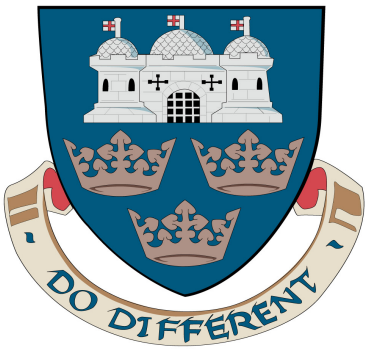
# Decoding visemes: improving machine lip-reading

Helen L. Bear

PhD Thesis

University of East Anglia
School of Computing Sciences




June 6, 2016

*Character is defined by what you do when no one is looking.*

John Wooden

*He who binds to himself a joy, does the winged life destroy;*
*But he who kisses the joy as it flies, lives in Eternity's sunrise.*

William Blake

# Abstract

This thesis is about improving machine lip-reading, that is, the classification of speech from only visual cues of a speaker. Machine lip-reading is a niche research problem in both areas of speech processing and computer vision.

Current challenges for machine lip-reading fall into two groups: the content of the video, such as the rate at which a person is speaking or; the parameters of the video recording for example, the video resolution. We begin our work with a literature review to understand the restrictions current technology limits machine lip-reading recognition and conduct an experiment into resolution affects. We show that high definition video is not needed to successfully lip-read with a computer.

The term "viseme" is used in machine lip-reading to represent a visual cue or gesture which corresponds to a subgroup of phonemes where the phonemes are indistinguishable in the visual speech signal. Whilst a viseme is yet to be formally defined, we use the common working definition 'a viseme is a group of phonemes with identical appearance on the lips'. A phoneme is the smallest acoustic unit a human can utter. Because there are more phonemes per viseme, mapping between the units creates a many-to-one relationship. Many mappings have been presented, and we conduct an experiment to determine which mapping produces the most accurate classification. Our results show Lee's [82] is best. Lee's classification also outperforms machine lip-reading systems which use the popular Fisher [48] phoneme-to-viseme map.

Further to this, we propose three methods of deriving speaker-dependent phoneme-to-viseme maps and compare our new approaches to Lee's. Our results show the

sensitivity of phoneme clustering and we use our new knowledge for our first suggested augmentation to the conventional lip-reading system.

Speaker independence in machine lip-reading classification is another unsolved obstacle. It has been observed, in the visual domain, that classifiers need training on the test subject to achieve the best classification. Thus machine lip-reading is highly dependent upon the speaker. Speaker independence is the opposite of this, or in other words, is the classification of a speaker not present in the classifier's training data. We investigate the dependence of phoneme-to-viseme maps between speakers. Our results show there is not a high variability of visual cues, but there is high variability in trajectory between visual cues of an individual speaker with the same ground truth. This implies a dependency upon the number of visemes within each set for each individual.

Finally, we investigate how many visemes is the optimum number within a set. We show the phoneme-to-viseme maps in literature rarely have enough visemes and the optimal number, which varies by speaker, ranges from 11 to 35. The last difficulty we address is decoding from visemes back to phonemes and into words. Traditionally this is completed using a language model. The language model unit is either: the same as the classifier, e.g. visemes or phonemes; or the language model unit is words. In a novel approach we use these optimum range viseme sets within hierarchical training of phoneme labelled classifiers. This new method of classifier training demonstrates significant increase in classification with a word language network.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| Abbreviation | Meaning |
|---|---|
| AAM | Active Appearance Model |
| AFR | Automatic Face Recognition |
| ALR | Automatic Lip-Reading |
| ASR | Automatic Speech Recognition |
| AV | Audio-Visual |
| AVSR | Audio-Visual Speech Recognition |
| BNW | Bigram Word Network |
| CF | Confusion Factor |
| CMU | Carnegie Mellon University |
| DCT | Discrete Cosine Transform |
| PDF | Probability Density Function |
| HCI | Human Computer Interaction |
| HD | High Definition |
| HMM | Hidden Markov Model |
| HTK | Hidden markov model Tool Kit |
| ISR | Intelligence, Surveillance and Reconnaissance |
| LRS | Lip-Reading System |
| P2V | Phoneme-to-Viseme |
| PCA | Principal Component Analysis |
| PDM | Point Distribution Model |
| SAM | Shape and Appearance Model |
| UWN | Unigram Word Network |
| VSR | Visual Speech Recognition |

# Acknowledgements

This is my opportunity to say thank you to some extraordinary people without whom I really couldn't have completed my PhD. My heartfelt thanks go to each and every one of you for so much more than I am capable of conveying in words. This list is in no way complete, but in true Oscar acceptance speech style...

Professor Richard Harvey (aka PhD supervisor extraordinare), you are my dream supervisor and friend. Your intelligence, patience, support and humour have been invaluable and I have loved working with you this past four years. To the rest of my supervisory team: Dr Barry-John Theobald, Dr Yuxuan Lan, Professor Stephen Cox, and Dr Anthony Bagnall, you are amazing. Thank you all for your patient education, support and guidance. I am also grateful for my examiners Professor Andy Day and Dr Naomi Harte for assessing my viva performance.

To my lab colleagues Mr Thomas Le Cornu and Mr Danny Websdale - you guys are the best lab buddies I could ever have dreamed of. Thank you for making coming in to work every day so good.

I want to pay special mention to a number of individuals:

Mrs Laura Barter, because oceans, babies and broken backs can't keep us apart. Mr Andrew Cadley, you are my rock who is never afraid to tell me when I am wrong. Miss Rebecca Clifford, your sage advice and kindness will always be remembered. Mrs Adelin Downing, because despite long absences we never lose each other. Miss Naomi Edwards, for being my confidence when I've lost my own. Mr Oliver Henderson, for the daily friendly thesis writing check-ups. Miss Kate Johnson, because you would follow me anywhere and always get me home safe. Miss Beth Judge,

who loves me exactly as I am and won't let me change. Mr Oliver Kirkland, for the silent study space, the endless cake, cookies, coffee and cocktails. Miss Gemma Maryan, you gave me a home and you taught me how to say 'no'. Mr Christopher Pantling, who is always on my side and has my back. Mr Michael Reed, your blind support of my PhD and your belief in me has been insurmountable. Mr David Reynolds, you are my Norwich brother who helped me make Norwich my second home. Miss Vanessa Schneider, for all the adventures and permitting me to live vicariously through you. Mr Russell Smith, thank you for more things over the past decade than I can possibly list here, please know I will be eternally grateful. Dr Sarah Taylor, my friend since undergraduate first year, I see you and all you have achieved and I am in awe, you are my inspiration. Mr Andrew Wood, for your candour, intelligence and diplomacy over the past four years.

You are all wonderful, thank you for everything.

Finally, thank you to my family, Barbara (aka Mum), Jeremy (aka Dad), Philip, Michelle, and Amelia, who know barely anything about what I've been doing for the past four years and understand it even less, but have supported me throughout this crazy endeavour, #proudtobeabear

Love and hugs,

Helen

# Statement of originality

Unless otherwise noted or referenced in the text, the work described in this thesis is that of the author. The following aspects are considered novel:

- A lower resolution limit for viseme lip-reading and demonstration of resolution resilience in machine lip-reading (Chapter 5.1).
- Differentiation between useful visemes for good classification (Chaper 6) .
- Evidence that consonant and vowel phonemes should not be mixed within sets (Chapter 7).
- Comparison of current phoneme-to-viseme mappings (Section 7.4).
- A new method of devising speaker-dependent viseme classes (Sections 7.5.2 & 7.5.3).
- Demonstration of how visual gestures themselves do not change, rather how a speaker uses them does (Chapter 8).
- A new similarity measure for comparing phoneme to viseme maps (Chapter 8).
- Presentation of comparable sets of speaker-dependent visemes (Chapter 9, section 9.2).
- An optimum range for viseme set sizes (Chapter 9).
- Visualised effects of homophones with viseme classes (Chapter 9).
- Differences in decoding between language units and classifier units (Chapter 9).
- Evidence to the nuanced hypothesis which is that there are intermediary units, visemes, that can provide superior classification in machine lip-reading (Chapter 9).
- Demonstration that weak learning of visemes can produce better phoneme labelled classifiers which improves machine lip-reading (Chapter 9).

# Publications from this thesis

1. Helen L Bear, Richard W Harvey, Yuxuan Lan, Barry-John Theobald, Resolution limits on visual speech recognition. IEEE International Conference on Image Processing (ICIP) 2014 [14].

2. Helen L Bear, Gari Owen, Richard W Harvey, Barry-John Theobald, Some observations on computer lip-reading: moving from the dream to the reality. Symposium of Photonics and Intelligent Engineering (SPIE) - Security & Defence 2014.[1] [17].

3. Helen L Bear, Richard W Harvey, Barry-John Theobald, Yuxuan Lan. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? Advances in Visual Computing (presented at the International Symposium for Visual Computing (ISVC)) 2014 [16].

4. Helen L Bear, Stephen J Cox, Richard W Harvey. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. 1st Joint Conference on Facial Analysis, Animation and Audiory-Visual Speech Processing (FAAVSP) 2015, satellite workshop of Interspeech 2015 [12].

5. Helen L Bear, Richard W Harvey, Yuxuan Lan. Finding phonemes: improving machine lip-reading. 1st Joint Conference on Facial Analysis, Animation and Audiory-Visual Speech Processing (FAAVSP) 2015, satellite workshop of Interspeech 2015 [15].

6. Helen L Bear, Richard Harvey. Decoding visemes: improving machine lip-reading. The 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 [13].

[1] 1st runner-up for best student paper

# Chapter 1

# Introduction

Speech is bimodal. This means there are two modes of information: acoustic and visual. Humans use both signals to understand the speech of others [101]. Given that acoustic recognition has been studied for over fifty years [38], it is not surprising that acoustic recognition is far more mature than visual-only recognition and there have been significant increases in performance in speech recognition systems, although they remain susceptible to noise [51]. Imagine trying to recognise a pilot's speech over the background noise of the aeroplane engine in a cockpit. In this case, the audio signal is severely deteriorated by the noise of the environment. However, this noise does not affect the visual signal. Thus, a desire to recognise speech from the visual signal alone is born. The visual signal can be used in combination with the acoustic signal, this is audio-visual speech recognition (AVSR) [120], or, there is the possibility of using the visual signal alone. This latter configuration is machine lip-reading which is the topic of this thesis.

Lip-reading is a challenging task. When researchers investigate AVSR, it is common for audio recognition to dominate any benefit from lip-reading, nevertheless, if we can make pure lip-reading successful there would be benefits for audio-visual recognition. Furthermore, there are a few scenarios where it is impractical or senseless to install a close microphone. An example might be an interactive booth in a busy station or airport where there is poor signal-to-noise ratio (SNR) or some

distance between the person and the screen. In practice however, a major use of a good machine lip-reading system would be as part of an AVSR system.

## 1.1 Applications of machine lip-reading

There are a range of scenarios where a machine lip-reading system would be beneficial. We discuss a few examples here.

During sports events there are often headlines about arguments between players, referees and even supporters. In the 2006 football World Cup Final between France and Italy, it was 19 minutes into extra time when Zinedine Zidane, on the opposite end of the pitch to the football, head-butted an Italian player without apparent justification. This action earned him a red card and consequently France went on to lose both the match and the world cup [103]. It later transpired, as admitted by Materazzi (the recipient of the head-butt), Zidane was provoked by a targeted insult of a late family member. In this case, if a machine lip-reading system had been present to confirm the provocation, whilst Zidane would have still been red carded, so would have Materazzi. Thus playing ten men against ten, the outcome of the match, and the World Cup, could have been different.

In history there are a great number of silent videos. Common examples are silent entertainment films and historical documentaries. In [136] we see a professional human lip-reader assist researchers comprehend what soldier's conversations were before they went into battle and during battle preparations. Similarly, in [3] we are shown how lip-readers used on the home movies of Hitler give historians an insight to an infamous figure of interest.

There has been long debate about if, in silent entertainment films of the era 1895-1927, films were ever scripted as the audio could not be captured with the video channel. In [6] we learn that, not only were these films in fact fully scripted, but in human lip-reading experiments, variation from the scripts were fully noticeable. Collectively, this human nature to be interested in history and learn from historical

evidence is a further motivator for achieving robust automatic lip-reading systems.

Theobald *et al.* [137] examine lip-reading for law enforcement. They note that in law-enforcement there are many departments who would benefit from an automatic lip-reading system. They present a new technique for improved lip-reading whereby the extracted features are modified to increase the classification performance. The modification is amplifying the feature parameters (they use Active Appearance Models which we explain fully in Chapter 2), to exaggerate the lip gestures recorded on camera. The technique was tested using a phonetically balanced corpus of syntactically correct sentences. The data set had very little contextual information [138] to remove effects of context network support. Machine lip-reading would help in law enforcement as robust lip-reading of filmed conversations during criminal acts, e.g. on CCTV could be evidence for the prosecution of offenders.

In the murder case of Arelene Fraser, Nat Fraser was caught and imprisoned. Evidence used by the prosecution included transcripts provided by professional lip-reader Jessica Rees [115]. Whilst the perpetrator thought he had committed the 'perfect' murder, and took steps to avoid any conversations being overheard, he had not thought about those who could read lips. With the transcripts of Fraser's conversations, prosecutors turned the co-conspirators into witnesses and Nat Fraser was prosecuted. However later, the reliability of lip-reading transcripts as evidence was successfully challenged, because human lip-readers are unreliable.

The reliability of human lip-readers is debatable. It has been said that this reliability varies not just between different pairings of speakers within a conversation, but also subject to the situation (context and environment) of the conversation (with the same speakers) [86]. This means that a good lip-reader on one day with a particular speaker could either misinterpret an alternative speaker or if lip-reading the same person in another place, fail to comprehend the speech uttered. Furthermore, human lip-readers are expensive, examples of Consuelo Gonsales [52] and Jessica Rees [122] operate on an as quoted basis. So we know that robust lip-readers are rare [86] and often we have no way of verifying the accuracy of the lip-reading per-

formance as a ground truth is rarely available. It is only in controlled experiments that a ground truth exists [20, 132, 63].

In [86] an investigation into the effect of likeability between individuals in a lip-read conversation, such as the status of their relationship, showed that a good relationship increases the accuracy of the lip-reading interpretation. To apply this observation to a real world scenario of introducing a lip-reader to someone they do not know personally, such as on a video documentary, deteriorates the confidence that their lip-reading ability will be robust. This idea is supported in Nichie's lip-reading and practice handbook [109] where in Chapter two it is suggested that the value of practicing lip-reading is rightly attached to the teacher's personality for success.

In [133] Summerfield describes some reasons which can distinguish poor from good lip-readers. This list is deduced from the results of a series of experiments ([61, 40, 92, 91, 148]) which show that the achievement rates in lip-reading tests can range from 10% to over 70%. These achievement rates vary due to the parameter selections for each experiment which are chosen for the specific task being addressed. In particular, the accuracy metric (some present word error rate, $w.e.r$ whereas others present percent true positive matches, %, others alternative metrics like the HTK correctness and accuracy scores (explained in full in section 2.5.1, Equations 2.7 & 2.8 respectively) and the classification unit (there are a number of options here - matching on phonemes, visemes or words) have a significant affect on how one should compare such investigations.

Some affects on human lip-reading performance are:

- intelligence and verbal reasoning - McGrath [100] showed that a fundamental level of intelligence and verbal reasoning are essential to be able to lip-read at all, but beyond a limit these skills could not raise human comprehension further.

- Training - human lip-readers who have either self-studied or have been trained

in some manner to practice the skill of lip-reading are shown to be no better than those who have received no training [31, 40]. Also it has been shown that human lip-readers can actually get worse with training [21], and this effect is more present when humans lip read from videos rather than in the presence of the speaker [76].

- Low-level visual-neural processing - Summerfield [133] discusses the physiological matter of the processing speed of these neural processes in the human brain. The suggestion is that lip-reading is difficult to learn because it is dependent upon these low-level neural processes. This suggestion has however, not received reproducible results to support the proposition which comforts us that human lip-reading is possible, however challenging.

- Closeness between the conversation participants - studies show that a relationship of some description between those talking, or personable knowledge of the speaker by the interpreter can improve human lip-reading [86, 123, 46].

- Knowledge of conversation context - without the constraint that is the 'rules' of a language to limit what a probable utterance is, lip-reading becomes almost impossible, or akin to guessing [126]. In [125] experiments showed that recognising isolated sentences was as low scoring as simply guessing from the context alone.

In summary, the main application of a machine lip-reading system would be any situation where the audio signal in a video is either absent or too noisy to comprehend, or where the alternative, human lip-readers, are too expensive or too unreliable.

## 1.2 The research problem

A conventional lip-reading system consists of a sequence of tasks as shown in Figure 1.1. Our work focuses on the classification task. Currently we have to make

some assumptions by tracking a face in a video in order to extract some features before we can undertake machine lip-reading.

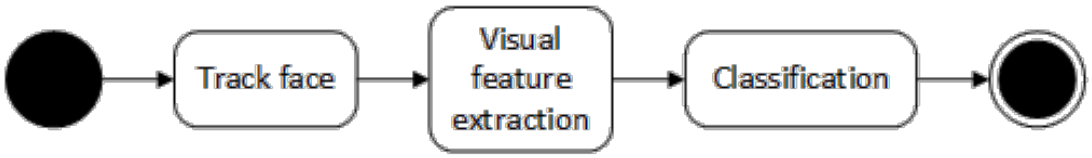



Figure 1.1: The three main functions in a traditional lip-reading system

The first task on the left hand side of Figure 1.1, is face tracking. This means to locate a face in an image (one frame of a video) and track it throughout the whole video sequence. By the end of the tracking process, often completed by fitting a model to each frame, we have a data structure containing information about the face through time. Examples of work showing face finding and tracking are in [128] and [139]. Example tracking methods are, with Active Appearance Models [33], or with Linear Predictors [112]. We discuss these two methods in Chapter 2. The second task, in the centre of Figure 1.1, is visual feature extraction. Using the fitted data parameters from task one, we can extract features which contain solely information pertaining to the speaker's lips. The third and final task on the right hand side of Figure 1.1 is classification. This is where we train some kind of classification model, using some visual features as training data, and use the classifiers to classify some unseen test data. Classification produces an output which can be compared with a ground truth to evaluate the accuracy of the classifiers.

There is a lot of literature on methods of feature extraction methods [111, 9, 65, 149, 119, 90] and tracking faces through images, [33, 146, 102, 83, 36] for lip-reading. However, to date, there is no one accepted method as the de facto method for extracting lip-reading features. In lieu of this, in [155], Zhou *et al.* ask two questions about feature extraction, specifically for lip reading: primarily, how to cope with the speaker identity dependency in visual data? But also, how to incorporate the temporal information of visual speech? The intent of this second question is for capturing co-articulation effects into features. Zhou *et al.* categorise a comprehensive range of feature extraction techniques into four groups:

- Image-based e.g. [53],
- Motion-based e.g. [93],
- Geometric-feature-based e.g. [107], or,
- Model-based e.g. [45].

This categorisation serves to show the breadth of current research into features. However, this attention on feature extraction does not address the only challenges in machine lip-reading. Improvements can still be made in the classification stage of lip-reading also. Therefore much of this thesis is focused on classification, rather than additional tasks such as tracking and feature extraction. That is not to say we are dismissive of the feature extraction and tracking requirements, rather that we wish focus our work to improve the classification methods.

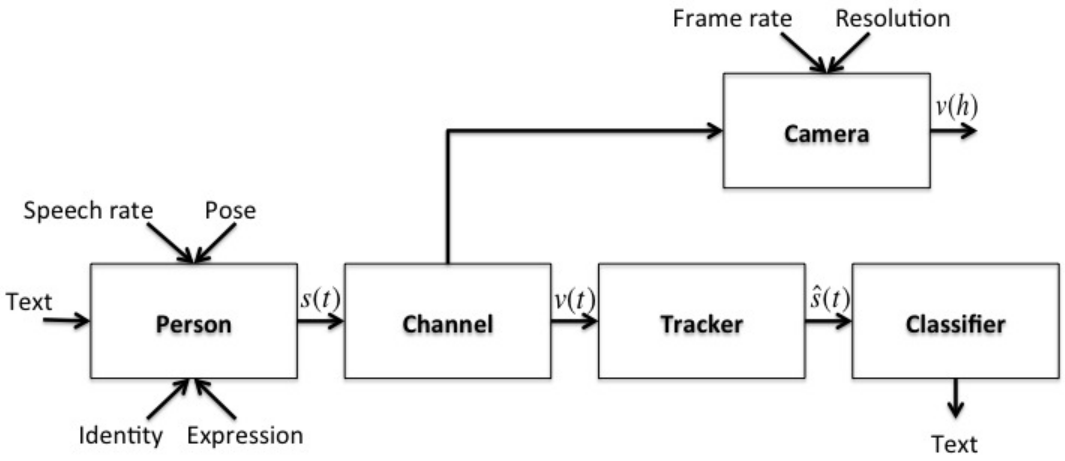



Figure 1.2: Sources of variability in computer lip-reading: affects on automatic lip-reading systems

Figure 1.2 shows the situation in which we are trying to recreate the text in the mind of the speaker. Each speaker articulates differently, and so the identity of the individual speaker is a significant affect on the efficacy of lip-reading. The visual signal is also affected by the speaker's pose, motion and expression. Cameras typically have many parameters that might affect lip-reading. Of these, we mention frame rate and resolution as highly probable to be significant.

Table 1.1: A list of affects on automatic lip-reading systems

| Evaluation | Previously studied, in | Likely sensitivity |
|---|---|---|
| Motion | Yes, [111, 97] | Low |
| Pose | Yes, [78] | Medium |
| Expression | Yes, [106] | Low |
| Frame rate | Yes, [22, 124] | Low |
| Resolution | No | Unknown |
| Colour | Yes, [71] | Low |
| Classifier unit choice | Yes, [28] | High |
| Feature type | Yes, [98, 77] | High |
| Classifier technology | Yes, [96, 150] | Medium |
| Multiple persons | Yes, [68] | Medium |
| Speaker identity | Yes, [89] | High |
| Rate of speech | Yes, [134] | High |

In Table 1.1 we have listed and assessed a number of environmental affects on machine lip-reading. There are a number of factors that can be difficult to control in machine lip-reading. These include, but are not limited to, lighting, identity, motion, emotion, and expression. Table 1.1 is an attempt at a systematic study of the affects. Considering initially the problem of speaker-dependent lip-reading, then three factors are of immediate interest: resolution because it does not appear to have been studied systematically, and unit choice, and feature type because they are likely to be highly significant to performance. For the time-being, speaker identity and rate of speech can be ignored since they are constant for a given speaker.

The choice of feature has been studied quite well and there have been a number of 'contests' between feature types (e.g. [77, 28]) which have led to the conclusion that state of the art Active Appearance Models (AAMs) are highly likely to give the best known performance. These are the features we use and the subject of the next chapter. However the choice of visual unit, the analogous quantity to a phoneme is more intriguing.

A phoneme is the smallest sound which can be uttered [5]. A viseme is not so precisely defined [30, 48, 58]. However, a working definition is that a viseme is a set of phonemes that have identical appearance on the lips. Therefore many phonemes

fall into one viseme class: a many-to-one mapping. There are alternative definitions of visemes in which the viseme is, for example, seen as a repeatable, visual gesture. In [27] two alternative definitions are explored: visemes based upon articulatory gestures or on similar visual appearance. The tentative conclusion is that visemes based upon the articulatory gestures definition perform better. This study only looks at recognition, in synthesis studies, visemes are considered as 'temporal units that describe distinctive speech movements of the visual speech articulators' [135]. As there are many definitions to choose from, we continue with the recognition working definition of 'a viseme is a group of phonemes with identical appearance on the lips'. Thus, our study starts with two key problems: resolution which has not been systematically studied before in isolation from observing the effects of noise, and unit selection because it is likely to be highly significant. But, before we can study these items, it is necessary to discuss the third affect to which classification is highly sensitive: feature selection.

## 1.3 Our research question

We ask, 'can we augment or replace the current lip-reading classifiers to improve machine lip-reading?'

# Chapter 2

# Features and classification methods

In the previous chapter it was asserted that feature choice was likely to be highly significant. In this chapter therefore, we examine the full processing chain in more detail from tracking to classification, dwelling on the methods of special relevance to this thesis.

## 2.1 Linear predictors

Linear Predictors (LPs) are a person-specific and data-driven facial tracking method. Devised primarily for observing visual changes in the face during speech, these make it possible to cope with facial feature configurations not present in the training data by treating each feature independently. For speech, this means isolating the lips from the eyes, outline of the face, etc.

The linear predictor itself is a part of the tracking mechanism. It is the central point around which support pixels are used to identify the change in position of the central point over time. The central point is visually seen as a landmark on the outline of a feature. A set of these landmarks represent the changing shape of

something (in our case lips) morphing over time. In this method both the shape (comprised of landmarks) and the pixel information surrounding the linear predictor position are intrinsically linked.

A single LP alone is not enough to provide robust and accurate tracking, so [111] explains how rigid flocks (a small group) of selected LPs are grouped around a central feature (not the linear predictor central point, but as an example, the feature mean position) restrict the motion of the LPs within a boundary and reduce their susceptibility to noise. These LPs have been successfully used to track objects in motion [95].

Further improvements to the LPs selection method are described in [111, 112], both of which show improvement of over original LP tracking accuracy.

An interactive LP tracking tool has been made at the University of Surrey. Its benefits are the real-time tracking and autonomous use, but a limitation of this tool is when a face is partially off-screen, the real-time tracking requires the user to guess in real time where the appropriate LP should be. This is not a simple task to perform with any accuracy or consistency and when tested with our Rosetta Raven dataset (see Chapter 3) we found the AAM features still outperformed the LP features.

## 2.2 Active shape and appearance models

An Active Appearance model (AAM) [33] is a combined shape and appearance trained model used in tracking a face throughout a video sequence. The model is constructed from a small training subset (Table 3.3) and is a type of Point Distribution Model (PDM) used to represent the shape of a face and how it varies during speech. The shape $s$ of an AAM is the coordinates of the $v$ vertices which make up a mesh,

$$s = (x_1, y_1, x_2, y_2, ..., x_v, y_v)^T \tag{2.1}$$

Training creates a mean model permitting deviations within a predetermined range of variance. Any of the training co-ordinate vectors used for model creation with 30% or more of occluded landmarks are omitted from the mean shape formation. Normalised meshes are built from the manually trained data (landmarks) for translation, scale and rotation (i.e. movement between the image frames).We now have a vector of $2n$ values for $n$ landmarks upon the face. Principal Component Analysis (PCA) provides us with eigenvectors so an independent shape model becomes a set of meshes,

$$s = s_0 + \sum_{i=1}^{n} p_i s_i \tag{2.2}$$

where $s_0$ is the mean shape, $p_i$ are coefficient shape parameters, and $s_i$ are the eigenvectors of the covariance matrix of the $n$ largest eigenvalues. We can assume $s_i$ is orthonormal because we can always perform a linear reparameterisation [97]. The landmarks are chosen to model the sub-shapes within the face such as: the outline of the hairline and jaw, eyes, nose or lips. We have to hand label these training images. The meshes constructed with our hand labelling are normalised by Procrustes analysis [54] before we apply PCA. An example of a full face shape model is shown in Figure 2.1. In this Figure there are 104 landmarks, the majority (44) of which are modelling the inner and outer lip contours.

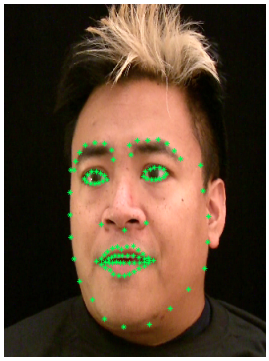


Figure 2.1: Example Active Appearance Model shape mesh.

An independent appearance AAM uses appearance data over the base mesh, $S_0$. This allows linear variation in the shape whilst maintaining a compact model. $S_0$

also denotes the set of pixels that lie inside the base mesh. Thus $A(x)$ (or AMM appearance) is an image defined over the pixels $x \in S_0$. This means pixels are mapped into the triangles of the shape model by Procrustes analysis [54] over the shape model vector (the aligned the set of points) to build the statistical model. Each training image is warped to match the mean shape to identify a shape-free area of the training image. This shape-free area is normalised with a linear transform before the texture model is built by eigen-analysis [33].

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \quad \forall x \in s_0 \tag{2.3}$$

In Equation 2.3 the coefficients $\lambda_i$ are the appearance parameters, $A_0$ is the base appearance, and $A_i(x)$ are the appearance image eigenvectors of the covariance matrix. Our appearance $A(x)$ is $A_0$ plus a combination of images $A_i(x)$. $A_0$ is the mean image, and $A_i$ are the $m$ eigenimages with the $m$ largest eigenvalues.

It has been demonstrated that the combination of appearance and shape models significantly improves lip-reading performance [96, 33] and we use these in the work presented here unless explicitly stated otherwise. The combination of these model types requires a single parameter set to represent the relationship between shape and appearance. In independent shape and appearance AAMs [97], the shape parameters, $p$, and appearance parameters $\lambda$, are distinct. In a combined model, we use one set of parameters, $C = (c_1, c_2, c_3, ..., c_n)^T$. This is shown in Equation 2.4 and Equation 2.5. This usage of a common parameter set, $c_i$, intrinsically ties the models together by warping the image over the shape model to represent both the appearance and shape variation in a face.

For shape

$$s = s_0 + \sum_{i=1}^{n} c_i s_i \tag{2.4}$$

and for appearance

$$A(x) = A_0(x) + \sum_{i=1}^{n} c_i A_i(x) \tag{2.5}$$

A combined AAM requires a third application of PCA on the weighted shape, $p$ and appearance, $\lambda$ parameters. The correlation between the shape and texture (appearance) model is learned and integrated into the combined model.

To initialise the AAM we use the shape parameters $p = (p_1, p_2, ..., p_n)^T$ in Equation 2.1 to generate the shape $s$, and the appearance parameters $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)^T$ to generate the appearance $A(x)$ in $s_0$. This AAM instance is built by a piecewise affine warp of $A$ from the base mesh $s_0$ to the AAM shape $s$.

Finally we fit the AAM using the Inverse Compositional algorithm [7] to all frames in the video sequence [97]. This algorithm uses the coordinate frame of the image $I$ and the coordinate frame of the AAM. To initiate the fit with the best starting position, the first image frame in a video sequence receives a manually labelled shape, $s$. Iterating through each frame of the video in turn, a backwards warp $W$ is used to warp each image $I$ onto the base mesh $s_0$ until the landmark positions converge into place to match corresponding pixels between frames. The more movement there is between frames, or the lower the frame rate, tracking is more difficult as these create greater variation between frame images.

## 2.3 Discrete cosine transforms

The Discrete Cosine Transform (DCT) [1] is a technique for converting a signal into elementary frequency components, or in other words, it transforms an image from a spatial to frequency domain by separating an image into parts of unequal importance. There are many variants of DCT and in lip-reading and AVSR authors use 2D-DCT (Equation 2.6) as it is applied too each two-dimensional frame image throughout a video. For example in [79, 28] and [104]. To create 2D-DCT features co-efficient vectors are extracted from the information from the region of interest in an image, for machine lip-reading, this is the lips.

$$q_{u,v} = W_u W_v \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} p_{i,j} \cos\left(\frac{u\pi(2i+1)}{2N}\right) \cos\left(\frac{v\pi(2j+1)}{2M}\right)$$

$$W_u = \begin{cases} \sqrt{1/N} & \text{if } u = 0 \\ \sqrt{2/N} & \text{otherwise} \end{cases} \quad W_v = \begin{cases} \sqrt{1/M} & \text{if } v = 0 \\ \sqrt{2/M} & \text{otherwise} \end{cases} \tag{2.6}$$

In Equation 2.6 we show that 2D-DCT is pixel-based, features are extracted from a region of interest matrix of size $M$ by $N$, where $P$ is the mouth centre. $p_{ij}$ is pixel intensity in row $i$ and column $j$. This creates $q_{uv}$.

## 2.4 Comparison of available feature types

Lan *et al.* present in [79] a comparison of different features first presented in [33]. Revisited in [97], AAM features are produced as either model-based (using shape information) or pixel-based (using appearance information). In [79] Lan *et al.* observed that state of the art AAM features with appearance parameters outperform other feature types like sieve features, 2D DCT, and eigen-lip features, suggesting appearance is more informative than shape. Also pixel methods benefit from image normalisation to remove shape and affine variation from region of interest (in this example, the mouth and lips). The method in [79] classified words with the RMAV dataset but recommended in future creating classifiers with viseme labels for lipreading, and advises that most information is from the inner of the mouth.

A comparison of two current key methods for fitting and extraction of facial features for computer lip-reading is summarised in Table 2.1.

For the work presented in this thesis, we chose to use AAMs. This is because whilst DCT features can outperform geometric features (as shown in [60]), a state of the art AAM can outperform DCT features. In [108] the results suggest that DCT features outperform AAMs because they complete most experiments with them after initial AAMs performed poorly, (65.9% *w.e.r* for AAMs compared to 61.80% *w.e.r* with DCT features). However, the authors also note that their AAMs

Table 2.1: A summary of shape and appearance models and linear predictors.

| Linear Predictors (LP) | Shape Appearance Model (SAM) |
|---|---|
| Data driven. | Face knowledge required from training for modelling. |
| Unsupervised. | Supervised. |
| Feature independent. | Feature dependence improves tracking. |
| Use only intensity information ie. grey scale images. | The fitted model can be either solely shape model, an appearance model (pixel information) or a combined model of shape and appearance where each pixel is related to a triangular section of the shape model. |
| Prior training shape models or temporal models for dynamics are not required or used. | An active appearance model is built from training data to fit new images. |
| Can cope with feature configurations not present in training data. | Training needs to encapsulate all variance in the video to be tracked. |
| Multiple LPs are grouped into flocks for robustness. | Primary landmarks are used for the important positions in training data. |

were not good ones and the reasons for this could be attributed to either; modeling or tracking errors. This is because insufficient training data can have two effects. First, that the AAM is not generalised enough from the training data to classify the test data, and secondly, an undertrained AAM will not fit well when tracking a face. It should be noted that in comparing DCT and AAM features, Neti *et al.* use different regions of interest for the feature types. For the DCT features, the ROI is the mouth, compared to the whole face for the AAMs [108].

In the work presented in Chapters 5 to 9, particularly for continuous speech experiments with newer datasets, we have confidence that our AAMs are state of the art, have tracked well between all frames (this is confirmed by producing a jpg image of each frame with the AAM landmarks plotted on and the fit is manually checked) and is achieved by using a higher number of landmarks, we use 104 [14] rather than the 68 in [108]).

## 2.5 Hidden Markov models

Hidden Markov Models (HMMs) have been used in speech classification for some time for acoustic, audio-visual and visual-only classification. Both channels of speech can be considered as a time series, i.e. they will produce data points in a causal manner. Other domains which have applied HMMs are sets of temporal data such as handwriting, DNA sequences and energy consumption.

A HMM has two stochastic processes: the first process is based around state transition probabilities, and the second, is based upon state emission probabilities.

A Markov model (also known as a Markov chain), is made of a number of states connected to all other states. Each connection has transition probabilities for moving between the states it is connected to. In a $n^{th}$-order Markov chain, an inherent assumption is that state transitions are dependent upon the $n$ previous states. In a Markov chain the stochastic process output is the sequence of states. Practically, in speech classification, a first order model is normally used. In a first order HMM the state transitions are dependent only on the current state. The probabilities of all possible actions (transitions) at time, $t$, are dependent upon the state the HMM is in at time $t$, not the value of $t$.

The second stochastic process is concerned with emission probabilities. Each HMM state has an associated Probability Density Function (PDF). A PDF used on feature vectors determines the emission probabilities of any particular feature vector being output (emitted) by the state, when the HMM is in that state. Whereas in a Markov chain the output is the sequence of states, in an HMM the PDF means the output is a feature vector. Because the emission probabilities are a function of the state, the knowledge of the state is hidden from the observer [64].

In a network of HMMs, each HMM is labelled by its representative unit. In visual speech, these units are referred to as visemes, in acoustic speech phoneme labels are used. In some simple speech classification tasks, or with limited datasets, words may be used as the HMM unit label. Additional HMMs can also be built to model

the silence at the start and end of utterances and the shorter silence pauses between words. In the work presented in this thesis, all HMMs are monophones.

### 2.5.1 HTK: an HMM toolkit

HTK provides a set of tools which enable users to build speech processing tools, including recognisers and estimators. The main algorithm used in HMM estimation is the Baum-Welch algorithm [10], and the algorithm used in classification is the Viterbi algorithm [142]. The HTK book [151] details the background of HTK in full, up to its current version for full information of its implementation and use.

The use of HTK is commonplace in acoustic speech classification [2, 120, 67, 98] and current lip-reading literature [78, 77, 68, 79, 63]. So using HTK for machine lip-reading allows very easy replication of our results. HTK has achieved ubiquity due to its generally high performance, so we can be confident that our results will be close to the best achievable performance when we adopt similar strategies as described in previous works.

In HTK recognition. performance of the HMMs can be measured by both correctness, $C$, and accuracy, $A$,

$$C = \frac{N - D - S}{N}, \tag{2.7}$$

$$A = \frac{N - D - S - I}{N} \tag{2.8}$$

where $S$ is the number of substitution errors, $D$ is the number of deletion errors, $I$ is the number of insertion errors and $N$ the total number of labels in the reference transcriptions [150].

We can explain these types of errors with an example. Suppose we have a ground truth utterance, “John wanted to visit the shop to buy groceries”. Our classifiers can produce different outputs. Possible output 1: “ John wanted visit the to groceries” has three words missing. ‘to’, ‘shop’, and ‘buy’. In this instance, these are deletion

errors. In another possible output: "John wanted to visit visit the shop to buy groceries", the word 'visit' is included twice. This is an insertion error. Finally, if we achieved a classifier output of "John wanted to shop the shop to buy groceries". The word 'shop' has been identified where the word 'visit' should be. This is a substitution error.

Common tools used for a classification task in HTK are: `HCompV`, `HERest`, `HHed`, `HVite` and `HResults`.

`HCompV` - used to flat start each HMM subject to a prototype file determining number of states and mixtures. It does this based upon the data within the whole dataset so all states are equal. It uses a prototype HMM definition, some training data and initialises each new HMM where every local HMM mean is the same as the global mean across the whole set. Only the covariances are updated.

`HERest` - is the Balm-Welsh re-estimation of each HMM using the training fold samples and a transcription using the HMM labels. HERest uses embedded training to simultaneously updated all HMMs within a systems using all training data available within a fold. This is particularly important for systems where the HMM labels are sub word models as HERest ignores boundary information in transcripts of training samples.

`HHEd` - permits the tying together of states within an HMM model to allow fast transitions between states and shorter Markov chains. This is particularly useful for similar or short models such as silence (at the start and end of utterances) and short pauses between words.

`HVite` - is commonly used for both forced alignment of HMMs using the ground truth transcription, and also for the crucial classification task. Using the trained HMMs, `HVite` attempts to recognise test samples and produces a classification output.

`HResults` - compares the classification output to the ground truth, HResults provides statistics about how accurate the HMM recognisers have been, primarily

correctness (Equation 2.7) at both the unit and network level, and also includes model-level accuracy (Equation 2.8).

# Chapter 3

# Datasets

This chapter summarises the datasets used in the work presented throughout this thesis. Note that while this thesis is about machine lip-reading (visual speech recognition), audio-visual datasets are commonplace since researchers often wish to compare visual-only performance to audio and audio-visual performance for the purposes of audio-visual integration such as in [108]. A summary of the most common AVSR databases is presented in Table 3.1. The result values listed are those from the original presented papers referenced in column 1. The results vary based upon the specific experiments, content, classification units (e.g. words, visemes, or phonemes), and original intent of each dataset. Other databases are available, such as those in [85, 4, 129] but these are non-English (Mandarin, Arabic and French respectively) and therefore not considered here.

Table 3.1: Common databases available for machine lip-reading research.

| Name | Speakers | Content | Results |
|---|---|---|---|
| AVLetters [96] | 10 | Alphabet letters | $< 27\%$ |
| AVLetters2 [35] | 5 | High definition alphabet letters | $80\% >< 90\%$ |
| AV-TIMIT [58] | 223 | TIMIT sentences | 35% p.e.r |
| CUAVE [117] | 36 | Digits | 87% Acc |
| GRID [32] | 36 | Command sentences | $< 1.85\%$ w.e.r |
| IBM LVCSR (ViaVoice) [99] | 290 | Continuous speech | 58% w.e.r |
| OuluVS [154] | 20 | 10 everyday phrases | 70% Acc |
| RMAV (LILIR) [79] | 20 | Context dependent sentences | $20\% >< 60\%$ |
| Rosetta Raven [14] | 2 | E. A. Poe's The Raven | $20\% >< 60\%$ |
| TCD-TIMIT [56] | 62 | 98 sentences | $> 55\%$ Acc |

For the work presented in this thesis, the Rosetta Raven database was selected for the resolution robustness experiment in Chapter 5 because it is both continuous and structured speech. This means that there is a good quantity of data but also that the speech itself is constrained meaning that the task is simpler than that of say AV-TIMIT, this is better for a controlled experiment to measure the affects of a single parameter. Note that AusTalk, AV-TIMIT and IBM LVCSR are proprietary and thus not freely available.

We have confidence that the larger (in regards to number of speakers) continuous speech datasets have a good phoneme coverage and so, subject to the viseme mapping selected, will also have good viseme coverage, however the smaller datasets, including those with limited vocabularies, the quantity of visemes (and the consequential volume of training samples per viseme class) will be at risk of inter-class skew. Therefore preliminary experiments in later chapters were undertaken first with AVLetters2 for proof of concept and confirmation that hypotheses were sound, before repeating experiments with RMAV. RMAV has sentences selected from the resource management data [49] which ensures a good phoneme coverage in its content. RMAV was selected as extracted features were available which enabled focusing on the classification task rather than that of tracking and extracting features.

## 3.1 Pronunciation dictionaries

To accommodate the breadth of possible pronunciations, a number of dictionaries are available for use in machine lip-reading. These dictionaries map words to phoneme sequences subject to the pronunciation habits of the speaker. Two are described here: firstly, CMU [29], has been used in conjunction with the Rosetta Raven data, and secondly, BEEP [130], is used in later chapters with AVLetters2 and RMAV.

The Carnegie Mellon University North American Pronunciation Dictionary [29], known as CMU, uses 39 phonemes and also encodes whether vowels carry levels of lexical stress [62] of either 0-None, 1-Primary or 2-Secondary. Lexical stress

is the relative emphasis placed upon certain syllables within a word. Including lexical stress representations, this dictionary has 57 phonemes. Containing over 125,000 words, it is based on the ARPAbet symbol set (which relates to the standard IPA symbol set) developed for speech recognition uses. This dictionary is used for American speakers speaking English i.e. American English.

The Cambridge University British English Pronunciation dictionary, known as BEEP, [130] has 49 phonemes mapped to over 250,000 words allowing for duplicate pronunciations of the same word. For example, the word 'read' phonetically can be, '/r/ /eh/ /d/' as in 'I read my book last night' or, '/r/ /ɪ/ /d/' as in 'I like to read'. This dictionary is used for British speakers of English.

## 3.2 AVLetters2 - an isolated word dataset

AVLetters 2 (AVL2) [35] is an HD version of the AVLetters dataset [98]. It is a single word dataset of four British English speakers (all male) each reciting the 26 letters of the alphabet seven times. We can not present the quantity of visemes in the data set at this stage as it is dependent upon the viseme set being used (see Section 7). The speakers in this dataset can be seen in Figure 3.1. AVL2 has 28 videos of between $1,169$ and $1,499$ frames between $47s$ and $58s$ in duration. As the dataset provides isolated words of single letters, it lends itself to controlled experiments without needing to address matters such as co-articulation.

(a) Speaker 1 (b) Speaker 2 (c) Speaker 3 (d) Speaker 4

Figure 3.1: Example faces from the AVLetters2 videos (four speakers).

There are 30 unique British English phonemes in AVL2, the occurrence frequency of these is shown in Figure 3.2. Therefore, the data set is missing 19 phonemes found in spoken British English.

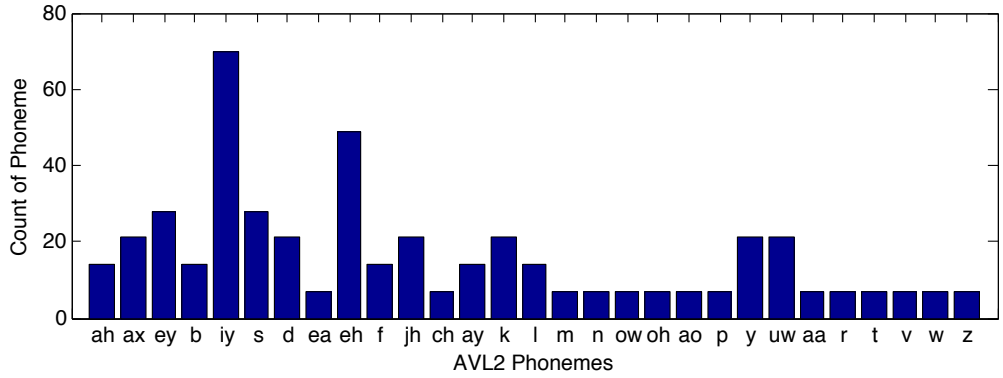



Figure 3.2: Occurrence frequency of phonemes in the AVLetters2 dataset.

Table 3.2 describes the features extracted from the AVL2 videos. These features have been derived after tracking a full-face Active Appearance Model throughout the video before extracting features containing only the lip area. Therefore, they contain information representing only the speaker's lips and none of the rest of the face. Speakers 2, 3 and 4 are similar in number of parameters contained in the features. The combined features are the concatenation of the shape and appearance features [97]. All features retain 95% variance of facial shape and appearance information.

Table 3.2: The number of parameters in shape, appearance and combined shape & appearance AAM features for each speaker in the AVLetters2 dataset for each speaker. Features retain 95% variance of facial information.

| Speaker | Shape | Appearance | Combined |
|---|---|---|---|
| S1 | 11 | 27 | 38 |
| S2 | 9 | 19 | 28 |
| S3 | 9 | 17 | 25 |
| S4 | 9 | 17 | 25 |

This dataset is used for comparing visemes, testing new speaker-dependent visemes (Chapter 7) and for evaluating the robustness of speaker-dependent phoneme-to-viseme maps in Chapter 8.

## 3.3 Rosetta Raven - a stylised continuous speech dataset

This dataset was recorded at UCLA in January 2012 by Dr Eamon Keogh and was formulated as an attempt to provide a standardised audio-visual machine learning problem [14]. It comprises four videos which consist of two North American untrained speakers (one male, one female, seen in Figure 3.3) each reciting E.A.Poe's 'The Raven'. The poem was published in 1845 and the linguistic content of the Raven make this an interesting dataset as the narrative uses a stylised language including internal rhyme and alliteration. The poem is described as being generally trochaic octameter [121].

Trochaic octameter is a rarely used meter in poetry. Within each line of a trochaic octametric poem, there are eight trochaic metrical feet. Each of these eight feet consist of two syllables, the first of the two is stressed, the latter unstressed giving rise to an 'up and down' effect to a professional recitation. This pairing of a stressed and an unstressed syllable (or poetic foot) is trochaic [18]. However, this does not appear to have been followed by the speakers in this dataset.

(a) Speaker 1 (b) Speaker 2

Figure 3.3: Example faces from the Rosetta Raven videos (two speakers).

Table 3.3: Summary of video content in the Rosetta Raven dataset.

| Video | AAM train frames | AAM fit frames | Duration |
|---|---|---|---|
| Speaker1_v1 | 11 | 31,858 | 00:08:52 |
| Speaker1_v2 | 11 | 33,328 | 00:09:17 |
| Speaker2_v1 | 10 | 21,648 | 00:06:01 |
| Speaker2_v2 | 10 | 21,703 | 00:06:02 |

In linguistic terms the the videos have 56 phonemes present with minor variation on their occurrences in each video (Figure 3.4). It is noted some phonemes namely /ɔ0/, /uw0/, /ɑʊ2/, /ʊ2/, /ae0/, /eh0/, /ey2/, /ɑ2/, /ʌ2/, /ɑ0/ and /əʊ2/ have less than ten instances within the whole data set. These phonemes all have lexical stress shown by the numbers in their naming convention, this comes from the American English set of phonemes used in the CMU pronunciation dictionary. Again, we can not quantify the viseme counts in this dataset as it varies with the viseme set used in any particular experiment.

For these data to be used in a machine lip-reading system, we need to extract features. The training images from each speaker video (Table 3.3) were used together to make a single AAM model for tracking the rest of the video. A full face AAM was used to track the face for a robust fitting, whereas a lip-only AAM was used to extract lip-only feature. These features retained 95% of the speakers face shape and appearance variance throughout the video and are used in the resolution work described in Chapter 5 and for assessing the contribution of individual visemes within a set in Chapter 6.

