensional sieves, in: ECCV, Springer, 1996, pp. 187–198.

[48] E.-J. Ong, R. Bowden, Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors, Trans on Pattern Analysis and Machine Intelligence 33 (9) (2011) 1844–1859.

[49] I. Matthews, S. Baker, Active appearance models revisited, IJCV 60 (2) (2004) 135–164.

[50] Cambridge University, UK. BEEP pronounciation dictionary [online] (1997) [cited Jan 2013].

[51] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, The DARPA speech recognition research database, in: Proc. DARPA Workshop on speech recognition, 1986, pp. 93–99.

[52] J. Matas, K. Zimmermann, T. Svoboda, A. Hilton, Learning efficient linear predictors for motion estimation, in: ICVGIP, Springer, 2006, pp. 445–456.

[53] T. Sheerman-Chase, E.-J. Ong, R. Bowden, Non-linear predictors for facial feature tracking across pose and expression, in: Automatic Face and Gesture Recognition, IEEE, 2013, pp. 1–8.

[54] E. Ong, R. Bowden, Robust lip-tracking using rigid flocks of selected linear predictors, in: IEEE Int Conf Automatic Face & Gesture Recognition, 2008, pp. 247–254.

[55] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, The American Statistician.

[56] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchec, P. Woodland, The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.
URL `http://htk.eng.cam.ac.uk/docs/docs.shtml`

[57] I. Matthews, T. Cootes, J. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, Trans. Pattern Analysis and Machine Intelligence 24 (2) (2002) 198 –213.

[58] F. E. Joumun, P. Gnanayutham, J. George, Multimedia Interfaces for BSL Using Lip Readers, 2008, pp. 663–669.

[59] K. Thangthai, H. L. Bear, R. Harvey, Comparing phonemes and visemes with DNN-based lipreading, in: BMVC Deep learning for machine lip reading workshop, BMVA, 2017.

[60] H. L. Bear, R. Harvey, Alternative visual unit building for a phoneme-based lipreading system, IEEE Trans on Circuits and Systems for Video Technology.

[61] H. L. Bear, R. Harvey, B.-J. Theobald, Y. Lan, Finding phonemes: improving machine lip-reading, in: Int Conf on Facial Analysis, Animation and Audio-Visual Speech Processing (FAAVSP), ISCA, 2015, pp. 190–195.

[62] H. L. Bear, R. Harvey, Decoding visemes: improving machine lip-reading, in: Int Conf Acoustics, Speech and Signal Processing (ICASSP), 2016.

[63] Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, Lipnet: Sentence-level lipreading, CoRR abs/1611.01599. `arXiv:1611.01599`.
URL `http://arxiv.org/abs/1611.01599`