Table 3.4: The number of parameters in shape, appearance, and combined shape and appearance AAM features for the Rosetta Raven dataset speakers. Features retain 95% variance of facial information.

| Speaker | Shape | Appearance | Combined |
|---|---|---|---|
| S1 | 6 | 14 | 20 |
| S2 | 7 | 14 | 21 |

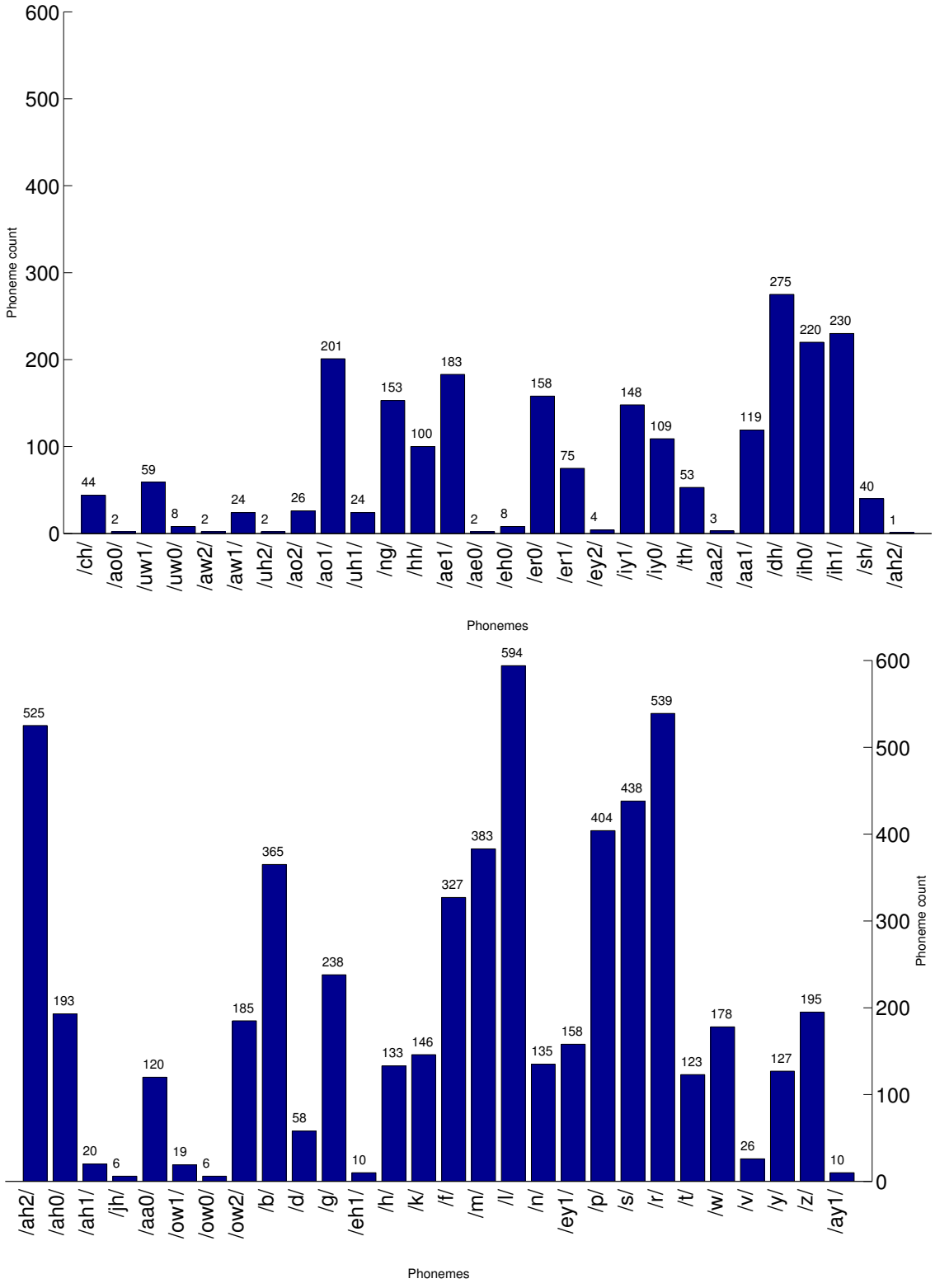



Figure 3.4: Occurrence frequency of phonemes in the Rosetta Raven dataset.

## 3.4 RMAV - a context-independent continuous speech dataset

Formerly known as LiLIR, the RMAV dataset consists of 20 British English speakers (we use 12, seven male and five female), 200 utterances per speaker of the Resource Management (RM) context independent sentences from [49] 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 by a phonetician to maintain as much coverage of all phonemes as possible. The original videos were recorded in high definition and in a full-frontal position. Individual speakers are tracked using Active Appearance Models [97] and AAM features of concatenated shape and appearance information have been extracted.

Table 3.5: The number of parameters of shape, appearance, and combined shape and appearance AAM features for the RMAV dataset speakers. Features retain 95% variance of facial information.

| Speaker | Shape | Appearance | Combined |
|---|---|---|---|
| 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 |

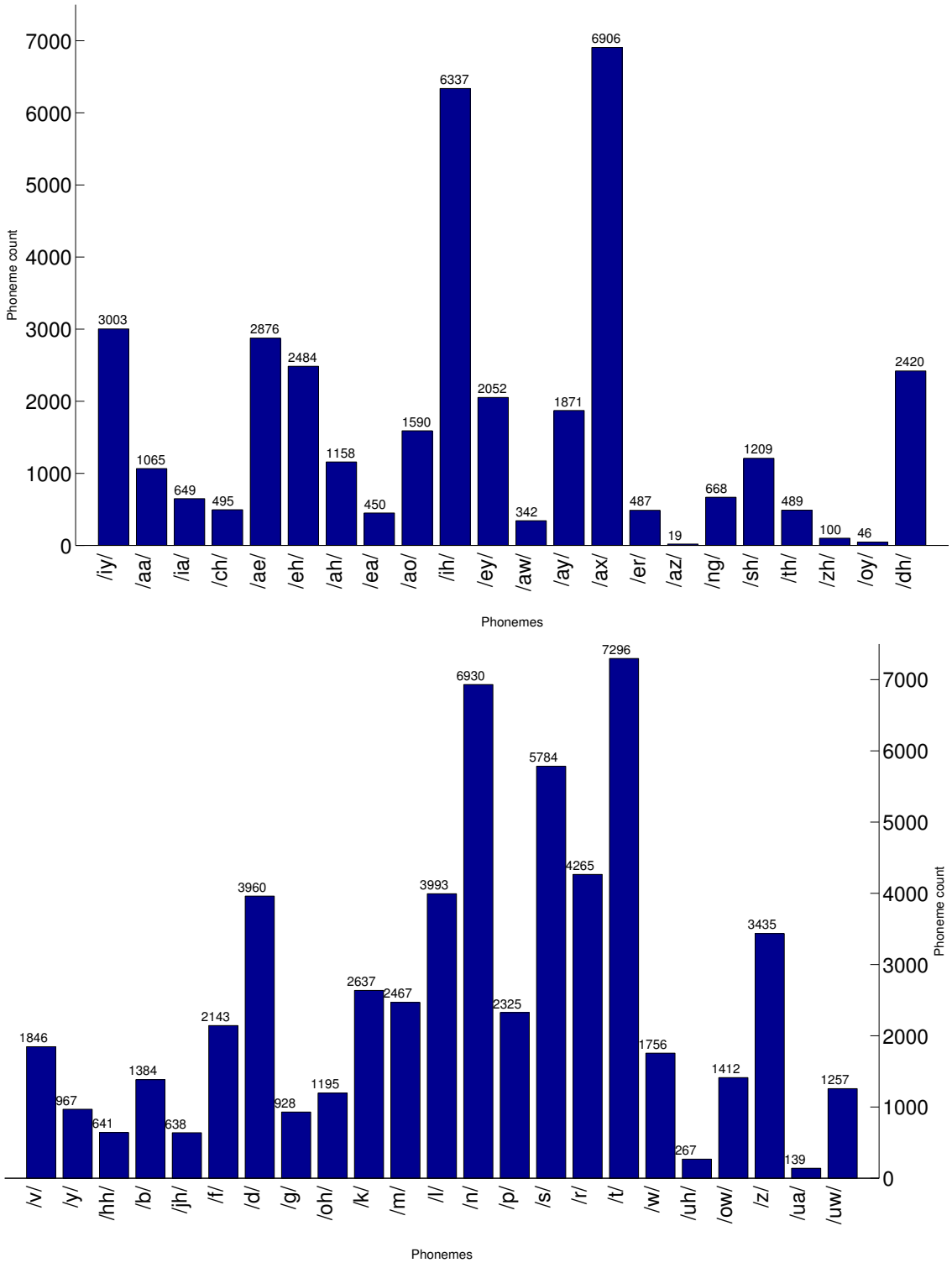
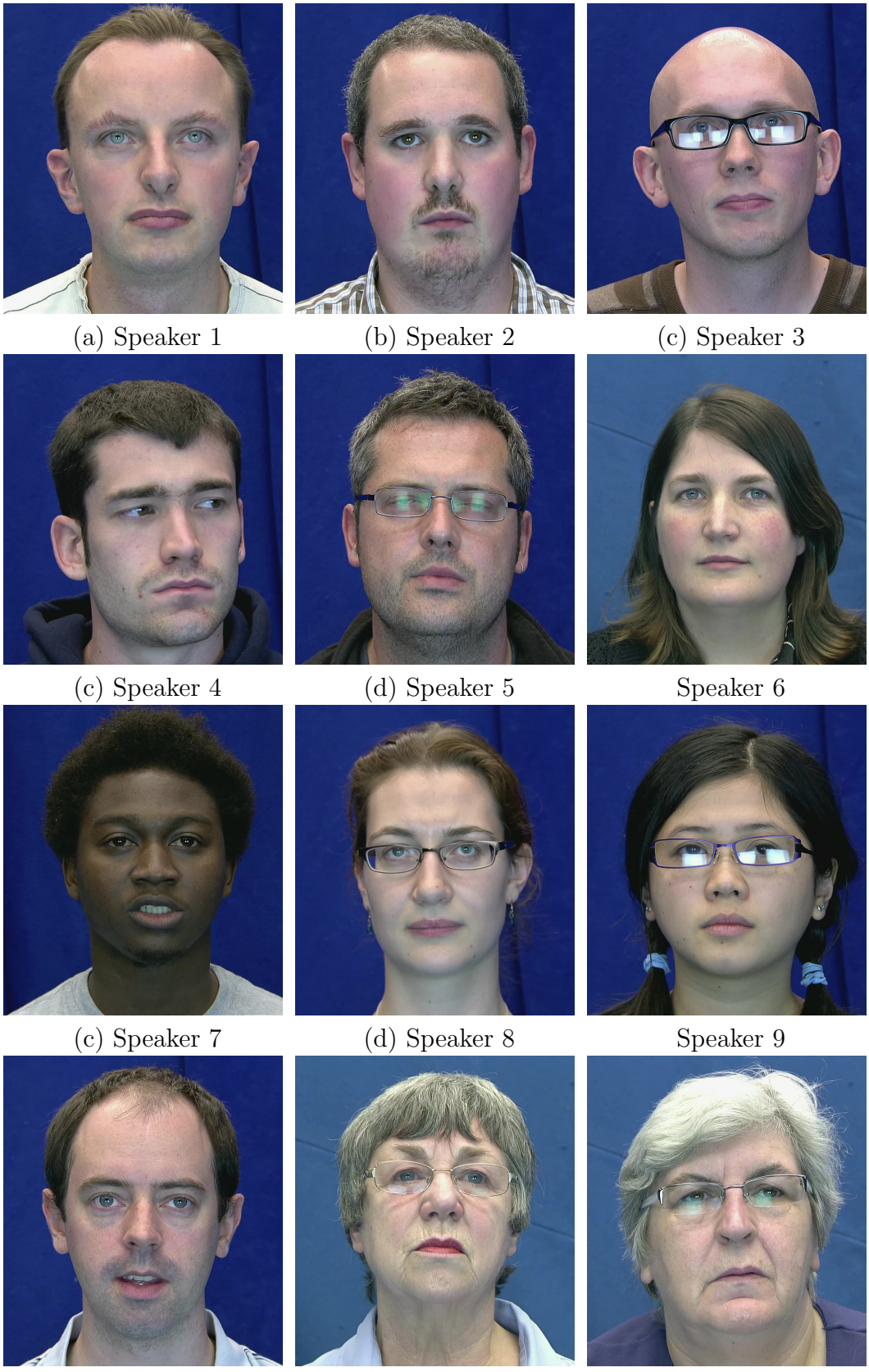



Figure 3.5: Occurrence frequency of phonemes in the RMAV dataset.



(a) Speaker 1 (b) Speaker 2 (c) Speaker 3

(c) Speaker 4 (d) Speaker 5 Speaker 6

(c) Speaker 7 (d) Speaker 8 Speaker 9

(c) Speaker 10 (d) Speaker 11 Speaker 12

Figure 3.6: Example faces from the RMAV videos (12 speakers).

# Chapter 4

# Current difficulties in machine lip-reading

In Chapter 1, we identified a number of factors, or affects, in machine lip-reading which are often difficult to control such as lighting, pose, identity, motion, emotion, linguistic content and expression. We now address these challenges in turn.

## 4.1 Motion

The ability to recognise lip gestures throughout a video is addressed in the tracking part of the lip-reading task. There are two systems most commonly used for tracking faces in videos for machine lip-reading. These systems are Active Appearance Models (which can be shape, appearance or shape and appearance models) [33] and Linear Predictors [112]. Both of these systems are effective, even on low quality videos, for tracking the motion of a face during speech. Chapter 2 has described these two systems in full. Both methods make some assumptions about motion within videos, LPs are locally affine whereas AAMs are globally affine. Therefore the only minor issue that remains is for non-affine transformations.

## 4.2 Pose

There is literature about the effects of pose on computer lip-reading. Some look at expression recognition for Human Computer Interactions (HCI) [106] and present an improvement in expression recognition by computers and humans when the pose is rotated to 45°. Others by Kumar *et al.* and Kaucic *et al.* [74, 71], look at visual speech classification and suggest that the profile view gives a better classification. However, they also processed the visual features over a longer time period than the duration marked by the endpoints of each speech utterance to consider co-articulation within their tests and so can not isolate which of the longer time window or the pose improved classification.

When considering lip-reading, the study in [11] examines the effects of human sentence perception across three viewing angles in relation to the camera position: full-frontal view (0°), angled view (45°), and side view (90°). The performance of a female adult with post-lingual hearing loss was measured for accuracy at each angle. This study used a single-subject, with alternating treatment design where three treatment angles were randomly presented in every session. The accuracy for each session was compared to determine the most effective viewing angle of the speaker. The results indicated that the side-view angle was most effective, as the percentage gain of improvement was greatest in combination with the consistent upward trend of the data points across treatment sessions. The performance of frontal-view and angled-view angles were also successful but not significantly more so than full-frontal. The results of this preliminary effort indicate the value of treatment for visual sentence perception at all three angles, including the non-traditionally targeted side view for human lip-reading.

Preliminary studies into non-frontal pose affects in lip-reading can be found in [87] & [75]. In both a small vocabulary is used in order to simplify the recognition task for measuring the effects of features extracted from non-frontal camera positions. In [87] the classifiers were trained on frontal features and tested on non-frontal features and the results showed that the greater the off-frontal angle became, then the word

error rate increased. However, the frontal view features provided inferior recognition to off-angle features in [75]. The key distinction between these studies is the visual noise of image backgrounds in the original videos.

Most AVSR databases are recorded face frontal, an alternative idea of lip-reading non-frontal camera angles with frontal-trained classifiers using a mapping from the recorded angle to the estimated actual angle of the speaker to the camera is presented in [116]. In this work, we see a new dataset recorded for the specifically for the mapping technique and the results support the observations in [87] & [75] but add the observation that with the larger off-camera angles, then a smaller feature vector of only the higher order features is preferable.

These studies into the affect of pose on machine lip-reading are taken further by Lucey *et al.* [88] with a proprietary dataset. Here the authors undertake three activities with a small vocabulary (connected digit strings) on 38 speakers; comparing the frontal and profile view lip-reading performance (akin to the experiments in [87] & [75]), but they also take the challenge further by experimenting with concatenating both the frontal and profile view features into *multi-view* features, and attempting to lip-read using a single pose-invariant normalisation method. The results for task one support those seen in [87] whereby the frontal features outperformed the profile features. This is considered due to both datasets being recorded in controlled conditions with minimal noise.

The results for the *multi-view* features in [88], marginally better than frontal, and significantly better than profile features. The $w.e.r$ reduces from 38.88% for profile features, for 27.66% for frontal features and the best *multi-view* features achieved a 25.36% $w.e.r$. This was achieved by simply concatenating the two sets of features. This observation is important that it is important to not simply pick a pose for lip-reading, but rather, there are useful visual cues from all angles.

Finally, in the third test, Lucey *et al.* develop a single pose-invariant model for lip-reading, regardless of the pose of the test data. They compare different pairings of features over the training/testing split. For example, using frontal features $F$, for

training and testing with frontal features. Then using the same features $F$, to test profile features $P$ and vice versa. A third training model using a 50/50 split of $F$ and $P$ is included in the experiment setup. Also adopted is the projection of each set of features, $F$ and $P$ into the alternative feature space for new features $F'$ and $P'$ for alternative testing data for the three training options, $F$, $P$, and $[F_{50}, P_{50}]$. These tests showed best recognition where the training and test features matched. Where these didn't match the $w.e.r$ dramatically increased, for example for an ($F$,$F$) train/test pairing the $w.e.r$ was 29.18%. The train/test pair of ($F$,$P$) achieves a $w.e.r$ of 87.07%. However, the authors also show that this can be mitigated by the projection of the test profile data back into the frontal feature space where the train/test split ($F$,$P'$) recovers the $w.e.r$ back down to 54.85%. This transformation principle is also used in [78] by Lan *et al.* who presented an view-independent lip-reading system. This investigation uses a continuous speech corpus compared to the small vocabulary dataset in [88]. This later study acknowledges a human lip-readers preference for a non-frontal view and suggests it could be attributed to lip protrusion. A different approach for the feature transform is presented, (a linear mapping between poses) but the development of a such system shows computer lip-reading can be independent of speaker pose.

## 4.3 Multiple people

The challenge of machine lip-reading a video with more than one person, meaning to track their faces, has a number of solutions. [68] demonstrates multiple person tracking (albeit not lip-reading) and has also implemented this into a simple HCI system. Also, in [81] we see how a person can be re-identified between videos, either a second view of the same space at an alternate perspective or, as a person moves through a location. An example of a speaker identification method is detailed in [89], and [70, 84] detail lip-reading of multiple people, [70] recognises consonants, and [84] visual vowels. Whilst none of these papers have directly tested concurrent speech, it would be interesting to know what effect, if any, speakers talking in unison would

cause upon current lip-reading systems. [37] presents an audio-visual system for HCI which automatically detects a talking person (both spatially and temporally) using video and audio data from a single microphone. Until visual-only classifiers have improved, a robust visual-only system for machine lip-reading still needs to be developed and the classifiers are a essential part of the system.

## 4.4 Video conditions

Studies such as [22] on the effect of low video frame-rate on human speech intelligibility during video communications, suggest that lower frame rates encourage humans to over-articulate to compensate for the reduced visual information available, akin to a visual Lombard effect. (N.B. this is only when the speakers are aware of the low quality parameters e.g. during a video conference.) Therefore, it should be asked: does a computer need more information (higher frame rate/resolution) to lip-read a speaker in a recorded video sequence? The study in [22] observes in face-to-face human interactions, articulation is relaxed. So one could ask, in the instance where a computer needs extra visual information throughout the recording, (think of the example where a face-to-face conversation is being recorded incognito), how much does this lack of visual information impact on the classification performance? That is, how far does the lack of video recording quality affect classification?

Another study into frame rate in computer lip-reading, [124], tells us the greatest classification is achieved when the same frame rate is used for both training and testing data. This is perhaps unsurprising as it is shown that when both training and test data sets are at low frame rates, classification drops when the frame rate of the training data is lower than the test data. They show longer words are easier to classify. It would be interesting to see if this is the same for visemes. [124] also shows a dependency between frame rates and classification accuracy by speaker. When training and test data do not have the same, or very similar frame rates, it is recommended training data has a higher frame rate (for feature extraction) than the

test (fit) data. It observes word classification rates vary in a non-linear fashion as the frame rate is reduced which is caused by the particular words being recognised. The duration of an utterance does not have an effect on the classification rate in this paper.

## 4.5 Speech methods and rates

People have different speaking styles, accents and rates of speech. Some people talk fast, some slow, some talk out of the side of their mouth, others naturally over-articulate and others have facial hair which occludes the visibility of lip movement during speech. The rate of speech alters both an utterance duration and articulator positions. Therefore, both the sounds produced, but particularly, visible appearance are altered. In [134], the authors present an experiment which measures the effect of speech rate and shows the effect is significantly higher on visual speech than in acoustic.

Because of this variable, some people undertake elocution classes for a myriad of reasons. Examples include call centre employees undertaking 'accent neutralisation' courses to make them more approachable for their target customers [34]. This is supported in [55] where they state "Speakers of non-prestige dialects in some countries take elocution courses, or respond to newspaper adverts which promise to 'eliminate' their 'embarrassing' accents, and second language learners fret that they'll never sound like a native.".

## 4.6 Resolution

In this chapter we have reviewed the environmental affects of lip-reading classification. Whilst many can be controlled, and we have seen in the literature how some of the effects can be managed, we also note previously considered challenges such as, outdoor video, poor lighting, and agile motion can all be overcome [24].

In regards to studies about the affects of resolution, there is limited literature found at the time of writing which examines this. Some experiments touch on this area of interest with investigations into recognition from noisy images.

An investigation into the effects of compression artefacts, visual noise (simulated with white noise), localisation errors in training is presented in [59], and in [143] the authors undertake two experiments, of which the first includes some attention to spatial resolution (the number of pixels). This inclusion of features from three different resolutions is interesting but the resolutions selected have differing aspect ratios and as such it is not a controlled method of resolution variation. Also, the effect of this spatial resolution is not measured or presented, rather it is included as a property of tests on frame rate and contrast. Neither of these papers consider the simple removal of information from a smaller image compared to a larger one.

Therefore testing of this is necessary (see Chapter 5). Given that, up to this point, with a known speaker and reduced linguistic context, classification rates can be high, it is a fair bet the most sensitivity is to be found on the parameters associated with the left hand side of Figure 1.2 (identity, expression etc). Nevertheless, there has been surprisingly little attention paid to a systematic review of the cameras parameters. Therefore, in our first practical experiment we ask 'what is the lowest resolution at which a machine can lip-read?'.

# Chapter 5

# Resolution limits in lip-reading

We have discussed how machine lip-reading depends on factors which can be difficult to control, such as: lighting [131], identity [35], motion [77] and pose [71, 78, 74, 11], rate of speech [134], and expression [106]. But some factors, such as video resolution, are controllable. So it is surprising there is not yet a specific, systematic and complete study of the effect of resolution on lip-reading in non-noisy conditions. There is a tendency, without evidence, to assume a high resolution video will produce better classification results and so a study to measure the effect of resolution on classification is needed and this is undertaken in this chapter.

## 5.1 Image pre-processing for feature modification

For this work we use the Rosetta Raven dataset as already described in Section 3.3. Before feature extraction however, we undertake some image pre-processing. All four videos in the dataset were converted into a set of images (one per frame in PNG format) with ffmpeg [140] using image2 encoding at full high-definition resolution ($1440 \times 1080$).

To build an initial Active Appearance Model for tracking each video, we select the first frame and nine or ten others randomly. These *key frames* are hand-labelled

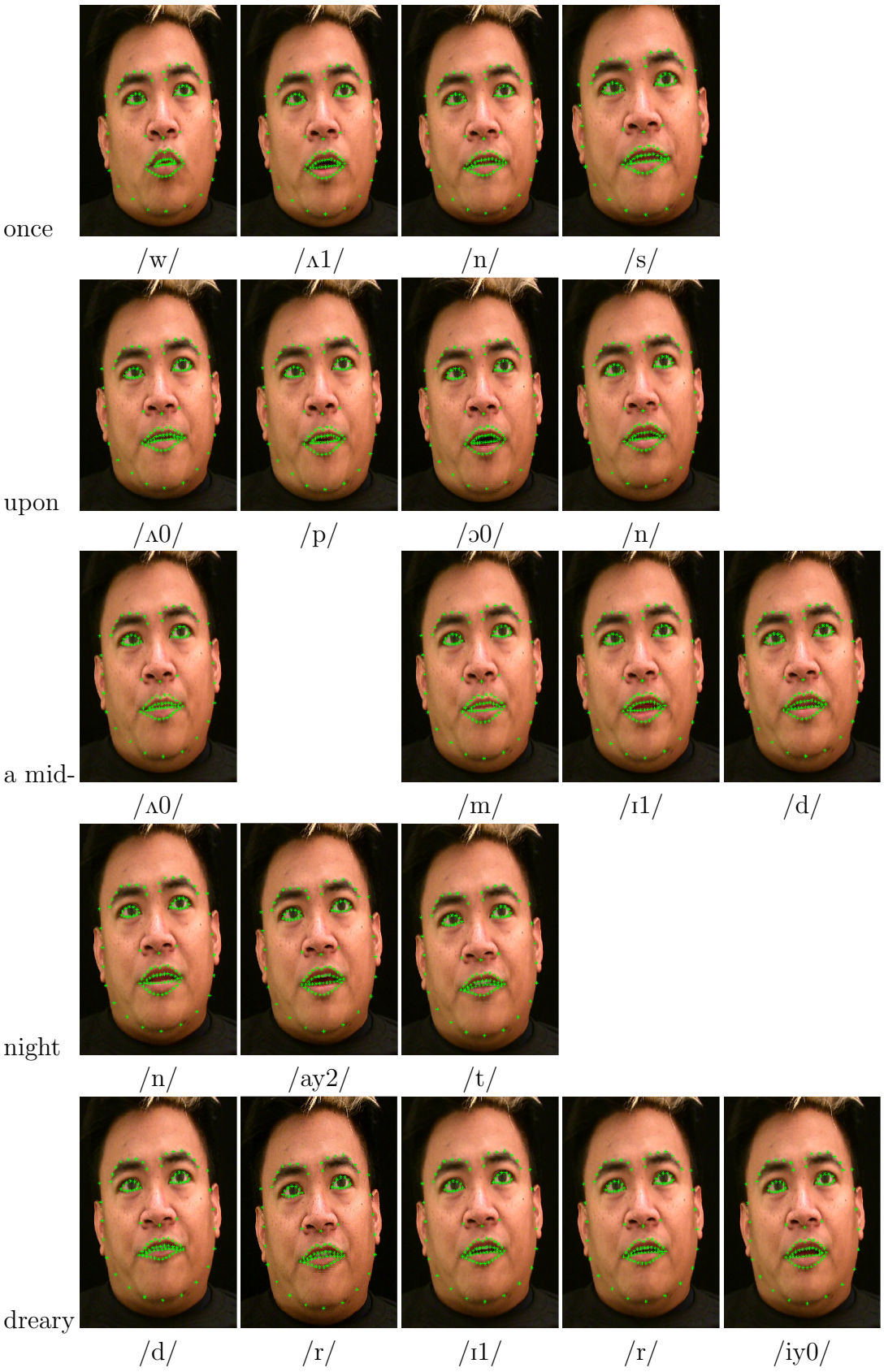



Figure 5.1: Tracking a Rosetta Raven speaker saying 'Once upon a midnight dreary' with a full-face Active Appearance Model.

with a model of a face including: facial outline (jaw and hairline, in front of ears), eyebrows, eyes, nose and lips. To track the face, this preliminary AAM is then fitted, via Inverse Composition fitting [7, 97] to the unlabelled frames (Table 3.3 in Chapter 3 gives the numbers of frames for each video). In Figure 5.1 we show, for Speaker 1, the tracked full-face AAM mesh (one frame per phoneme), for the first sentence of The Raven "Once upon a midnight dreary" used in tracking the speaker face.

At this stage full-face speaker dependent AAMs are tracked and fitted on all full resolution lossless PNG frame images as in Figure 5.2 (a) and (b) for both speakers in the Rosetta Raven dataset.

(a) S1 face AAM points (b) S2 face AAM points (c) S1 lips AAM points (d) S2 lips AAM points

Figure 5.2: Active Appearance Model shape landmarks for two Rosetta Raven speakers.

The AAMs used for tracking are now decomposed into sub-models for the eyes, eyebrows, nose, face outline and lips. The purpose of this is to allow us to obtain a robust fit from the full face model but extract features of only the lip information for use during classification. Both speaker lips sub-model can be seen in Figure 5.2 (c) and (d). There are 24 landmarks in the outer lip contour and 20 in the inner lip contour. Next, the video frames used in the high-resolution tracking were down-sampled to each of the required resolutions (listed below) by nearest neighbour sampling (Figure 5.3(b)) and then up-sampled via bilinear sampling (Figure 5.3(c)) to provide us with 18 sets of frames per original video. We use a different sampling method to upsample as this provided a more consistent visual degradation of information in the resulting images to show the reduction in resolution with minimum

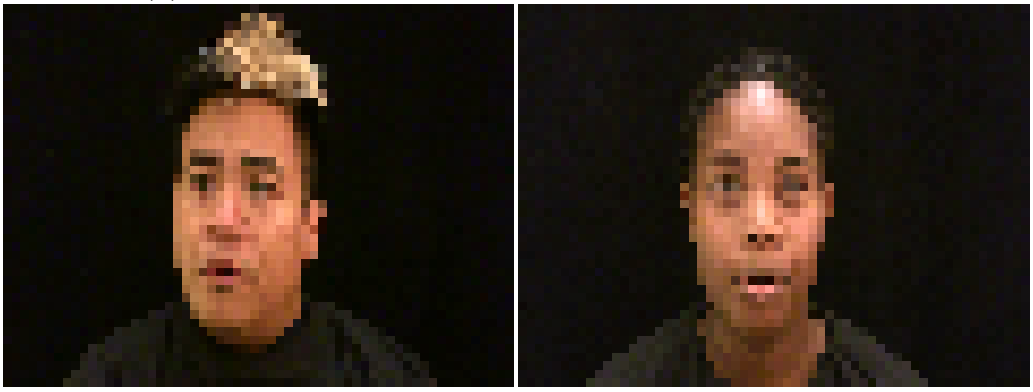
consistent processing artefacts compared to other sampling methods. These new frames are the same physical size as the original ($1440 \times 1080$) recordings but contain less information due to the downsampling i.e. only the information available at a lower resolution version of the original.

1. $1440 \times 1080$
2. $960 \times 720$
3. $720 \times 540$
4. $360 \times 270$
5. $240 \times 180$
6. $180 \times 135$
7. $144 \times 108$
8. $120 \times 90$
9. $90 \times 67$
10. $80 \times 60$
11. $72 \times 54$
12. $65 \times 49$
13. $69 \times 45$
14. $55 \times 42$
15. $51 \times 39$
16. $48 \times 36$
17. $45 \times 34$
18. $42 \times 32$

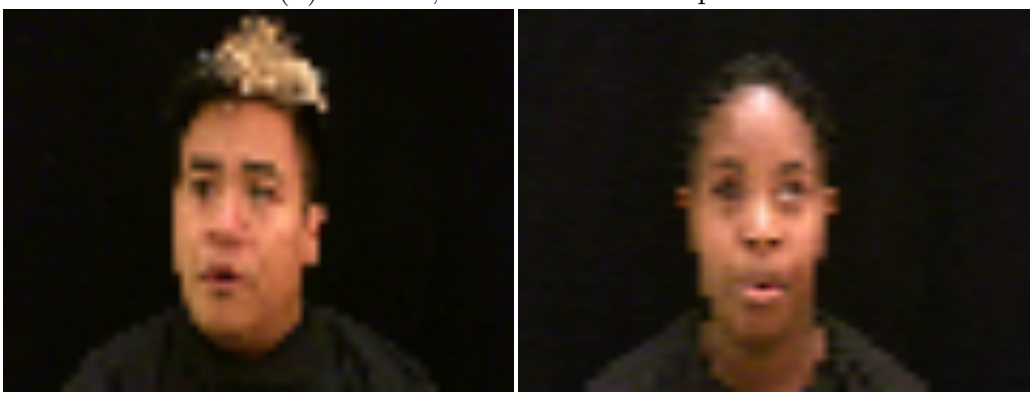
We remind the reader that our point of interest in this study, is the affect low resolution has on the loss of lip-reading information, rather than the affect it would also have on the AAM tracking process. Some AAM trackers lose track quite easily at low resolutions or on lossy images and we do not wish to be overwhelmed with catastrophic errors caused by tracking issues or artefacts which can often be solved in other ways [113]. Accordingly, this is why we have fitted at the original full resolution before the refitting of the lips sub model for feature extraction. Consequently the shape features in this experiment are unaffected by the downsampling process, whereas the appearance features vary. This will turn out to be a useful benchmark.

Our image processing method is specific to our research question, what are the limitations (if any) of resolution in achieving machine lip reading? We have minimised the effects of compression artefacts by using the most successful pair of algorithms for downsampling and upsampling respectively. By using a dataset recorded in laboratory controlled conditions we have no white noise or occlusions. There are of course other methods available to us, such as simply filling the feature vectors with zeros to represent the loss of data, or not resizing the smaller images back to

(a) 1440 × 1080, Original resolution image for S1 & S2



(b) 60 × 45, S1 & S2 downsampled



(c) 1440 × 1080, S1 & S2 restored

Figure 5.3: Downsampling of frame images in PNG format: (a) Original captured images, (b) nearest neighbour down-sampled images and (c) and their bilinear sampled restored pictures without original high definition information.

the original size. But the major advantage of our method is that it encourages good tracking with the AAM and with this good tracking, we can complete a direct A to B comparison of classification outputs from features derived from videos with varying resolution information.

For Speaker 1 (S1), six shape and 14 appearance parameters and for Speaker 2 (S2), seven shape and 14 appearance parameters are retained. This number of parameters was chosen to retain 95% variance in facial information in the usual way [33], see Table 3.4 presented in Chapter 3.

## 5.2 Classification method

Table 5.1: A phoneme-to-viseme mapping from combining Walden's consonant visemes with Montgomery's vowel visemes.

| vID | Phonemes | vID | Phonemes |
|---|---|---|---|
| v01 | /p/ /b/ /m/ | v10 | /i/ /ɪ/ |
| v02 | /f/ /v/ | v11 | /eh/ /ae/ /ey/ /ay/ |
| v03 | /θ/ /ð/ | v12 | /ɑ/ /ɔ/ /ʌ/ |
| v04 | /t/ /d/ /n/ /k/ /g/ /h/ /j/ /ŋ/ /y/ | v13 | /ʊ/ /ɜ/ /ax/ |
| v05 | /s/ /z/ | v14 | /u/ /uw/ |
| v06 | /l/ | v15 | /ɔɪ/ |
| v07 | /r/ | v16 | /iy/ /hh/ |
| v08 | /ʃ/ /ʒ/ /tʃ/ /dʒ/ | v17 | /ɑʊ/ /əʊ/ |
| v09 | /w/ | v18 | /sil/ /sp/ |

We listened to each recitation of the poem and produced a ground truth text (some recitations of the poem are not word-perfect to the original writing (see Appendix 10.2)). This word transcript is converted to an American English phoneme-level transcript using the CMU pronunciation dictionary [29] introduced in Chapter 3. Then, using the viseme mapping based upon Walden's consonants [144] and Montgomery *et al.*'s [94] vowel phoneme-to-viseme mapping (as in Table 5.1), a viseme transcript was created. Thus we have translated each recitation from words, to phonemes, and finally, to visemes. Viseme classification is selected over phonemes as, on a small data set, it has the benefits of reducing the number of classifiers needed and increasing the training data available for each viseme classifier. Note not all visemes are equally represented in the data as is shown by the viseme histogram in Figure 5.4, Chapter 3. Whist the volumes in this Figure are lower than an equivalent histogram for a continuous speech dataset, the distributions are similar.

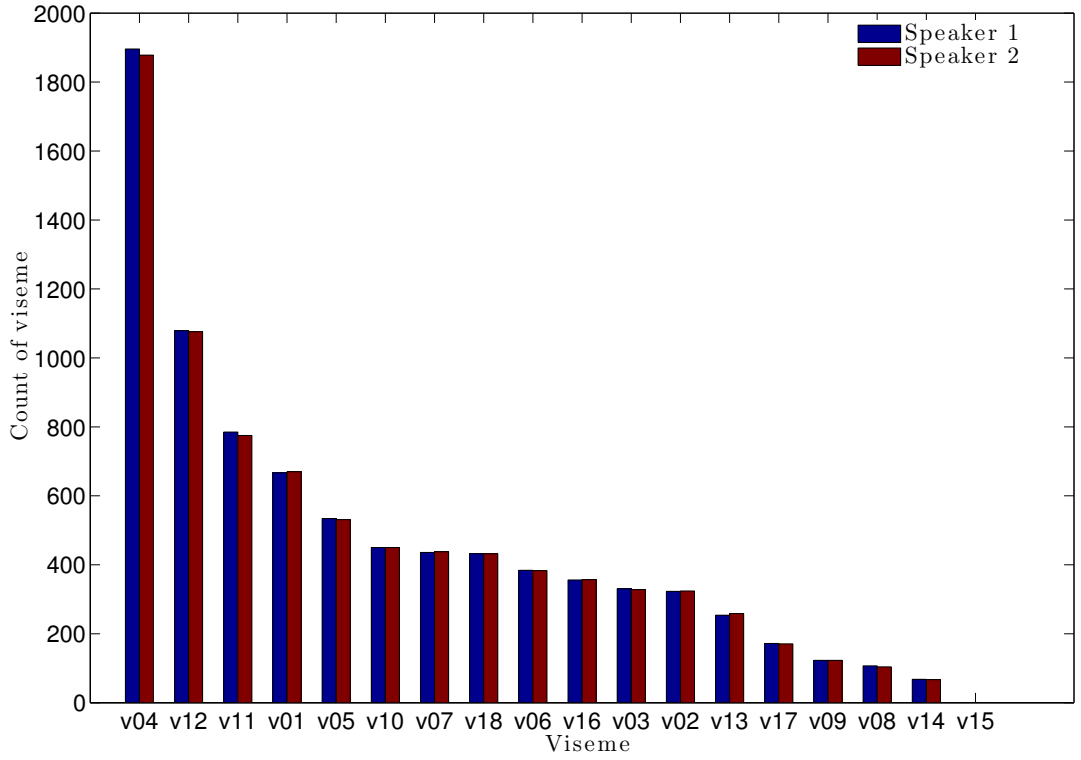



Figure 5.4: Occurrence frequency of visemes per speaker based upon ground truth transcripts of the Rosetta Raven dataset speakers using Walden's and Montgomery's visemes.

For each speaker, a test fold is randomly selected as 42 of the 108 lines (20% of data) in the poem. The remaining lines (80% of data) are used as the training fold. Repeating this five times gives five-fold cross-validation. Note visemes cannot be equally represented in all folds.

For classification Hidden Markov Models (HMMs) are built with the Hidden Markov Toolkit (HTK) [150] already introduced in Section 2.5.1. An HMM is initialised using the 'flat start' method (using `HCompV`), with a prototype of five states and five mixture components, and the information in the training samples. Five states and five mixtures are selected based upon the work in [96]. An HMM is defined for each viseme plus silence and short-pause labels (Table 5.1) and we re-estimate the HMM parameters four times with no pruning.

The HTK tool `HHEd` ties together the short-pause and silence models between

states two and three before re-estimating the HMMs a further two times. Then `HVite` is used with the `-m` flag to force-align the data using the word transcript. We create a viseme version of the CMU dictionary for word-to-viseme mapping (whereby the phonemes are replaced with their respective viseme characters from the phoneme-to-viseme map in Table 5.1) and use this viseme CMU dictionary to produce a time-aligned viseme transcription which includes natural breakpoints between words.

The HMMs are now re-estimated twice more. However, now the force-aligned viseme transcript replaces the original viseme transcript used in the previous HMM re-estimations. A word network is needed to complete the classification. `HLStats` and `HBuild` used together twice make both a Unigram Word-level Network (UWN) and a Bigram Word-level Network (BWN). Finally, `HVite` is used with the different network support for the classification task and `HResults` gives us the correctness and accuracy values. All HTK tools named here are described in Chapter 2.5.1.

## 5.3 Analysis of resolution affects on classification

Accuracy, $A$, (Equation 2.8), is selected as a measure rather than correctness, $C$, (Equation 2.7) since it accounts for all errors. Including insertion errors is important as they are notoriously common in lip-reading. An insertion error occurs when the recogniser output has extra words/visemes not present in the original transcript [150]. As an example one could say, "Once upon a midnight dreary",

but the recogniser outputs:

"Once upon upon midnight dreary dreary".

Here the recogniser has inserted two words which were never present,

"Once upon **upon** midnight dreary **dreary**"

and it has deleted one ('a'). The missing 'a' is a deletion error.

"Once upon **...** midnight dreary".

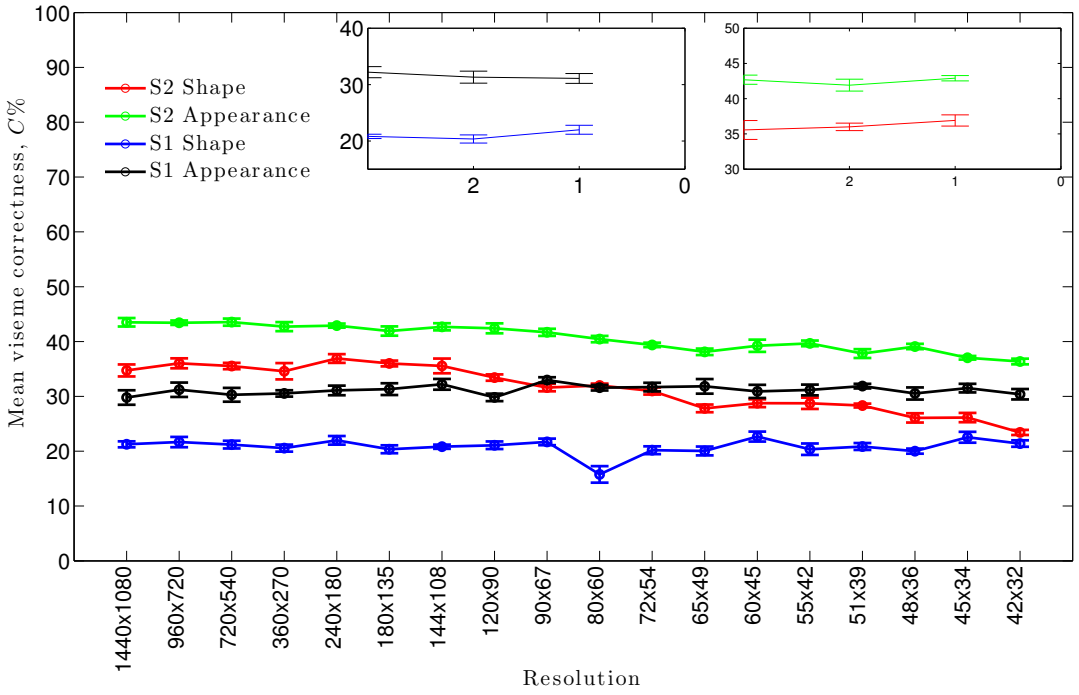



Figure 5.5: Viseme classification in Correctness, $C \pm 1\frac{\sigma}{\sqrt{5}}$, with a unigram word network (on the $y$-axis) at 18 degraded measured in pixels ($x$-axis).

In Figures 5.5 and 5.7 we have plotted, for our 18 different resolutions along the $x$-axis, the mean viseme correctness on the $y$-axis for each speaker. Supported by a unigram language network and bigram language network respectively. Speaker 1 shape classification is shown in blue and appearance classification in black. Speaker 2 shape and appearance classification is plotted in red and green respectively. The corresponding graphs of mean accuracy classification are shown in Figures 5.6 and 5.8. All four figures include one standard error over the five folds.

Figure 5.6 plots viseme accuracy with a unigram network on the $y$-axis and all points are negative values. This is worse than chance and demonstrates the debilitating effect of insertion errors where the language network is not strong enough to sieve them out of the classification output. Viseme correctness supported by a unigram word network is shown in Figure 5.5, where we see a slow but significant decrease in classification as the resolutions decrease in size along the $x$-axis. At no point do the appearance features drop below the shape features. This trend is

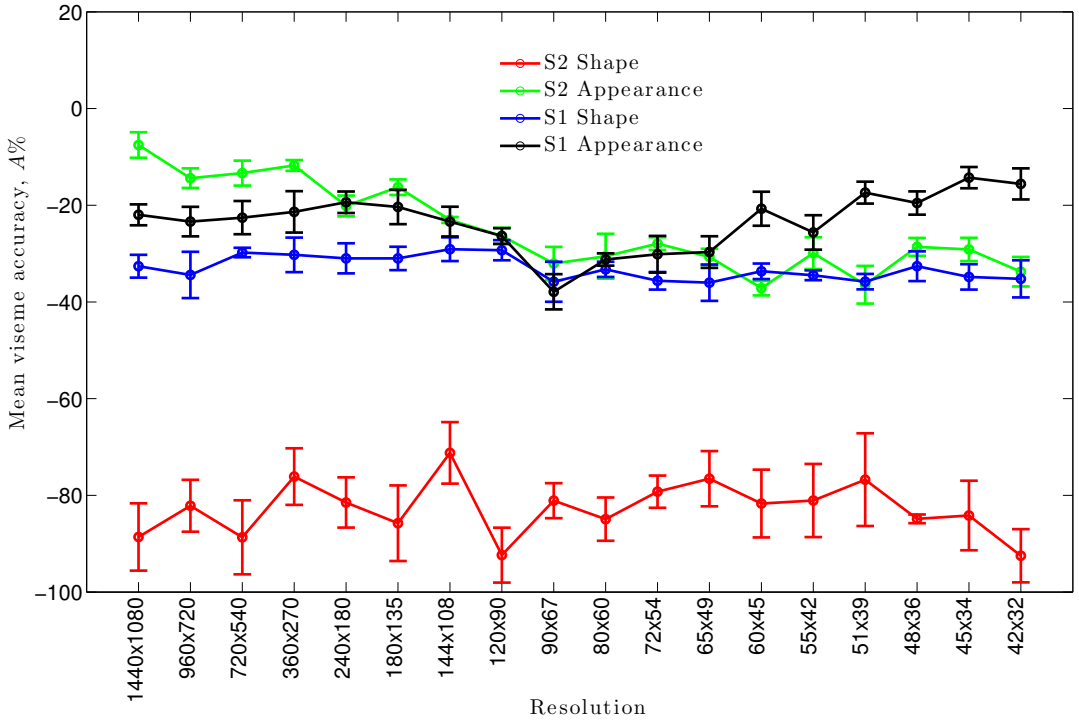



Figure 5.6: Viseme classification in Accuracy, $A \pm 1\frac{\sigma}{\sqrt{5}}$, with a unigram word network (on the $y$-axis) at 18 degraded resolutions in pixels ($x$-axis).

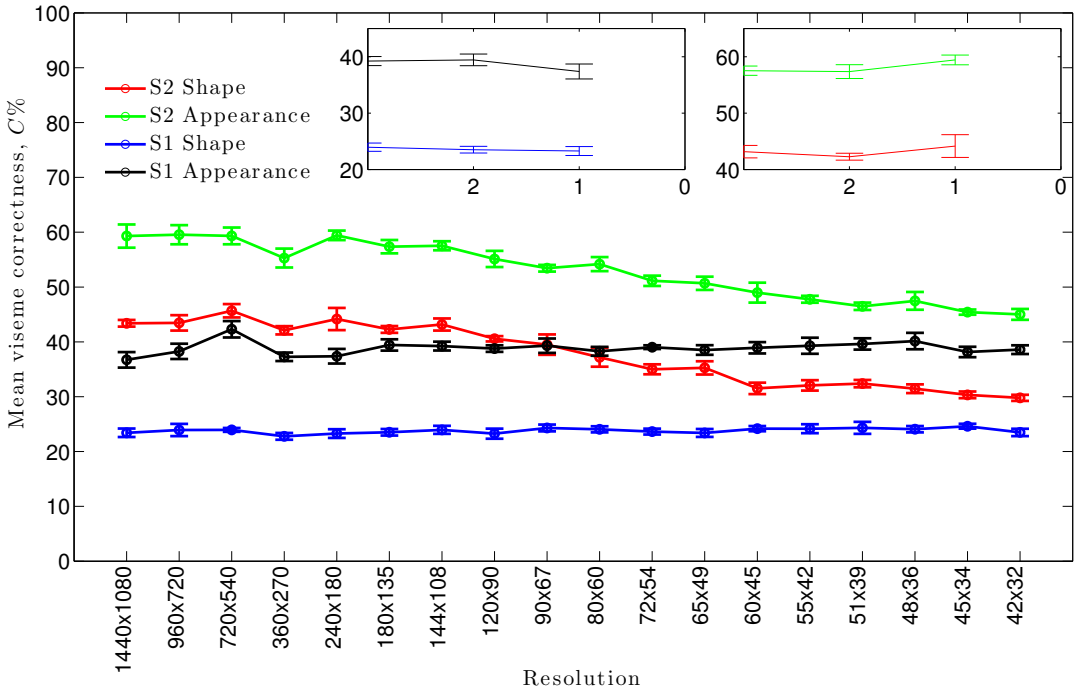



Figure 5.7: Viseme classification in Correctness, $C \pm 1\frac{\sigma}{\sqrt{5}}$, with a bigram word network (on the $y$-axis) at 18 degraded resolutions in pixels ($x$-axis).

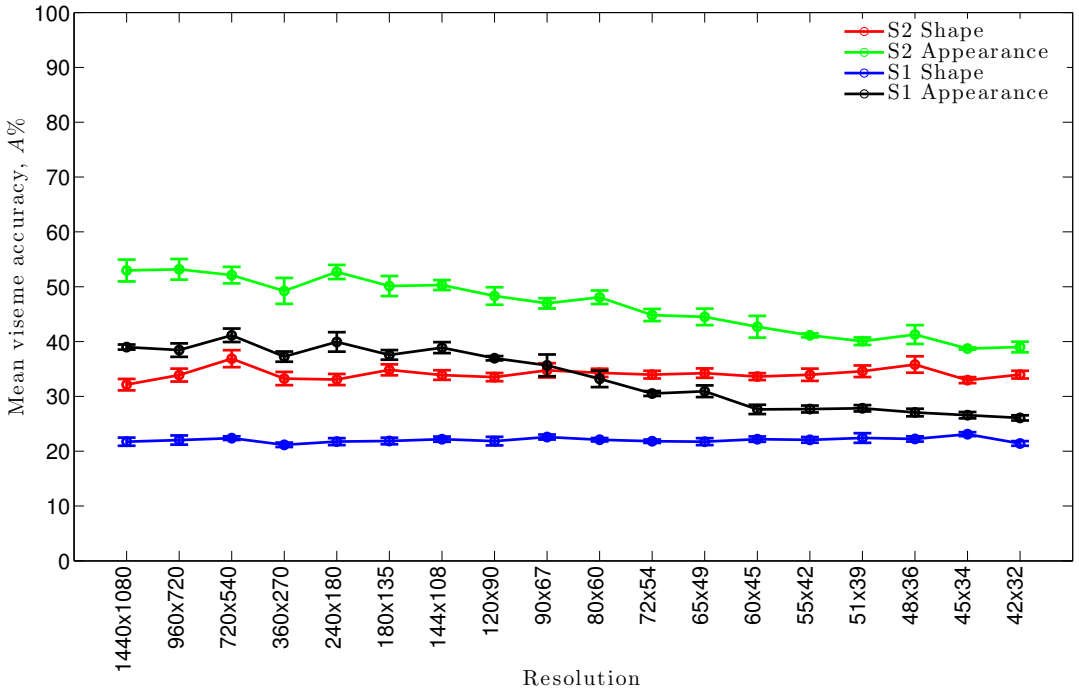



Figure 5.8: Viseme classification in Accuracy, $A \pm 1\frac{\sigma}{\sqrt{5}}$, with a bigram word network (on the $y$-axis) at 18 degraded resolutions in pixels ($x$-axis).

matched in our BWN experiments in Figures 5.7 and 5.8.

These Figures, however, are not normalised to account for the actual differences in information between resolutions. As we can see in our list of resolutions in Section 5.1, there is not an equal interval between each size. Therefore we replot these results by measuring the resting lip-pixels which cover the lip-shape. The resting lip pixel distance is shown in Figure 5.9 for our two speakers in the first $1080 \times 1440$ resolution image frame. This means, as there are less pixels per lip we can appropriately plot along our $x$-axis as we have done in Figures 5.10, 5.11, 5.12 and 5.13.

Figure 5.11 shows the accuracy, $A$, (on the $y$-axis) versus resolution (on the $x$-axis) for an UWN. The $x$-axis is calibrated by the vertical height of the lips of each speaker in their rest position (Figure 5.9). For example, at the maximum resolution of $1440 \times 1080$ speaker S1 has a lip-height of approximately 26 pixels in the rest position whereas S2 has a lip-height of approximately 17 pixels. The worst

(a) S1 (b) S2

Figure 5.9: Showing the resting lip-pixel distance measures for two Rosetta Raven speakers.

performance is from speaker S2 using shape-only features. The shape features do not vary with resolution so any variation in this curve is due to the cross-fold validation error (all folds do not contain all visemes equally). Nevertheless, the variation is within one standard error, and so not signifiant. This is not a surprise as AAM shape features are scale invariant. The poor performance is, as usual with lip-reading, dominated by insertion errors (hence the negative $A$ values in Figure 5.11). The usual explanation for this effect is shape data contain a few characteristic shapes (which are easily recognised) in a sea of indistinct shapes - it is easier for a classifier to insert garbage symbols than it is to learn the duration of a symbol which has an indistinct start and end shape due to co-articulation. We suggest that speaker S1 has more distinctive shapes so scores better on the shape feature as more distinctive shapes between classification models differentiate more definitively.

However, it is the appearance features which are of more interest since this varies as we downsample. At resolutions lower than four pixels it is difficult to be confident the shape information is effective. However, the basic problem is a very high error rate (shown in Figures 5.10 and 5.11) therefore a more supportive word model is required [67].

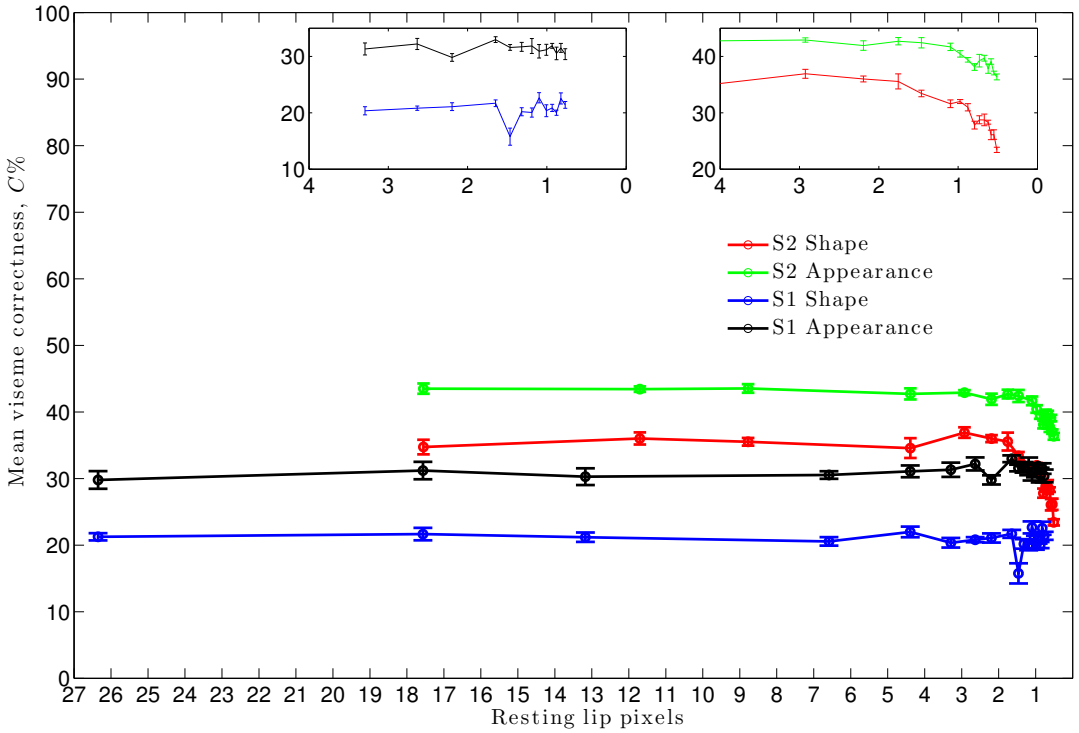



Figure 5.10: Viseme classification in Correctness, $C \pm 1\frac{\sigma}{\sqrt{5}}$, with a unigram word network (on the $y$-axis) by vertical resting lip height in pixels ($x$-axis).

Figures 5.12 and 5.13 shows the classification accuracy versus resolution (represented by the same $x$-axis calibration in Figures 5.10 and 5.11) for a BWN. It also includes two sub-plots which magnify the right-most part of the graph. Again, the shape models perform worse than the appearance models, but looking at the magnified plots, appearance never becomes as poor as shape performance even at very low resolutions. As with the UWN accuracies, there is a clear inflection point at around four pixels (at two pixels per lip), and by two pixels the performance has declined significantly.

In Table 5.2 we have listed the different error types (insertion, deletions and substitutions) which can occur during classification for resolutions just before our identified minimum lip pixel threshold as well as just after. The values are the total errors over all five folds of cross validation. For Speaker 1, both deletion and substitution errors increase when there is no longer have enough pixels to differentiate between the two lips. For Speaker 2, we see only the substitution errors increase

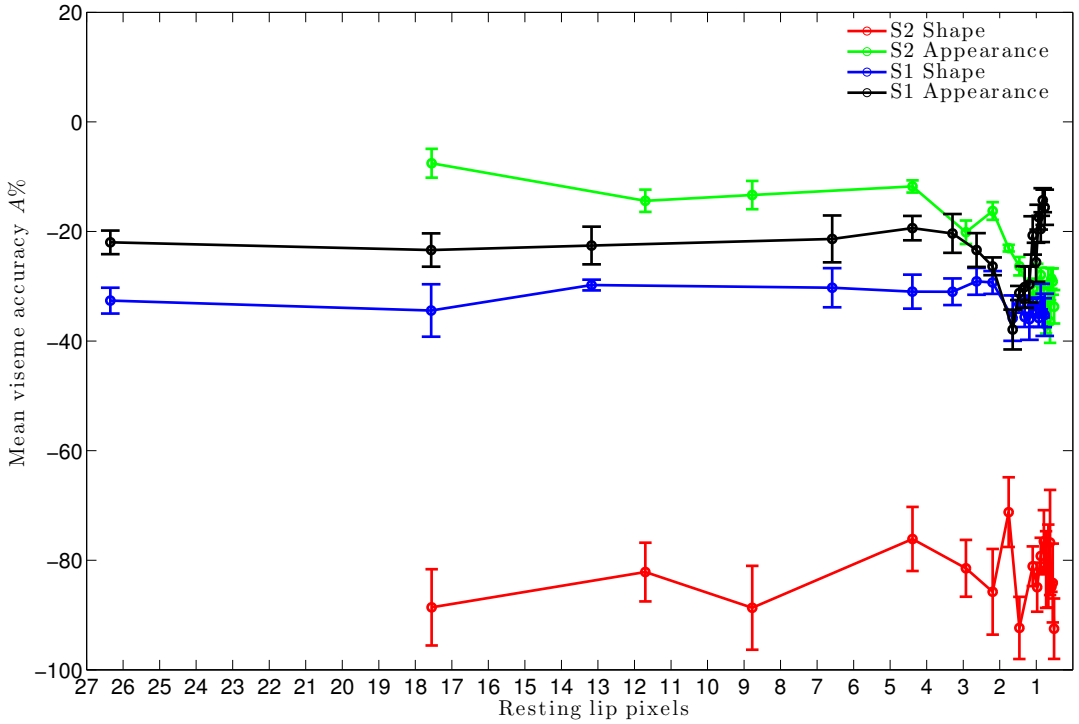



Figure 5.11: Viseme classification in Accuracy, $A \pm 1\frac{\sigma}{\sqrt{5}}$, with a unigram word network (on the $y$-axis) by vertical resting lip height in pixels ($x$-axis).

Table 5.2: Insertion, deletion and substitution error counts in classification transcripts at the smallest resolution above (before), and the largest resolution below (after), the minimum required lip pixel height of two pixels per lip. The values are the total sum over all five folds of cross validation.

| | Insertion | Deletion | Substitution |
|---|---|---|---|
| Speaker 1: | | | |
| Before | 348 | 3,385 | 1,298 |
| After | 305 | 3,646 | 1,355 |
| % change | −12% | +8% | +4% |
| Speaker 2: | | | |
| Before | 571 | 2,339 | 1,423 |
| After | 531 | 2,322 | 1,500 |
| % change | −7% | −1% | +5% |

but the deletion errors only decrease insignificantly at −1%.

It is interesting to see there are fewer insertion errors after our minimum lip-pixel threshold. In Chapter 2 we saw the difference between Accuracy (Equation 2.8) and Correctness (Equation 2.7) were the Insertion errors. Therefore, we can say we may

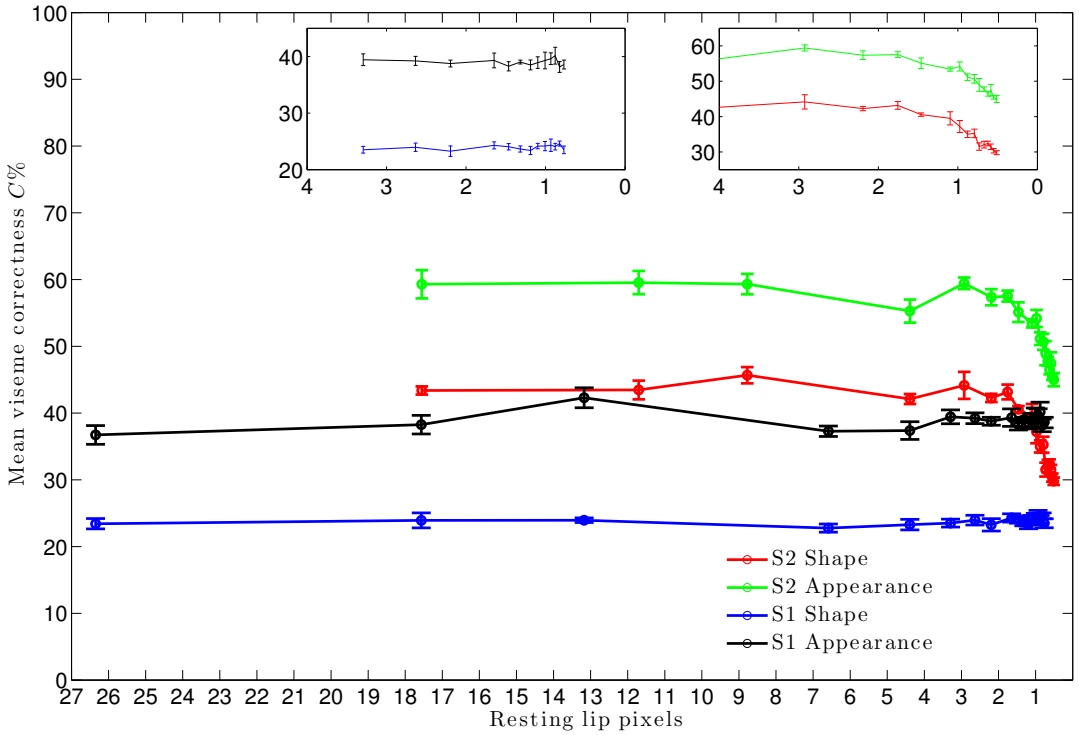



Figure 5.12: Viseme classification in Correctness, $C \pm 1\frac{\sigma}{\sqrt{5}}$, with a bigram word network (on the $y$-axis) by vertical resting lip height in pixels ($x$-axis).

need more visemes within a set to keep insertion errors down as these ensure more minor differences between classifiers are encapsulated within training.

## 5.4 The effect of resolution on lip-reading classifiers

In Chapter 4 we discussed the limitations in machine lip-reading. In this chapter we have added to this knowledge with our experiment into resolution.

Using the new Rosetta Raven data we have shown lip-reading HMM classifiers to have a threshold effect with resolution. We have trained and tested viseme classifiers and measured the effect on classification accuracy as we systematically reduced the resolution information in a video. The best recognition achieved was 59.55% accuracy with Speaker 2's appearance data with a bigram word level language model,

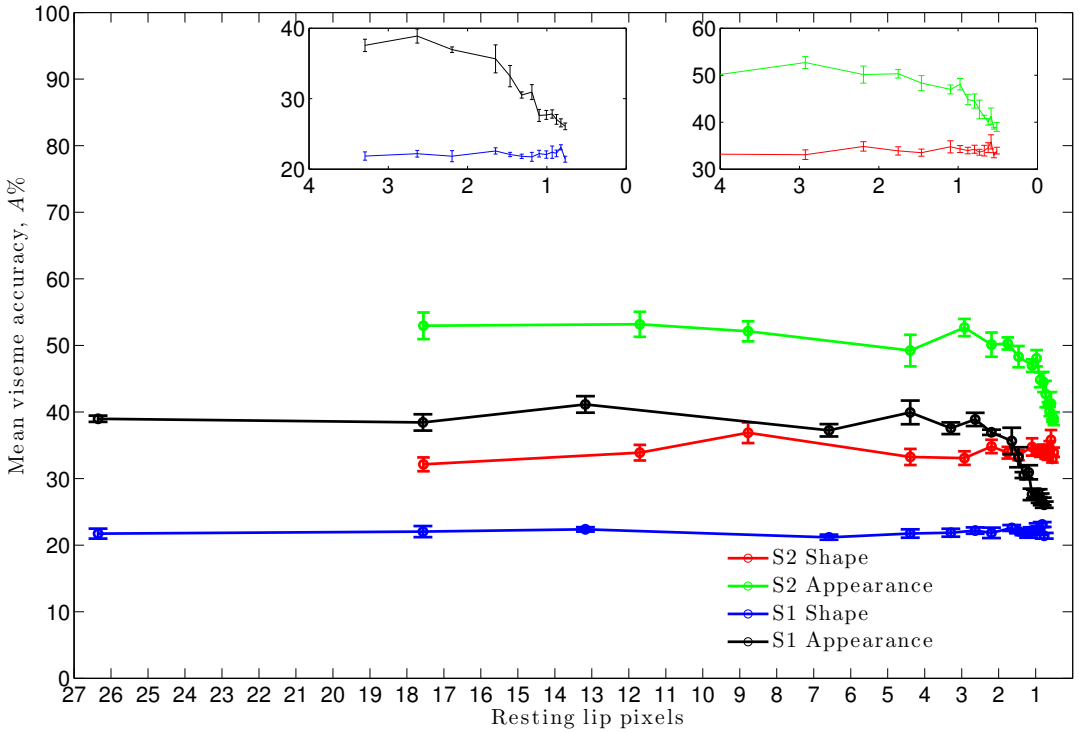



Figure 5.13: Viseme classification in Accuracy, $A \pm 1\frac{\sigma}{\sqrt{5}}$, with a bigram word network (on the $y$-axis) by vertical resting lip height in pixels ($x$-axis).

as this is the first time this dataset has been used this is the baseline for future uses.

Contrary to common assumption and practice, the unexpected observation here is the remarkable resilience to resolution in machine lip-reading. Given modern experiments in lip-reading usually take place with high-resolution video ([24] for example) the disparity between measured performance (shown here) and assumed performance is very remarkable. Our results show for successful lip-reading one needs a minimum of four pixels (two pixels per lip) across the closed lips.

The realisation of a minimum number of pixels per lip is a new piece of information in the area of machine lip-reading. Previous research in this area [143, 59] has focused on noisy images and the effect of noise on word error rates in audio-visual speech recognition system. In these experiments, we see corroborating results to support the premise that with less information then lip-reading is negatively affected, but also that there is an lower bound resolution which is essential for good

lip-reading.

It must not be forgotten a higher resolution video is beneficial for the tracking task but, as previous work demonstrates, other factors considered to negatively effect lip-reading classification such as off-axis views [78], actually have the ability to improve performance and here we see that a lower resolution video is not as detrimental as first assumed.

We therefore conclude that, for real situations, the limitations on lip-reading are not likely to come from factors to do with the environment. Rather, the poor performance of lip-reading is almost certainly to do with limitations in the signal - the lip-signal is very challenging to decode and what is needed is a better understanding of the visual signal, its components, and how they can be learnt. For this reason, we now turn to the problem of understanding visemes.

# Chapter 6

# A performance evaluation of visemes

This chapter is our first investigation into understanding visemes. Before we undertake complicated experiments and attempt to re-design or augment visemes, it is useful to understand what we can with what we have already tested. Currently we always use a whole set of visemes to include a large number of phonemes. But it would be nice to know:

- if all visemes contribute equally to the classification? If no, which of the visemes within the set are most useful?
- Are there any visemes which are not helpful, or in fact, detrimental? And,
- can we evaluate the performance of each viseme in isolation to understand more about the set of classes as a whole?

Therefore, this chapter describes an investigation into the difference in the contribution to accuracy of each viseme within a set. An analysis of the confusion matrices produced during viseme classification, obtained by comparing the classification output with the ground truth transcript, both of which are time-aligned,

provides us with measurements of viseme contributions to classification. This enables us to compare each viseme within a set to all others and determine which contributes the most for accurate machine lip-reading.

Additionally the balance between shape and appearance viseme probabilities are reviewed to see which type of feature (shape or appearance) contributes most to classification. We can also compare visual classification to audio using the same viseme classifier labels on audio features (we use MFCCs). This demonstrates a relationship between viseme classification accuracy and the spread of individual viseme contribution to classification.

## 6.1 Measuring the contribution of individual visemes

The point of interest in this chapter is in the contribution of each viseme to the classification performance. This work searches for any particular viseme (or subgroup of phonemes) which contributes more to the classification accuracy.

This study continues with the Rosetta Raven features extracted in Section 5.1. Short datasets, such as these, may not provide adequate training examples of all visemes. So we group the untrainable visemes into a single garbage viseme. In this case we estimate 150 samples as the minimum threshold (the mean training samples per viseme minus 1.5 standard error) to mitigate the bias caused by variation in training samples per classifier. Thus, visemes /$v$08/, /$v$09/, /$v$14/ and /$v$15/ are grouped giving Table 6.1. We have already reviewed the original dataset in Chapter 3, and Figure 5.4 shows the occurrence of visemes listed in the original phoneme-to-viseme map (see Table 5.1).

The classification method used is identical to the method in Chapter 5, the methodology varies in the analysis of the classification outputs.

Values from the `HResults` confusion matrices are extracted for analysis. For each viseme we have calculated the probability of its classification $\Pr\{v|\hat{v}\}$.

Table 6.1: Modified phoneme-to-viseme mapping due to lack of training data per viseme available in the Rosetta Raven dataset.

| vID | Phonemes | vID | Phonemes |
|---|---|---|---|
| v01 | /p/ /b/ /m/ | v11 | /eh/ /ae/ /ey/ /ay/ |
| v02 | /f/ /v/ | v12 | /ɑ/ /ɔ/ /ʌ/ |
| v03 | /θ/ /ð/ | v13 | /ʊ/ /ɜ/ /ax/ |
| v04 | /t/ /d/ /n/ /k/ /g/ /h/ /j/ /ŋ/ /y/ | v16 | /iy/ /hh/ |
| | | v17 | /ɑʊ/ /əʊ/ |
| v05 | /s/ /z/ | v18 | silence |
| v06 | /l/ | gar | /u/ /uw/ /ɔɪ/ /w/ /ʃ/ /ʒ/ /tʃ/ /dʒ/ |
| v07 | /r/ | | |
| v10 | /i/ /ɪ/ | | |

## 6.2 Analysis of viseme contribution

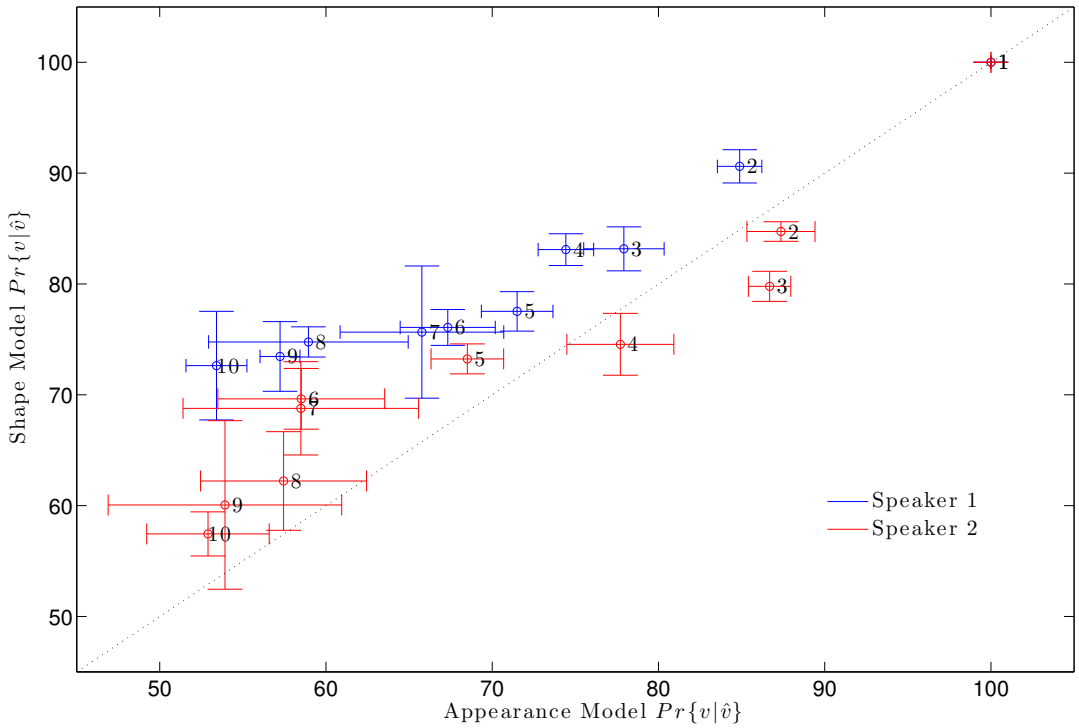



Figure 6.1: Relationship between shape and appearance model features for Speaker 1 and Speaker 2.

Figure 6.1 shows the mean $\Pr\{p|\hat{p}\}$ for the top 10 visemes over all five folds $\pm 1\frac{\sigma}{\sqrt{5}}$. The $x$-axis is the probability of correct classification when the viseme is trained on an appearance only model, the $y$-axis is the probability of correct classification when

the viseme is trained on a shape only model. Red are the results for Speaker 1, and the blue are Speaker 2. As the visemes are plotted by their rank, they do not always match for each speaker. For example, the second position for Speaker 1 is /$v$12/ whereas for Speaker 2 is /$v$04/. All ranked visemes are listed in Table 6.2. The fifth most useful viseme gives superior classification for both speakers. The conventional wisdom is appearance features give the best results but only in studio-type conditions with good tracking, whereas here shape features are more robust than appearance.

Table 6.2: Ranked visemes for separate shape and appearance features for each Rosetta Raven speaker.

| | Shape | | Appearance | |
|---|---|---|---|---|
| Rank | Speaker 1 | Speaker 2 | Speaker 1 | Speaker 2 |
| 1 | /$v$18/ | /$v$18/ | /$v$18/ | /$v$18/ |
| 2 | /$v$12/ | /$v$04/ | /$v$04/ | /$v$04/ |
| 3 | /$v$04/ | /$v$12/ | /$v$12/ | /$v$12/ |
| 4 | /$v$11/ | /$v$11/ | /$v$01/ | /$v$01/ |
| 5 | /$v$07/ | /$v$01/ | /$v$11/ | /$v$02/ |
| 6 | /$v$01/ | /$v$05/ | /$v$07/ | /$v$11/ |
| 7 | /$v$06/ | /$v$07/ | /$v$02/ | /*gar*/ |
| 8 | /$v$05/ | /*gar*/ | /$v$05/ | /$v$05/ |
| 9 | /$v$02/ | /$v$02/ | /*gar*/ | /$v$10/ |
| 10 | /*gar*/ | /$v$10/ | /$v$10/ | /$v$06/ |

Note the top right-hand point is the visual silence viseme, /$v$18/, for both Speaker 1 and Speaker 2. In general, visual silence can be quite variable compared to audio silence because speakers breathe and show emotion. However, because the source text is a poem, which has structure and natural pauses within its style, there are well-defined visual silence periods at the start of each line.

Figures 6.2 and 6.3 show, for the Speaker 1 and Speaker 2 shape and appearance models, the probability of correctly recognising the top ten visemes, $\Pr\{v|\hat{v}\}$. They also show the audio (MFCC) performance measured on visemes. The $x$-axis varies by performance, the best performing viseme is on the left hand side which for visual shape and appearance features is silence for all features. The next best viseme varies but is either /$v$4/, /$v$5/ or /$v$12/. /$v$4/ is a phonetically indistinct viseme (it is the

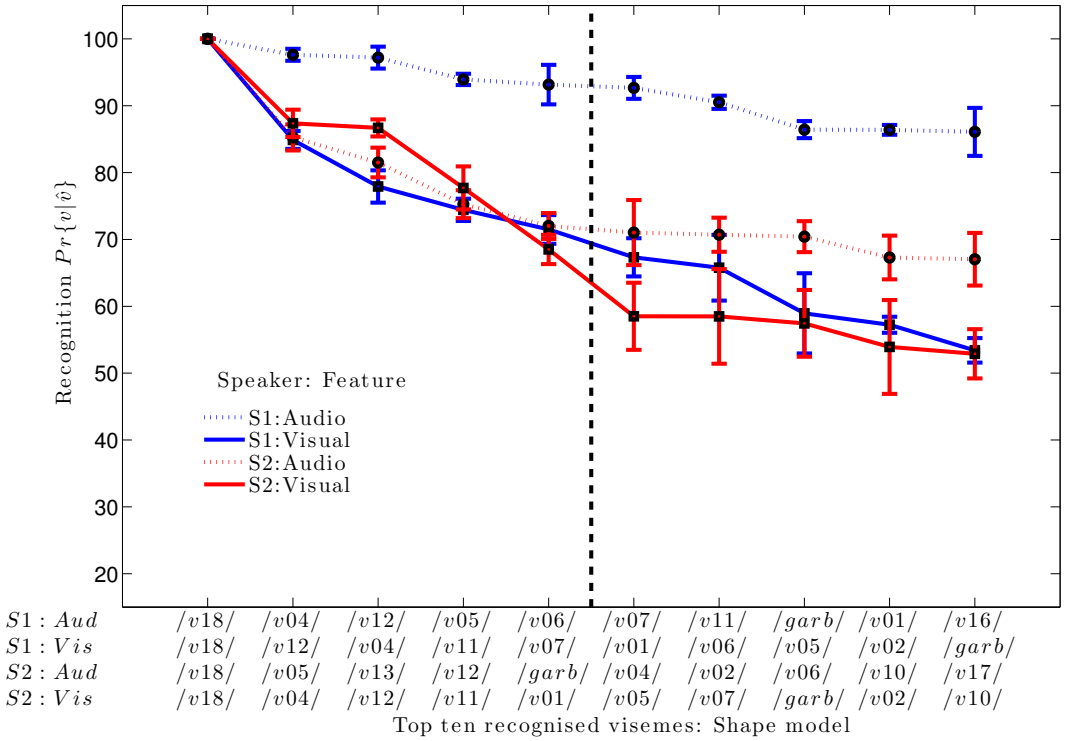



Figure 6.2: Classification probability $\Pr\{p|\hat{p}\}$ with a shape model for the top ten visemes in descending order. A threshold is plotted in a black vertical line to show the point at which the usefulness of each viseme significantly decreases (after five visemes) in the visual channel.

biggest cluster of phonemes) so appears as a "filler" viseme.

It has been observed in human lip-reading that there are few reliable visual cues and humans use these combined with rich contextual information to interpret or 'fill in the gaps' of what a speaker is saying [44, 132]. Therefore, the hypothesis is that robust audio classification is based upon a large spread of recognised phonemes and the resilience in classification is due to the number of phonemes contributing to the accuracy. Visually, as with human lip-readers, it is anticipated fewer visemes would perform the equivalent classification and, as such, the graph would demonstrate a steeper decline in $\Pr\{v|\hat{v}\}$ over the top performing visemes (from left to right along the $x$-axis).

In Figure 6.2 there is a greater decline from left to right over the top ten visemes for visual features than for audio for both speakers. Additionally, the error bars after

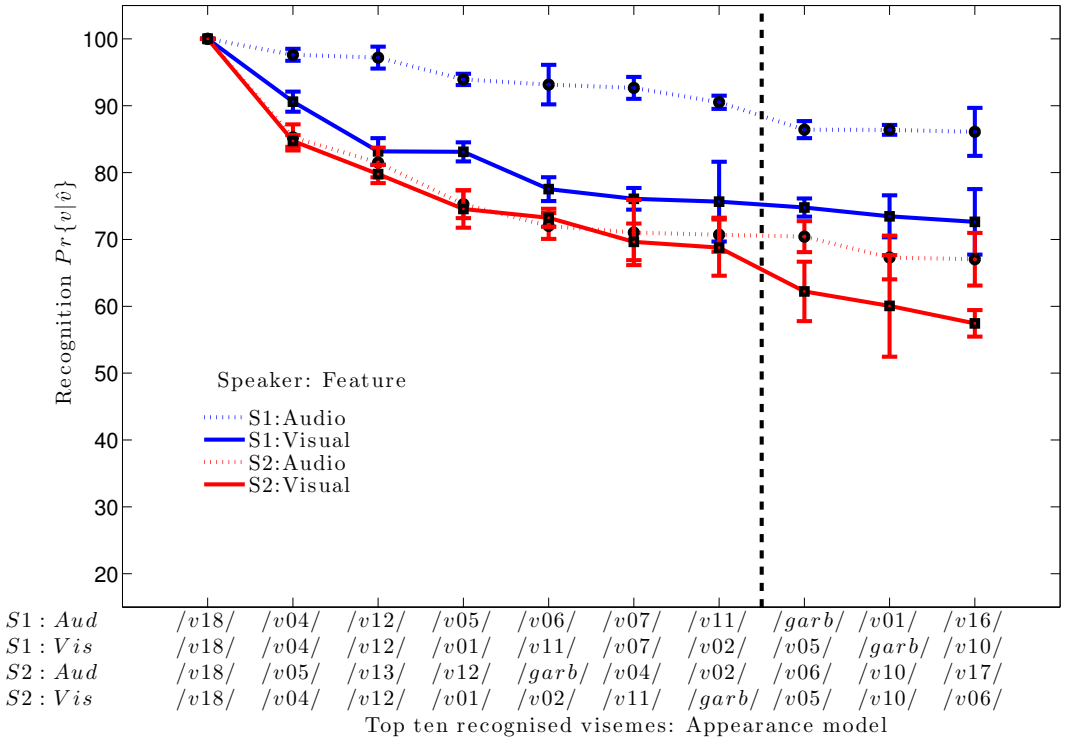



Figure 6.3: Classification probability $\Pr\{p|\hat{p}\}$ with an appearance model for the top ten visemes in descending order. A threshold is plotted in a black vertical line to show the point at which the usefulness of each viseme significantly decreases (after seven visemes) in the visual channel.

the $5^{th}$ position viseme increase for Speaker 2 (marginally so for Speaker 1), which provides evidence to support the hypothesis of audio classification is spread over more visemes to be correct. The top visemes (after silence /$v$18/) are /$v$04/, /$v$12/, /$v$11/ and /$v$01/. These are vowels (/$v$12/, /$v$11/) and front-of-mouth consonant visemes (/$v$04/, /$v$01/).

Figure 6.3, with appearance features, demonstrates a shallower decline from left to right than the shape graph in Figure 6.2 but still there is a greater decline for visual features than for audio. The error bars here increase after the $7^{th}$ position viseme. Note the order of the audio viseme ordering is identical in both Figures 6.2 and 6.3 as this is the same experiment.

The shape of the graph in Figure 6.3 is similar between audio and video which implies appearance-based classification is similar to noisy acoustic classification for

Table 6.3: Ranked mean viseme $\Pr\{p|\hat{p}\}$ for shape, appearance, Speaker 1, Speaker 2 and over all variables.

| Shape | Appearance | Speaker 1 | Speaker 2 | Overall |
|---|---|---|---|---|
| /v18/ | /v18/ | /v18/ | /v18/ | /v18/ |
| {/v04/ /v12} | /v04/ | {/v04/ /v12/} | {/v04/ /v12/} | /v04/ |
| /v11/ | /v12/ | /v11/ | /v11/ | /v12/ |
| /v01/ | /v11/ | /v01/ | /v01/ | /v01/ |
| /v07/ | /v01/ | /v07/ | /v07/ | /v11/ |
| /v05/ | /v07/ | {/v02/ /v05/} | {/v02/ /v05/} | /v07/ |
| {/v02/ /v06/ /gar/} | {/v02/ /v05/} | /v06/ | {/v06/ /gar/} | {/v02/ /v05/} |
| /v10/ | /v06/ | {v10/ /gar/} | /v10/ | /v19/ |
| {/v03/ /v13/} | /gar/ | /v03/ | /v03/ | /v06/ |
| /v16/ | /v10/ | /v13/ | /v13/ | /v10/ |
| /v17/ | /v03/ | /v16/ | /v16/ | /v13/ |
| | {/v13/ /v16/} | /v17/ | /v17/ | /v03/ |
| | /v17/ | | | /v16/ |
| | | | | /v17/ |

both speakers and hence is less fragile. The top visemes in Figure 6.3 (not including silence /$v$18/) are: /$v$04/, /$v$12/, /$v$11/, /$v$01/, and /$v$7/ i.e. identical for shape-only in the first six positions.

Where the error bars increase, this may be due to the few data available, which makes classification more unreliable due to less well trained HMM classifiers. This means our estimated threshold for minimum training samples per classifier was not high enough. The impact of this is reduced with the /$gar$/ viseme, but note with Figure 5.4 there are similarities between our top performing visemes and those with the most training samples.

Table 6.3 lists the mean ranking of visemes of both speakers shape models for all visemes in the tested mapping and both speaker's appearance models. Table 6.3 also gives mean viseme ranks for each speaker and over all speakers and models. The rankings are similar between all pairings.

Tables 6.4, 6.5 and 6.6 summarise the similarities between feature types and speakers by using Spearman rank correlation, $r$, [153] between the ranked viseme

Table 6.4: Comparing Speaker 1 and Speaker 2 viseme ordering with Spearman correlation.

| Speaker 1 | Speaker 2 | r | p |
|---|---|---|---|
| Audio | Audio | 0.43 | $1.63 \times 10^{-2}$ |
| Shape | Shape | <u>0.92</u> | 0.00 |
| Appearance | Appearance | <u>0.93</u> | 0.00 |

Table 6.5: Speaker 1 Spearman correlations of viseme performance ordering with different features: acoustic, shape, and appearance.

| Speaker 1 | Speaker 1 | r | p |
|---|---|---|---|
| Shape | Appearance | <u>0.90</u> | 0.00 |
| Audio | Shape | <u>0.85</u> | $2.39 \times 10^{-5}$ |
| Audio | Appearance | <u>0.74</u> | $9.2 \times 10^{-3}$ |

Table 6.6: Speaker 2 Spearman correlations of viseme performance ordering with different features: acoustic, shape, and appearance.

| Speaker 2 | Speaker 2 | r | p |
|---|---|---|---|
| Shape | Appearance | <u>0.92</u> | 0.00 |
| Audio | Shape | 0.42 | 0.12 |
| Audio | Appearance | 0.48 | 0.07 |

outputs. Those which are significant at the 5% threshold are underlined. This confirms a strong relation between shape-only and appearance-only classification. In lab conditions, appearance features outperform shape [14] but in real world conditions the shape information is more robust in the absence of non-noisy appearance data [79]. This strong coupling, and previous work, [77], shows the two modes of information are complimentary and we recommend the use of both, without forgetting that in the real world, artefacts such as motion blur significantly deteriorate appearance information. We also note for Speaker 1 (in Figure 6.4) the audio ranking is similar to the video ranking although as we have previously noticed there is a more rapid drop-off for video.

## 6.3 Viseme contribution observations

To summarise this chapter, we have shown that, with the assumption classifiers are trained with sufficient data, the order of single viseme performances are fairly consistent across the feature modes of audio, shape and appearance. It is also noted the visual classifiers depend more highly on a select few visemes performing well (between five for shape modes and seven for appearance mode out of a possible 15) than the audio classifiers.

The observation of how fragile machine lip-reading is, is re-enforced by this work. If these critical five or seven visemes cannot be built as sufficiently trained classifiers then lip-reading is impossible. When a human is trained in how to lip-read, many follow the method of recognising a small number of key gestures which we then process using our own sophisticated knowledge of language and context to create a classification output or transcript [63].

In audio it is surprisingly rare to see this effect measured, even though a good acoustic unit will have accuracies which are at least 10% higher than an average unit (the mean audio viseme performance on Speaker 2 is 76% for the all visemes).

We acknowledge most work in this field focuses on improving mean accuracies over the set of all visemes which can conceal the real source of overall performance. A system which achieves a mean viseme accuracy of, say, 53%, may be one which contains a few supremely accurate viseme classifiers or it maybe a system with a set of a large number of classifiers which all achieve a more modest performance. In our work we have seen a correlation between the spread of viseme contributions to classification and viseme classification performance, so we can now say higher classification is achieved with a set of equally useful visemes rather than a set of visemes where their usefulness ranges from poor to excellent.

This chapter, therefore, suggests two different strategies for improving future lip-reading systems; option one: one makes the select few best viseme classifiers better or, option two: one focuses upon improving the worst, which at this stage do not

contribute at all. We can not comment at this time which approach is likely to be more successful but our observations will allow future work to focus attention where it is likely to do the most good.

This work suggests five of the visemes are largely responsible for accurate classification, whereas for appearance there are seven visemes and for audio there are at least ten. This means there appears to be fewer recognizable shapes than there are distinguishable appearances, and in turn, sounds. This relates to the overall viseme classification of the set where audio results are better than appearance, which in turn are better than shape.

We suggest that a good threshold of viseme training samples, is not more or less than 1 standard error away from the mean number of training samples for all visemes in a set. This is stricter than the threshold we used and will ensure there is no bias towards any one particular viseme class which could then dominate the classification accuracy of the set.

Now we have a deeper understanding of visemes and their individual capabilities, we move onto investigating how they relate to phonemes, the acoustic units of speech. We are reminded of our viseme working definition, "a viseme is the visual equivalent of a phoneme" so we move on to a review of a number of the phoneme-to-viseme (P2V) mappings which have been presented in literature in order to assess which is optimal for machine lip-reading.

# Chapter 7

# Bear speaker-dependent visemes

In computer lip-reading literature there is debate over the mapping of phonemes to visemes. In this chapter the AVLetters2 dataset (Section 3.2) is used to train and test classifiers using 120 phoneme-to-viseme (P2V) mappings and the effect on word classification accuracy is measured. This chapter also presents and tests a new data-driven method for devising speaker-dependent phoneme to viseme maps using phoneme confusions. Our method is not influenced by perception bias since our confusions are based on machine observations, and not human perception. We compare word classification achieved with these new maps against the best performing previously published phoneme-to-viseme mapping. We demonstrate that whilst there are differences between each viseme map previously suggested, the best mapping over all speakers is from Lee [82]. This mapping is used as a benchmark to compare the performance of new data-derived speaker-dependent visemes.

A summary of published P2V maps is provided in [138] Tables 2.3 and 2.4. This list is not exhaustive and these mappings vary by: a focus on just consonants [21, 48, 50, 144], are speaker-dependent [73], or have an ordering [114]. These are useful starting points, but for the purpose of this study we would like the phoneme-to-viseme mappings to include all phonemes in the transcript of the dataset to accurately reflect the range of phonemes used in a full vocabulary. Therefore, some mappings used here are a pairing of two mappings suggested in literature, e.g. one

map for the vowels and one map for the consonants. A full list of the mappings used is in Tables 7.4 and 7.5. In total, 15 consonant maps and eight vowel maps are identified here and all of these are paired with each other to provide 120 P2V maps to test. The questions we ask are; does conventional a machine lip-reading system use the correct viseme mappings for machine lip-reading? And, is it possible to find a method for selecting better phoneme-to-viseme mappings?

## 7.1 Current viseme studies

There are many viseme classifications present in literature, the most common viseme classifications are: 'the Disney 12' [80], the 'lip-reading 18' by Nichie [110], and Fisher's [48]. Full phoneme to viseme mappings of these classes can be found in Tables 7.1, 7.2 and 7.3. The differences in these classifications are based around different groupings of phonemes, and in the literature we know of a number of recent attempts to compare these, such as [28] and as part of [138]. In [138] the following list of reasons are given for discrepancies between classifier sets.

- Variation between speakers - i.e. speaker identity.
- Variation between viewers - indicating lip-reading ability varies by individuals, those with more practise are better able to identify visemes.
- The context of the speech presented - context has an influence on how consonants appear on the lips. In real tasks the context will enable easier distinction between indistinguishable phonemes in syllable only tests.
- Clustering criteria - the grouping methods vary between authors. For example, 'phonemes are said to belong to a viseme if, when clustered, the percent correct identification for the viseme is above some threshold, which is typically between 70 - 75% correct. A stricter grouping criterion has a higher threshold, so more visemes are identified.'.

Table 7.1: The "Disney twelve" phoneme-to-viseme map.

| Viseme | Phonemes |
|---|---|
| /v01/ | /p/ /b/ /m/ |
| /v02/ | /w/ |
| /v03/ | /f/ /v/ |
| /v04/ | /θ/ |
| /v05/ | /l/ |
| /v06/ | /d/ /t/ /z/ /s/ /r/ /n/ |
| /v07/ | /s/ /ʃ/ /tʃ/ /j/ |
| /v08/ | /y/ /g/ /k/ /ŋ/ |
| /v09/ | /ʊ/ /ɦ/ |
| /v10/ | /ɛə/ /ɪ/ /ai/ /e/ /ʌ/ |
| /v11/ | /u/ |
| /v12/ | /ʊə/ /ɔ/ /ɔə/ |

Table 7.2: Fisher's phoneme-to-viseme map.

| Viseme | Phonemes |
|---|---|
| /v01/ | /k/ /g/ /ŋ/ /m/ |
| /v02/ | /p/ /b/ |
| /v03/ | /f/ /v/ |
| /v04/ | /ʃ/ /ʒ/ /tʃ/ /dʒ/ |
| /v05/ | /t/ /d/ /n/ /th/ /dh/ /z/ /s/ /r/ /l/ |

## 7.2 Data preparation

The AVLetters2 (AVL2) dataset [35] is used to train and test HMM classifiers based upon our 120 P2V mappings. AAM features are used as they are known to outperform other feature methods in machine lip-reading [28]. Tables 7.4 and 7.5 show all phonemes in each original P2V map. As each utterance is very short in our data set (each is a one word sentence of a single letter) there is no need to implement $\Delta$s within our features to address co-articulation.

In Table 7.6 we have described the sources and derivation methods for all of the phoneme-to-viseme maps used in our comparison study. We see the majority are constructed using human perception testing with few test subjects, e.g. Finn [47] only used 1 and Kricos [73] 12. Data-driven methods are most recent, e.g. Lee's [82] visemes were presented in 2002 and Hazen's [58] in 2006. The remaining visemes

Table 7.3: Nichie's "Lip-reading 18" phoneme-to-viseme map.

| Viseme | Phonemes |
|---|---|
| /v01/ | /p/ /b/ /m/ |
| /v02/ | /f/ /v/ |
| /v03/ | /ɯ/ /w/ |
| /v04/ | /r/ |
| /v05/ | /s/ /z/ |
| /v06/ | /ʃ/ /ʒ/ /tʃ/ /dʒ/ |
| /v07/ | /ð/ |
| /v08/ | /l/ |
| /v09/ | /t/ /d/ /n/ |
| /v10/ | /y/ |
| /v11/ | /k/ /g/ /ŋ/ |
| /v12/ | /ɦ/ |
| /v13/ | /uw/ |
| /v14/ | /ʊ/ /əʊ/ |
| /v15/ | /ɑʊ/ |
| /v16/ | /i/ /ay/ /ɪ/ |
| /v17/ | /u/ |
| /v18/ | /ʌ/ |
| /v19/ | /iy/ /ɛ/ |
| /v20/ | /e/ /ɪə/ |
| /v21/ | /ə/ /ei/ |

Table 7.4: Vowel phoneme-to-viseme maps previously presented in literature.

| Classification | Viseme phoneme sets |
|---|---|
| Bozkurt [25] | {/ei/ /ʌ/} {/ei/ /e/ /æ/} {/ɜ/} {/i/ /ɪ/ /ə/ /y/} {/ɑʊ/} {/ɔ/ /ɑ/ /ɔɪ/ /əʊ/} {/u/ /ʊ/ /w/} |
| Disney [80] | {/ʊ/ /h/} {/ɛə/ /i/ /ai/ /e/ /ʌ/} {/u/} {/ʊə/ /ɔ/ /ɔə/} |
| Hazen [58] | {/ɑʊ/ /ʊ/ /u/ /əʊ/ /ɔ/ /w/ /ɔɪ/} {/ʌ/ /ɑ/} {/æ/ /e/ /ai/ /ei/} {/ə/ /ɪ/ /i/} |
| Jeffers [69] | {/ɑ/ /æ/ /ʌ/ /ai/ /e/ /ei/ /ɪ/ /i/ /ɔ/ /ə/ /ɪ/} {/ɔɪ/ /ɔ/} {/ɑʊ/} {/ɜ/ /əʊ/ /ʊ/ /u/} |
| Lee [82] | {/i/ /ɪ/} {/e/ /ei/ /æ/} {/ɑ/ /ɑʊ/ /ai/ /ʌ/} {/ɔ/ /ɔɪ/ /əʊ/} {/ʊ/ /u/} |
| Montgomery [105] | {/i/ /ɪ/} {/e/ /æ/ /ei/ /ai/} {/ɑ/ /ɔ/ /ʌ/} {/ʊ/ /ɜ/ /ə/}{/ɔɪ/} {/i/ /hh/} {/ɑʊ/ /əʊ/} {/u/ /u/} |
| Neti [108] | {/ɔ/ /ʌ/ /ɑ/ /ɜ/ /ɔɪ/ /ɑʊ/ /ɦ/} {/u/ /ʊ/ /əʊ/} {/æ/ /e/ /ei/ /ai/} {/ɪ/ /i/ /ə/} |
| Nichie [110] | {/uw/} {/ʊ/ /əʊ/} {/ɑʊ/} {/i/ /ʌ/ /ay/} {/ʌ/} {/iy/ /æ/} {/e/ /ɪə/} {/u/} {/ə/ /ei/} |

Table 7.5: Consonant phoneme-to-viseme maps previously presented in literature.

| Classification | Viseme phoneme sets |
|---|---|
| Binnie [21] | {/p/ /b/ /m/} {/f/ /v/} {/θ/ /ð/} {/ʃ/ /ʒ/} {/k/ /g/} {/w/} {/r/} {/l/ /n/} {/t/ /d/ /s/ /z/} |
| Bozkurt [25] | {/g/ /ɦ/ /k/ /ŋ/} {/l/ /d/ /n/ /t/} {/s/ /z/} {/tʃ/ /ʃ/ /dʒ/ /ʒ/} {/θ/ /ð/} {/r/} {/f/ /v/} {/p/ /b/ /m/} |
| Disney [80] | {/p/ /b/ /m/} {/w/} {/f/ /v/} {/θ/} {/l/} {/d/ /t/ /z/ /s/ /r/ /n/} {/ʃ/ /tʃ/ /j/} {/y/ /g/ /k/ /ŋ/} |
| Finn [47] | {/p/ /b/ /m/} {/θ/ /ð/} {/w/ /s/} {/k/ /h/ /g/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/y/} {/z/} {/f/} {/v/} {/t/ /d/ /n/ /l/ /r/} |
| Fisher [48] | {/k/ /g/ /ŋ/ /m/} {/p/ /b/} {/f/ /v/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/t/ /d/ /n/ /θ/ /ð/ /z/ /s/ /r/ /l/} |
| Franks [50] | {/p/ /b/ /m/} {/f/} {/r/ /w/} {/ʃ/ /dʒ/ /tʃ/} |
| Hazen [58] | {/l/} {/r/} {/y/} {/b/ /p/} {m} {/s/ /z/ /h/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/t/ /d/ /θ/ /ð/ /g/ /k/} {/ŋ/} {/f/ /v/} |
| Heider [61] | {/p/ /b/ /m/} {/f/ /v/} {/k/ /g/} {/ʃ/ /tʃ/ /dʒ/} {/θ/} {/n/ /t/ /d/} {/l/} {/r/} |
| Jeffers [69] | {/f/ /v/} {/r/ /q/ /w/} {/p/ /b/ /m/} {/θ/ /ð/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/s/ /z/} {/d/ /l/ /n/ /t/} {/g/ /k/ /ŋ/} |
| Kricos [73] | {/p/ /b/ /m/} {/f/ /v/} {/w/ /r/} {/t/ /d/ /s/ /z/} {/k/ /n/ /j/ /h/ /ŋ/ /g/} {/l/} {/θ/ /ð/} {/ʃ/ /ʒ/ /tʃ/ /dʒ/} |
| Lee [82] | {/d/ /t/ /s/ /z/ /θ/ /ð/} {/g/ /k/ /n/ /ŋ/ /l/ /y/ /ɦ/} {/dʒ/ /tʃ/ /ʃ/ /ʒ/} {/p/ /b/ /m/} {/f/ /v/} {/r/ /w/} |
| Neti [108] | {/l/ /r/ /y/} {/s/ /z/} {/t/ /d/ /n/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/p/ /b/ /m/} {/θ/ /ð/} {/f/ /v/} {/ŋ/ /k/ /g/ /w/} |
| Nichie [110] | {/p/ /b/ /m/} {/f/ /v/} {/ɯ/ /w/} {/r/} {/s/ /z/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/θ/} {/l/} {/k/ /g/ /ŋ/} {/ɦ/} {/t/ /d/ /n/} {/y/} |
| Walden [144] | {/p/ /b/ /m/} {/f/ /v/} {/θ /ð/} {/ʃ/ /ʒ/} {/w/} {/s/ /z/} {/r/} {/l/} {/t/ /d/ /n/ /k/ /g/ /j/} |
| Woodward [148] | {/p/ /b/ /m/} {/f/ /v/} {/w /r/ /ɯ/} {/t/ /d/ /n/ /l/ /θ/ /ð/ /s/ /z/ /tʃ/ /dʒ/ /ʃ/ /ʒ/ /j/ /k/ /g/ /h/} |

are based around linguistic/phonemic rules.

As an example, the clustering method of Hazen [58] involved bottom-up clustering using maximum Bhattacharyya distances to measure similarity between the phoneme-labelled models. The models were represented by Gaussian distributions. Before clustering, some phonemes were manually merged, /*em*/ with /*m*/, /*en*/ with /*n*/, and /*Z*/ with /*S*/.

Table 7.6: A comparison of literature phoneme-to-viseme maps.

| Author | Year | Inspiration | Description | Test subjects |
|---|---|---|---|---|
| Binnie | 1976 | Human testing | Confusion patterns | unknown |
| Bozkurt | 2007 | Subjective linguistics | Common tri-phones | 462 |
| Disney | — | Speech synthesis | Observations | unknown |
| Finn | 1988 | Human perception | Montgomerys visemes and /ﬁ/ | 1 |
| Fisher | 1986 | Human testing | Multiple-choice intelligibility test | 18 |
| Franks | 1972 | Human perception | Confusions among sounds produced in similar articulatory positions | unknown 275 |
| Hazen | 2006 | Data-driven | Bottom-up clustering | 223 |
| Heider | 1940 | Human perception | Confusions post-training | unknown |
| Jeffers | 1971 | Linguistics | Sensory and cognitive correlates | unknown |
| Kricos | 1982 | Human testing | Hierarchical clustering | 12 |
| Lee | 2002 | Data-driven | Merging of Fisher visemes | unknown |
| Neti | 2000 | Linguistics | Decision tree clusters | 26 |
| Nichie | 1912 | Human observations | Human observation of lip movements | unknown |
| Walden | 1977 | Human testing | Hierachical clustering | 31 |
| Woodward | 1960 | Linguistics | Language rules and context | unknown |

Figure 3.5 (Chapter 3) shows the occurrence frequency of the 29 phonemes in AVL2 which details the volume of training samples available. Note, AVL2 does not include all phonemes in the British English phonetic alphabet [5]. It is a known problem in visual speech research that one limitation is the lack of sufficiently large datasets available [28]. This motivates the drive to find better P2V mappings to potentially avoid the need, and associated cost, in obtaining large audio-visual speech

datasets.

A P2V map introduces confusion in machine lip-reading. In an attempt to measure the level of this confusion, a simple ratio metric of the proportion of phonemes to visemes is shown in (Equation 7.1), where $CF_s$ is the compression factor for a set of visemes, $s$, $\#V$ is the number of visemes, and $\#P$ is the number of phonemes. The compression factors for the P2V maps are described in Table 7.7. The ideal ratio is a 1:1 phoneme to viseme mapping as this would mean we are identifying each phoneme uniquely. However, we still need to cluster the phonemes due to the lack of visual distinction between some phonemes. Thus the higher a Compression Factor (CF) (closer to one) the better it is as this means there is less dependency upon the language network for decoding of visemes back to phonemes. Silence and garbage visemes are not included in CFs.

$$CF_s = \frac{\#V}{\#P} \tag{7.1}$$

Table 7.7: Compression factors for viseme maps previously presented in literature.

| Consonant Map | V:P | CF | Vowel Map | V:P | CF |
|---|---|---|---|---|---|
| Woodward | 4:24 | 0.16 | Jeffers | 3:19 | 0.16 |
| Disney | 6:22 | 0.18 | Neti | 4:20 | 0.20 |
| Fisher | 5:21 | 0.23 | Hazen | 4:18 | 0.22 |
| Lee | 6:24 | 0.25 | Disney | 4:11 | 0.36 |
| Franks | 5:17 | 0.29 | Lee | 5:14 | 0.36 |
| Kricos | 8:24 | 0.33 | Bozkurt | 7:19 | 0.37 |
| Jeffers | 8:23 | 0.35 | Montgomery | 8:19 | 0.42 |
| Neti | 8:23 | 0.35 | Nichie | 9:15 | 0.60 |
| Bozkurt | 8:22 | 0.36 | - | - | - |
| Finn | 10:23 | 0.43 | - | - | - |
| Walden | 9:20 | 0.45 | - | - | - |
| Binnie | 9:19 | 0.47 | - | - | - |
| Hazen | 10:21 | 0.48 | - | - | - |
| Heider | 8:16 | 0.50 | - | - | - |
| Nichie | 18:33 | 0.54 | - | - | - |

Deliberate omission of the following phonemes from some mappings is required: */si/* (Disney [80]), */axr/ /en/ /el/ /em/* (Bozkirt [25]), */axr/ /em/ /epi/ /tcl/*

*/dcl/ /en/ /gcl/ kcl/*(Hazen [58]), and */axr/ /em/ /el/ /nx/ /en/ /dx/ /eng/ /ux/* (Jeffers [69]), because these are American diacritics which are not appropriate to a British English phonetic dataset. Moreover, Kricos provides speaker-dependent visemes [73]. These have been generalised for our tests using the most common mixtures of phonemes as the method is not reproducible. Where a viseme map does not include phonemes present in the ground truth transcript these are grouped into a garbage viseme (*/gar/*) to measure only the performance of the viseme sets previously prescribed in literature. Note that all phonemes in the each P2V map are in the dataset but no mapping includes all 29 phonemes in the AVL2 vocabulary.

## 7.3 Classification method

The method for these speaker-dependent classification tests on our combined shape and appearance features uses HMM classifiers built with HTK [150]. The features selected are from the AVL2 dataset described in Chapter 3. The videos are tracked with a full-face AAM and the features extracted consist of only the lip information. The classifiers are based upon viseme labels within each P2V map. A ground truth for measuring correct classification is a viseme transcription produced using the BEEP British English pronunciation dictionary [26] and a word transcription. The classification output is a viseme level script mapped to sentence (word) level classification. Working in British English the phonetic transcript is converted to a viseme transcript assuming the visemes in the mapping being tested (Tables 7.4 and 7.5). We test using a leave-one-out seven-fold cross validation. Seven folds are selected as we have seven utterances of the alphabet per speaker in AVL2. The HMMs are initialised using 'flat start' training and re-estimated eight times and then force-aligned using HTK's `HVite`. Training is completed by re-estimating the HMMs three more times with the force-aligned transcript.

## 7.4 Comparison of current phoneme to viseme maps

In this section, classification performance of the HMMs is measured by correctness, $C$ (Equation 2.7), as there are no insertion errors to consider [150]. It is acknowledged word classification is not as high performing as viseme classification. However, as each viseme set being tested has a different number of phonemes and visemes, a common comparator, here words, are used as they can compare different viseme sets. It is the difference between each set, rather than the individual performance, which is of interest in this investigation. Word level correctness rather than viseme level correctness normalises over all sets for a fair comparison. (Each viseme set has a different number of visemes in it and in turn a varying level of training samples per viseme).

We compare our values of accuracy to those in the literature, namely [35] & [118]. In [35] we see that speaker-dependent results with AVL2 are significantly higher than the values we have achieved. However, in this paper the experiments are designed to measure the efficacy of multi-speaker classifiers and thus the authors have permitted different HMM parameters between speakers. For example, the number of HMM states ranges between five and nine. In our work these values are constant to ensure any effects observed are the result of the viseme selection only.

In [118] AVLetters2 data achieves 91.8% with an unsupervised random forest classification technique. This out performs both [35] and our results here. However, this unsupervised method inhibits the option of knowing the visual units used by the forest. As our priority in this comparison study is to measure the effects of viseme selection rather than optimising a classification method for each individual speaker, we bare the cost to overall classification for the learning gained from the observations by comparing viseme sets.

Figures 7.1 and 7.3 show the word correctness percentage aggregated over all speakers, $\pm 1 \frac{\sigma}{\sqrt{7}}$. Respective heat maps for all phoneme-to-viseme maps are in Fig-

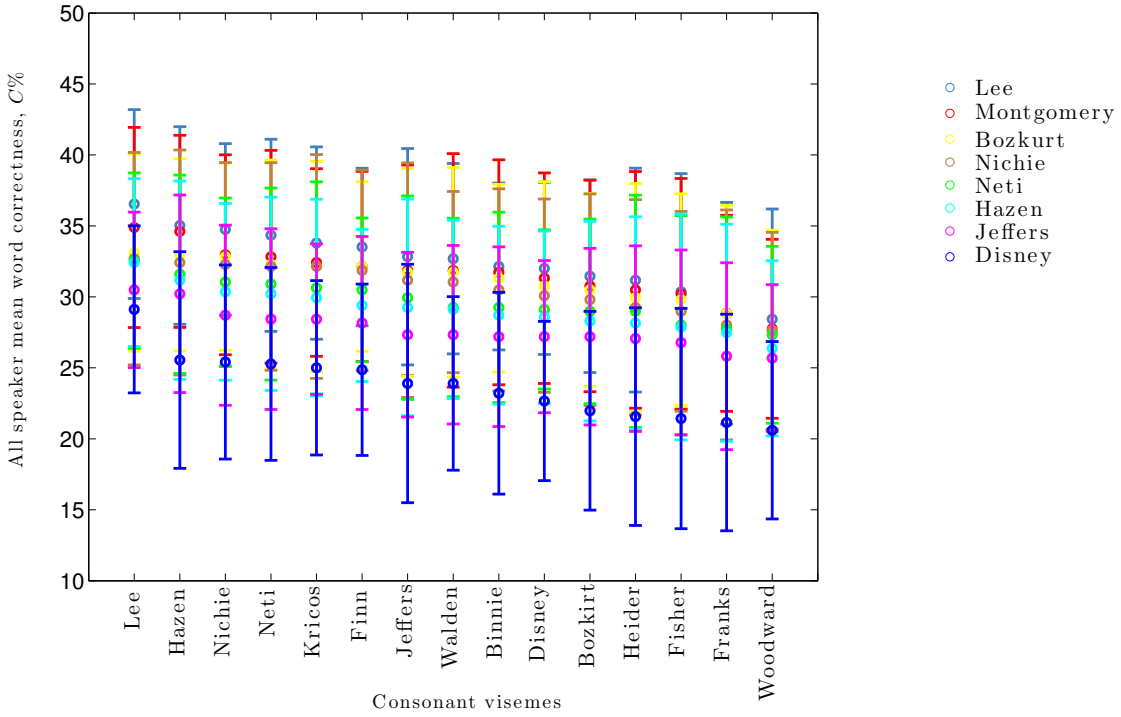



Figure 7.1: Speaker-dependent all-speaker mean word classification, $C \pm 1\frac{\sigma}{\sqrt{7}}$, over all four speakers comparing consonant P2V maps. For a given consonant mapping ($x-$axis) the performance is measured after pairing with all vowel mappings.

ures 7.2 & 7.4. Figure 7.1 shows all consonant maps along the $x$-axis and, for each consonant map, a pairing with a vowel map has been plotted at the respective consonant map position on the $x$-axis. This shows the differences between each consonant map and the effect of the vowel maps on each consonant map. Figure 7.3 is vice versa. The black line is the mean word classification grouped by all paired maps. Both $x-$axies are ordered by the map's mean rank over all speakers. This demonstrates the 'best' performing map for both consonants and vowels are from Lee (as this is left-most on the $x-$axis) for all speakers. Therefore, Lee's visemes [82] become the benchmark in the next piece of work in this chapter.

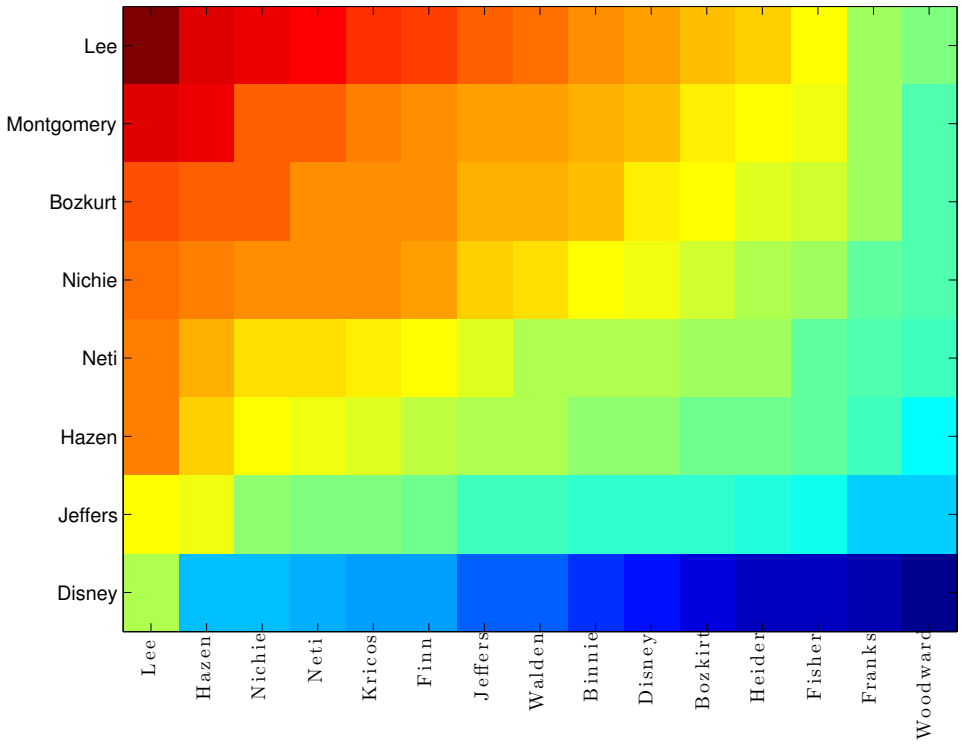



Figure 7.2: Speaker-dependent all-speaker mean word classification, $C$, heatmap.

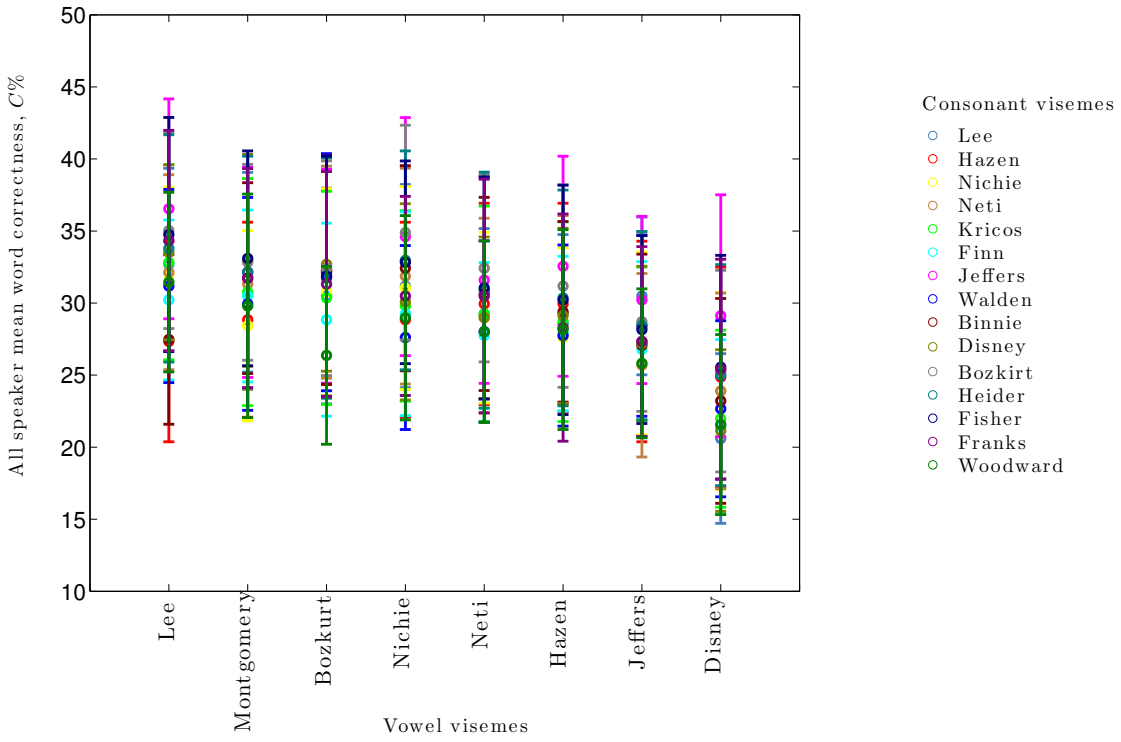



Figure 7.3: Speaker-dependent all-speaker mean word classification, $C \pm 1\frac{\sigma}{\sqrt{7}}$, over all four speakers comparing vowel P2V maps. For a given vowel mapping ($x-$axis) the performance is measured after pairing with all consonant mappings.

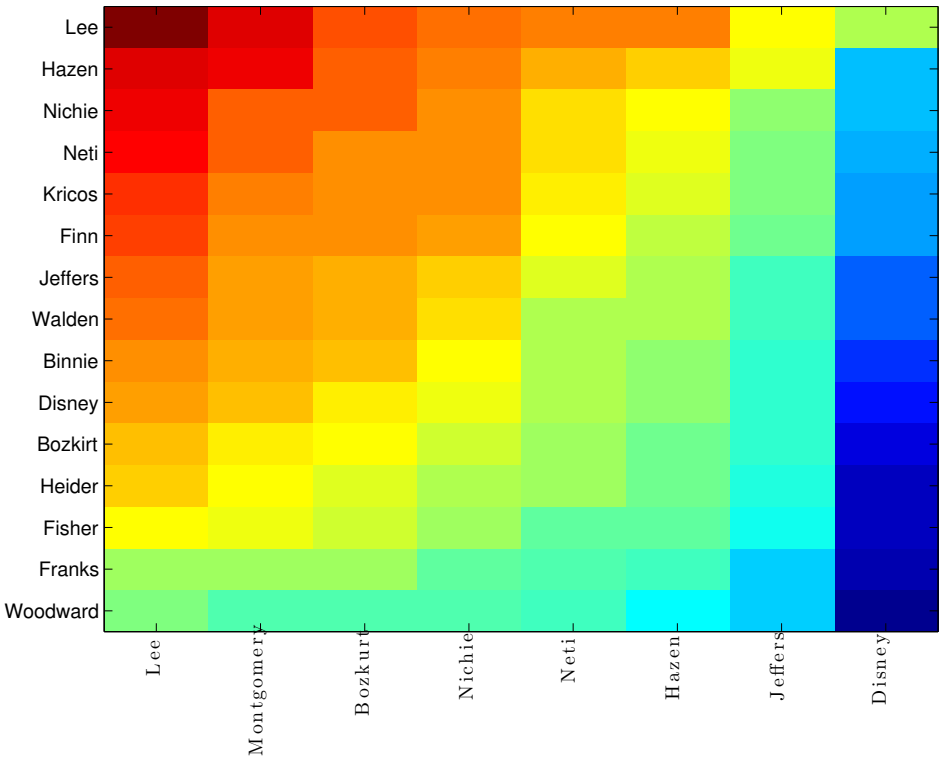



Figure 7.4: Speaker-dependent all-speaker mean word classification, $C$, heatmap.

Comparing the consonant P2V maps in Figure 7.1 shows the Disney vowels are significantly worse than all others when paired with all consonant maps. Over the other vowels there is overlap with the majority of error bars suggesting little significant difference over the whole group, although Lee [82] and Bozkurt [25] vowels are consistently above the mean and above the upper error bar for Disney [80], Jeffers [69] and Hazen [58] vowels. In comparing the vowel P2V maps in Figure 7.3 Lee [82] and Hazen [58] are the best consonants by a margin above the mean whereas Woodward [148] and Franks [50] are the bottom performers. Figures 7.1 and 7.3 show the performance of the viseme maps averaged across speakers, there is a significant difference between the 'best' visemes for individual speakers which arises from the unique way in which everyone articulates their speech.

These observations are confirmed in heatmaps in Figures 7.2 & 7.4.

Figures 7.5 and 7.6 are critical difference plots between the viseme class sets

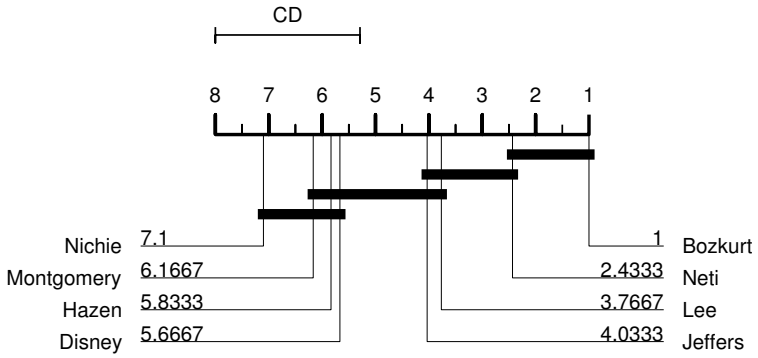



Figure 7.5: Critical difference of all vowel phoneme-to-viseme maps independent of consonant phoneme-to-viseme map pair partner.

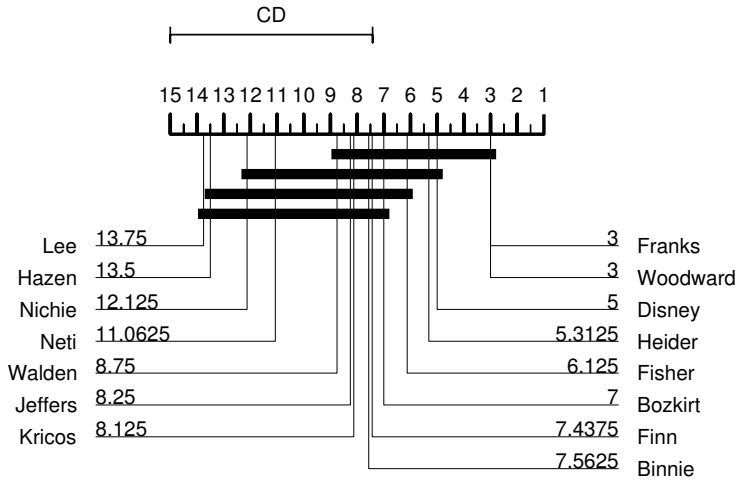



Figure 7.6: Critical difference of all consonant phoneme-to-viseme maps independent of vowel phoneme-to-viseme pair partner.

based upon their classification performance [39]. Critical difference is a measure of our confidence intervals between different machine learning algorithms. Two assumptions within critical difference are: all measured results are 'reliable', and all algorithms are evaluated using the same random samples [39]. As we use the HTK standard metrics [152], and use results with consistent random sampling across folds, these assumptions are not a concern. We have selected critical differences here as these evaluate the performance of multiple classifiers, and previous studies, such as [23, 19], do not consider the applicability of statistics when tested over more than one dataset [39]. As our HMM classifiers are speaker dependent, we can safely consider the data of each speaker as an isolated dataset within AVL2.

Figure 7.5 is the comparison of the vowel labelled viseme sets. Starting on the left-hand side of the figure, it shows that Nichie, Montgomery, Hazen, and Disney vowels are not critically different from each other signified by the black horizontal bar crossing their respective lines on the left side of the figure. Likewise, Montgomery, Hazen, Disney, Jeffers, and Lee vowels are also not critically different from each other. These two bars alone demonstrate that Nichie's vowels are critically different from Jeffers, Lee, Neti, and Bozkurt's. On the right hand side of the graph we can see that Bozkurt's vowels are critically different from all bar Neti's vowels. This is interesting as in Figure 7.3 they do not appear to perform significantly differently to any other vowel visemes. In fact, whilst Bozkurt and Nichie vowels are the most critically different from each other, they are adjacent in classification performance. This gives us hope that an optimal set of visemes is possible as the effect of clusters of phonemes varies by the specific phonemes being clustered.

Figures 7.5 and 7.6 demonstrate a significant difference between some sub-sets of viseme sets (the bars do not overlap all classifier maps). This is based upon insignificant variation within each sub-set. This suggests there could be dependency between some viseme sets as the groupings align with the derivation method of the P2V mappings.

The mean word classification for all speakers and all folds for each map is plotted

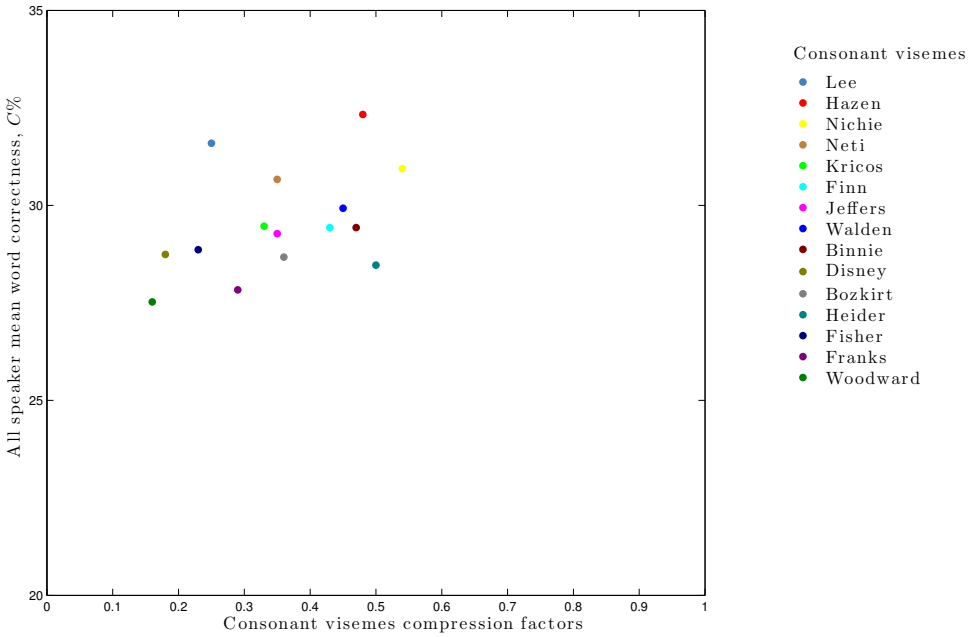



Figure 7.7: Scatter plot showing the relationship between compression factors and word correctness, $C$, classification with consonant phoneme-to-viseme maps.

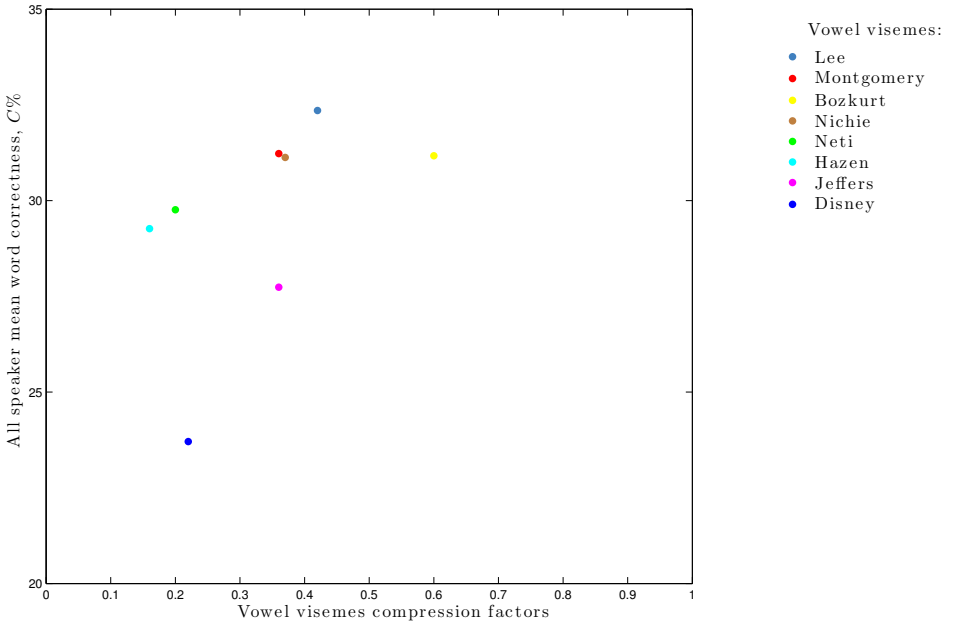



Figure 7.8: Scatter plot showing the relationship between compression factors and word correctness, $C$, classification with vowel phoneme-to-viseme maps.

in Figures 7.7 and 7.8. Looking at our confusion factors for the best performing P2Vs of each speaker (Figure 7.7 and Figure 7.8), this suggests a good preparation of phonemes to visemes is ideally around 0.45 or approximately ~2 phonemes per viseme. This also is the CF for Lee. Lee has the highest performing word classification map for both consonants and vowels displayed in Figure 7.9 and interestingly, not the highest number of visemes (the $x$-axis in Figure 7.9).

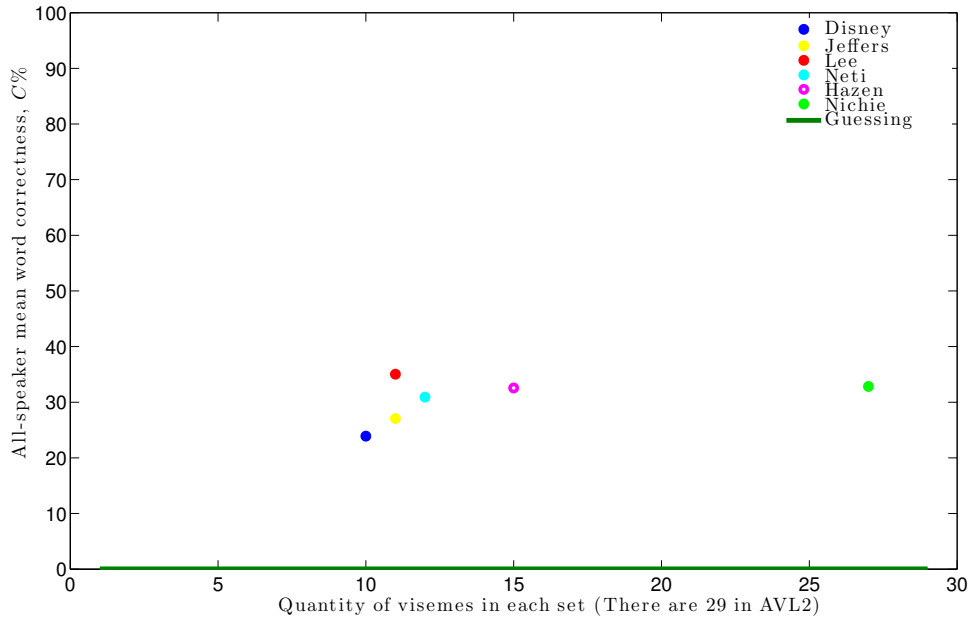



Figure 7.9: For previously presented phoneme-to-viseme maps which include both vowel and consonant phonemes, word correctness, $C$ is plotted against the count of visemes in each phoneme-to-viseme map.

## 7.5 New phoneme to viseme maps

In the second part of our phoneme-to-viseme mapping study, three approaches are used to find a better method of mapping phonemes to visemes. The first approach uses the most common pairs of phonemes from existing mappings.

A comparison of previously presented P2V maps shows subgroups of phonemes which are regularly grouped together into visemes [28, 138]. The most popular of these phoneme-subgroups have a high occurrence across sets. Our first new ap-

proach uses the number of occurrences and the size of the subgroup as a weighting for grouping together phonemes, i.e. the highest weighted phoneme-subgroup will be grouped into a viseme first, without duplicating phonemes into more than one viseme. The P2V maps used in this clustering process have been devised for different reasons (for example, based upon linguistic rules or upon human lip-reader observations, see Table 7.6). This set helps us to understand that if what we currently assume to be good groups really are the best groups of phonemes for optimal classification.

The second and third approaches are both speaker-dependent and data-driven from phoneme classification. Two cases are considered:

1. a strictly coupled map, where a phoneme can be grouped into a viseme only if it has been confused with *all* the phonemes within the viseme, and

2. a relaxed coupled case, where phonemes can be grouped into a viseme if it has been confused with *any* phoneme within the viseme.

With all new P2V mappings each phoneme can only be allocated to one viseme class. These new P2V maps are tested on the AVL2 dataset using the same classification method as described in Section 7.3. The results from the best performing P2V map from our comparison study (Lee [82]) is the benchmark to measure improvements.

### 7.5.1 Common phoneme-pair visemes

The first approach for finding a new speaker-independent P2V map uses the most commonly coupled phonemes to build new visemes. In detail, all visemes in the previous maps are searched to make a full dictionary of unique pairs of phonemes. Associated with each dictionary entry is a count of how many times they appear in *any* defined P2V map from those in the comparison study in Section 7.4 with HTK. This phoneme pair list is sorted by descending occurrence count. On passing through

this list the next phoneme pair is assigned to a viseme class based upon matching phonemes (whilst not duplicating the presence of a phoneme within one viseme). A phoneme is not permitted to be added to more than one viseme. Priority is given to the pairings with a higher count. If a particular phoneme was never coupled with other phonemes, that phoneme forms a unique viseme of its own.

Table 7.8: Visemes derived using most-common phoneme pairings in previously presented phoneme-to-viseme mappings.

| | |
|---|---|
| Common-pair Visemes (CF:0.28) | {/d/ /l/ /n/ /t/} {/b/ /m/ /p/} {/g/ /h/ /ɦ/ /k/ /ŋ/ /y/} {/f/ /v/} {/ɔ/ /ʊə/ /ɔə/} {/ɛə/ /i/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/ɑ/ /ʌ/ /ao} {/s/ /z/} {/dh/ /θ/} {/r/ /w/ /ɯ/} {/æ/ /e/ /ei/ /ɪə/} {/a/ /ai/ /ai/ /e/ /i/ /ɪ/} {/ɑʊ/ /ə/ /ɜ/ /əʊ/ /u/ /ʊ/ /u/} |

### 7.5.2 Viseme classes with strictly confusable phonemes

The second and third approaches for identifying visemes are speaker-dependent, data-driven and based on phoneme confusions within the classifier. The first undertaking in this work is to complete classification using phoneme labelled HHM classifiers. The classifiers are built in HTK with flat-started HMMs and force aligned training data for each speaker. The HMMs are re-estimated 11 times in total over seven folds of leave-one-out cross validation. This overall classification task does not perform well (see Table 7.9) particularly for an isolated word dataset. However, the HTK tool `HResults` is used to output a confusion matrix for each fold detailing which phoneme labels confuse with others and how often. Appendix 10.2, Figure 2 is an example confusion matrix. For both data-driven speaker-dependent approaches, this first step of completing phoneme classification is essential to create the data to derive the P2V maps from.

Table 7.9: Mean per speaker Correctness, $C$, of phoneme-labelled HMM classifiers.

| | Speaker 1 | Speaker 2 | Speaker 3 | Speaker 4 |
|---|---|---|---|---|
| Phoneme $C$ | 24.72 | 23.63 | 57.69 | 43.41 |

Now, let us use a smaller seven-unit confusion matrix example to explain our

clustering method in full. Our demonstration confusion matrix is in Figure 7.10.

| | /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 |

Figure 7.10: Demonstration (theoretical) confusion matrix showing confusions between phoneme-labelled classifiers to be used for clustering to create new speaker-dependent visemes. 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.

For the 'strictly-confused' viseme set (remember there is one per speaker), the second step of deriving the P2V map is to check for single-phoneme visemes. Any phonemes which have only been correctly recognised as themselves and have no false positive/negative classifications are permitted to be single phoneme visemes. In Figure 7.10 we have highlighted the true positive classifications in red and both false positives and false negative classifications in blue which shows /$p6$/ is the only phoneme to fit our 'single-phoneme viseme' definition. /$p6$/ has a true positive value of +4 and zero false classifications. Therefore this is our first viseme. /$v1$/ = {/$p6$/}.

This action is followed by defining all combinations of remaining phonemes which can be grouped into visemes and identifying the grouping that contains the largest number of confusions by ordering all the viseme possibilities by descending size (whole list shown in Figure 7.11).

Our grouping rule states that phonemes can be grouped into a viseme class only if all of the phonemes within the candidate group are mutually confusable. This means each pair of phonemes within a viseme must have a total false positive and false negative classification greater than zero. Once a phoneme has been assigned to a viseme class it can no longer be considered for grouping, and so any possible phoneme combinations that include this viseme are discarded. This ensures phonemes can

| | | |
|---|---|---|
| {/p1/, /p2/, /p3/, /p4/, /p5/, /p7/} | {/p1/, /p2/, /p3} | {/p1/, /p2/} |
| {/p1/, /p2/, /p3/, /p4/, /p5/} | {/p1/, /p2/, /p4/} | {/p1/, /p3/} |
| {/p1/, /p2/, /p3/, /p4/, /p7/} | {/p1/, /p2, /p5/} | {/p1/, /p4/} |
| {/p1/, /p2/, /p3/, /p5/, /p7/} | {/p1/, /p2, /p7/} | {/p1/, /p5/} |
| {/p1/, /p2/, /p4/, /p5/, /p7/} | {/p2/, /p3, /p4/} | {/p1/, /p7/} |
| {/p1/, /p3/, /p4/, /p5/, /p7/} | {/p2/, /p3, /p5/} | {/p2/, /p3/} |
| {/p2/, /p3/, /p4/, /p5/, /p7/} | {/p2/, /p3, /p7/} | {/p2/, /p4/} |
| {/p1/, /p2/, /p3/, /p4/} | {/p3/, /p4/, /p5/} | {/p2/, /p5/} |
| {/p1/, /p2/, /p3/, /p5/} | {/p3/, /p4/, /p7/} | {/p2/, /p7/} |
| {/p1/, /p2/, /p3/, /p7/} | {/p1/, /p3/, /p4/} | {/p3/, /p4/} |
| {/p2/, /p3/, /p4/, /p5/} | {/p4/, /p5/, /p7/} | {/p3/, /p5/} |
| {/p2/, /p3/, /p4/, /p7/} | {/p1/, /p4/, /p5/} | {/p3/, /p7/} |
| {/p3/, /p4/, /p5/, /p7/} | {/p2/, /p4/, /p5/} | {/p4/, /p5/} |
| {/p1/, /p3/, /p4/, /p5/} | {/p1/, /p5/, /p7/} | {/p4/, /p7/} |
| {/p1/, /p4/, /p5/, /p7/} | {/p2/, /p5/, /p7/} | {/p5/, /p7/} |
| {/p2/, /p4/, /p5/, /p7/} | {/p3/, /p5/, /p7/} | |
| | {/p1/, /p3/, /p5/} | |
| | {/p1/, /p3/, /p7/} | |
| | {/p1/, /p4/, /p7/} | |
| | {/p2/, /p4/, /p7/} | |

Figure 7.11: List of all possible subgroups of phonemes with an example set of seven phonemes

belong to only a single viseme.

By iterating though our list of all possibilities in order, we check if all the phonemes are mutually confused. This means all phonemes have a positive confusion value (a blue value in Figure 7.10) with all others.

The first phoneme possibility in our list where this is true is $\{/p1/, /p3/, /p7/\}$. This is confirmed by the Figure 7.10 values:

$\Pr\{/p1/|/p3/\}+\Pr\{/p3/|/p1/\} = 0 + 1 = 1$ which is $> 0$

also, $\Pr\{/p1/|/p7/\}+\Pr\{/p7/|/p1/\} = 4 + 1 = 5$ which is $> 0$

and $\Pr\{/p3/|/p7/\}+\Pr\{/p7/|/p3/\} = 1 + 3 = 4$ which is $> 0$.

This becomes our second viseme and thus our current viseme list looks like Table 7.10.

We now only have three remaining phonemes to cluster, $p2, p4$ and $p5$. This

Table 7.10: Demonstration example 1: first-iteration of clustering, a phoneme-to-viseme map for strictly-confused phonemes.

| Viseme | Phonemes |
|---|---|
| /v1/ | {/p6/} |
| /v2/ | {/p1/, /p3/, /p7/} |

{/p2/, /p4/, /p5/}
{/p2/, /p4/}
{/p2/, /p5/}
{/p4/, /p5/}

Figure 7.12: List of all possible subgroups of phonemes with an example set of seven phonemes after the first viseme is formed.

reduces our list of possible combinations substantially, see Figure 7.12.

The next iteration of our clustering algorithm identifies the combination of remaining phonemes which correspond to the next largest number of confusions, and so on, until no phonemes can be merged. This leaves us with the final visemes in Table 7.11.

Table 7.11: Demonstration example 2: final phoneme-to-viseme map for strictly-confused phonemes.

| Viseme | Phonemes |
|---|---|
| /v1/ | {/p6/} |
| /v2/ | {/p1/, /p3/, /p7/} |
| /v3/ | {/p2/, /p4/} |
| /v4/ | {/p5/} |

Our original phoneme classification has produced confusion matrices which permit confusions between vowel and consonant phonemes. We can see in Section 7.2 (Tables 7.4 and 7.5), previously presented P2V maps that vowel and consonant phonemes are not commonly mixed within visemes. Therefore, we make two types of P2V maps: one which permits vowels and consonant phonemes to be mixed within the same viseme, and a second which restricts visemes to be vowel or consonant only by putting an extra condition in when checking for confusions greater than zero.

It should be remembered that not all phonemes present in the ground truth transcripts will have been recognised and included in the phoneme confusion matrix. Any of the remaining phonemes which have not been assigned to a viseme are grouped into a single garbage */gar/* viseme. This approach ensures any phonemes which have been confused are grouped into a viseme and we do not lose any of the 'rarer', and less common visual phonemes. For example, */ea/*, */oh/*, */ao/*, and */r/* are not in the original transcript and so can be placed into */gar/*. But for Speaker 2, */gar/* also contains */ay/* and */p/*, and for Speaker 4 */gar/* also contains */p/* and */z/*, as these do not show up in the speaker's phoneme classification outputs. This task has been undertaken for all four speakers in our dataset. The final P2V maps are shown in Table 7.12.

Table 7.12: Strictly-confused phoneme speaker-dependent visemes. The score in brackets is the ratio of visemes to phonemes.

| Classification | P2V mapping - permitting mixing of vowels and consonants |
|---|---|
| Speaker1 (CF:0.48) | {/ʌ/ /ai/ /i/ /n/ /əʊ/} {/b/ /e/ /ei/ /y/ } {/d/ /s/} {/tʃ/ /l/} {/ə/ /v/} {/w/} {/f/} {/k/} {/ə/ /v/} {/dʒ/ /z/} {/ɑ/ /u/} {/t/} |
| Speaker2 (CF: 0.44) | {/ə/ /ai/ /ei/ /i/ /s/} {/e/ /v/ /w/ /y/} {/l/ /m/ /n/} {/b/ /d/ /p/} {/z/} {tʃ/} {/t/} {/ɑ/} {/dʒ/ /k/} {/ʌ/ /f/} {/əʊ/ /u/} |
| Speaker3 (CF: 0.68) | {/ei/ /f/ /n/} {/d/ /t/ /p/} {/b/ /s/} {/l/ /m/} {/ə/ /e/} {/i/} {/u/} {/ɑ/} {/dʒ/} {/əʊ/} {/z/} {/y/} {/tʃ}/ {/ai/} {/ʌ/} {/ɑ/} {/dʒ/} {/əʊ/} {/k/ /w/} {/v/} {/z/} |
| Speaker4 (CF: 0.64) | {/ʌ/ /ai/ /i/ /ei/ } {/m/ /n/} {/ə/ /e/ /p/} {/k/ /w/} {/d/ /s/} {/dʒ/ /t/} {/f/} {/v/} {/ɑ/} {/z/} {/tʃ/} {/b/} {/əʊ/} {/əʊ/} {/l/} {/u/} {/b/} |
| Classification | P2V mapping - restricting mixing of vowels and consonants |
| Speaker1 (CF:0.50) | {/ʌ/ /i/ /əʊ/ /u/} {/ɑ/ /ei/} {/ə/ /e/ /ei/} {/d/ /s/ /t/ } {/tʃ/ /l/ } {/k/} {/z/} {/w/} {/f/} {/m/ /n/} {/dʒ/ /v/} {/b/ /y/} |
| Speaker2 (CF: 0.58) | {/ai/ /ei/ /i/ /u/} {/əʊ/} {/ə/} {/e/} {/ʌ/} {/ɑ/} {/v/ /w/} {/dʒ/ /p/ /y/} {/d/ /b/} {/t/} {/k/} {/tʃ/} {/l/ /m/ /n/} {/f/ /s/} |
| Speaker3 (CF: 0.68) | {/ei/ /i/} {/ai/} {/ə/ /e/} {/ʌ/} {/d/ /p/ /t/} {/l/ /m/} {/k/ /w/} {/v/} {/tʃ/} {/əʊ/} {/y/} {/u/} {/ɑ/} {/z/} {/f/ /n/} {/b/ /s/} {/dʒ/} |
| Speaker4 (CF: 0.65) | {/ʌ/ /ai/ /i/ /ei/} {/ə/ /e/} {/m/ /n/} {/k/ /l/} {/dʒ/ /t/} {/d/ /s/} {/tʃ/} {/əʊ/} {/y/} {/u/} {/ɑ/} {/w/} {/f/} {/v/} {/b/} |

### 7.5.3 Viseme classes with relaxed confusions between phonemes

A disadvantage of the strictly confusable viseme set is that it contains some spurious single-phoneme visemes where the phoneme cannot be grouped because it is not confused with *all* other phonemes in the viseme. These types of phonemes are likely to be either: borderline cases at the extremes of a viseme cluster, i.e. they have subtle visual similarities to more than one phoneme cluster, or they do not occur frequently enough in the training data to be differentiated from other phonemes.

To address this we complete a second pass-through of the strictly-confused visemes listed in Table 7.11. We begin with the visemes as they currently stand (in our demonstration example containing four classes) and relax the condition requiring confusion with all of the phonemes. Now any single phoneme viseme (in our demonstration, $/v4/$) can be allocated to a previously existing viseme if it has been confused with any phoneme in the viseme. In Figure 7.10 we see $/p5/$ was confused with $/p1/$, $/p3/$, and $/p4/$. Because $/p4/$ is not in the same viseme as $/p1/$ and $/p3/$ we use the value of confusion to decide which to allocate it to as follows.

$\Pr\{/p1/|/p5/\}+\Pr\{/p5/|/p1/\} = 0 + 3 = 3$

$\Pr\{/p3/|/p5/\}+\Pr\{/p5/|/p3/\} = 0 + 1 = 1$

$\Pr\{/p4/|/p5/\}+\Pr\{/p5/|/p4/\} = 2 + 1 = 3$

Therefore; for $p5$ the total confusion with $/v2/$ is $3 + 1 = 4$, whereas the total confusion with $/v3/$ is 3. We select the viseme with most confusion to incorporate the unallocated phoneme $/p5/$. This reduces the number of viseme classes by merging single-phoneme visemes from Table 7.11 to form a second set shown in Table 7.13. This has the added benefit that we have also increased the number of training samples for each classifier.

Table 7.13: Demonstration example 3: final phoneme-to-viseme map for relaxed-confused phonemes.

| Viseme | Phonemes |
|---|---|
| $/v1/$ | $\{/p6/\}$ |
| $/v2/$ | $\{/p1/, /p3/, /p5/, /p7/\}$ |
| $/v3/$ | $\{/p2/, /p4/\}$ |

Remember, as we have two versions of Table 7.11 - one with mixed vowel and consonant phonemes and a second with divided vowels and consonant phonemes - the same still applies to our relaxed-confused visemes sets. This means we end up with four types of speaker-dependent phoneme-to-viseme maps, described in Table 7.14. For our strictly-confused P2V maps in Table 7.12, these become the relaxed P2V maps in Table 7.15.

| | |
|---|---|
| Mixed vowels and consonants<br>+<br>Strict-confusion of phonemes | Split vowels and consonants<br>+<br>Strict-confusion of phonemes |
| Mixed vowels and consonants<br>+<br>Relaxed-confusion of phonemes | Split vowels and consonants<br>+<br>Relaxed-confusion of phonemes |

Table 7.14: The four variations on speaker-dependent phoneme-to-viseme maps derived from phoneme confusion in phoneme classification.

Table 7.15: Relaxed-confused phoneme speaker-dependent visemes. The score in brackets is the ratio of visemes to phonemes.

| Classification | P2V mapping - permitting mixing of vowels and consonants |
|---|---|
| Speaker1 | {/b/ /e/ /ei/ /p/ /w/ /y/ /k/} {/ʌ/ /ai/ /f/ /i/ /m/ /n/ /əʊ/} |
| (CF:0.28) | {/dʒ/ /z/} {/ɑ/ /u/} {/d/ /s/ /t/} {/tʃ/ /l/} {/ə/ /v/}{/ə/ /v/} |
| Speaker2 | {/ɑ/ /ə/ /ai/ /ei/ /i/ /s/ /tʃ/} {/e/ /t/ /v/ /w/ /y/} {/l/ /m/ /n/} |
| (CF: 0.32) | {/ʌ/ /f/} {/z/} {/b/ /d/ /p/} {/əʊ/ /u/} {/dʒ/ /k/} |
| Speaker3 | {/ʌ/ /ai/ /ei/ /f/ /i/ /n/} {/ə/ /e/ /y/ /tʃ/} {/b/ /s/ /v/} {/l/ /m/ /u/} |
| (CF: 0.40) | {/dʒ/} {/əʊ/} {/z/} {/d/ /p/ /t/} {/k/ /w/} {/ɑ/} |
| Speaker4 | {/ʌ/ /ai/ /tʃ/ /i/ /ei/ } {/ɑ/ /m/ /u/ /n/} {/ə/ /e/ /p/ /v/ /y/} |
| (CF: 0.32) | {/dʒ/ /t/} {/k/ /l/ /w/} {/əʊ/} {/d/ /f/ /s/} {/b/} |
| Classification | P2V mapping - restricting mixing of vowels and consonants |
| Speaker1 | {/ʌ/ /i/ /əʊ/ /u/} {/ɑ/ /ai/} {/ə/ /e/ /ei/} {/b/ /w/ /y/} {/d/ /f/ /s/ /t/} |
| (CF:0.47) | {/k/} {/z/} {/m/} {/l/} {/tʃ/} {/dʒ/ /k/ /v/ /z/} |
| Speaker2 | {/ɑ/ /ʌ/ /ə/ /ai/ /ei/ /i/ /əʊ/ /u/} {/k/ /t/ /v/ /w/} {/tʃ/ /l/ /m/ /n/} |
| (CF: 0.29) | {/f/ /s/} {/dʒ/ /p/ /y/} {/b/ /d/} {/z/} |
| Speaker3 | {/ʌ/ /ai/ /i/ /ei/} {/ə/ /e/} {/b/ /s/ /v/} {/d/ /p/ /t/} {/l/ /m/} |
| (CF: 0.56) | {/y/} {/dʒ/} {/əʊ/} {/z/} {/u/} {/ə/ /e/} {/k/ /w/} {/f/ /n/} {/ɑ/} {/tʃ/} |
| Speaker4 | {/ʌ/ /ai/ /i/ /ei/} {/tʃ/ /k/ /l/ /w/} {/d/ /f/ /s/ /v/} {/m/ /n/} |
| (CF: 0.50) | {/f/} {/ɑ/} {/dʒ/ /t/} {/əʊ/} {/u/} {/y/} {/b/} |

Now, and this is why these visemes are defined as relaxed, any remaining phonemes which have confusions, but are so far not assigned to a viseme, the phoneme-pair

confusions are used to map the remaining phonemes to an appropriate viseme, even though it does not confuse with all phonemes already in it. Any remaining phonemes which are not assigned to a viseme are grouped into a new garbage /gar/ viseme. This approach ensures any phonemes which have been confused with any other are grouped into a viseme.

## 7.6 Bear speaker-dependent visemes

Figure 7.13 shows word correctness of the common phoneme-pair visemes against Lee's benchmark. It is no surprise the common-pair visemes are all worse than Lee, as Lee gave the maximum performance of the original P2V mappings used to deduce the new map. However, the overlap in error bars shows that for two speakers this is not a significant reduction. Unfortunately, no particular viseme, or group of visemes, particularly contribute to the set correctness.

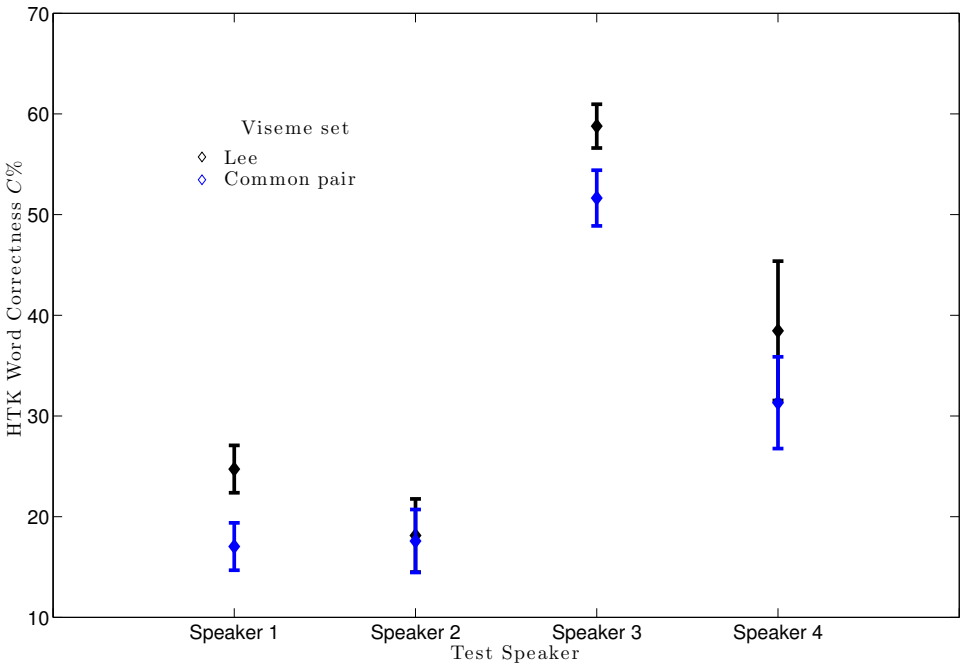



Figure 7.13: Word classification correctness $C \pm 1\frac{\sigma}{\sqrt{7}}$, using the common phoneme-pairs phoneme-to-viseme map. Lees benchmark is in black.

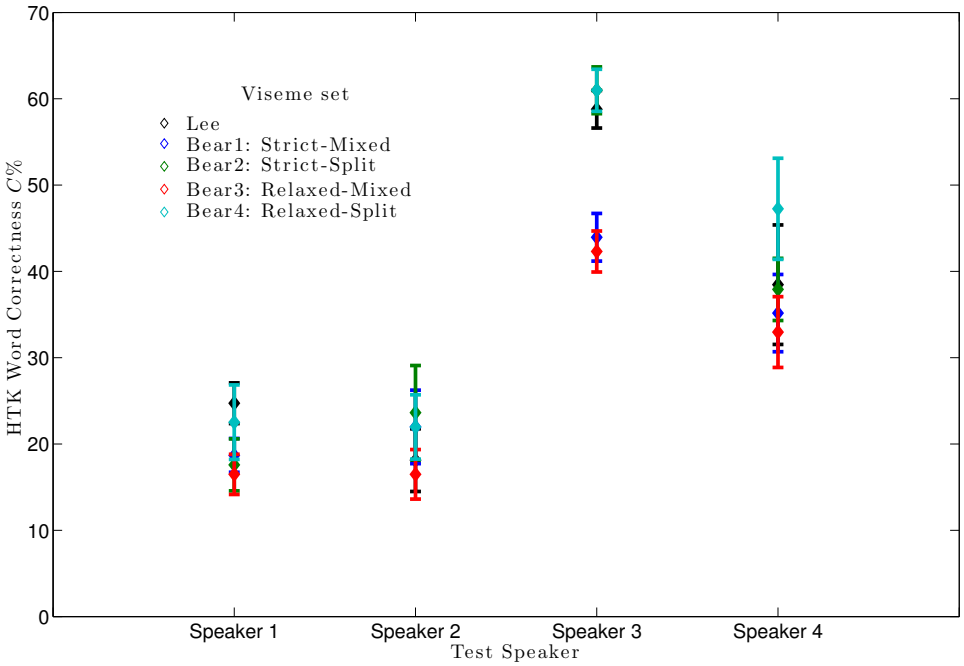



Figure 7.14: Word classification correctness $C \pm 1\frac{\sigma}{\sqrt{7}}$, using all four new methods of deriving speaker dependent visemes. Lees benchmark is in black.

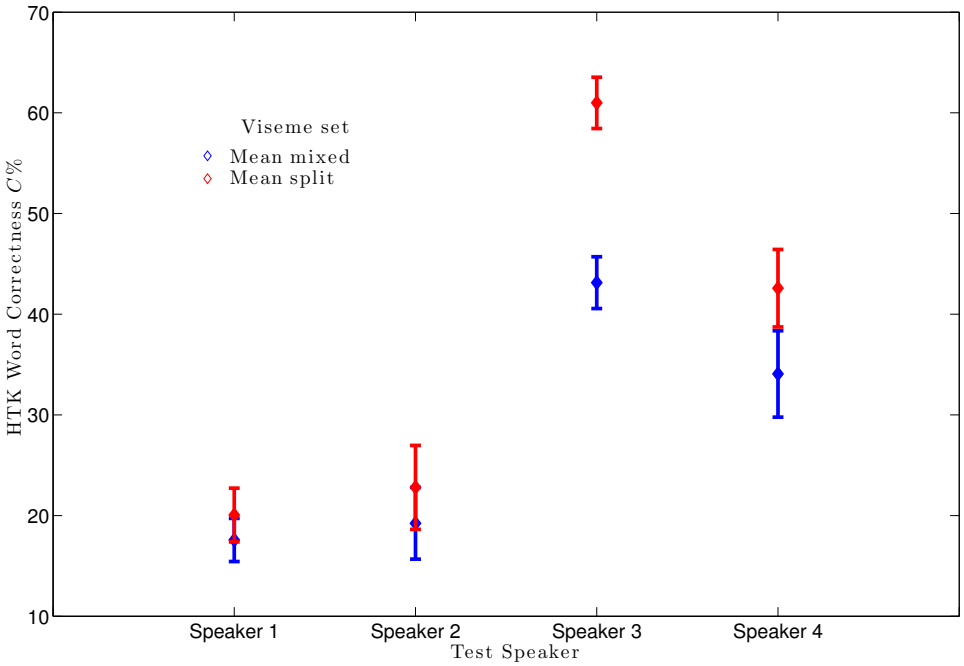



Figure 7.15: A comparison of the split vowel and consonant phoneme visemes and the mixed vowel and consonant phoneme visemes with AVLetters2 speakers.

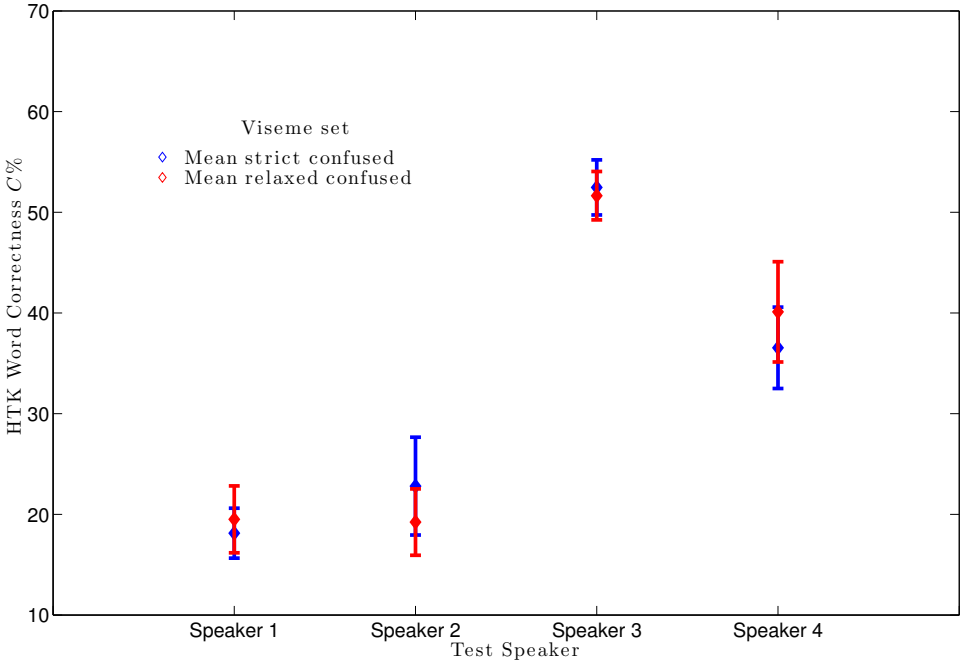



Figure 7.16: A comparison of the strict mutually confusable phoneme viseme classes and the relaxed confused phoneme visemes with AVLetters2 speakers.

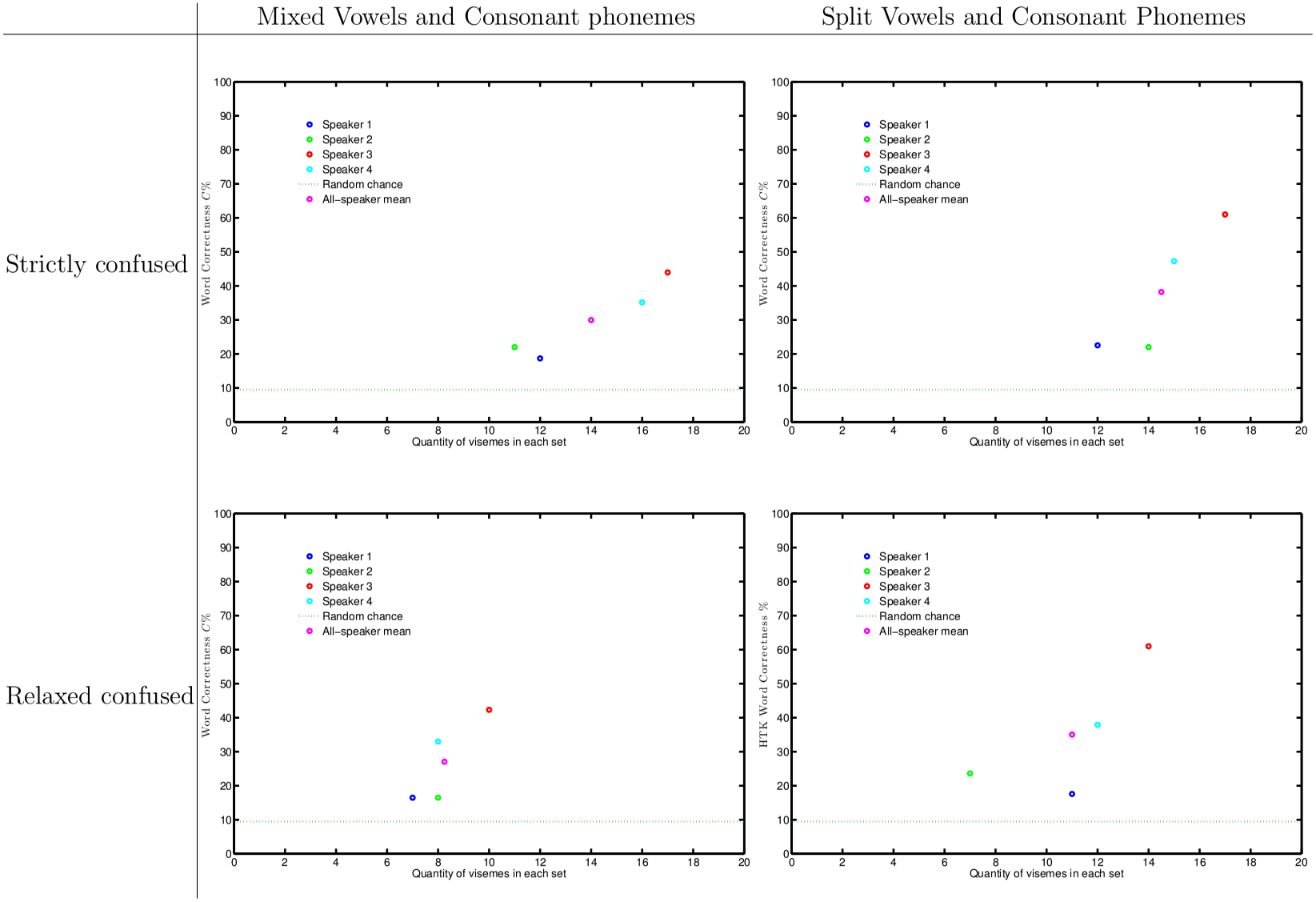
In Figure 7.14 all four speaker-dependent maps tested on each speaker are plotted on the $x$-axis to compare the difference in word classification (shown on the $y$-axis). The benchmark from the comparison study, Lee, is in black. For Speaker 1 and Speaker 3, no new viseme map significantly improves upon Lee's performance although we do see improvements for both Speaker 2 and Speaker 4. The strictly-confused and split viseme map improves upon Lee's previous best word classification.

Figure 7.15 compares the mixed consonant and vowel maps against split consonant and vowel maps, also measured in word correctness, $C$, on the $y$-axis. The split P2V maps are always better than mixed for all speakers. Figure 7.16 shows the comparison of strictly-confused and loosely confused viseme classes. The strict confusions are better for two out of four speakers. These are speakers with the highest ratio of phonemes to visemes (Tables 7.15 and 7.12).

In Figure 7.17, all four variants of our new P2V maps are plotted for each speaker and an all-speaker mean against the number of visemes in each set. Splitting vowel and consonant phonemes gives a greater number of classifiers, which reduces the

number of training samples per class, but results in higher correctness for all speakers. This shows that having the *right* training samples is more important than having simply 'more data'. Whilst showing a smaller effect, the two graphs on the left hand side of Figure 7.17 shows the relaxing of confusable phonemes has a negative influence, even though this reduces the number of visemes and increases training samples per class, they are not good training samples to include for the class.

Figure 7.17: How Correctness varies with quantity of visemes in each set. All four variants on a speaker-dependent data-driven approach to finding visemes plotted against the count of visemes within each set.



In Figures 7.18, 7.19, 7.20, and 7.21, the contribution of each viseme has been listed in descending order along the $x-$axis for each speaker in AVL2. The contribution of each viseme is measured as the inverse probability of each class, $\Pr\{v|\hat{v}\}$. These values have been calculated from the `HResults` confusion matrices. There is no significant step when a viseme contribution is no longer needed, that is in speaker-dependent visemes we need all class labels within a set. This analysis of visemes within a set is also used in [17], which proposes a threshold subject to the information in the features. Using combined shape and appearance features here removes the threshold as these figures show irrespective of which method of phoneme-clustering is used for devising visemes, the greater the number of visemes in a set, the higher the overall classification. More important to see is the overall classification $C$ is higher when there is less range between individual viseme $\Pr\{v|\hat{v}\}$ values within a set of visemes. The difference values between the highest and least contributing visemes for each method and speaker are listed in Table 7.16.

Table 7.16: Viseme variation in $\Pr\{v|\hat{v}\}$ showing the best and worst classifiers within each set of visemes for each derivation method per speaker.

| Method | Speaker 1 | Speaker 2 | Speaker 3 | Speaker 4 |
|---|---|---|---|---|
| Relaxed: mixed | 76.97 | 32.86 | 14.29 | 7.14 |
| Relaxed: split | 100.00 | 19.05 | 7.14 | 14.29 |
| Strict: mixed | 56.35 | 37.50 | 14.26 | 14.29 |
| Strict: split | 65.54 | 42.86 | 8.57 | 5.71 |

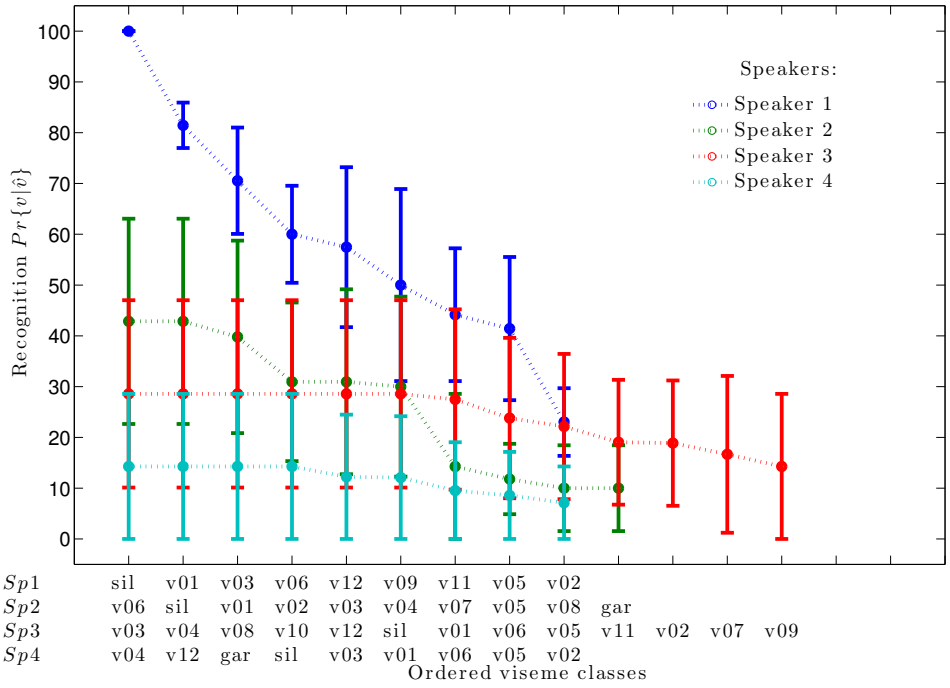



Figure 7.18: Individual viseme classification, $\Pr\{v|\hat{v}\}$ with the relaxed, mixed vowels and consonant Bear visemes.

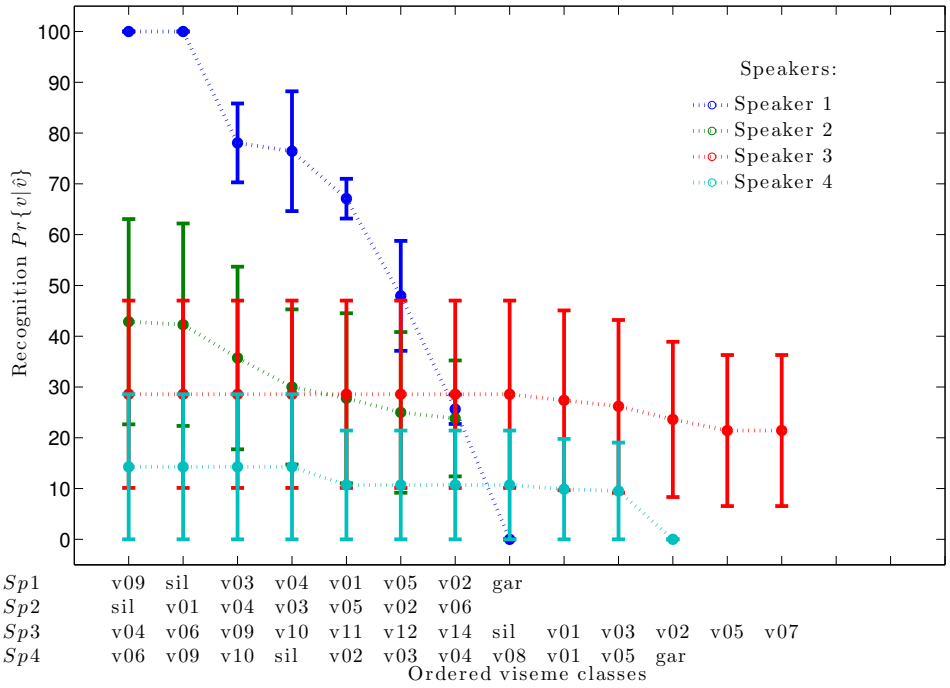



Figure 7.19: Individual viseme classification, $\Pr\{v|\hat{v}\}$ with the relaxed, split vowels and consonant Bear visemes.

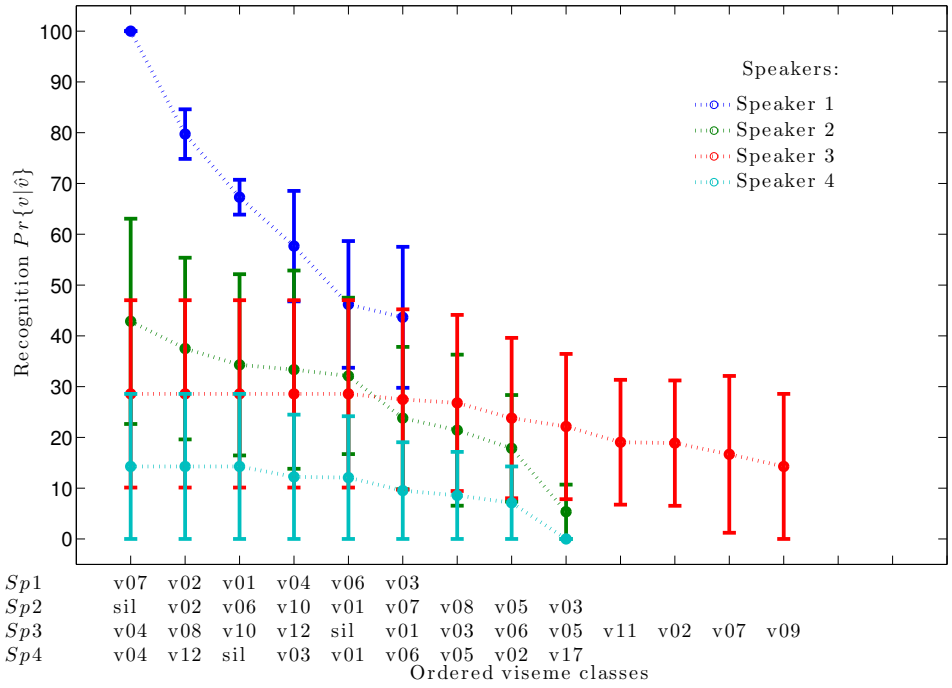



Figure 7.20: Individual viseme classification, $\Pr\{v|\hat{v}\}$ with the strictly confused, mixed vowels and consonant Bear visemes.

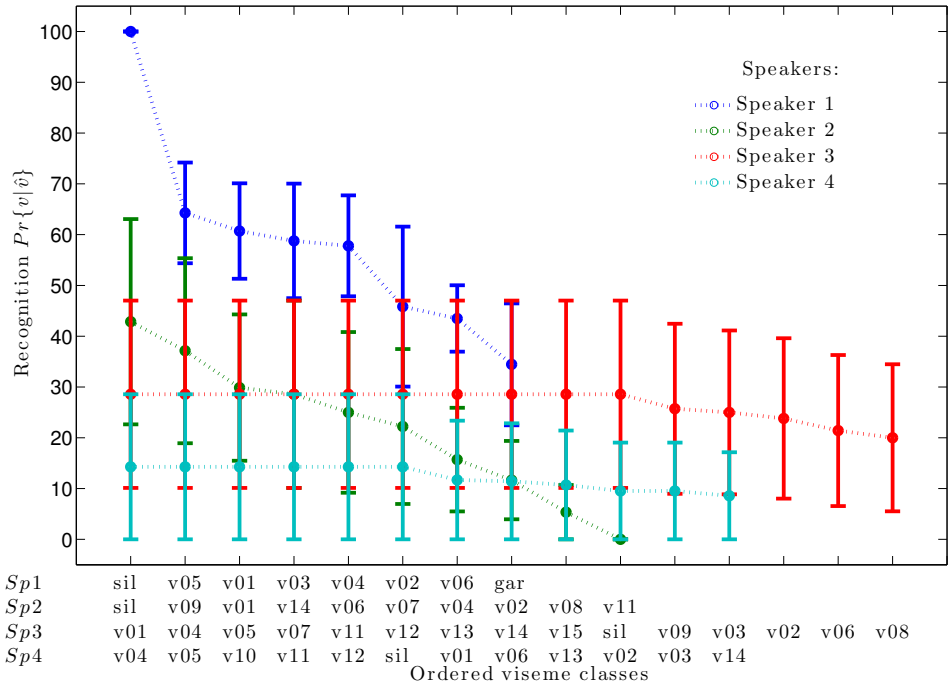



Figure 7.21: Individual viseme classification, $\Pr\{v|\hat{v}\}$ with the strictly confused, split vowels and consonant Bear visemes.

## 7.7 Improving lip-reading with speaker-dependent phoneme-to-viseme maps

This chapter has described a comprehensive study of previously suggested P2V maps and shown Lee's [82] is the best of the previously published P2V maps. The new data-driven approach respects speaker individuality in speech and uses this to demonstrate our second data-driven method tested, a strictly-confused viseme derivation with split vowel and consonant phonemes, can improve word classification. We call these speaker-dependent visemes 'Bear visemes' after the author's surname and show how these fit into the conventional lip-reading system in Figure 7.22, our new steps are highlighted with dash-edged boxes.

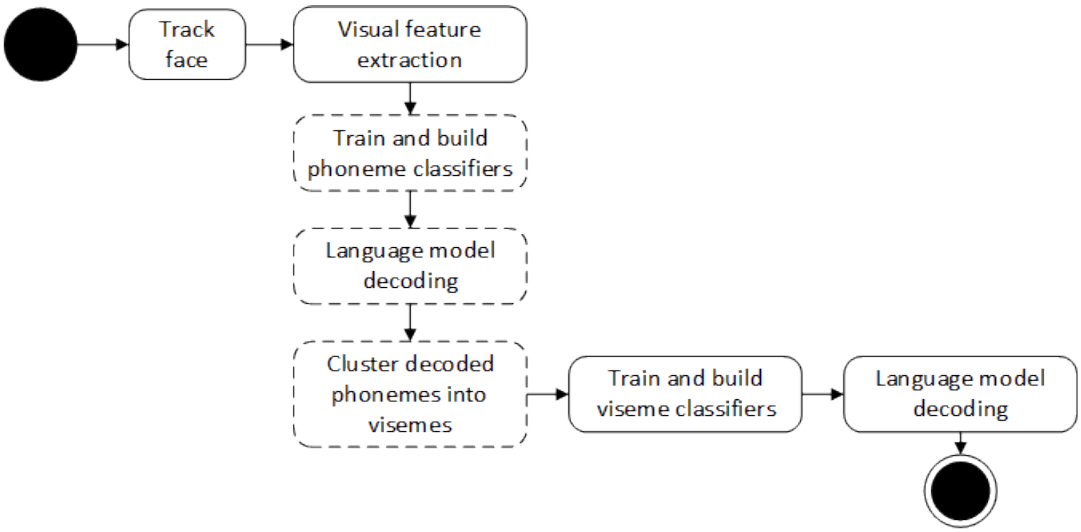



Figure 7.22: First augmentation to the conventional lip-reading system to include speaker-dependent visemes.

For phoneme confusion driven visemes, it is possible AVL2 contains insufficient samples to fairly identify confusion. So whilst improving performance, the classifiers still need more data for training. The reduction in word correctness by the data-driven confused mixed visemes is attributed to the mixing of vowels and consonants as this work shows when keeping these separate an improvement is possible.

The ratio of phonemes to visemes is useful, but secondary to confusions between phonemes, and does not help to discriminate phonemes within visemes for improved

word classification. To discriminate between words which are visually similar we still need to be able to reverse any P2V mapping.

This work highlights bad training samples are worse than less training samples and the boundary between good and bad samples is blurred. We have designed and implemented a new method of producing speaker-dependent visemes, in doing so showing speaker identity is important for good machine lip-reading classification. As speaker dependence is so prevalent in machine lip-reading systems, we need to cast our eyes towards how difficult a task speaker-independent classification is, and this is what our next chapter begins to investigate.

# Chapter 8

# Speaker-independence in phoneme-to-viseme maps

More than in audio speech, in machine lip-reading speaker identity is important for accurate classification [35]. We know a major difficulty in visual speech is the labelling of classifier units so we need to address the questions; to what extent such maps are independent of the speaker? And if so, how might speaker independent P2V maps be examined? Alongside of this, it would be useful to understand the interactions between the model training data and the classes. Therefore in this chapter we will use both the the AVL2 dataset [35] and the RMAV dataset to train and test classifiers based upon a series of P2V mappings.

## 8.1 Speaker independence

At the current time, good machine lip-reading performances are achieved with speaker dependent classification models, this means the test speaker must be included within the classifier training data. Speaker independent machine lip-reading is less successful [35]. Only a few large scale investigations have shown speaker independence to be viable. Neti *et al.* in [108] state that they created multi-speaker classifiers as contingency should speaker independent models fail to generalise well to

unseen speakers. After preliminary experiments these multi-speaker classifiers were considered not needed. However, this is achieved with state-of-the-art modelling, a permitted increase in word error rate and with a lot of speakers (IBM's via voice has 290 speakers [145]) (a currently unavailable dataset) which implies that with enough data and speakers that the speaker independence obstacle is surmountable by achieving generalisation on a large scale. In the majority of papers referenced in this thesis for example, speaker-dependent experiments are still used for greater results as speaker independence is rare and difficult to achieve.

On the continuous speech datasets, it is interesting to note that most still use speaker-dependent tests [76, 58, 147, 28]. We note that some are single speaker-dependent, others multi-speaker dependent, the crux of the point is that test speaker samples are included in the training data. In contrast only AVICAR [72] and IBM's LVCSR [108] achieve speaker-independent success. The former is a specific AV dataset for in car speech, and the latter is not available [28] so we suffice with the best datasets we have available to us.

Thus we understand speaker independence in visual speech to be the ability to classify a speaker who is not involved in the classifier training. This is a difficult, and as yet, unsolved problem. From this we are confident that, in visual speech, the identification of the person speaking is important. 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.

An example of a study into speaker independence in machine lip-reading is [35], here the authors use AVL2 and compare single speaker, multi-speaker and speaker independent classification using two types of classifiers (HMMs & Sieves [8]). However, this investigation uses word labels for classifiers and we are interested to know if the results could be improved using either phonemes or speaker-dependent visemes.

## 8.2 Method overview

We use the phoneme clustering approach described in Chapter 7 (or [16]) to produce a series of speaker-dependent P2V maps. This series of maps is made up of the following:

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

So we have nine phoneme-to-viseme 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). AVL2 P2V maps are constructed using separate training and test data over seven fold cross-validation [42]. RMAV maps from ten fold cross-validation. The variation in folds is due to the volume of data in each dataset.

With the HTK toolkit [150] HMM classifiers are built with the viseme classes in each P2V map. HMMs are flat-started with `HCompV`, re-estimated 11 times over (`HERest`) with forced alignment between seventh and eighth re-estimates. The final steps are classification using `HVite` and output of results with `HResults`. The models are three state HMMs each having an associated Gaussian mixture of five components to keep our results comparable to previous work.

To measure our performance of AVL2 speakers we note the classification network restricts the output to be one of the 26 letters of the alphabet. Therefore, our simplified measure of accuracy is; $\frac{\#letterscorrect}{\#lettersclassified}$.

For RMAV a bigram word network is built with `HBuild` and `HLStats`, and classification is measured as Correctness (Equation 2.7). The BEEP pronunciation dictionary used throughout these experiments is in British English [26] for all speakers.

## 8.3 Experiment design

The P2V maps formed in these experiments are designated as:

$$M_n(p, q) \tag{8.1}$$

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.

### 8.3.1 Baseline: Same Speaker-Dependent (SSD) maps

For our experiments we need a baseline for comparison. We select our same speaker-dependent P2V maps as based on previous literature [16], these provide the best results. The baseline tests involved are: $M_1(1, 1)$, $M_2(2, 2)$, $M_3(3, 3)$ and $M_4(4, 4)$ (for the four speakers in AVL2), additional tests for RMAV are: $M_5(5, 5)$, $M_6(6, 6)$, $M_7(7, 7)$ and $M_8(8, 8)$, $M_9(9, 9)$, $M_{10}(10, 10)$, $M_{11}(11, 11)$ and $M_{12}(12, 12)$. Remember, we now have AVL2 speakers 1 to 4, and RMAV speakers 1 to 12. Speakers 1 to 4 are not the same in AVL2 and RMAV. These tests are Same Speaker-Dependent (SSD) because the same speaker is used to create the map, to train the models and for the testing data. Tables 8.1 & 8.2 depict how these tests are constructed.

Table 8.1: Same Speaker-Dependent (SSD) experiments for AVLetters2 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 | $M_4(4, 4)$ |

The resulting AVL2 four speakers SSD P2V maps are listed in Table 8.3 and the

Table 8.2: Same Speaker-Dependent (SSD) experiments for RMAV 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 | $M_4(4,4)$ |
| Sp5 | Sp5 | Sp5 | $M_5(5,5)$ |
| Sp6 | Sp6 | Sp6 | $M_6(6,6)$ |
| Sp7 | Sp7 | Sp7 | $M_7(7,7)$ |
| Sp8 | Sp8 | Sp8 | $M_8(8,8)$ |
| Sp9 | Sp9 | Sp9 | $M_9(9,9)$ |
| Sp10 | Sp10 | Sp10 | $M_{10}(10,10)$ |
| Sp11 | Sp11 | Sp11 | $M_{11}(11,11)$ |
| Sp12 | Sp12 | Sp12 | $M_{12}(12,12)$ |

RMAV speakers SSD maps are in Appendix 10.2, Tables 3, 4, 5, 6, 7 & 8. We also permit a garbage, /*garb*/, viseme which is a cluster of phonemes in the ground truth which did not appear at all in the output from the phoneme classifier. Every viseme is listed with its associated mutually-confused phonemes e.g. for AVL2 Speaker 1, $M_1$, we see /$v$01/ is made up of phonemes {/ʌ/, /iy/, /əʊ/, /uw/}. We know from our clustering method in Chapter 7 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 from Chapter 7 with split vowel and consonant groupings as these achieved the most accurate classification.

### 8.3.2 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

Table 8.3: Speaker-dependent phoneme-to-viseme mapping derived from phoneme classification confusions for each speaker in AVLetters2.

| Speaker 1 $M_1$ | | Speaker 2 $M_2$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ʌ/ /iy/ /əʊ/ /uw/ | /v01/ | /ay/ /ey/ /iy/ /uw/ |
| /v02/ | /ə/ /eh/ /ey/ | /v02/ | /əʊ/ |
| /v03/ | /ɑ/ /ay/ | /v03/ | /ə/ |
| /v04/ | /d/ /s/ /t/ | /v04/ | /eh/ |
| /v05/ | /tʃ/ /l/ | /v05/ | /ʌ/ |
| /v06/ | /m/ /n/ | /v06/ | /ə/ |
| /v07/ | /dʒ/ /v/ | /v07/ | /dʒ/ /p/ /y/ |
| /v08/ | /b/ /y/ | /v08/ | /l/ /m/ /n/ |
| /v09/ | /k/ | /v09/ | /v/ /w/ |
| /v10/ | /z/ | /v10/ | /d/ /b/ |
| /v11/ | /w/ | /v11/ | /f/ /s/ |
| /v12/ | /f/ | /v12/ | /t/ |
| | | /v13/ | /k/ |
| | | /v14/ | /tʃ/ |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ /p/ | /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ /z/ |
| Speaker 3 $M_3$ | | Speaker 4 $M_4$ | |
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ey/ /iy/ | /v01/ | /ʌ/ /ay/ /ey/ /iy/ |
| /v02/ | /ə/ /eh/ | /v02/ | /ə/ /eh/ |
| /v03/ | /ay/ | /v03/ | /ə/ |
| /v04/ | /ʌ/ | /v04/ | /əʊ/ |
| /v05/ | /ə/ | /v05/ | /uw/ |
| /v06/ | /əʊ/ | /v06/ | /m/ /n/ |
| /v07/ | /uw/ | /v07/ | /k/ /l/ |
| /v08/ | /d/ /p/ /t/ | /v08/ | /dʒ/ /t/ |
| /v09/ | /l/ /m/ | /v09/ | /d/ /s/ |
| /v10/ | /k/ /w/ | /v10/ | /w/ |
| /v11/ | /f/ /n/ | /v11/ | /f/ |
| /v12/ | /b/ /s/ | /v12/ | /v/ |
| /v13/ | /v/ | /v13/ | /tʃ/ |
| /v14/ | /dʒ/ | /v14/ | /b/ |
| /v15/ | /tʃ/ | /v15/ | /y/ |
| /v16/ | /y/ | | |
| /v17/ | /z/ | | |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ | /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ /p/ /z/ |

Table 8.4: Different Speaker-Dependent maps and Data (DSD&D) experiments with the four AVLetters2 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 | $M_3(3,4)$ |

construct $M_2(2,1)$, $M_3(3,1)$ and $M_4(4,1)$ and so on for the other speakers, this is depicted in Table 8.4.

Table 8.5: Different Speaker-Dependent maps and Data (DSD&D) experiments for one of the 12 RMAV speakers (speaker one).

| 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 | $M_4(4,1)$ |
| Sp5 | Sp5 | Sp1 | $M_5(4,1)$ |
| Sp6 | Sp6 | Sp1 | $M_6(4,1)$ |
| Sp7 | Sp7 | Sp1 | $M_7(4,1)$ |
| Sp8 | Sp8 | Sp1 | $M_8(4,1)$ |
| Sp9 | Sp9 | Sp1 | $M_9(4,1)$ |
| Sp10 | Sp10 | Sp1 | $M_{10}(10,1)$ |
| Sp11 | Sp11 | Sp1 | $M_{11}(11,1)$ |
| Sp12 | Sp12 | Sp1 | $M_{12}(12,1)$ |

For the RMAV speakers, we undertake this for all 12 speakers using the maps of the 11 others. We show the tests for a single speaker (Speaker 1) in Table 8.5 as an example. The other speakers are in Appendix 10.2.

### 8.3.3 Different Speaker-Dependent maps (DSD)

Now we wish to isolate the effects of the HMM classifier from the effect of using different viseme P2V 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 8.6).

Table 8.6: Different Speaker-Dependent maps (DSD) experiments for AVLetters2 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 | $M_3(4,4)$ |

These are the same P2V maps as in Table 8.3 but trained and tested differently. In Table 8.7 we show the equivalent DSD tests for Speaker 1 of RMAV as an example. For the tests conducted on all speakers, speakers 2 to 12 are listed in Appendix 10.2.

### 8.3.4 Multi-Speaker maps (MS)

A multi-speaker (MS) P2V map forms the viseme classifier labels in our third set of experiments. This map is constructed using phoneme confusions produced by *all* speakers in each data set and is shown in Table 8.8, for the four AVL2 speakers, and Table 8.9 for the 12 RMAV speakers.

For our 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

Table 8.7: Different Speaker-Dependent maps (DSD) for one of the 12 RMAV speakers (Speaker one).

| 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 | $M_4(1,1)$ |
| Sp5 | Sp1 | Sp1 | $M_5(1,1)$ |
| Sp6 | Sp1 | Sp1 | $M_6(1,1)$ |
| Sp7 | Sp1 | Sp1 | $M_7(1,1)$ |
| Sp8 | Sp1 | Sp1 | $M_8(1,1)$ |
| Sp9 | Sp1 | Sp1 | $M_9(1,1)$ |
| Sp10 | Sp1 | Sp1 | $M_{10}(1,1)$ |
| Sp11 | Sp1 | Sp1 | $M_{11}(1,1)$ |
| Sp12 | Sp1 | Sp1 | $M_{12}(1,1)$ |

Table 8.8: Multi-Speaker (MS) phoneme-to-viseme mapping for AVLetters2 speakers.

| Viseme | Phonemes |
|---|---|
| /v01/ | /ʌ/ /ay/ /ey/ /iy/ /əʊ/ /uw/ |
| /v02/ | /ə/ /eh/ |
| /v03/ | /ɑ/ |
| /v04/ | /d/ /s/ /t/ /v/ |
| /v05/ | /f/ /l/ /n/ |
| /v06/ | /b/ /w/ /y/ |
| /v07/ | /dʒ/ |
| /v08/ | /z/ |
| /v09/ | /p/ |
| /v10/ | /m/ |
| /v11/ | /k/ |
| /v12/ | /tʃ/ |
| /sil/ | /sil/ |
| /gar/ | /ɛ/ /ɒ/ /ɔ/ /r/ |

as follows: $M_{[all]}(1,1)$, $M_{[all]}(2,2)$, $M_{[all]}(3,3)$ and $M_{[all]}(4,4)$: this is explained in Table 8.10 and the RMAV MS map is tested as: $M_{[all]}(1,1)$, $M_{[all]}(2,2)$, $M_{[all]}(3,3)$, $M_{[all]}(4,4)$, $M_{all]}(5,5)$, $M_{[all]}(6,6)$, $M_{[all]}(7,7)$, $M_{[all]}(8,8)$, $M_{[all]}(9,9)$, $M_{[all]}(10,10)$, $M_{[all]}(11,11)$, $M_{[all]}(12,12)$, as shown in Table 8.11.

Table 8.9: Multi-Speaker (MS) phoneme-to-viseme mapping for RMAV speakers.

| Viseme | Phonemes |
|---|---|
| /v01/ | /ɑ/ /æ/ /ʌ/ /ɔ/ /ə/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ |
| /v02/ | /ɔə/ /ʊ/ /ɔə/ |
| /v03/ | /ɑʊ/ |
| /v04/ | /ɔɪ/ |
| /v05/ | /ə/ |
| /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /ɦ/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /sil/ | /sil/ |
| /sp/ | /sp/ |
| /gar/ | /ʒ/ /c/ |

Table 8.10: Multi-Speaker (MS) experiments for AVLetters2 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 8.11: Multi-Speaker (MS) experiments for RMAV 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 | $M_{all}(4,4)$ |
| Sp[all] | Sp5 | Sp5 | $M_{all}(5,5)$ |
| Sp[all] | Sp6 | Sp6 | $M_{all}(6,6)$ |
| Sp[all] | Sp7 | Sp7 | $M_{all}(7,7)$ |
| Sp[all] | Sp8 | Sp8 | $M_{all}(8,8)$ |
| Sp[all] | Sp9 | Sp9 | $M_{all}(9,9)$ |
| Sp[all] | Sp10 | Sp10 | $M_{all}(10,10)$ |
| Sp[all] | Sp11 | Sp11 | $M_{all}(11,11)$ |
| Sp[all] | Sp12 | Sp12 | $M_{all}(12,12)$ |

### 8.3.5 Speaker-Independent maps (SI)

Finally, our last set of tests looks at speaker independence in P2V maps themselves. Here we use maps which are derived using all speakers confusions bar the test speaker. This time we substitute the symbol '$!x$' in place of a list of speaker identifying numbers, meaning 'not including speaker $x$'. The tests for these maps are as follows $M_{!1}(1,1)$, $M_{!2}(2,2)$, $M_{!3}(3,3)$ and $M_{!4}(4,4)$ as shown in Tables 8.13 & 8.14 for AVL2 and RMAV speakers respectively. Speaker independent P2V maps for AVL2 speakers are shown in Table 8.12. SI maps for RMAV speakers are in Appendix 10.2, Tables 9 to 14.

## 8.4 The homophone risk factor

P2V maps are a many-to-one mapping. This creates the possibility of creating visual homophones when translating a phonetic transcript into a viseme transcript. For example, in the AVL2 data (isolated words are the letters of the alphabet) the phonetic realisation 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 tokens ($T$) for each map after each word has been translated to visemes are listed in Table 8.15. More homophones means a greater the chance of substitution errors and a reduced correct classification.

## 8.5 Measuring similarity between phoneme-to-viseme maps

In Table 8.16 and 8.17 we present a similarity score for comparing each pair of phoneme-to-viseme maps for AVL2 speakers and RMAV speakers respectively.

Table 8.12: Phoneme-to-viseme mapping derived from phoneme classification confusions of the three other speakers in AVLetters2.

| Speaker 1 $M_{234}$ | | Speaker 2 $M_{134}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ʌ/ /ə/ /ay/ /ey/ /iy/ | /v01/ | /ʌ/ /ay/ /ey/ /iy/ |
| /v02/ | /əʊ/ /uw/ | /v02/ | /ɑ/ /əʊ/ /uw/ |
| /v03/ | /eh/ | /v03/ | /ə/ /eh/ |
| /v04/ | /ɑ/ | /v04/ | /d/ /s/ /t/ |
| /v05/ | /d/ /s/ /t/ /v/ | /v05/ | /tʃ/ /l/ |
| /v06/ | /l/ /m/ /n/ | /v06/ | /b/ /dʒ/ |
| /v07/ | /dʒ/ /p/ /y/ | /v07/ | /v/ /y/ |
| /v08/ | /k/ /w/ | /v08/ | /k/ /w/ |
| /v09/ | /f/ | /v09/ | /p/ |
| /v10/ | /tʃ/ | /v10/ | /z/ |
| /v11/ | /b/ | /v11/ | /m/ |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ /z/ | /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ /f/ /n/ |

| Speaker 3 $M_{124}$ | | Speaker 4 $M_{123}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ʌ/ /ay/ /ey/ /iy/ /əʊ/ /uw/ | /v01/ | /ʌ/ /ay/ /ey/ /iy/ /əʊ/ /uw/ |
| /v02/ | /ɑ/ | /v02/ | /ɑ/ |
| /v03/ | /ə/ /eh/ | /v03/ | /ə/ /eh/ |
| /v04/ | /d/ /s/ /t/ /v/ | /v04/ | /dʒ/ /s/ /t/ /v/ |
| /v05/ | /l/ /m/ /n/ | /v05/ | /f/ /l/ /n/ |
| /v06/ | /b/ /w/ /y/ | /v06/ | /b/ /d/ /p/ |
| /v07/ | /dʒ/ | /v07/ | /w/ /y/ |
| /v08/ | /z/ | /v08/ | /z/ |
| /v09/ | /p/ | /v09/ | /m/ |
| /v10/ | /k/ | /v10/ | /k/ |
| /v11/ | /f/ | /v11/ | /tʃ/ |
| /v12/ | /tʃ/ | | |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /garb/ | /ɛ/ /ɒ/ /ɔ/ /r/ /iy/ | /garb/ | ea/ /ɒ/ /ɔ/ /r/ |

Table 8.13: Speaker-Independent (SI) experiments with AVLetters2 speakers.

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

Table 8.14: Speaker-Independent (SI) experiments with RMAV speakers.

| Speaker-Independent (SI) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp[!1] | Sp1 | Sp1 | $M_{!1}(1,1)$ |
| Sp[!2] | Sp2 | Sp2 | $M_{!2}(2,2)$ |
| Sp[!3] | Sp3 | Sp3 | $M_{!3}(3,3)$ |
| Sp[!4] | Sp4 | Sp4 | $M_{!4}(4,4)$ |
| Sp[!5] | Sp5 | Sp5 | $M_{!5}(5,5)$ |
| Sp[!6] | Sp6 | Sp6 | $M_{!6}(6,6)$ |
| Sp[!7] | Sp7 | Sp7 | $M_{!7}(7,7)$ |
| Sp[!8] | Sp8 | Sp8 | $M_{!8}(8,8)$ |
| Sp[!9] | Sp9 | Sp9 | $M_{!9}(9,9)$ |
| Sp[!10] | Sp10 | Sp10 | $M_{[10}(10,10)$ |
| Sp[!11] | Sp11 | Sp11 | $M_{[11}(11,11)$ |
| Sp[!12] | Sp12 | Sp12 | $M_{!12}(12,12)$ |

Table 8.15: Count of visual homophones by each phoneme-to-viseme map, allowing for variation in pronunciation in AVLetters2 speakers.

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

Table 8.16: Similarity scores between all AVLetters2 phoneme-to-viseme maps.

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_{[all]}$ | $M_{!1}$ | $M_{!2}$ | $M_{!3}$ | $M_{!4}$ |
|---|---|---|---|---|---|---|---|---|---|
| $M_1$ | 0.000 | 0.327 | 0.322 | 0.247 | 0.199 | 0.244 | 0.048 | 0.112 | 0.222 |
| $M_2$ | 0.327 | 0.000 | 0.410 | 0.303 | 0.333 | 0.266 | 0.256 | 0.254 | 0.253 |
| $M_3$ | 0.322 | 0.410 | 0.000 | 0.157 | 0.465 | 0.400 | 0.394 | 0.398 | 0.396 |
| $M_4$ | 0.247 | 0.303 | 0.157 | 0.000 | 0.301 | 0.298 | 0.172 | 0.246 | 0.378 |
| $M_{[all]}$ | 0.199 | 0.333 | 0.465 | 0.301 | 0.000 | 0.311 | 0.220 | 0.098 | 0.136 |
| $M_{!1}$ | 0.244 | 0.266 | 0.400 | 0.298 | 0.311 | 0.000 | 0.086 | 0.160 | 0.218 |
| $M_{!2}$ | 0.048 | 0.256 | 0.394 | 0.172 | 0.220 | 0.086 | 0.000 | 0.155 | 0.160 |
| $M_{!3}$ | 0.112 | 0.254 | 0.398 | 0.246 | 0.098 | 0.160 | 0.155 | 0.000 | 0.222 |
| $M_{!4}$ | 0.222 | 0.253 | 0.396 | 0.378 | 0.136 | 0.218 | 0.160 | 0.222 | 0.000 |

Table 8.17: Similarity scores between all RMAV phoneme-to-viseme maps.

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ | $M_7$ | $M_8$ | $M_9$ | $M_{10}$ | $M_{11}$ | $M_{12}$ | $M_{ad}$ | $M_{11}$ | $M_{12}$ | $M_{13}$ | $M_{14}$ | $M_5$ | $M_6$ | $M_{17}$ | $M_8$ | $M_9$ | $M_{10}$ | $M_{111}$ | $M_{112}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $M_1$ | 0.000 | 0.174 | 0.289 | 0.099 | 0.009 | 0.313 | 0.201 | 0.237 | 0.095 | 0.256 | 0.006 | 0.258 | 0.653 | 0.564 | 0.564 | 0.660 | 0.564 | 0.571 | 0.564 | 0.658 | 0.564 | 0.660 | 0.564 | 0.471 | 0.564 |
| $M_2$ | 0.174 | 0.000 | 0.196 | 0.031 | 0.159 | 0.235 | 0.194 | 0.175 | 0.194 | 0.192 | 0.026 | 0.304 | 0.544 | 0.449 | 0.449 | 0.550 | 0.449 | 0.456 | 0.449 | 0.550 | 0.449 | 0.550 | 0.449 | 0.347 | 0.449 |
| $M_3$ | 0.289 | 0.196 | 0.000 | 0.128 | 0.044 | 0.357 | 0.228 | 0.333 | 0.271 | 0.188 | 0.024 | 0.260 | 0.633 | 0.543 | 0.543 | 0.645 | 0.543 | 0.566 | 0.543 | 0.655 | 0.543 | 0.645 | 0.543 | 0.572 | 0.543 |
| $M_4$ | 0.099 | 0.031 | 0.128 | 0.000 | 0.072 | 0.133 | 0.126 | 0.264 | 0.151 | 0.185 | 0.172 | 0.108 | 0.598 | 0.505 | 0.505 | 0.605 | 0.505 | 0.512 | 0.505 | 0.605 | 0.505 | 0.605 | 0.505 | 0.417 | 0.505 |
| $M_5$ | 0.009 | 0.159 | 0.044 | 0.072 | 0.000 | 0.256 | 0.046 | 0.213 | 0.130 | 0.237 | 0.086 | 0.301 | 0.738 | 0.649 | 0.649 | 0.746 | 0.649 | 0.658 | 0.649 | 0.746 | 0.649 | 0.746 | 0.649 | 0.556 | 0.649 |
| $M_6$ | 0.313 | 0.235 | 0.357 | 0.133 | 0.256 | 0.000 | 0.213 | 0.279 | 0.340 | 0.357 | 0.291 | 0.220 | 0.715 | 0.621 | 0.621 | 0.751 | 0.621 | 0.679 | 0.621 | 0.714 | 0.621 | 0.751 | 0.621 | 0.572 | 0.621 |
| $M_7$ | 0.201 | 0.194 | 0.228 | 0.126 | 0.046 | 0.213 | 0.067 | 0.334 | 0.263 | 0.327 | 0.379 | 0.324 | 0.631 | 0.544 | 0.544 | 0.638 | 0.544 | 0.550 | 0.544 | 0.637 | 0.544 | 0.638 | 0.544 | 0.466 | 0.544 |
| $M_8$ | 0.237 | 0.175 | 0.333 | 0.264 | 0.213 | 0.279 | 0.334 | 0.000 | 0.229 | 0.340 | 0.230 | 0.310 | 0.667 | 0.569 | 0.569 | 0.635 | 0.569 | 0.575 | 0.569 | 0.671 | 0.569 | 0.635 | 0.569 | 0.475 | 0.569 |
| $M_9$ | 0.095 | 0.194 | 0.271 | 0.151 | 0.130 | 0.340 | 0.263 | 0.229 | 0.000 | 0.276 | 0.209 | 0.324 | 0.741 | 0.656 | 0.656 | 0.748 | 0.656 | 0.672 | 0.656 | 0.790 | 0.656 | 0.748 | 0.656 | 0.570 | 0.656 |
| $M_{10}$ | 0.256 | 0.192 | 0.188 | 0.185 | 0.237 | 0.357 | 0.327 | 0.340 | 0.276 | 0.000 | 0.181 | 0.272 | 0.694 | 0.598 | 0.598 | 0.626 | 0.598 | 0.544 | 0.598 | 0.636 | 0.598 | 0.626 | 0.598 | 0.463 | 0.598 |
| $M_{11}$ | 0.006 | 0.026 | 0.024 | 0.172 | 0.086 | 0.291 | 0.379 | 0.230 | 0.209 | 0.181 | 0.000 | 0.167 | 0.610 | 0.517 | 0.517 | 0.618 | 0.517 | 0.548 | 0.517 | 0.642 | 0.517 | 0.618 | 0.517 | 0.515 | 0.517 |
| $M_{12}$ | 0.258 | 0.304 | 0.260 | 0.108 | 0.301 | 0.220 | 0.324 | 0.310 | 0.324 | 0.272 | 0.167 | 0.008 | 0.664 | 0.561 | 0.561 | 0.589 | 0.561 | 0.528 | 0.561 | 0.561 | 0.561 | 0.589 | 0.561 | 0.496 | 0.561 |
| $M_{ad}$ | 0.653 | 0.544 | 0.633 | 0.598 | 0.738 | 0.715 | 0.631 | 0.667 | 0.741 | 0.694 | 0.610 | 0.664 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.035 | 0.000 | 0.076 | 0.000 | 0.005 | 0.000 | 0.070 | 0.000 |
| $M_{11}$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |
| $M_{12}$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |
| $M_{13}$ | 0.660 | 0.550 | 0.645 | 0.605 | 0.746 | 0.751 | 0.638 | 0.635 | 0.748 | 0.626 | 0.618 | 0.589 | 0.005 | 0.057 | 0.057 | 0.000 | 0.057 | 0.070 | 0.057 | 0.030 | 0.057 | 0.000 | 0.057 | 0.018 | 0.057 |
| $M_{14}$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |
| $M_5$ | 0.571 | 0.456 | 0.566 | 0.512 | 0.658 | 0.672 | 0.550 | 0.575 | 0.672 | 0.544 | 0.548 | 0.528 | 0.035 | 0.033 | 0.033 | 0.070 | 0.033 | 0.000 | 0.033 | 0.099 | 0.033 | 0.070 | 0.033 | 0.100 | 0.033 |
| $M_6$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |
| $M_{17}$ | 0.658 | 0.550 | 0.655 | 0.605 | 0.746 | 0.714 | 0.637 | 0.671 | 0.790 | 0.636 | 0.642 | 0.496 | 0.070 | 0.067 | 0.067 | 0.018 | 0.067 | 0.100 | 0.067 | 0.018 | 0.067 | 0.018 | 0.067 | 0.000 | 0.067 |
| $M_8$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |
| $M_9$ | 0.660 | 0.550 | 0.645 | 0.605 | 0.746 | 0.649 | 0.621 | 0.714 | 0.751 | 0.621 | 0.618 | 0.589 | 0.005 | 0.057 | 0.057 | 0.000 | 0.057 | 0.070 | 0.057 | 0.030 | 0.057 | 0.000 | 0.057 | 0.018 | 0.057 |
| $M_{10}$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |
| $M_{111}$ | 0.471 | 0.347 | 0.572 | 0.417 | 0.556 | 0.572 | 0.466 | 0.475 | 0.570 | 0.463 | 0.515 | 0.496 | 0.070 | 0.067 | 0.067 | 0.018 | 0.067 | 0.100 | 0.067 | 0.018 | 0.067 | 0.018 | 0.067 | 0.000 | 0.067 |
| $M_{112}$ | 0.564 | 0.449 | 0.543 | 0.505 | 0.649 | 0.621 | 0.544 | 0.569 | 0.656 | 0.598 | 0.517 | 0.561 | 0.000 | 0.000 | 0.000 | 0.057 | 0.000 | 0.033 | 0.000 | 0.020 | 0.000 | 0.057 | 0.000 | 0.067 | 0.000 |

The score addresses the phonemes within each viseme, the total number of phonemes clustered, the number of visemes within each set and ignores the ordering of the visemes within the set. As an example to explain our similarity algorithm, imagine we have the two phoneme-to-viseme maps shown in Figure 8.1.

| Map | Viseme | Phonemes |
|---|---|---|
| Map 1 | /v01/ | {/p1/ /p2/ /p3/} |
| | /v02/ | {/p4/ /p5/} |
| | /v03/ | {/p6/} |
| | /v04/ | {/p7/ /p8/} |
| Map 2 | /v01/ | {/p1/ /p3/} |
| | /v02/ | {/p2/ /p4/} |
| | /v03/ | {/p5/} |
| | /v04/ | {/p6/} |
| | /v05/ | {/p7/ /p8/ /p9/} |

Figure 8.1: Similarity algorithm: example phoneme-to-viseme maps.

Our first step is to attribute a weight to each phoneme within each viseme. This is; $\frac{1}{\#phonemes}$ and is shown in Figure 8.2.

| | Viseme | Phonemes | Phoneme weight |
|---|---|---|---|
| Map 1 | /v01/ | {/p1/ /p2/ /p3/} | 0.3r |
| | /v02/ | {/p4/ /p5/} | 0.5 |
| | /v03/ | {/p6/} | 1.0 |
| | /v04/ | {/p7/ /p8/} | 0.5 |
| Map 2 | /v01/ | {/p1/ /p3/} | 0.5 |
| | /v02/ | {/p2/ /p4/} | 0.5 |
| | /v03/ | {/p5/} | 1.0 |
| | /v04/ | {/p6/} | 1.0 |
| | /v05/ | {/p7/ /p8/ /p9/} | 0.3r |

Figure 8.2: Phoneme-to-viseme map similarity algorithm step 1: Example phoneme-to-viseme maps with weighted phonemes.

Now we use these values to compare all visemes of one map with another.

| | | Map 2 | | | | |
|---|---|---|---|---|---|---|
| | | /v1/ | /v2/ | /v3/ | /v4/ | /v5/ |
| Map 1 | /v1/ | /p1/, /p3/ | /p2/ | - | - | - |
| | /v2/ | - | /p4/ | /p5/ | - | - |
| | /v3/ | - | - | - | /p6/ | - |
| | /v4/ | - | - | - | - | /p7/, /p8/ |

Figure 8.3: Phoneme-to-viseme map similarity algorithm step 2: phoneme in viseme matches.

Where two visemes $V_i$ and $V_j$ contain the same phonemes (see Figure 8.3), the $V_{ij}$ score is the sum of the matched phoneme weights (Figure 8.4). The final values are in Figure 8.5.

| | | Map 2 | | | | |
|---|---|---|---|---|---|---|
| | | /v1/ | /v2/ | /v3/ | /v4/ | /v5/ |
| Map 1 | /v1/ | $/p1/ = 0.3r + 0.5$ | $/p2/ = 0.3r + 0.3r$ | 0 | 0 | 0 |
| | | $/p3/ = 0.3r + 0.5$ | - | - | - | - |
| | /v2/ | 0 | $/p4/ = 0.5 + 0.5$ | $/p5/ = 0.5 + 1.0$ | 0 | 0 |
| | /v3/ | 0 | 0 | 0 | $/p6/ = 1.0 + 1.0$ | 0 |
| | /v4/ | 0 | 0 | 0 | 0 | $/p7/ = 0.5 + 0.3r$ |
| | | - | - | - | 0 | $/p8/ = 0.5 + 0.3r$ |

Figure 8.4: Phoneme-to-viseme map similarity algorithm step 3: summing the phoneme weights.

| | | Map 2 | | | | |
|---|---|---|---|---|---|---|
| | | /v1/ | /v2/ | /v3/ | /v4/ | /v5/ |
| Map 1 | /v1/ | $1.6r$ | $0.6r$ | 0 | 0 | 0 |
| | /v2/ | 0 | 1.0 | 1.5 | 0 | 0 |
| | /v3/ | 0 | 0 | 0 | 2.0 | 0 |
| | /v4/ | 0 | 0 | 0 | 0 | $1.6r$ |

Figure 8.5: Phoneme-to-viseme map similarity algorithm step 4: total phoneme weights.

Finally we need to sum of all values in the upper triangle $U\forall_{ij} i > j$, minus the sum of all values in the lower triangle $L\forall_{ij} i < j$, normalised by dividing by the total number of matched phonemes, $N_p$, (in our example, eight) to give the value $0.73'$ (8.2). $S$ is the similarity score.

$$S = U - L \tag{8.2}$$

This similarity measure is calculated to compare all the P2V maps used in our experiments in pairs and the results are shown in Tables 8.16 and 8.17. The values closest to zero show the most similar maps, thus the closer to 1, the more different the maps are. We have not compared the maps between datasets due to biased effects caused by the disparity between word content and data size. Unsurprisingly, with the RMAV dataset, the MS and SI P2V maps are all very similar because of the volume of speakers and folds of phoneme classification, there is more chance of unique phonemes being confused. There is at most 3 phonemes different between them all.

If we compare all the P2V maps in Tables 8.8 & 8.12, there are similarities. Mostly because there is only one speaker at a time removed from within SI P2V maps. However, if these are compared to the speaker-dependent maps in Table 8.3, a different picture can be seen. Speaker 4 is significantly affected by the introduction of /əʊ/ and /uw/ into viseme /$v$01/. Where Speaker 1 has these in $M_1(1,1)$, his SD word classification of 15.9% is less than half of Speaker 4's 38.4% (Figure 8.11).

## 8.6 Analysis of speaker independence in phoneme-to-viseme maps

Figure 8.6 shows the word correctness of AVL2 speaker-dependent viseme classes on the $y$-axis. 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, $M_1(1,2)$, $M_3(3,2)$,

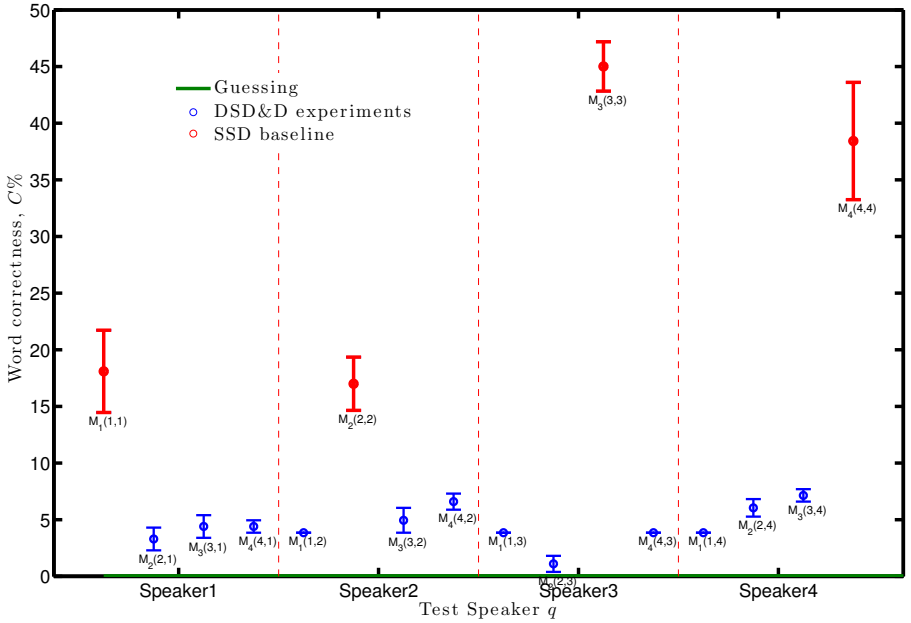



Figure 8.6: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{7}}$, of the DSD&D tests where HMM classifiers are tested on all three other speakers in AVLetters2. Baseline is the SSD maps.

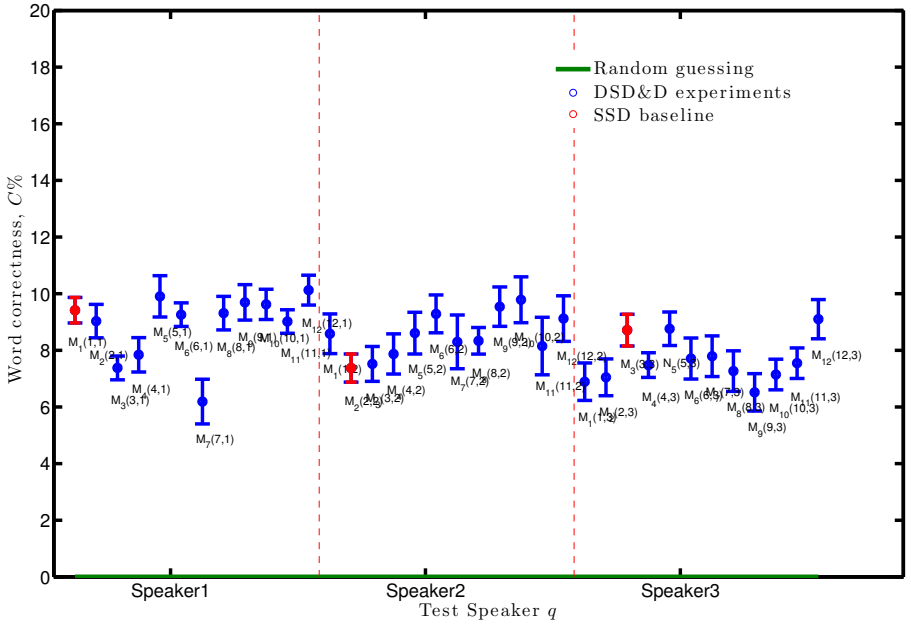



Figure 8.7: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD&D tests where HMM classifiers are tested on all eleven other speakers in RMAV. Baseline is SSD maps (red) - Speakers 1-3.

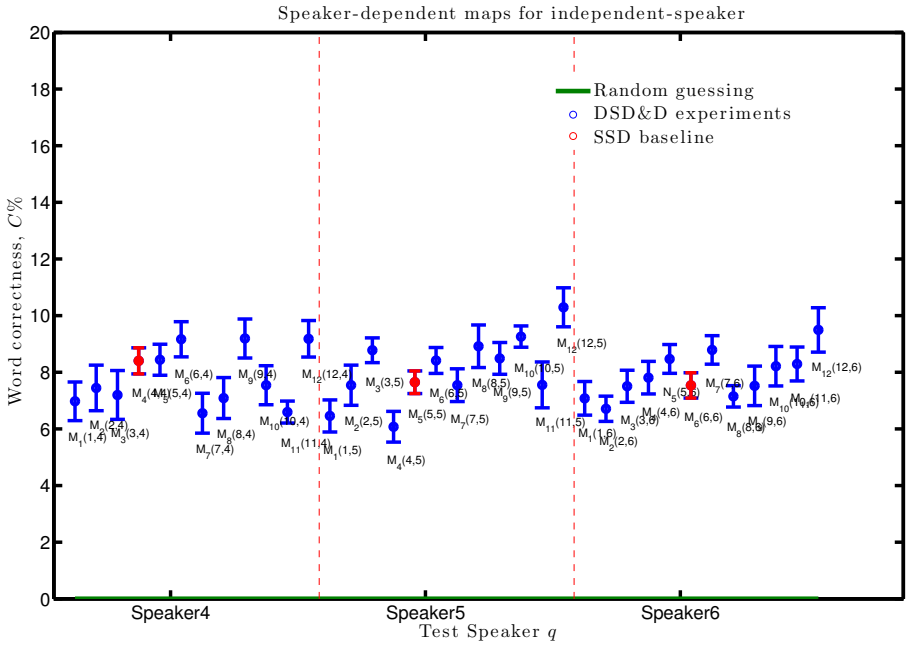



Figure 8.8: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD&D tests where HMM classifiers are tested on all eleven other speakers in RMAV. Baseline is SSD maps (red) - Speakers 4-6.

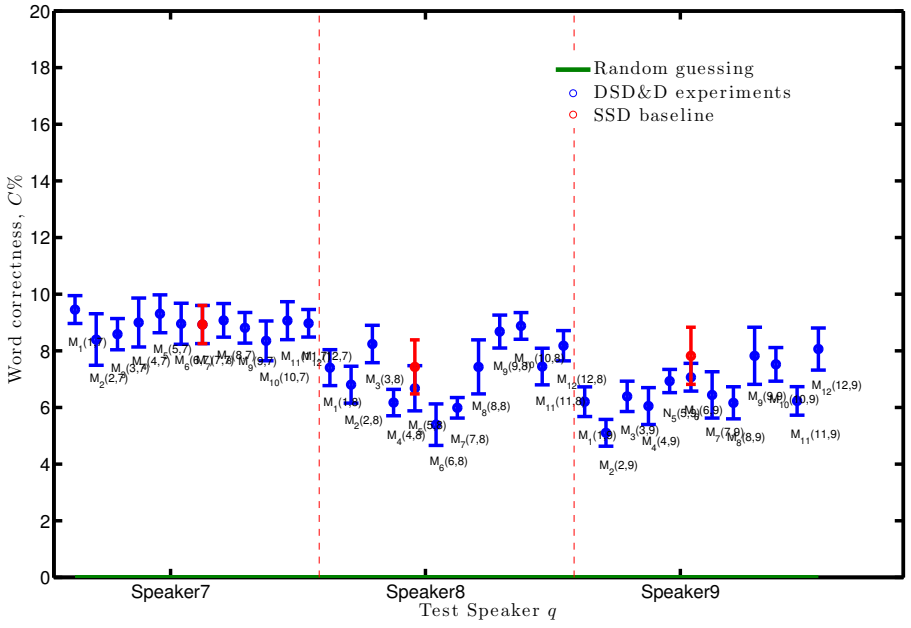



Figure 8.9: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD&D tests where HMM classifiers are tested on all eleven other speakers in RMAV. Baseline is SSD maps (red) - Speakers 7-9.

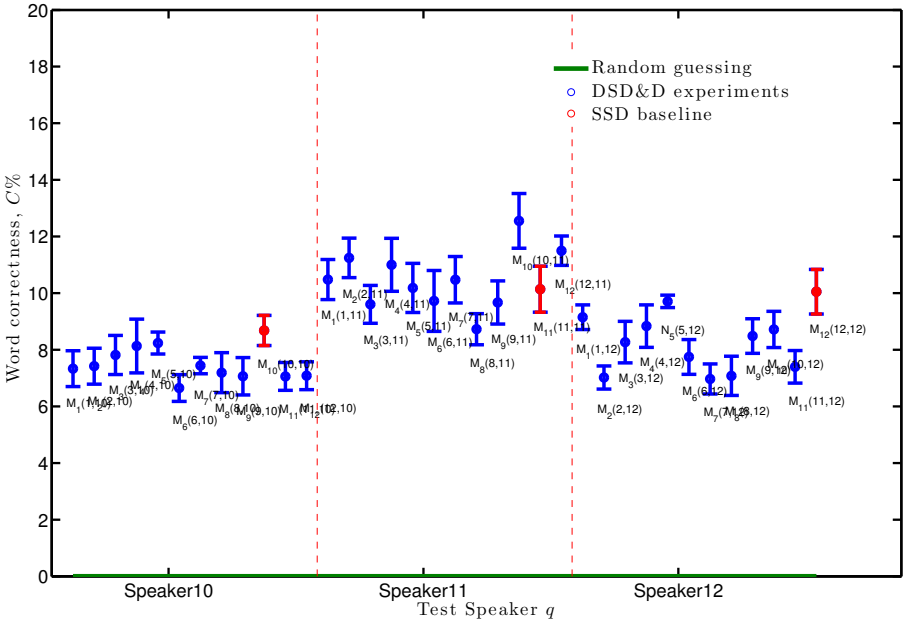



Figure 8.10: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD&D tests where HMM classifiers are tested on all eleven other speakers in RMAV. Baseline is SSD maps (red) - Speakers 10-12.

$M_4(4,2)$ 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 8.4. We also plot guessing (calculated as $1/N$, where $N$ is the total number of words in the dataset. For AVL2 this is 26, for RMAV speaker this ranges between 1362 and 1802). DSD 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 HMM's in [35]. This gap is attributed to two possible effects, either - the visual units are incorrect, or they are trained on the incorrect speaker.

Figures 8.7, 8.8, 8.9, & 8.10 show the same tests but on the continuous speech data. It is reassuring to see some speakers significantly deteriorate the classification rates when the speaker used to train the classifier is not the same as the test speaker. As an example we look at Speaker 1 on the leftmost side of Figure 8.7. Here the test speaker is Speaker 1. The speaker-dependent maps for all 12 speakers have been used to build HMMs classifiers. But when tested on Speaker 1, only maps and models for speakers 3, 7 and 12 show a significant reduction in word correctness.

All eight other speakers are within one standard error.

Figure 8.8, for the RMAV speakers four to six, we see a similar trend with Speaker 4 showing the most variation of these three speakers. To lip-read Speaker 4 we actually see a significant improvement by using the map and model of Speaker 6 and less significant improvements by speakers 3, 5 and 11. In Figure 8.9 we see Speaker 11's SD map and models majorly improve the classification of Speaker 8. However, whilst these are all signs of possibly making strides towards speaker independent classification, Speaker 12 in Figure 8.10 shows the most common trend is there is a lot of overlap between our continuous speech speakers and this natural variation is attributed to the speaker identity.

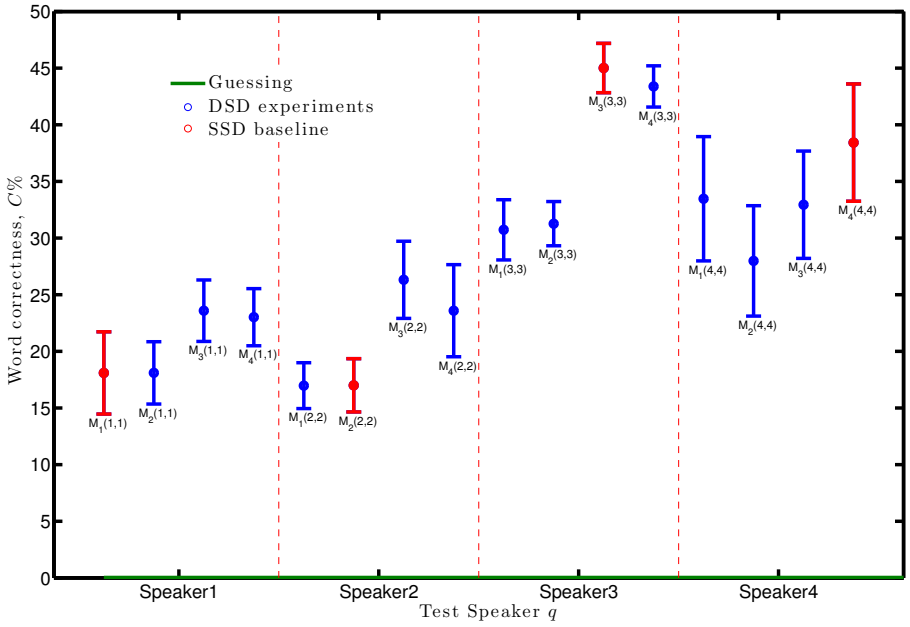



Figure 8.11: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{7}}$, of the DSD tests where HMM classifiers are constructed with single-speaker dependent phoneme-to-viseme maps for all four speakers in AVLetters2. Baseline is the SSD maps.

Figure 8.11 shows our AVL2 DSD experiments from Table 8.6. Our results in word correctness, $C$, are plotted on the $y$-axis and we also plot the same benchmark as in Figure 8.6 ($n = p = q$). In our DSD tests, the HMM 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 8.6 was not due to the choice of visemes but rather, the badly trained HMMs.

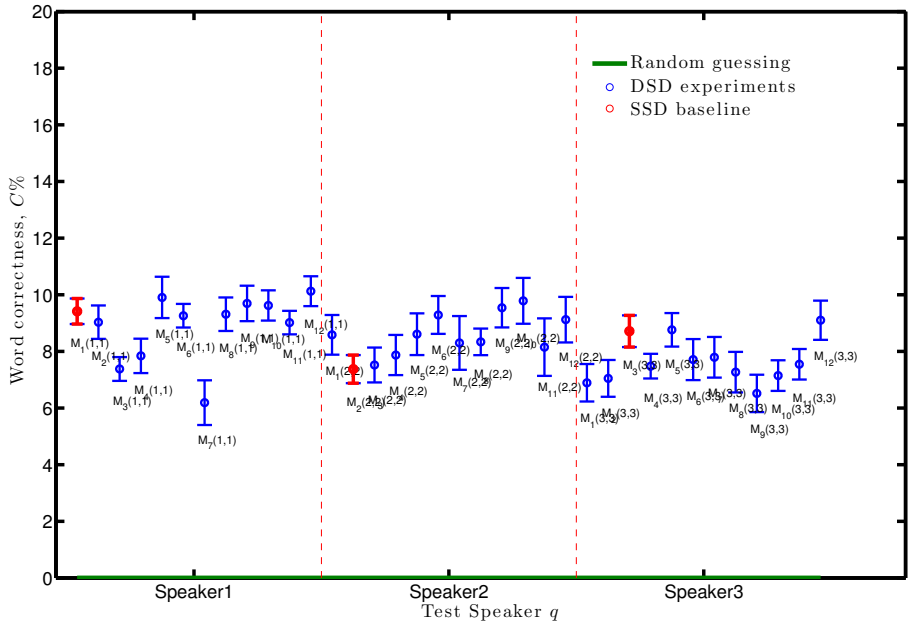



Figure 8.12: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD tests where HMM classifiers are constructed with single-speaker dependent phoneme-to-viseme maps for all speakers in RMAV and tested on others. Baseline is SSD maps (red), results shown for HMMs trained on speakers 1-3.

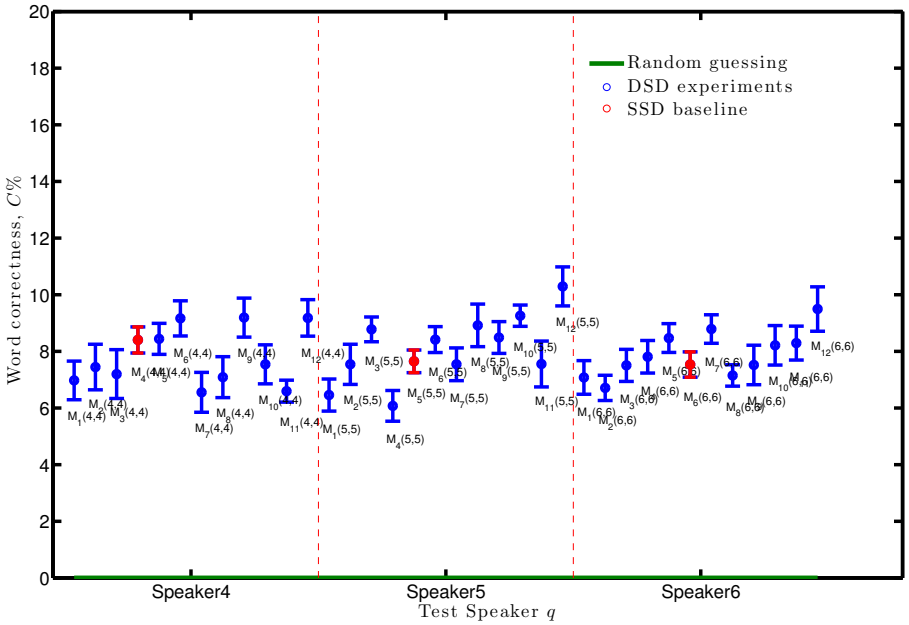



Figure 8.13: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD tests where HMM classifiers are constructed with single-speaker dependent phoneme-to-viseme maps for all speakers in RMAV and tested on others. Baseline is SSD maps (red), results shown for HMMs trained on speakers 4-6.

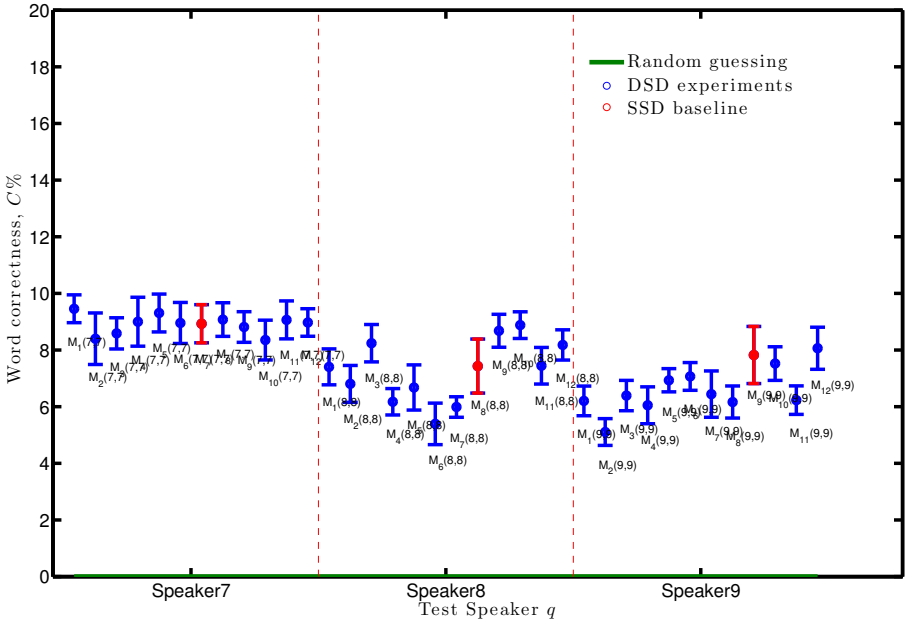



Figure 8.14: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD tests where HMM classifiers are constructed with single-speaker dependent phoneme-to-viseme maps for all speakers in RMAV and tested on others. Baseline is SSD maps (red), results shown for HMMs trained on speakers 7-9.

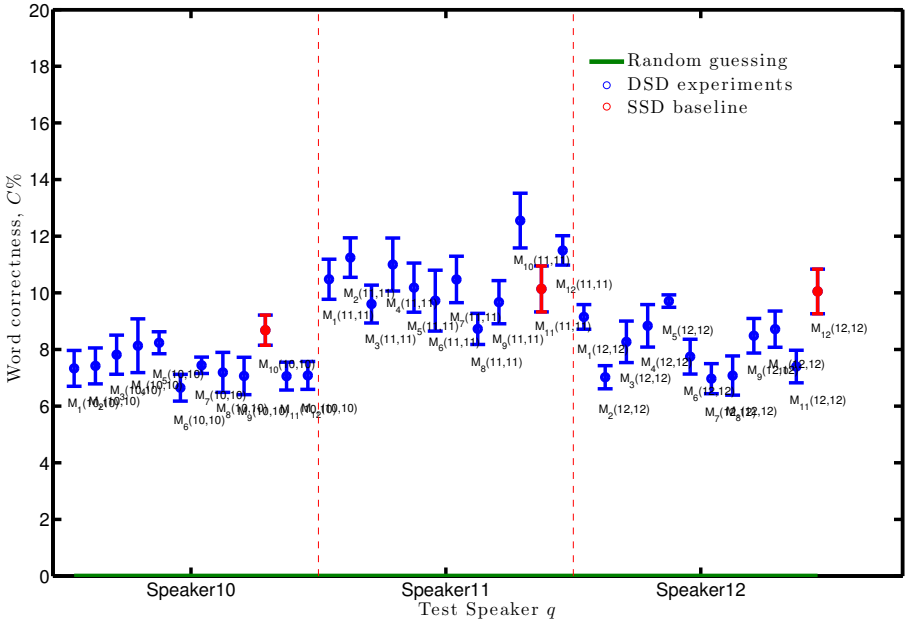



Figure 8.15: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the DSD tests where HMM classifiers are constructed with single-speaker dependent phoneme-to-viseme maps for all speakers in RMAV and tested on others. Baseline is SSD maps (red), results shown for HMMs trained on speakers 10-12.

The equivalent graphs for the 12 RMAV speakers are in Figures 8.12, 8.13, 8.14 and 8.15. Now we can see the effects of the unit selection. Using Speaker 1 for example, in Figure 8.12 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 1, 4, 5, 6, 9 and 11 all significantly improve the classification of Speaker 2. This suggests its 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 (for reasons as yet unknown) or there are similarities between certain speakers which when learned properly on one speaker are able to better classify the rarer visual distinctions between phonemes on similar other speakers.

In Figure 8.14 we see Speaker 7 is particularly robust to visual unit selection for the classifier labels. Conversely Speakers 5 (Figure 8.13) and 12 (Figure 8.15) 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. Again, the number of visual classifiers really does vary subject to the speaker identity.

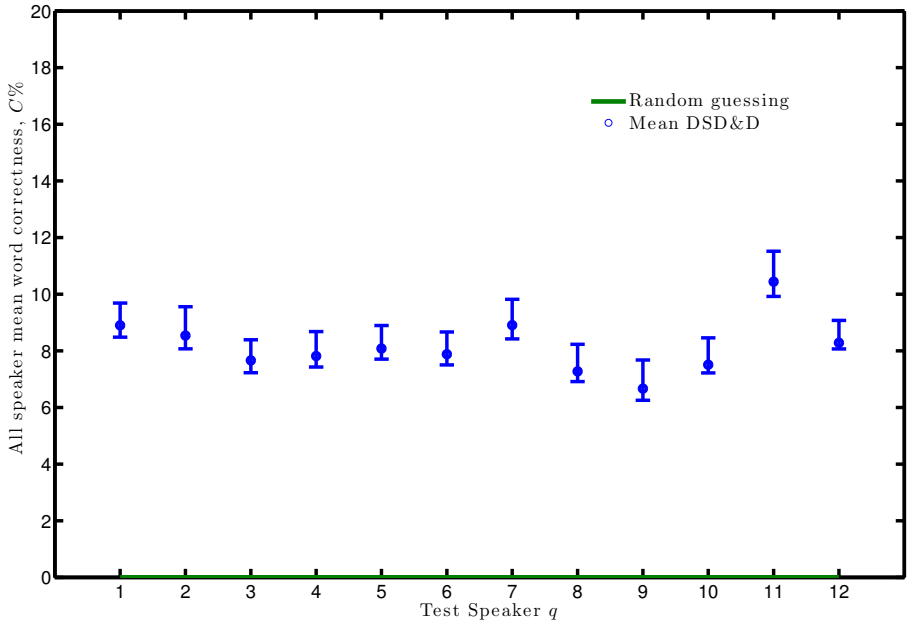



Figure 8.16: All-speaker mean word classification correctness, $C$, of the DSD classifiers constructed with single-speaker dependent phoneme-to-viseme maps for twelve speakers in RMAV and tested on others. Baseline is SSD maps (red) and error bars show $\pm 1 \frac{\sigma}{\sqrt{10}}$.

Figure 8.16 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 have also plotted random guessing and one standard error over the ten folds. 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.

The performance of each viseme set is ranked by speaker by weighting the effect of the DSD tests. Each map scores as in Table 8.18. If a map increases on SSD performance within error bar range this scores +1 or outside error bar range scores +2. If a map decreases classification on SSD performance, these values are negative.

Table 8.18: Weighted ranking scores from comparing the use of speaker-dependent maps for *other* speaker lip-reading in isolated word speech (AVLetters2 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** |

Therefore these values show $M_3$ is the best 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. Note this order matches a decreasing order of quantity of visemes in the speaker-dependent viseme sets i.e. the more similar to phoneme classes visemes are, then the better the classification performance. This ties in with Table 8.15, where the larger P2V maps create less homophones.

In Table 8.3, which lists our 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 present three times, {/*f*/}, {/*k*/}, {/*w*/} and {/*z*/} twice. The lesson from Figure 8.11, is the selection of incorrect units, whilst detrimental, is not as devastating as training classification classes on alternative speakers.

The same measure has been listed in Table 8.19 for our 12 RMAV speakers. The key observation in this table is Speaker 12 on the far right column. The speaker dependent map of Speaker 12 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 8.19), and crucially, $M_{12}$ only has one speaker (Speaker 10) for whom the visemes in $M_{12}$ does 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 our assertion visual speech is dependent upon the individual but we also now have evidence there are exceptions to the rule. In order the RMAV P2Vs are:

Table 8.19: Weighted scores from comparing the use of speaker-dependent maps for *other* speaker lip-reading in continuous speech (RMAV speakers).

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ | $M_7$ | $M_8$ | $M_9$ | $M_{10}$ | $M_{11}$ | $M_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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** |

1. $M_{12}$
2. $M_5$
3. $M_9$
4. $M_{10}$
5. $M_6$
6. $M_3$
7. $M_4$
8. $M_7$ and $M_{11}$
9. $M_1$ and $M_8$
10. $M_2$

Figure 8.17 shows the correctness of both the MS viseme class set and the SI tests (Tables 8.10 and 8.13) against our SSD baseline for AVL2 speakers. Word correctness, $C$ is plotted on the $y$-axis. For the multi-speaker classifiers, these are all built on the same map $M_{all}$, and tested on the same 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 $M_{!4}(4,4)$. Again the same baseline is repeated where $n = p = q$ for reference.

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 ($M_{!3}$)

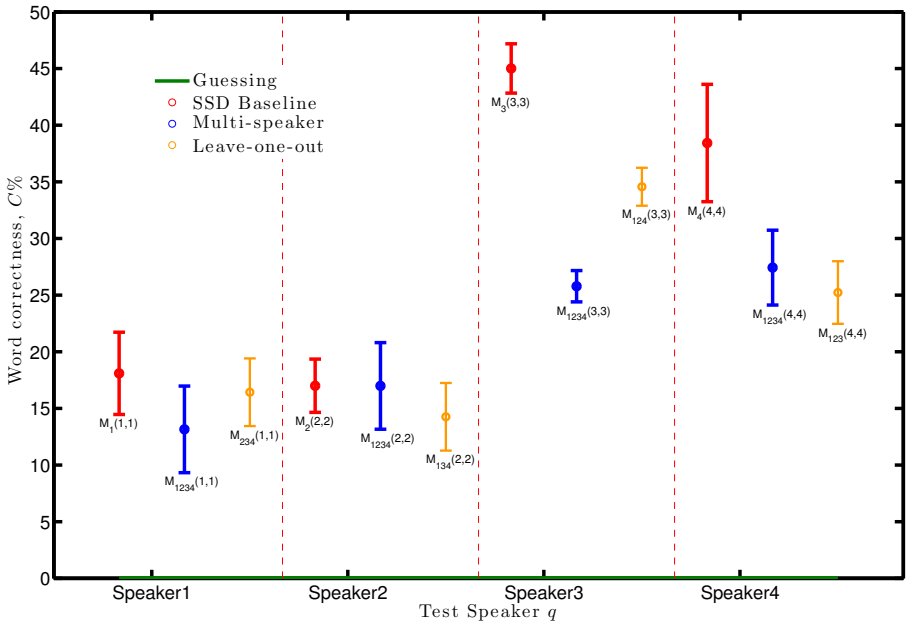



Figure 8.17: Word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{7}}$, of the classifiers using MS and SI phoneme-to-viseme maps on AVLetters2 speakers. Baseline is SSD maps (red).

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 phonemes for speaker-dependent visemes as we are trying to maximise on the speaker individuality to make better viseme classes.

We have plotted the same MS & SI experiments on RMAV speakers in Figures 8.18 and 8.19 (six speakers in each figure). In continuous speech, all bar Speaker 2 are significantly negatively affected by using generalised multi-speaker visemes, whether the visemes include the test speakers phoneme confusions or not. This reminds us of the dependency on speaker identity in machine lip-reading but we do see the scale of this effect depends on which two speakers are being compared. For our exception speaker (Speaker 2 in Figure 8.18) there is only a insignificant decrease in correctness when using MS and SI visemes. Therefore it could be possible with making multi-speaker visemes based upon groupings of visually similar

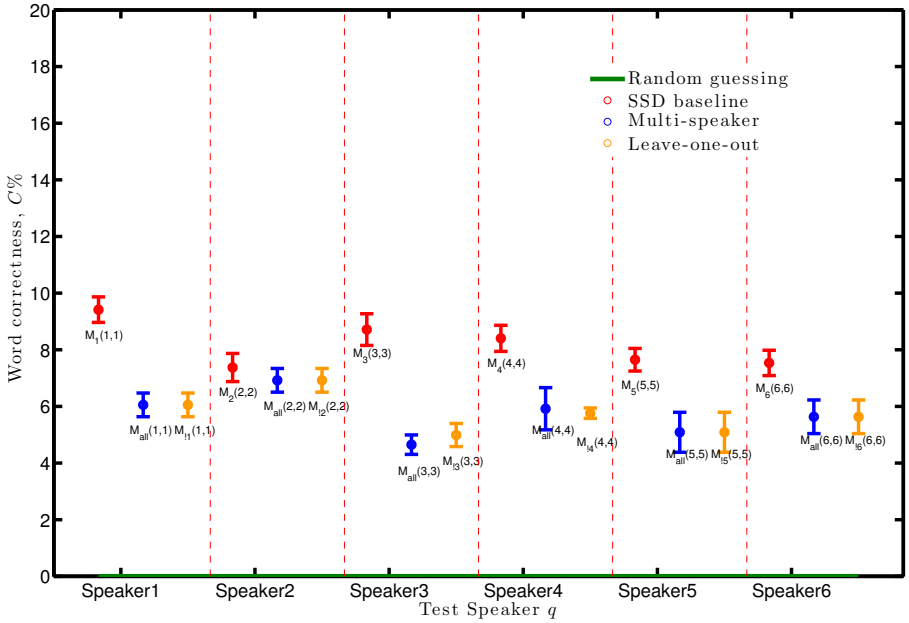



Figure 8.18: Mean word correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the classifiers using MS and SI phoneme-to-viseme maps on RMAV speakers. Baseline is SSD maps (red) - Speakers 1-6.

speakers, even better visemes could be created. The challenge remains in knowing which speakers should be grouped together before undertaking P2V map derivation.

## 8.7 Speaker independence between sets of visemes

For isolated word classification our main conclusion of this chapter is shown by comparing Figures 8.11 & 8.17 with Figure 8.6. The reduction in performance in Figure 8.6 is when the system classification models are trained on a speaker who is not the test speaker. This raised the question if 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 lip-reading, is it not the choice of phoneme-to-viseme map which causes significant degradation to accurate classification, 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.

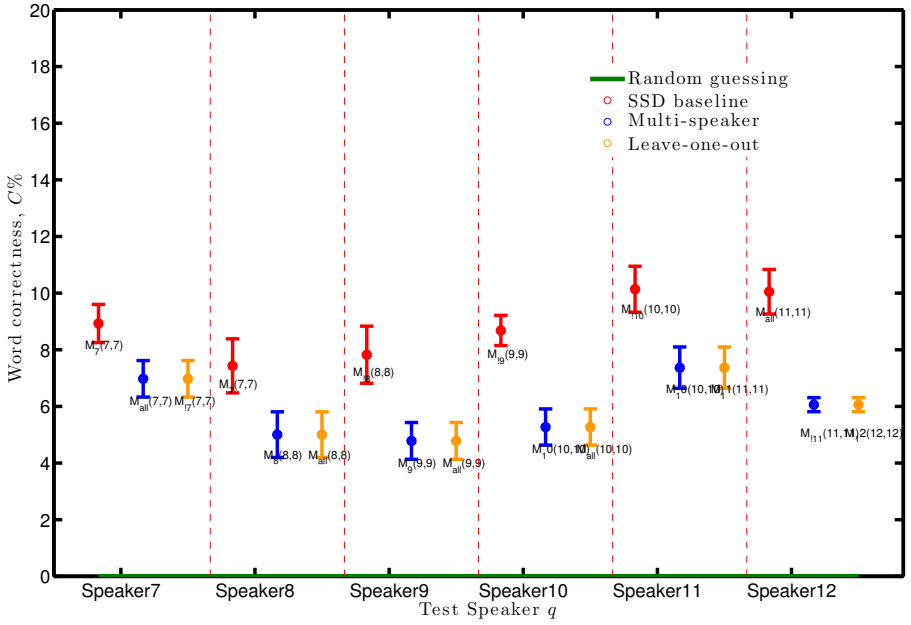



Figure 8.19: Mean word correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$, of the classifiers using MS and SI phoneme-to-viseme maps on RMAV speakers. Baseline is SSD maps (red) - Speakers 7-12.

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. This is further reinforced by Tables 8.3, 8.8 & 8.12. There are differences between speakers, but not significant ones. However, we have seen some exceptions within our 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 talker? 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 our multi-speaker and SI maps, there are 11-12 visemes per set

whereas in our 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 phoneme-to-viseme mappings. This is something we have seen with our 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 [17]. 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.

We compare our speaker independent results to the AAM results of Neti *et al.* [108] and we see our results are inferior overall. Neti *et al.* achieved a $w.e.r$ of 64% compared to our accuracy of around 5% (with AVL2) and between 6-10% with RMAV. We attribute this to the training data volumes in each dataset. The IBM via voice dataset [99] used in [108] is not publicly available but as it has 290 speakers and 10,500 word vocabulary, compared to RMAV which has 12 speakers and 1000 words per speaker.

It is often said in machine lip-reading 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 this. This implies a dependency upon the number of visemes within each set for individuals so this is what we investigate in the next chapter.

# Chapter 9

# Finding phonemes

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 generalisation between different articulated sounds. To find an optimal set of viseme classes, we need to minimise the generalisation to maintain good classification but also to maximise the training data available.

In Chapter 7 we have shown how P2V maps can be derived automatically from phoneme confusions. A by-product of clustering phonemes from classification data is the option to control how many visemes a set contains within the phoneme clustering algorithm. This allows precision when answering questions about the optimal number of visemes. We ask how many visemes is the optimum number? And does this optimum vary by speaker in visual speech?

For this work we use the RMAV dataset [79] and BEEP pronunciation dictionary [26]. Figure 9.1 shows a high level overview of the experiment. It begins by performing classification using phoneme-labelled classifiers. This provides a set of speaker-dependent confusion matrices which are used to cluster together single phonemes (monophones) into subgroups, or as we call them, visemes.

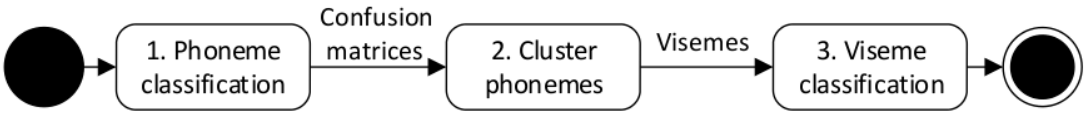



Figure 9.1: Three-step high-level process for viseme classification where the visemes are derived from phoneme confusions.

This time around, we adopt a different phoneme clustering process (described in subsection 9.2). By this process, a new P2V mapping is derived for every time a pair of classes is re-classified in to a new class grouping. There are a maximum of 45 phonemes in the phonetic transcript of the RMAV speakers. This means we can create up to 45 P2V maps per speaker. The actual number of maps produced is subject to the number of phonemes matched during the phoneme classification (step 1 of Figure 9.1). This first step produces the phoneme confusion matrices from which we create new phoneme clusters into visemes. If a phoneme has not been classified, either incorrectly or correctly, then it is not included in the resulting confusion matrix from which our visemes are created. Thus, we now have up to 45 sets of viseme labels to use for labelling our HMMs when repeating the word classification task.

We continue with analysing the word classification rather than visemes as we do not wish our results to be affected by the variance in training samples for each set of classifiers. It is not the performance itself which is relevant here, rather it is any improvement a variance in classes can provide. It is important the reader remember the presentation of this new method is *not* a suggestion this particular clustering algorithm will deliver the optimum visemes, but rather address the need in this case for a method to enable a controlled comparison of the phoneme to viseme distributions as the number of classes reduces.

## 9.1 Step One: phoneme classification

Step 1 implements 10-fold cross-validation with replacement [42], of 200 sentences per speaker, 20 are randomly selected as test samples and are not included in the training folds. Using the HTK toolkit [150] to implement HMM classifiers, the HMMs are initialised by the flat-start method, and re-estimated 11 times with forced alignment between seventh and eighth estimates. The prototype HMM is based upon a Gaussian mixture of five components and three state HMMs. Included is a single-state tied short-pause, or 'sp' HMM for short silences between words in the sentence utterances. A bigram word network is used to support classification.

## 9.2 Step Two: phoneme clustering

The phonemes are clustered into new viseme classes for each speaker as follows; step 1 produces ten confusion matrices for each speaker (one from each fold), these are summed together to form one confusion matrix representing all confusions for that speaker. Clustering begins with this phoneme confusion matrix:

$$[K_m]_{ij} = N(\hat{p}_j|p_i) \tag{9.1}$$

where the $ij^{th}$ element is the count of the number of times phoneme $i$ is classified as phoneme $j$. This algorithm works with the column normalised version,

$$[P_m]_{ij} = Pr\{p_i|\hat{p}_j\} \tag{9.2}$$

the probability that, given a classification of $p_j$ that the phoneme really was $p_i$. The subscript $m$ in $K_m$ and $P_m$ indicates $K_m$ and $P_m$ have $m^2$ elements ($m$ phonemes). Merging of phonemes is done by looking for the two most confused phonemes and hence create a new class with confusions $K_{m-1}, P_{m-1}$.

Table 9.1: An example phoneme-to-viseme map, this is the phoneme-to-viseme map for RMAV Speaker 1 with ten visemes.

| Viseme | Phonemes |
|---|---|
| /*v*01/ | /ax/ |
| /*v*02/ | /v/ |
| /*v*03/ | /ɔɪ/ |
| /*v*04/ | /f/ /ʒ/ /w/ |
| /*v*05/ | /k/ /b/ /d/ /θ/ /p/ |
| /*v*06/ | /l/ /dʒ/ |
| /*v*07/ | /g/ /m/ /z/ /y/ /tʃ/ /ð/ /s/ /r/ /t/ /ʃ/ |
| /*v*08/ | /n/ /hh/ /ŋ/ |
| /*v*09/ | /ɛ/ /ae/ /ɔ/ /uw/ /ɒ/ /ɪə/ /ey/ /ua/ /ɜ/ |
| /*v*10/ | /ay/ /ɑ/ /ʌ/ /ɑʊ/ /ʊ/ /əʊ/ /ɪ/ /iy/ /ə/ /eh/ |

Specifically for each possible merged pair, $Pr, Ps$ score is calculated by:

$$q = [P_m]_{rs} + [P_m]_{sr} = Pr\{\hat{P}r|Ps\} + Pr\{\hat{P}s|Pr\} \tag{9.3}$$

Phonemes are assigned to one of two classes, $V\&C$, vowels and consonants. Vowels and consonants can not be mixed. The pair with the highest $q$ is merged. Equal scores are broken randomly. This process is repeated until $m = 2$. Each intermediate step, $M = 45, 44, 43...2$ forms a possible set of visual units.

This is a more controlled approach than the method used in Chapter 7 and [16], and incorporates our conclusions vowel and consonant phonemes should not be clustered together when devising phoneme-to-viseme mappings. An example P2V mapping is shown in Table 9.1.

## 9.3 Step Three: viseme classification

Similar to step 1, step 3 involves implementation of 10-fold cross-validation with replacement [42], of 200 sentences per speaker, 20 are randomly selected as test samples and these are not included in the training folds. Using the HTK toolkit [150] to use Hidden Markov Model (HMM) classes, viseme labelled HMMs are flat-

started, re-estimated 11 times over with forced alignment between seventh and eighth estimates. The same HMM prototype is used and a bigram word network supports classification along with the application of a grammar scale factor of 1.0 (shown to be optimum in [67]) and a transition penalty of 0.5.

The important difference this time around are the viseme classes being used as classification labels. By using these sets of classes which have been shown in step 1 to be confusing on the lips, we now perform classification for each class set. In total this is 45 sets, where the smallest set is of two classes (one with all the vowel phonemes and the other all the consonant phonemes), and the largest set is of 45 classes with one phoneme in each - thus the largest set for each speaker is a repeat of the phoneme classification task but using only phonemes which were originally recognised (either correctly or incorrectly) in step 1.

## 9.4 Searching for an optimum

In Figures 9.2 - 9.13, we show the word correctness, plotted on the $y$-axis for all 12 speakers. Each of the viseme sets, identified by the number of visemes within the set, are plotted in increasing order along the $x$-axis. We have also plotted, in green, guessing weighted by the visual homophones in the transcripts. This has been calculated by:

$$\sum_{i=1}^{i=N}(\frac{TC_i}{W}) * (\frac{1}{N}) \tag{9.4}$$

where $TC$ is the total individual token count for that speaker (for each token), $W$ is the total words for that speaker, and $N$ is the number of tokens. $i$ is for all each token where a token is a unique word.

Viseme sets containing fewer visemes produce viseme strings which represent more than one word: homophones. The effect of homophones can be seen on the left side of the graphs in Figures 9.2 - 9.13 with viseme sets with fewer than 11 visemes.

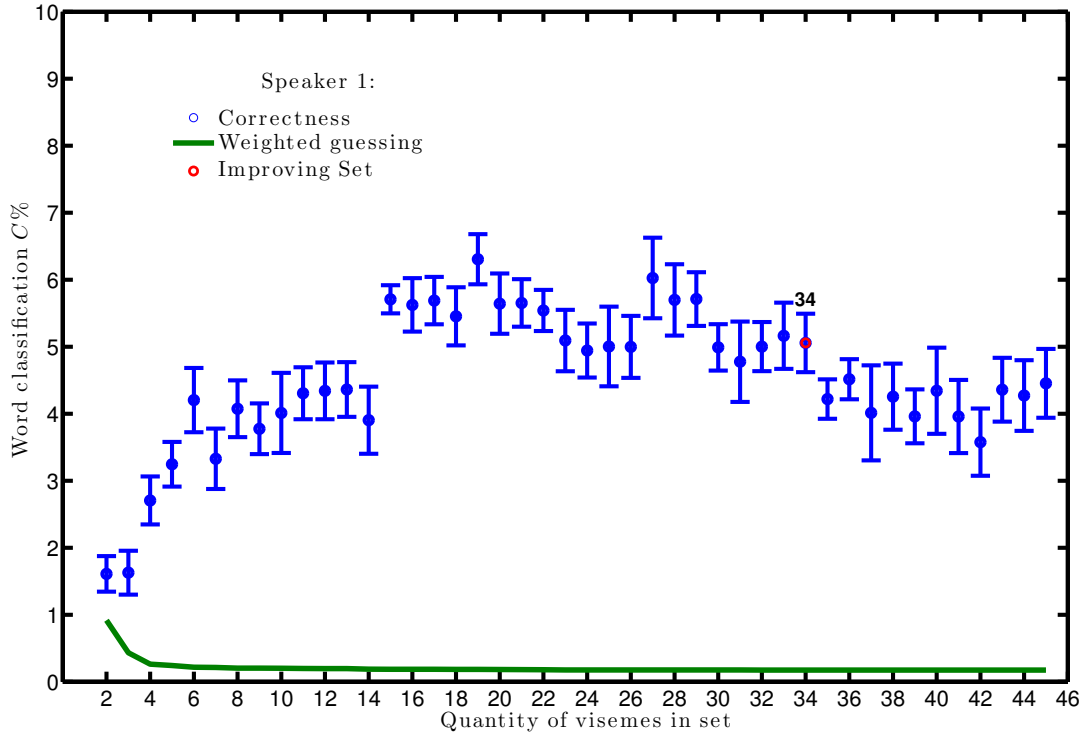



Figure 9.2: Speaker 1: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

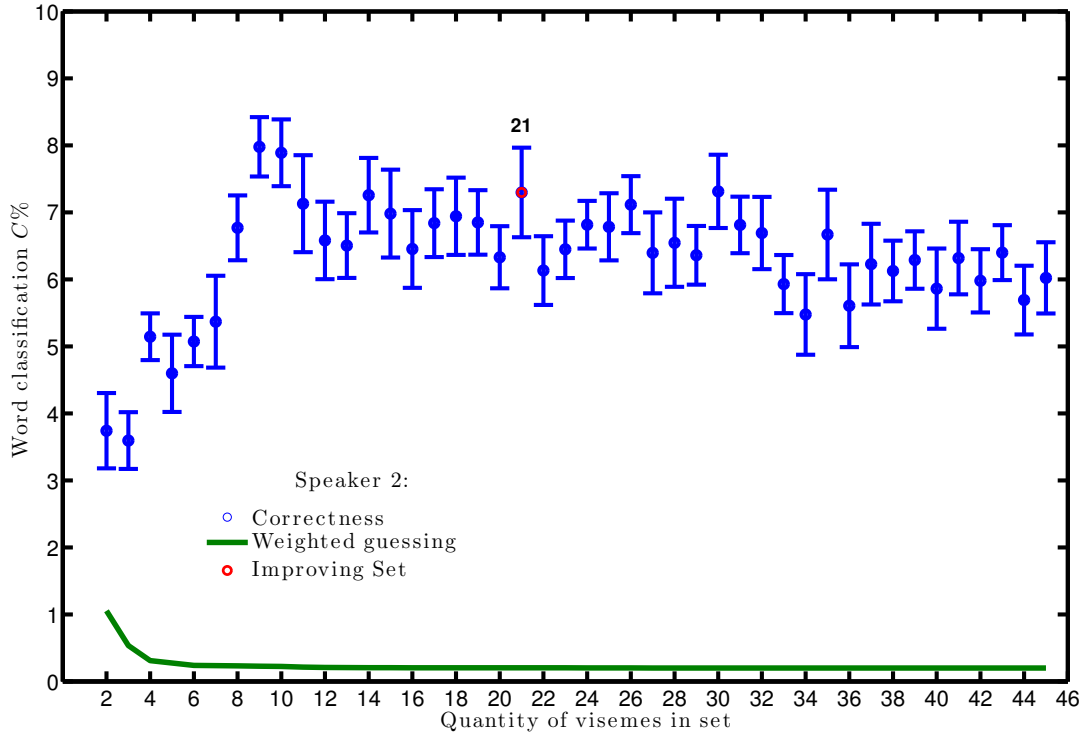



Figure 9.3: Speaker 2: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

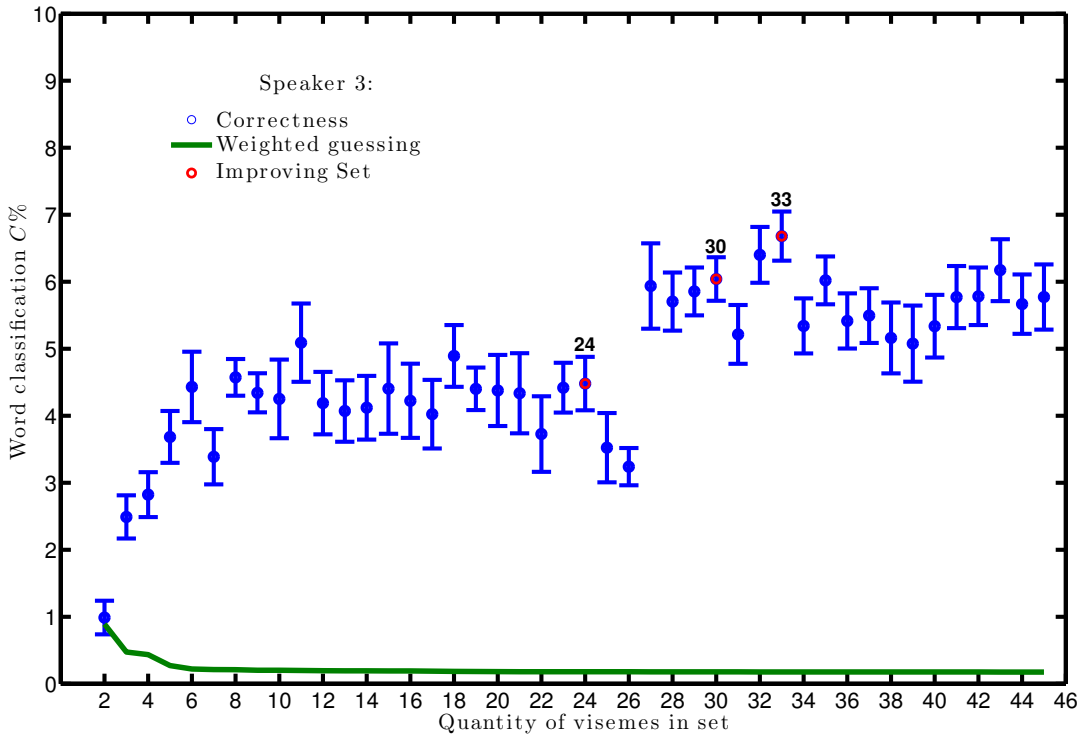



Figure 9.4: Speaker 3: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

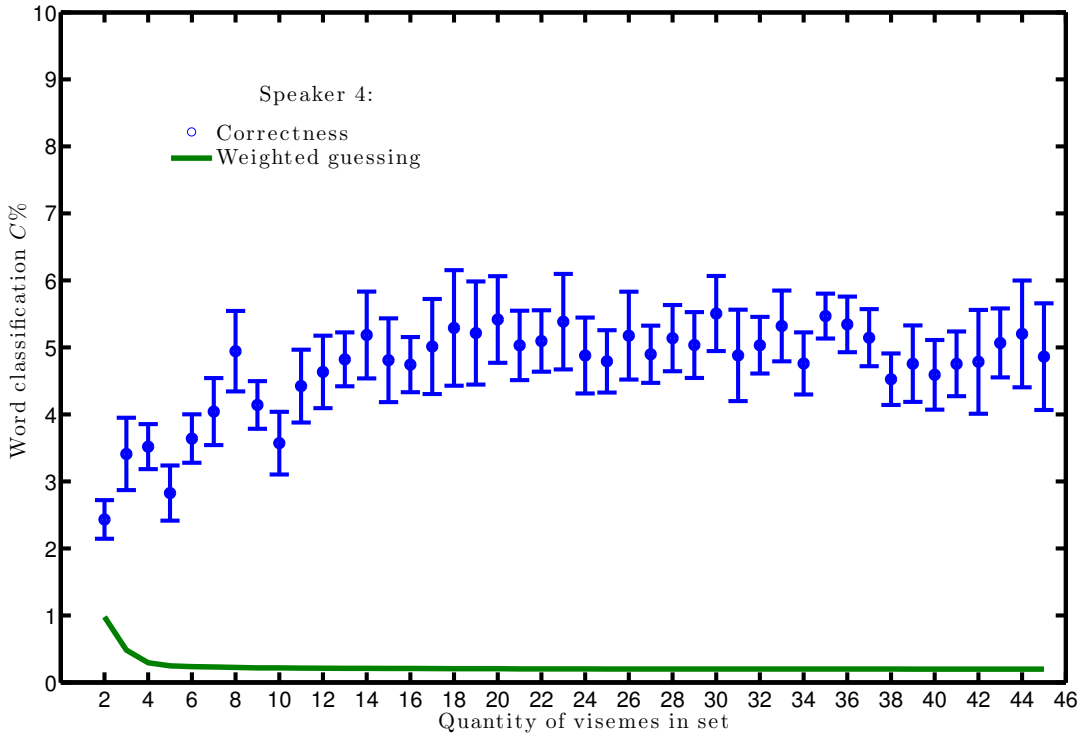



Figure 9.5: Speaker 4: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

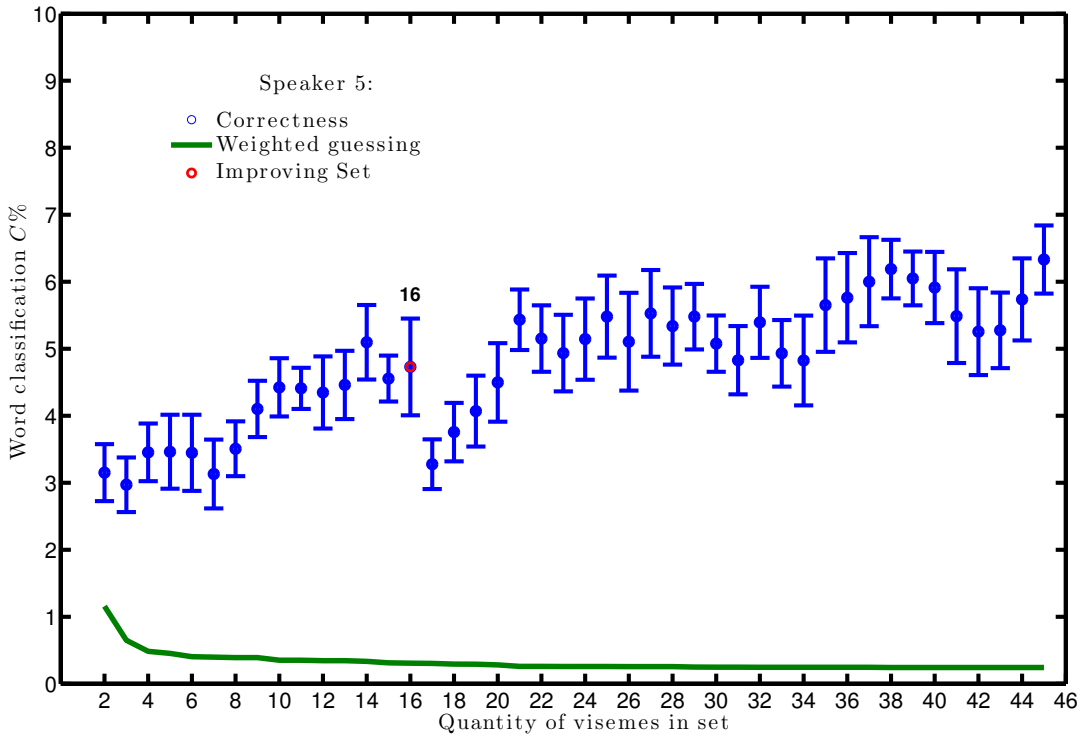



Figure 9.6: Speaker 5: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

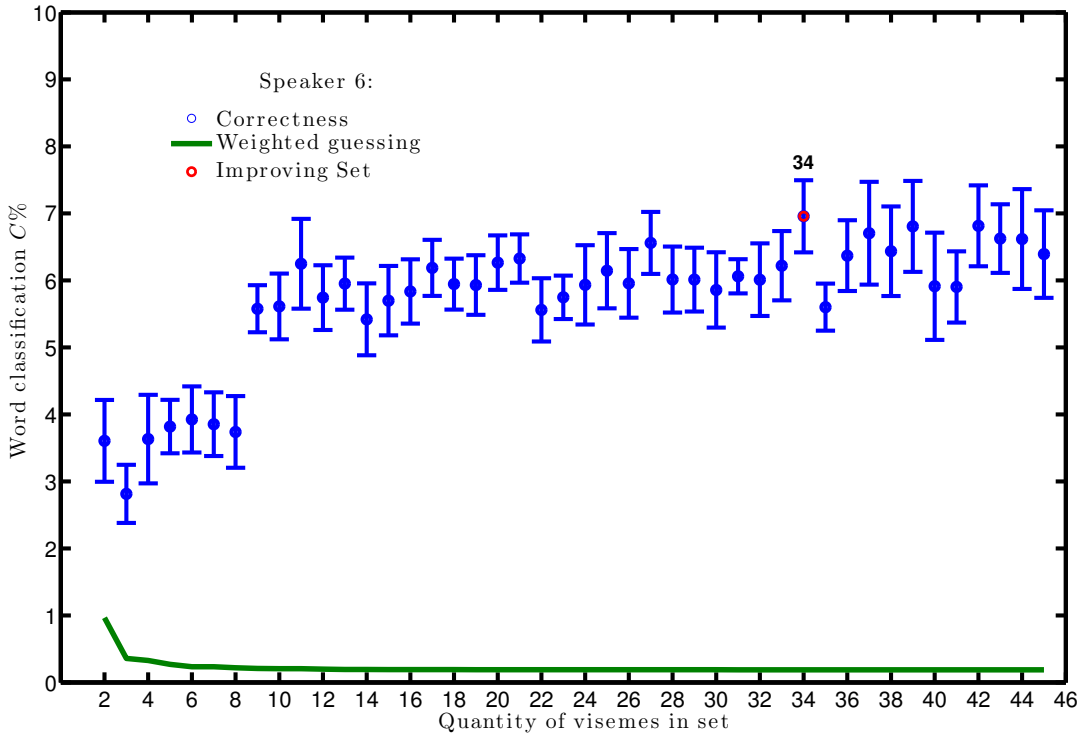



Figure 9.7: Speaker 6: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

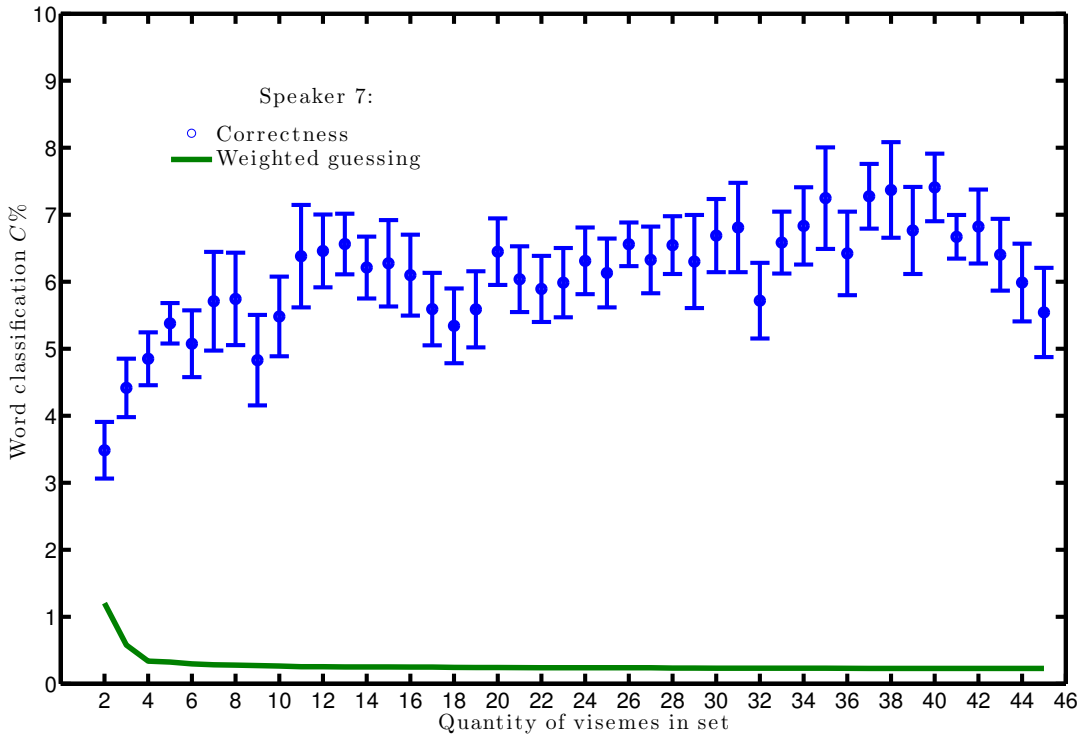



Figure 9.8: Speaker 7: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-45.

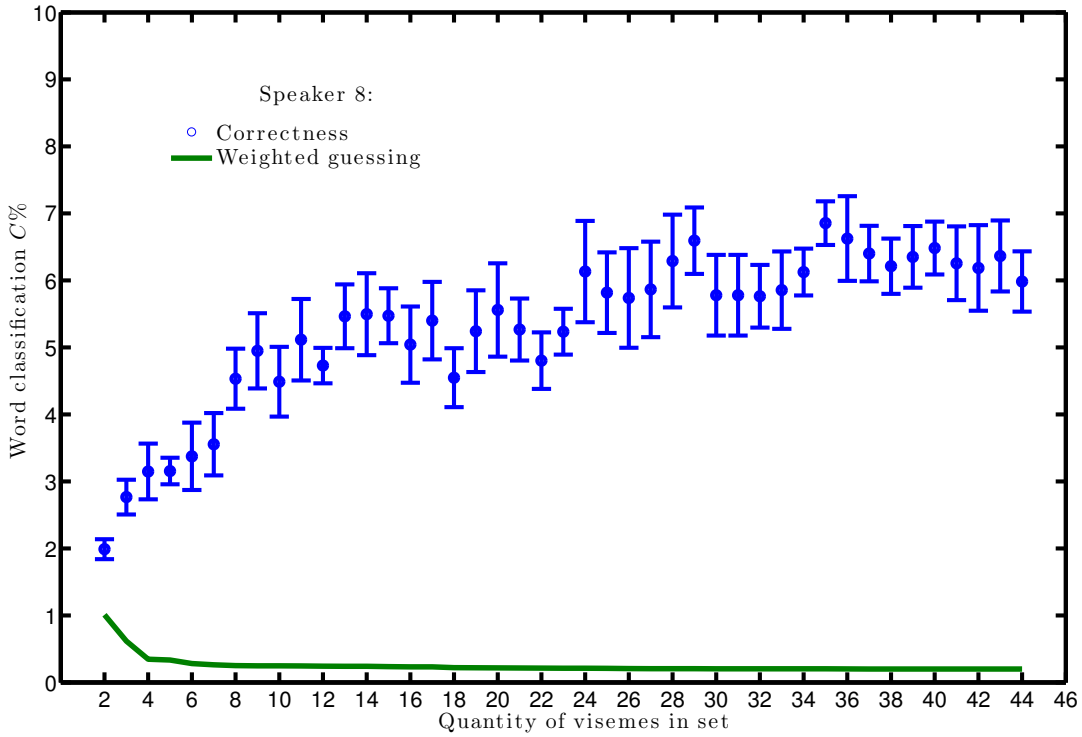



Figure 9.9: Speaker 8: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-44.

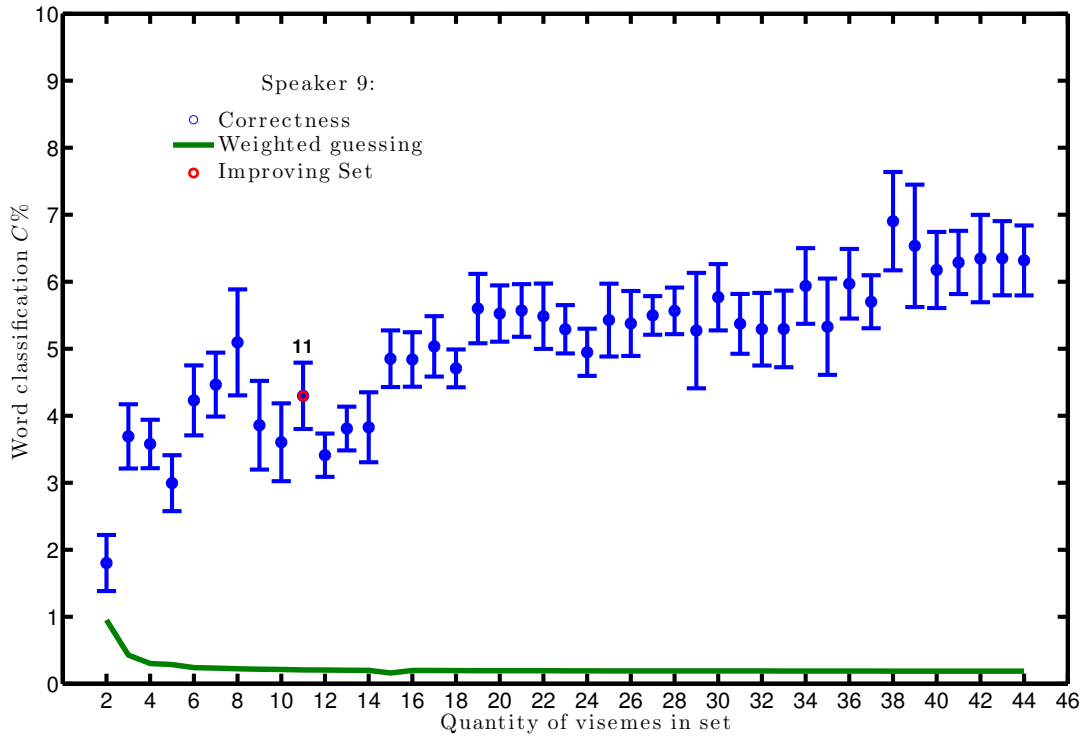



Figure 9.10: Speaker 9: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-44.

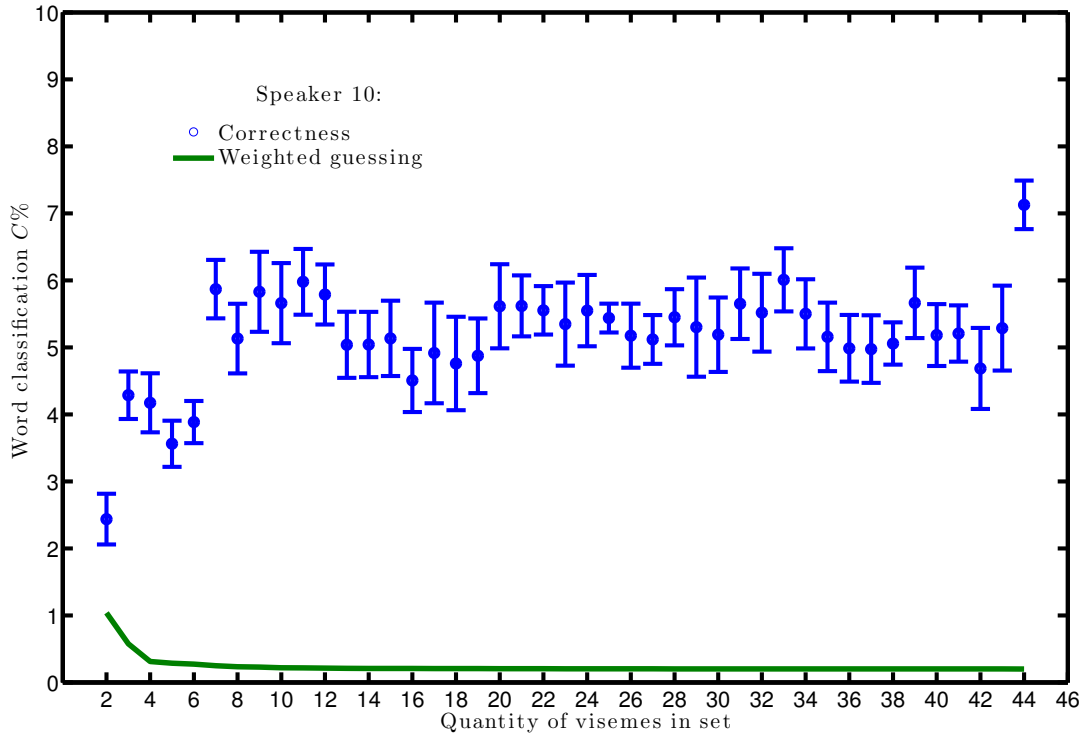



Figure 9.11: Speaker 10: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-44.

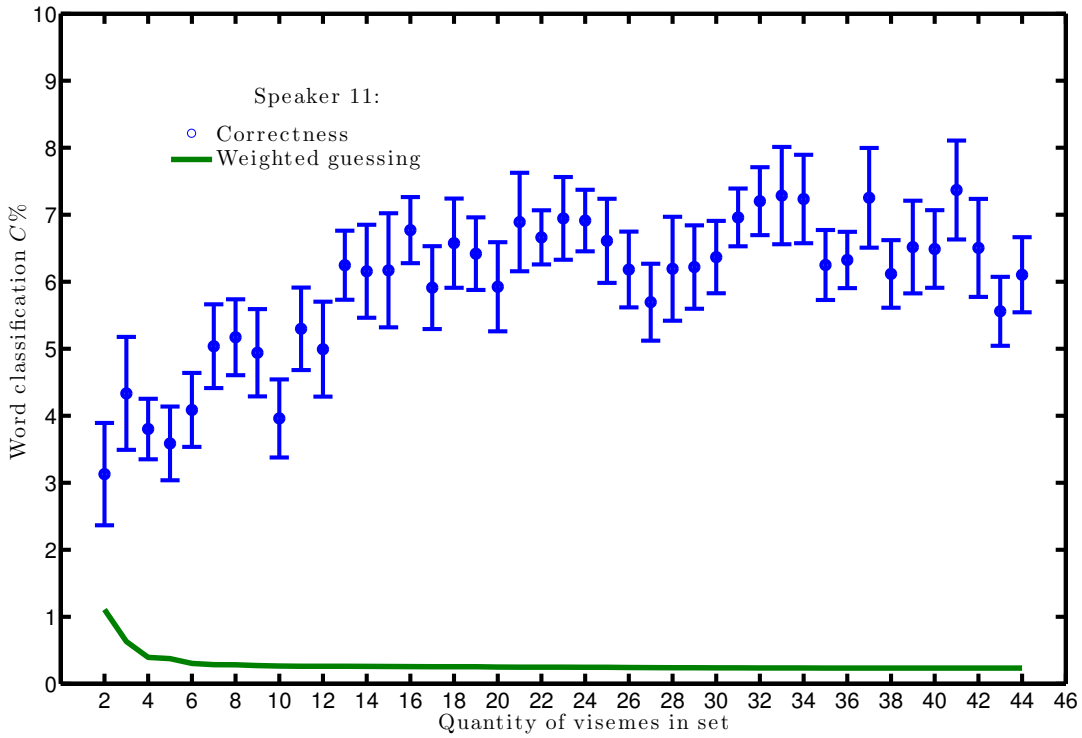



Figure 9.12: Speaker 11: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-44.

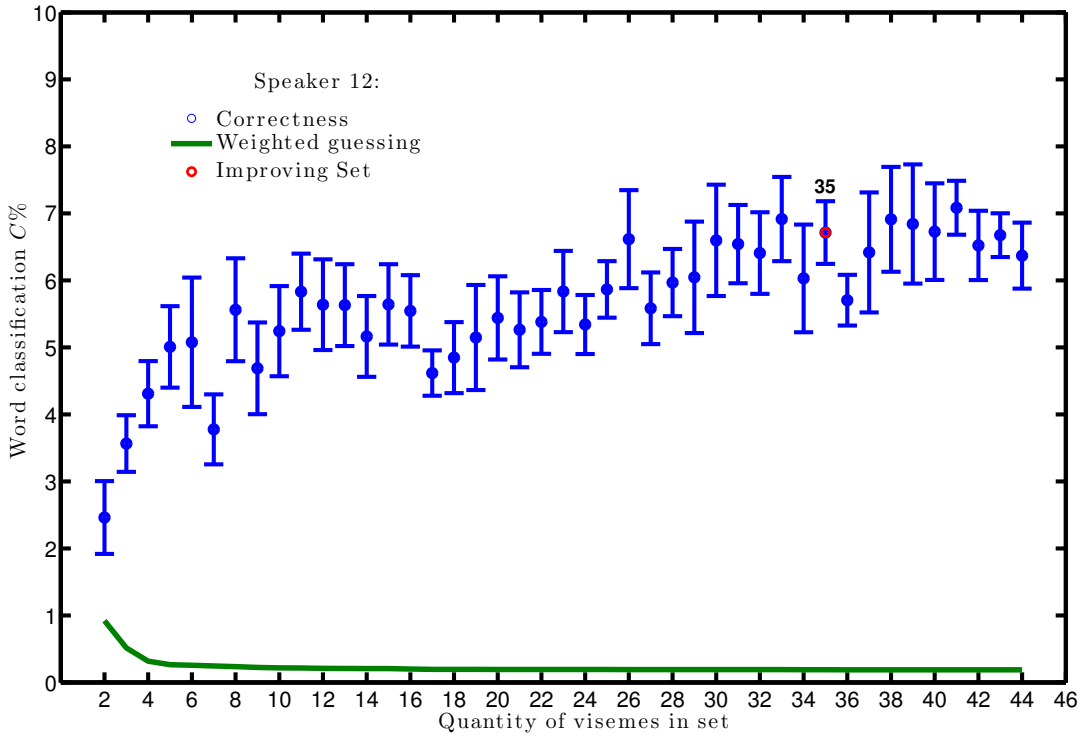



Figure 9.13: Speaker 12: word classification correctness, $C \pm 1\frac{\sigma}{\sqrt{10}}$ for phoneme-to-viseme map sizes 2-44.

All correctness scores are significantly above chance albeit still low. There is variation between speakers, which is expected but there is a very clear overall trend. Superior performances are to be found with larger numbers of visemes. An important point is some authors report word accuracy as viseme performance when using a word unit language network. This is unhelpful as it masks the effect of homophones by using the network level unit rather than the accuracy of the viseme models themselves. Had we reported this then the effect of needing larger numbers of visemes would not be visible.

Also in Figures 9.2 - 9.13, we have highlighted the class sets in red where these show a significant improvement in classification over the adjacent set of units on its right side along the $x$-axis. This is where we can identify the pairs of classes which, when merged into one class, significantly improve classification. Table 9.2 lists these special viseme combinations. Referencing back to speaker demographics (such as gender or age), there is no apparent pattern through these viseme combinations. So we have further evidence to reinforce the knowledge that all speakers are visually unique and we are reminded of how difficult finding a set of cross-speaker visemes is when phonemes require alternative groupings for each individual.

Table 9.2: Viseme class merges which improve word classification in correctness; $V_n = V_i + V_j$.

| Speaker | Set No | $V_i$ | $V_j$ | Set No | $V_n$ |
|---|---|---|---|---|---|
| Sp01 | 35 | /s/ /r/ | /ð/ | 34 | /s/ /r/ /ð/ |
| Sp02 | 22 | /d/ | /z/ /y/ | 21 | /d/ /z/ /y/ |
| Sp03 | 34 | /b/ /tʃ/ | /ʒ/ | 33 | /b/ /tʃ/ /ʒ/ |
| Sp03 | 31 | /ʒ/ /b/ /tʃ/ | /z/ | 30 | /ʒ/ /b/ /tʃ/ /z/ |
| Sp03 | 25 | /p/ /r/ | /ŋ/ | 24 | /p/ /r/ /ŋ/ |
| Sp05 | 17 | /ae/ | /eh/ | 16 | /ae/ /eh/ |
| Sp06 | 35 | /ae/ /ʌ/ | /iy/ | 34 | /ae/ /ʌ/ /iy/ |
| Sp09 | 12 | /b/ /w/ /v/ | /dʒ/ /hh/ | 11 | /b/ /w/ /v/ /dʒ/ /hh/ |
| Sp12 | 36 | /ʌ/ | /ɔ/ | 34 | /ʌ/ /ɔ/ |

The conventional wisdom, that visemes are needed for lip-reading, (in [57] for example), is countered in our experiments as our phoneme classification is not significantly different from viseme classification. It is however an over simplification to

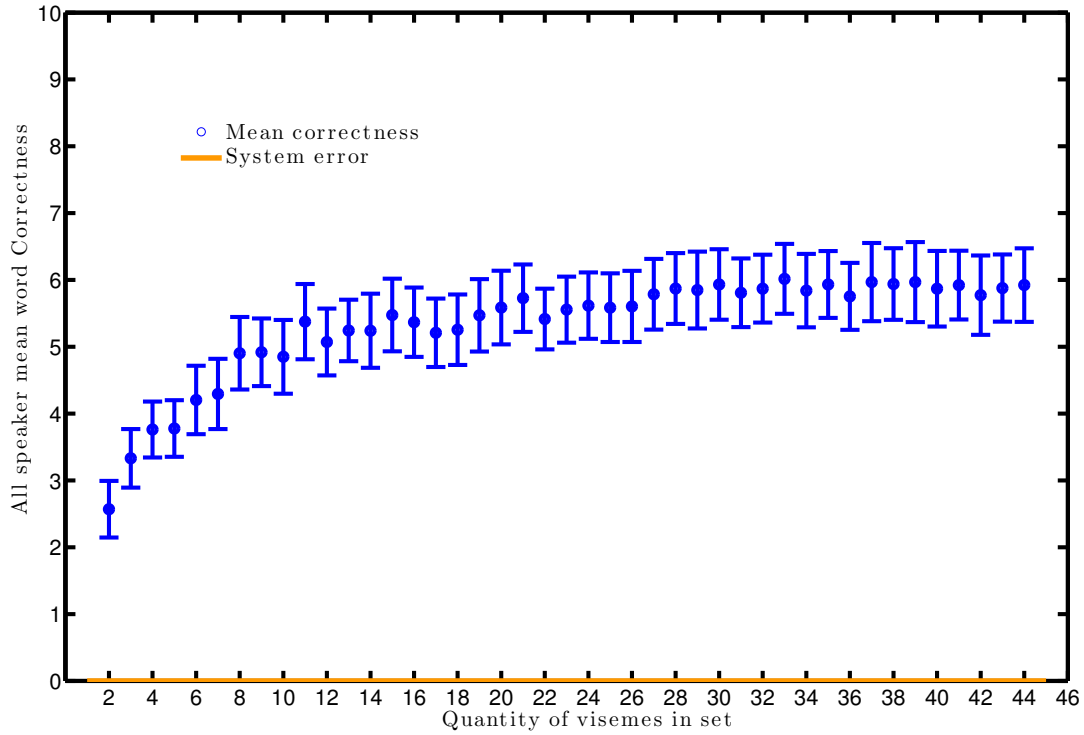



Figure 9.14: All-speaker mean word classification correctness $C \pm 1\frac{\sigma}{\sqrt{10}}$.

assert better lip-reading can be achieved with phonemes than visemes as this has not been shown here with any significance. Generally speaking, larger numbers of visemes out-perform smaller numbers. However, when classification is aggregated in Figure 9.14, which is the mean word correctness, $C$, classification over all speakers, there is, within an error bar, a monotonic trend. In Figure 9.14 we have also plotted the system error instead of guessing. System error is calculated by using the ground truth transcript of the test data in place of the classifiers output in `HResults`, in doing so we obtain any errors caused by the system rather than the classifiers. Fortunately, this is zero, demonstrating the robustness of an HMM lip-reading system.

In the literature we have already reviewed a number of proposed phoneme-to-viseme maps, typically these generate between 10 and 20 visemes (see subsection 7.4 for a summary) - the Lee set has six consonant visemes and five vowel visemes [82]; Jeffers eight & three [69] respectively and so on. Figures 9.2-9.13 & 9.14 show a definite rapid drop-off in performance for sets which contain fewer than

ten visemes but the region between 11 and 20 contains the optimum viseme set for three out of the 12 speakers which is more than chance. This mean, for each speaker we have shown an optimal number of visual units (shown by the best performing result in Figures 9.2-9.13) but the optimal number is not related to any of the conventional viseme definitions, neither is the number of phonemes. Table 9.3 shows the correctness of each speakers phoneme classification.

Table 9.3: Phoneme correctness $C$ for each speaker, these are plotted on the right hand side in Figures 9.2 to 9.13 as the largest set of visemes (either 44 or 45, subject to the speaker).

| Speaker | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Phoneme $C$ | 0.045 | 0.060 | 0.058 | 0.049 | 0.063 | 0.063 |
| Speaker | 7 | 8 | 9 | 10 | 11 | 12 |
| Phoneme $C$ | 0.055 | 0.090 | 0.063 | 0.071 | 0.061 | 0.064 |

The implication is that, for a few speakers, it is possible to conclude a small number of visemes are optimal. However, when considering all speakers, it is much more likely phonemes provide a better set of classifier labels for classification.

The two factors at play in these graphs are, the underlying accuracy with which the visual units represent the mouth shape and appearances versus the introduction of homophones. For large numbers of visemes these are close to phonetic classification, (with fewer homophones) but they run the risk of visual units which are not visually distinctive - several of the HMM models will "match" on a particular subsequence. This latter problem creates a decoding lattice in which there are several near equal probability paths which, in turn, implies state-of-the-art language models would improve results still further.

## 9.5 Hierarchical training for weak-learned visemes

Some recent work presents evidence viseme labels may not be needed because with enough data, classifiers based upon phoneme labels can outperform viseme classification [67, 57]. Additionally, we have now seen there are challenges with using

viseme/phoneme labelled classifiers including; the homophone effect, not enough training data per class, and the consequential lack of differentiation between classes when we get too many classes to distinguish between them. These can be seen in Figure 9.15 where we have replotted word correctness for 12 speakers from Section 9.4 onto one graph.

Figure 9.15 shows our previous results [15], derived using the algorithm described in [16]. We were able to generate viseme sets of varying size. Here the $x$-axis runs from 2 to 45. The $y$-axis shows the word correctness of HMM classifiers trained on each viseme in the viseme set. There are 12 lines for the 12 RMAV speakers.

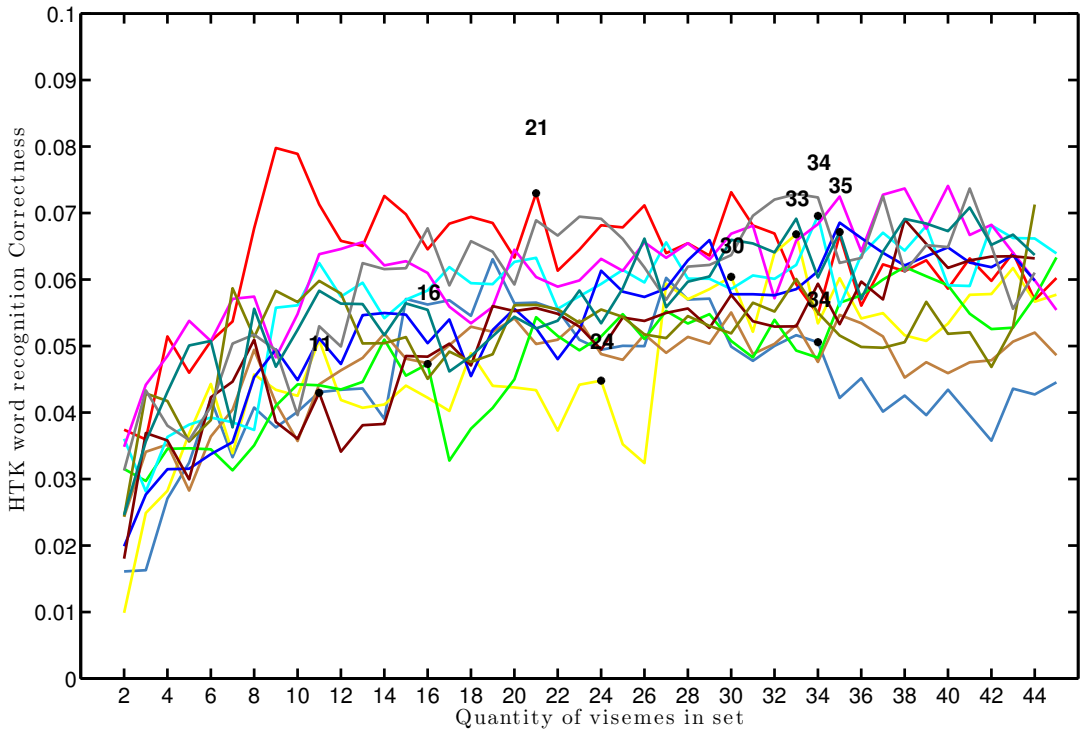



Figure 9.15: Viseme correctness as the quantity of visemes decreases in a set of classifiers for 12 RMAV speakers. Results from [15].

Figure 9.15 also shows for each of our 12 speakers the significantly improving viseme sets listed in Table 9.2. So we know there are sometimes units between traditional visemes and phonemes which are better for classification of the visual speech signal. Our evidence is pointing towards a larger number of visual units than was previously thought sensible. In the extreme example, if we assume one visual

unit per phoneme then there is the problem that identical lip gestures may appear in two separate visemes.

To examine this, we propose the concept of adopting weak learning for hierarchical classifier training. Our intention is to test if this method can improve phoneme classification without the need for more training data as this approach shares training data across models. This premise avoids the negative effects of introducing more homophones but will assist the identification of the more subtle but important differences in visual gestures representing alternative phonemes. Crucially, this method means we are increasing our valid training data without needing to create or record it. We remember from Chapter 7 using the wrong clusters of phonemes is worse than using none.

Weak learning [127] is an alternative approach to training classification models in lip-reading. Weak learning is traditionally applied in ensembles of classifiers where the sum of the classifiers produces a stronger classifier than that of the independently-weak-trained classifiers [41]. By acknowledging the poor performance of our viseme labelled classifiers we can assume that they are weakly trained. That is, that whilst they outperform guessing, they are not strongly trained classifiers (confirmed by our dependence on the language model to improve results). Thus, we if we can adopt a method which boosts these weakly trained viseme classifiers into strongly trained phoneme classifiers we hope to achieve significantly higher classification rates. This also encourages use of more training data for the weak-learning phase [43], and specialised training of specific phoneme samples for the phoneme classifier training phase.

Therefore our last investigation in this thesis is an attempt to modify the lip-reading process in which we apply weak learning during classifier training, to test if the visual signal can be better translated from visemes to phonemes to better train classifiers with the same volume of visual data, whilst improving the classification. In doing so, our method addresses the challenges identified in this chapter thus far.

An additional benefit of the the revised classification process is because weak

learning in the model training phase is before phoneme classification, we no longer need to consider post-classification-processing such as weighted finite state transducers [66] to reverse the phoneme-to-viseme mapping in order to get the real phoneme recognised.

In Figure 9.15 the performance of classifiers with small numbers of visemes ($< 10$) is poor due to the large number of homophones. Large numbers of visemes ($> 35$) do not appear to noticeably improve the correctness: many phonetic variations look similar on the lips. The set numbers printed in black are the significantly improving viseme sets identified by the number of visemes in the set. Therefore we focus on viseme sets in the range 11 to 35 with the same speakers for our experiments using weak learning.

## 9.6 Classifier training adaptation

The basis of the new training approach is to hierarchically train HMM classifiers. Figure 9.16 shows a stylised illustration in which we have five phonemes (in reality there are 45) and two visemes (in reality there will be between 11 and 35). Each phoneme has been assigned to a viseme as in [15] but here we are going to learn intermediate HMMs which are identical to those in [15]. These are the viseme HMMs. We now create models for the phonemes. In this example /$p1$/, /$p2$/ and /$p4$/ are associated with /$v1$/, so are initialised as replicas of HMM /$v1$/. Likewise /$p3$/ and /$p5$/ are initialised as replicas of /$v2$/. We now retrain the phoneme models using the same training data.

In full; we initialise *viseme* HMMs with `HCompV`, the HTK tool `HCompV` used for initialising HMMs defines all models equal [152]. Our prototype HMM is based upon a Gaussian mixture of five components and three states. These are trained 11 times over, including both short pause model state tying (between re-estimates 3 & 4), and forced alignment between re-estimates 7 & 8 (this is steps 1 & 2 in Figure 9.16). But before classification, these viseme HMM definitions are used as initialised def-

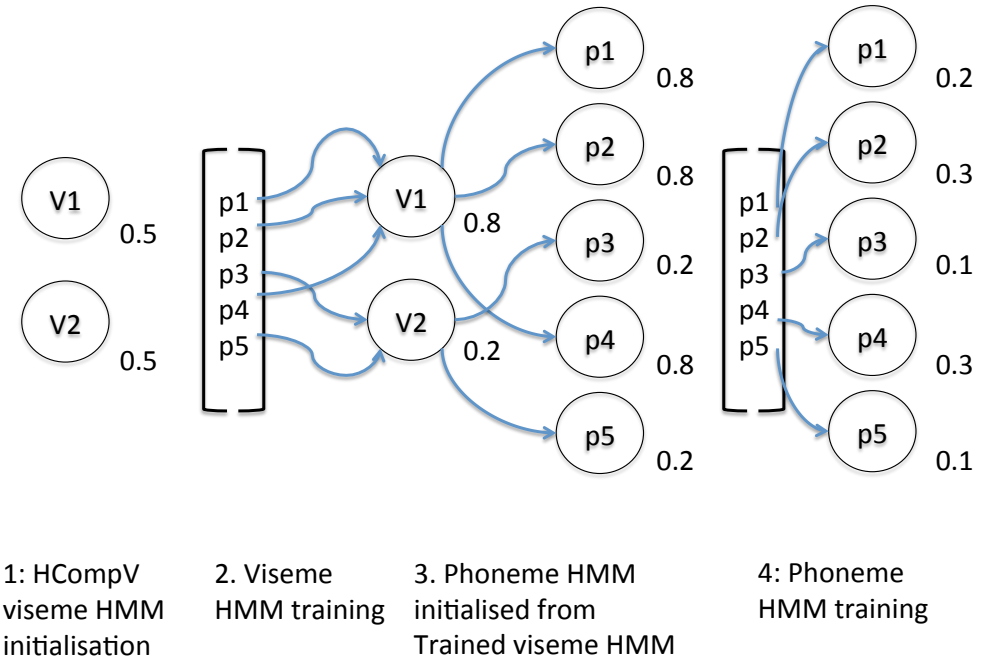



Figure 9.16: Hierarchical training strategy for weak learning of visemes HHMs into phoneme labelled HMM classifiers.

initions for phoneme labelled HMMs (Figure 9.16 step 3). The respective viseme HMM definition is used for all the phonemes in its relative phoneme-to-viseme map. These phoneme HMMs are retrained and used for classification. As part of the classification, we use a bigram network, apply a grammar scale factor of 1.0 and apply a transition penalty of 0.5 (based on [67]). This is implemented using 10-fold cross-validation with replacement [42].

The advantage of this approach is the phoneme classifiers have seen mostly positive cases therefore have good mode matching, the disadvantage is they are limited in their exposure to negative cases, less than the visemes.

### 9.6.1 Language network units

As we are investigating the correct unit selection for our classifiers, we must not forget about the unit selection for the language network which is used to decode classification transcripts. This means we need to review any effect of the language network unit choice before our final experiment. Using the common process previously described for lip-reading, we perform classification using speaker-dependent visemes [16], phonemes and word HMMs with the optional unit networks as listed in Table 9.4. This means we can answer the question 'is there any dependency between the unit choice for the classifier labels and the unit of supporting language network?'.

Table 9.4: Unit selection pairs for HMMs and language network combinations.

| Classifier units | Network units | $C$ |
|---|---|---|
| Viseme | Viseme | 0.0231 |
| Viseme | Phoneme | 0.1914 |
| Viseme | Word | 0.0851 |
| Phoneme | Phoneme | 0.1980 |
| Phoneme | Word | 0.1980 |
| Word | Word | 0.1874 |

### 9.6.2 Linguistic content

The linguistic content of any dataset has an impact on a computer lip-reading classification performance. Stylised texts have more structure and restrictions on how a speech or utterance can be organised therefore classification becomes a simpler task. In our case, with the RMAV dataset we have the challenge of lip-reading continuous speech, this is much more difficult as the complexity of the task grows with the size of the variability in what is being said, in what order and how.

As part of the classification task, we ask where does the error rate come from? Which phonemes/visemes are currently recognisable? By this we mean, are there some phonemes which help the classification task more than others, can a classifier place more weight on these phonemes to improve their classification performance?

Within HTK classification, grammar networks built on probability statistics of the training data have a priori knowledge of linguistic content at a word or phoneme level to improve lip-reading classification. But when considering natural continuous speech, this makes a word or phoneme/viseme network exceptionally large in order to permit any order combination of utterances. Likewise, a higher-order N-gram language model may improve classification rates but the cost of this model is disproportionate to our intention to develop better classifiers.

Dictionaries help to define the vocabulary to be recognised but in natural speech what happens when a word is uttered which is not previously known? A new slang term for example. A new entry is required to be made up of phonemes which already exist.

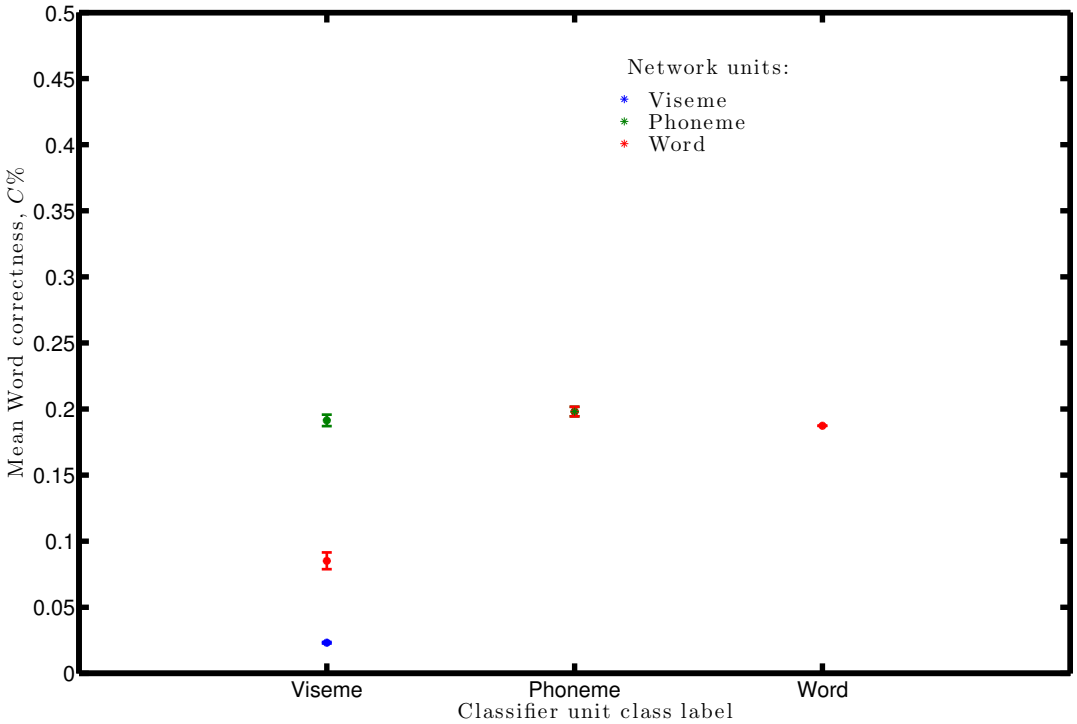



Figure 9.17: Effects of support network unit choice with varying HMM classifier units (along the $x$-axis) measured in all speaker mean correctness, $C$. Units supported by a viseme network are shown in blue, phoneme networks are in green and word networks in red. All {HMM,network} pairings are shown in Table 9.4.

The effects of the network units are shown in Figure 9.17 which plots the HMM units on the $x$-axis against the classification in Correctness $C$ (defined in [152]).

Table 9.5: All-speaker error counts for different combinations of units for HMM classifiers with bigram support networks. HMM units run vertically and network units run horizontally through the table.

| | Viseme | Phoneme | Word |
|---|---|---|---|
| Viseme | 0.0005 | 0.0043 | 0.0063 |
| Phoneme | - | 0.0036 | 0.0036 |
| Word | - | - | 0.0 |

Error bars show one standard error. Using a viseme network shows the worst classification. This can be attributed to the volume of homophones introduced by translating from words to phonemes to visemes. We no longer consider this option. More interesting are the word and phoneme networks. The phoneme network greatly improves classification for viseme HMMs, more so than a word network. When we use phoneme HMMs, there is no difference at all between an phoneme or word network and the standard error is identical. Thus we use both phoneme and word networks in our final method.

## 9.7 Effects of weak learning in viseme classifier training

In analysing our results, it must be remembered whilst our HMM training is hierarchical, our testing is not. Figure 9.18 shows the mean Correctness, $C$, for all speakers $\pm 1\frac{\sigma}{\sqrt{10}}$ over 10 folds. There are four lines plotted subject to the pairings of our HMM unit labels and the language network unit.

The $x$-axis of Figure 9.18 is the size of the viseme sets from Figure 9.15 from 11 to 36. We remind the reader this is the range of optimal number of visemes where phoneme label classifiers do not improve classification. The baseline of viseme classification with a word network from [15] is shown in blue and is not significantly different from conventionally learned phoneme classifiers. Based on our unit selection for language network study in section 9.6.1, it is not a surprise to see just by

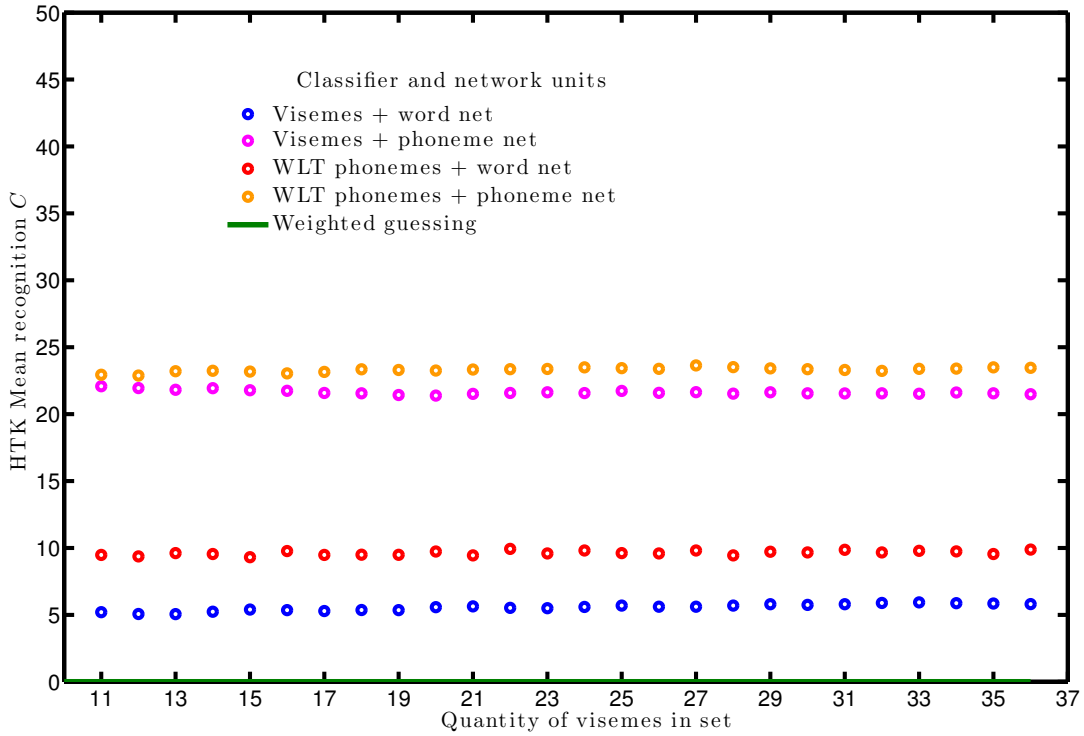



Figure 9.18: HTK Correctness $C$ for viseme classifiers with either phoneme or word language models and weak learned phoneme classifiers with either phoneme or word language models averaged over all 12 speakers.

using a phoneme network instead of a word network to support viseme classification we significantly improve our mean correctness score for all viseme set sizes for all speakers (shown in pink). Guessing is repeated as per our first previous experiments in this chapter.

Table 9.6: Minimum and maximum all speaker mean correctness, $C$, showing the effect of weak learning on phoneme labelled HMM classification.

| | Min | Max | Range |
|---|---|---|---|
| Visemes + word net | 0.0274 | 0.0601 | 0.0327 |
| Phonemes + word net | 0.0905 | 0.0995 | 0.0090 |
| Effect of WLT | 0.0631 | 0.0394 | – |
| Visemes + phoneme net | 0.2036 | 0.2214 | 0.0179 |
| Phonemes + phoneme net | 0.2253 | 0.2367 | 0.0114 |
| Effect of WLT | 0.0217 | 0.0153 | – |

More interesting to see is our new weakly-trained phoneme HMMs are significantly better than the viseme HMMs. In the original work of [15] phoneme HMMs

gave an all-speaker mean $C = 0.059$ . Here, regardless of the size of the original viseme set, $C$ is almost double. Weakly learnt phoneme classifiers with a word network gain 0.0313 to 0.0403 in mean $C$, and when these phoneme classifiers are supported with a phoneme network we see a correctness gain range from 0.1661 to 0.1775. These gains are supported by the all speaker mean minimum and maximums listed in Table 9.6. These gain scores are from over all the potential viseme-to-phoneme mappings and show there is little difference in which phoneme-to-viseme map is best for knowing which set of visemes to initialise our phoneme classifiers. All results, including the baseline, are significantly better than guessing (shown in green).

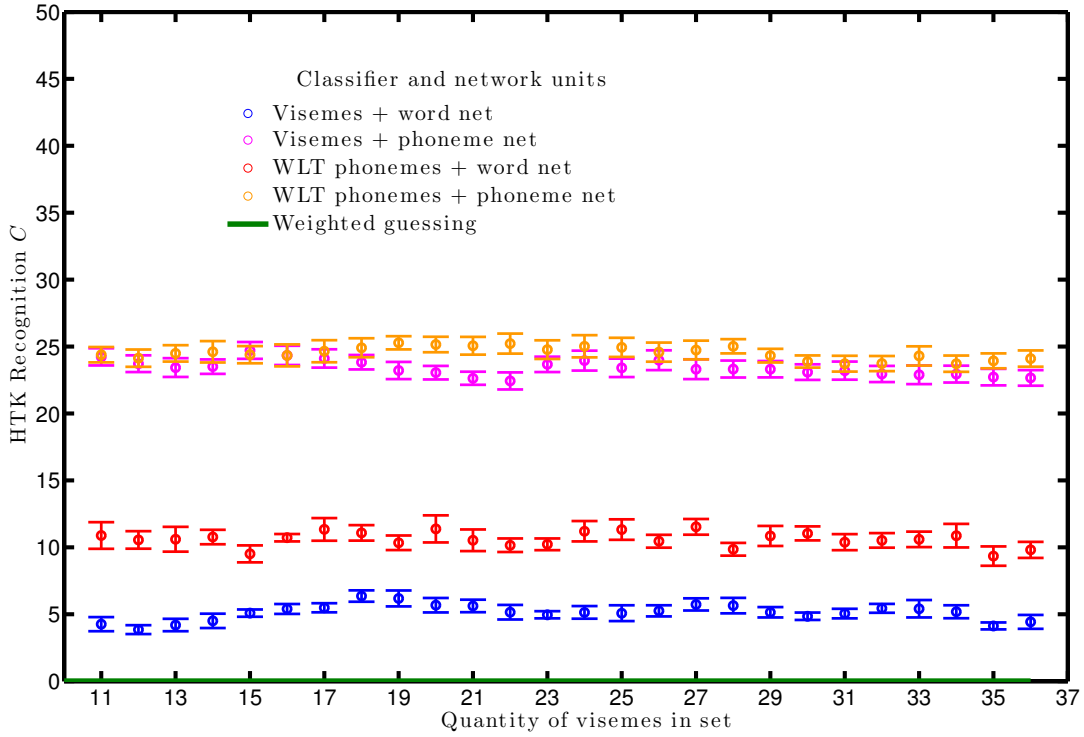



Figure 9.19: Speaker 1 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

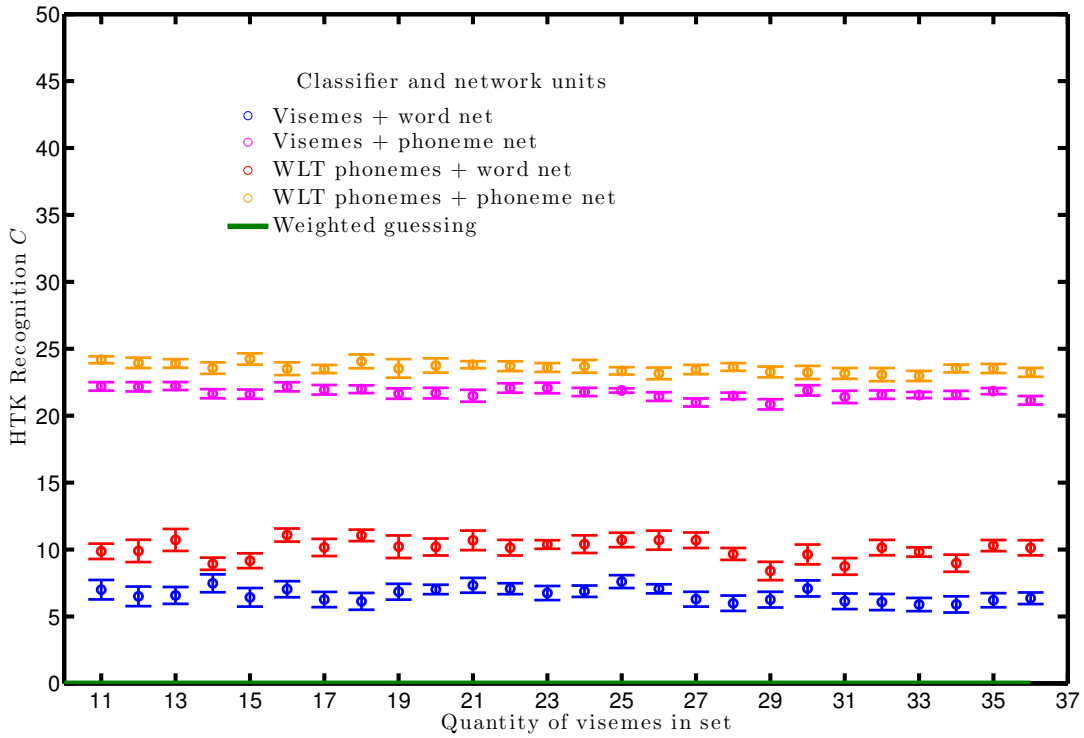



Figure 9.20: Speaker 2 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

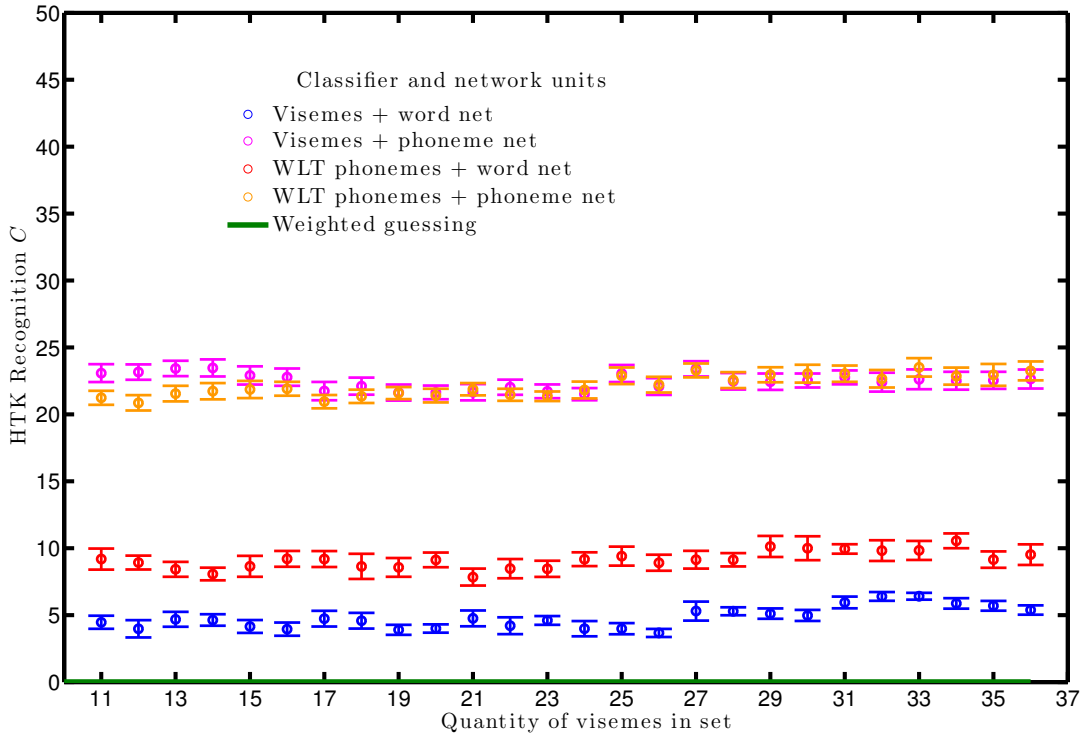



Figure 9.21: Speaker 3 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

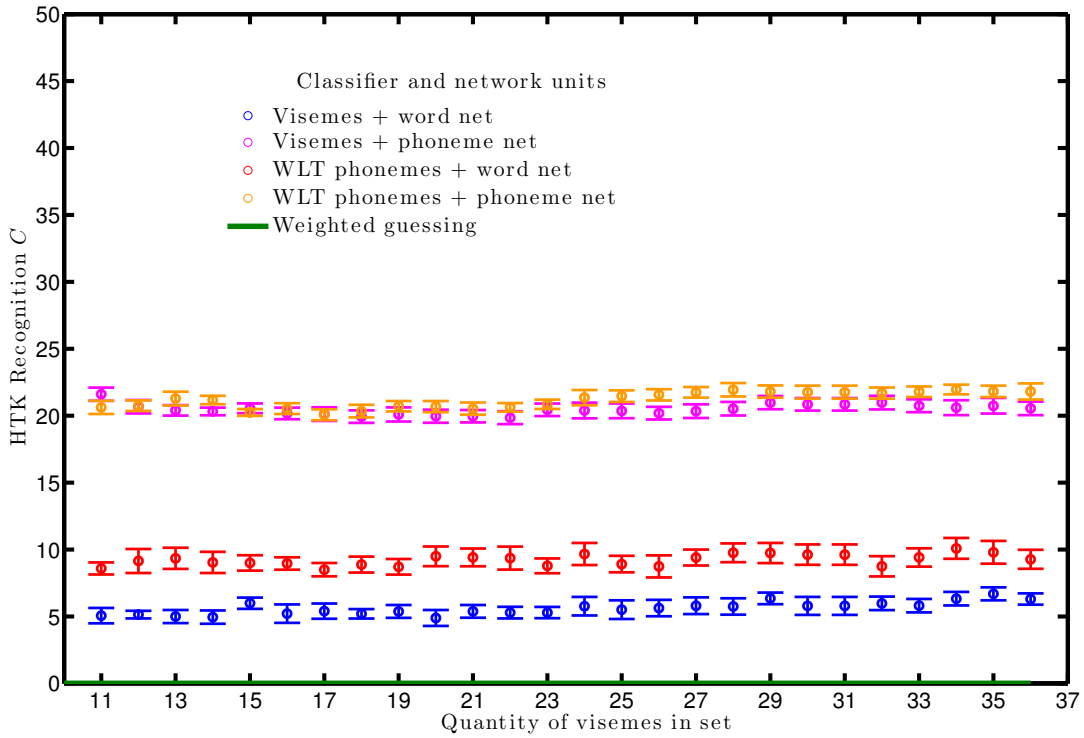



Figure 9.22: Speaker 4 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

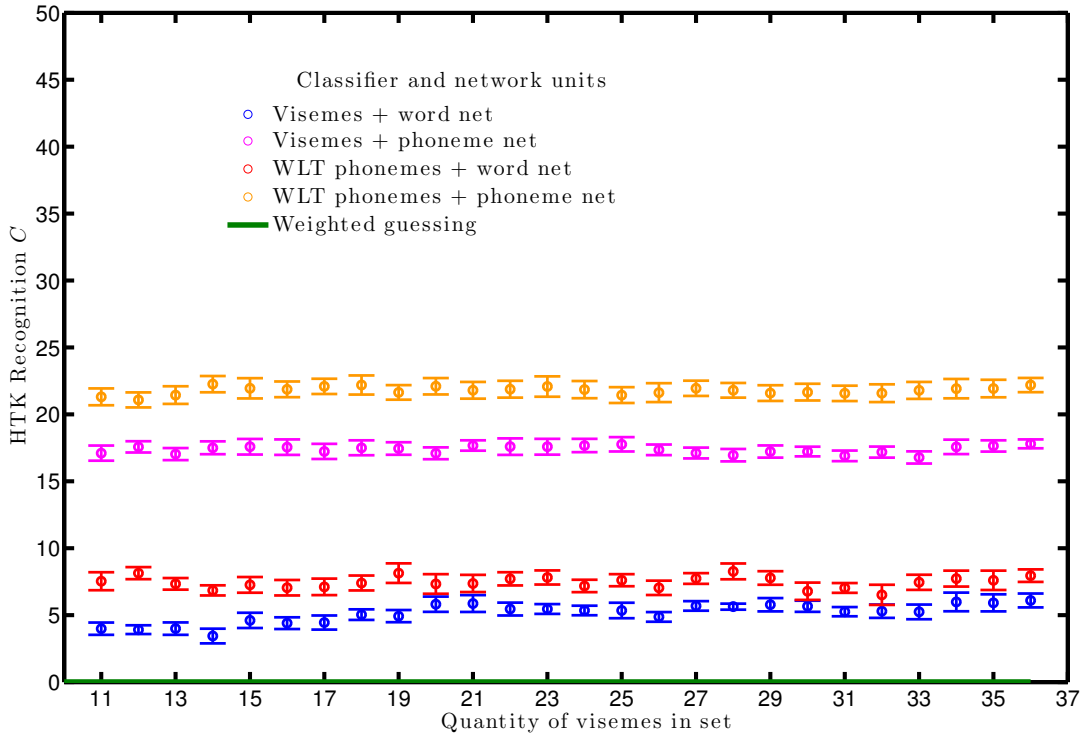



Figure 9.23: Speaker 5 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

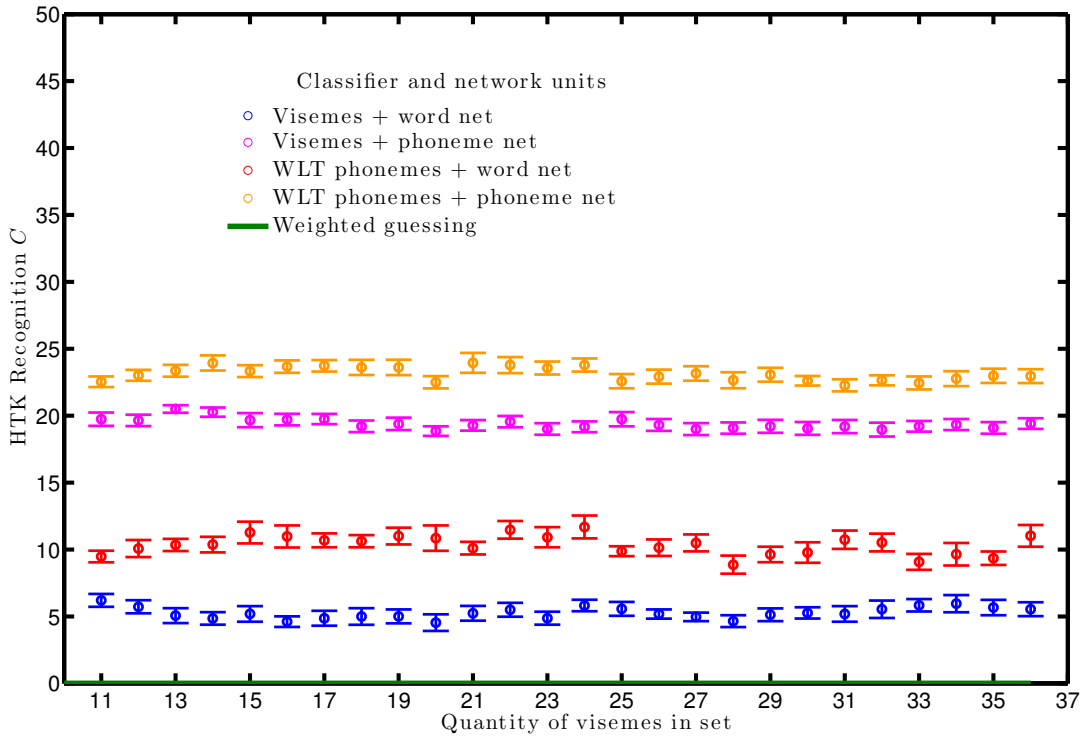



Figure 9.24: Speaker 6 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

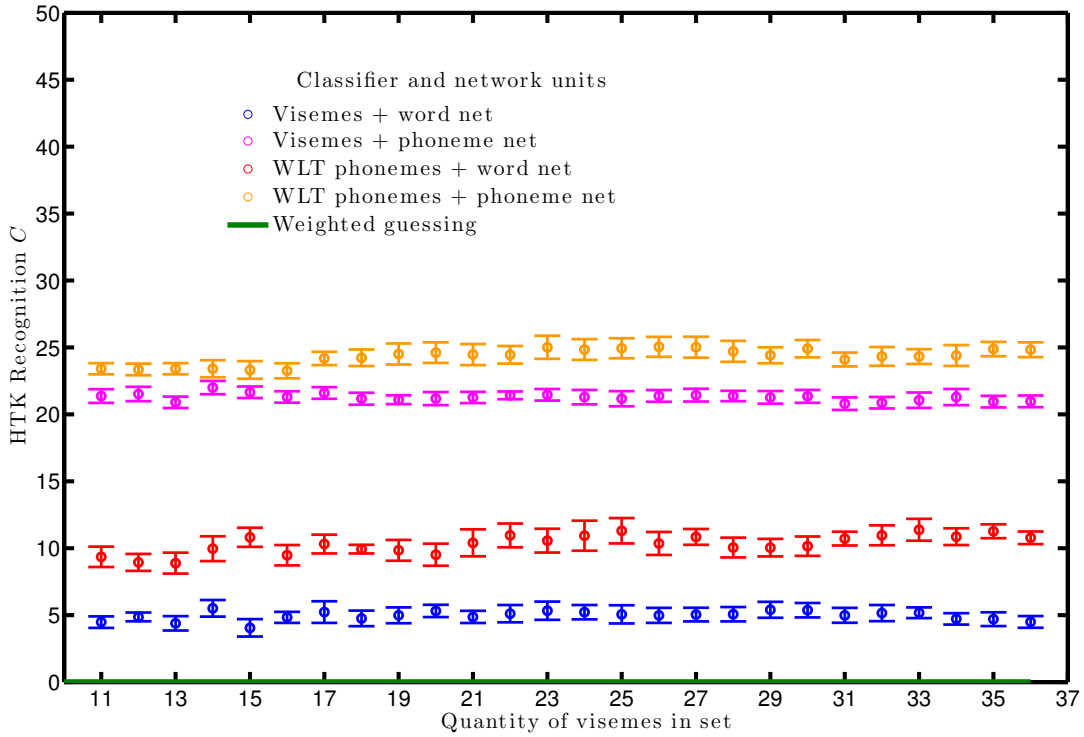



Figure 9.25: Speaker 7 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

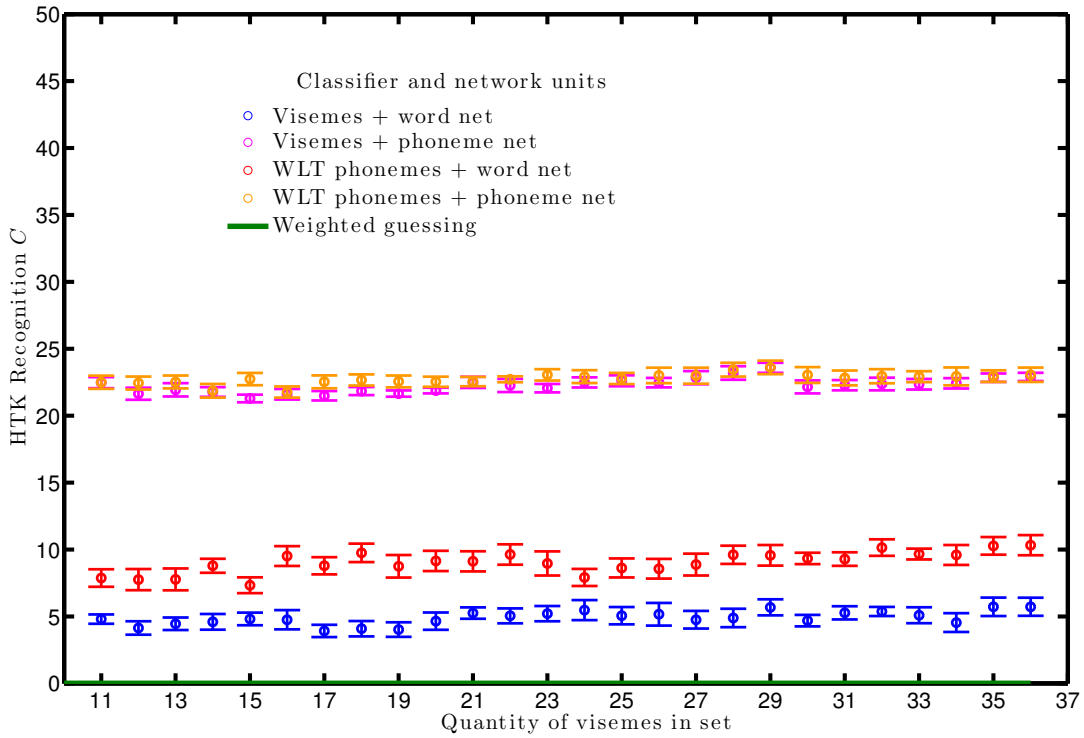



Figure 9.26: Speaker 8 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

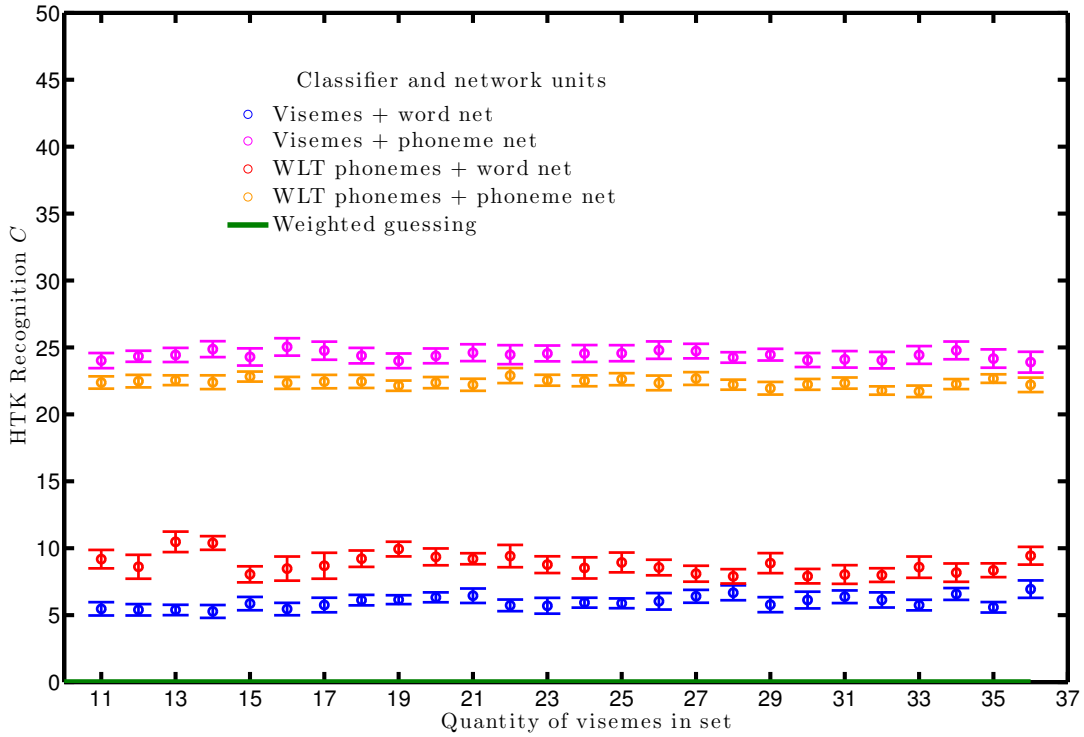



Figure 9.27: Speaker 9 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

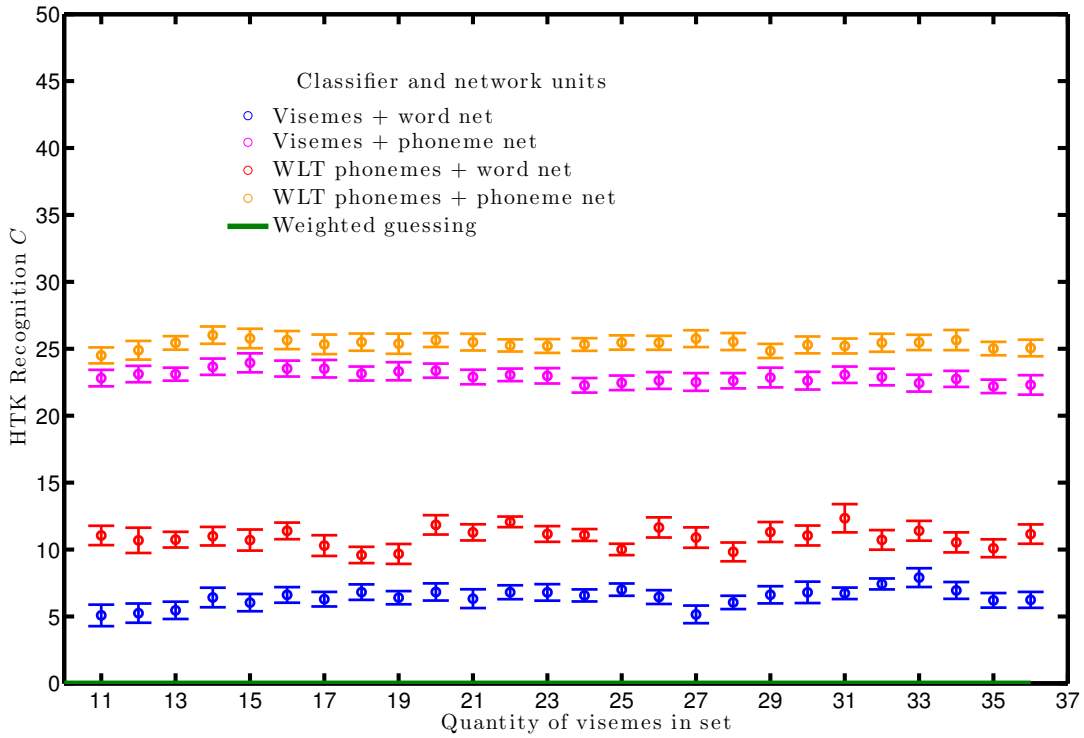



Figure 9.28: Speaker 10 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

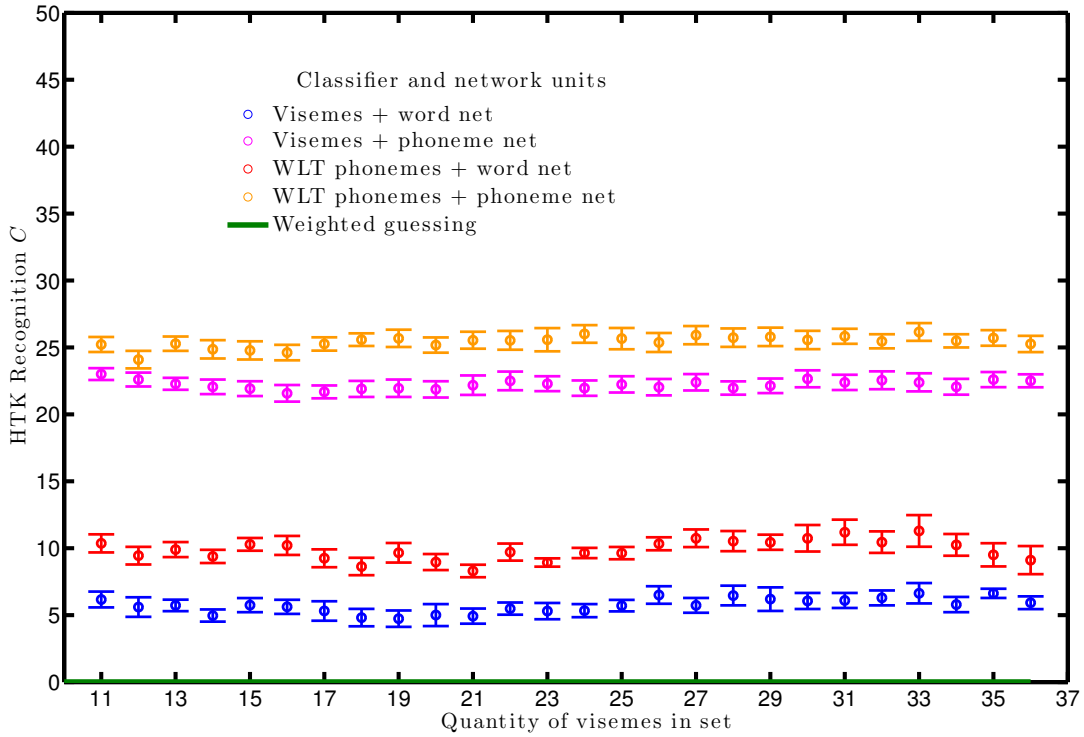



Figure 9.29: Speaker 11 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

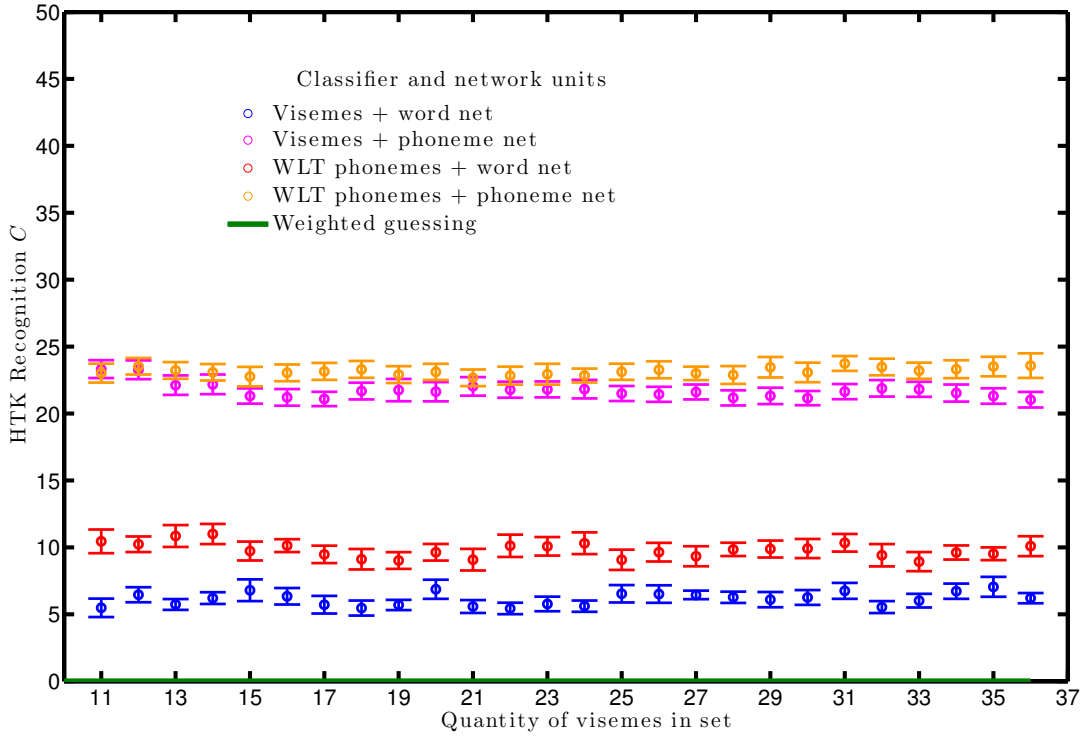



Figure 9.30: Speaker 12 correctness of viseme sets with a word language model (blue) and the weak learned phoneme classifiers with a phoneme or word network.

In Figures 9.19 - 9.30, we have plotted for our 12 speakers non-aggregated results showing $C \pm 1s.e.$ Whilst not monotonic, these graphs are much smoother than the speaker-dependent graphs shown in [15]. The significant differences between viseme set sizes shown in Figure 9.15 have now disappeared because the learning of differences between visemes, has been incorporated into the training of phoneme classifiers, which in turn are now better trained (plotted in red and orange which improve on blue and pink respectively).

Our speaker-dependent results with hierarchical learning are intriguing as in [76], with RMAV and published at the start of this thesis, showed an average viseme accuracy of ~46%. Here, we have presented a word accuracy (which we have previously shown to be weaker than viseme accuracy but more useful) of ~10%. We can not present viseme accuracy as our hierarchical training method has transformed the viseme classifiers into phoneme labelled classifiers, but reporting phoneme accuracy provides us with ~25% classification. This is beneficial as phoneme transcripts are both more comprehensible due to less homophones, and reduces our dependency on the language model for comprehension.

An intriguing observation is comparing the use of a phoneme network for visemes and for weakly taught phonemes. For some speakers, the weakly learned phonemes are not always as important as having the right network unit. This is seen in Figures 9.19, 9.21, 9.22, 9.26, and 9.30 for Speaker's 1, 3, 4, 8 and 12. By rewatching the original videos to estimate the age of our speakers, we categorise them as either an 'older' or 'younger' speaker. The speakers with less significant difference in the effect of weak learning are younger. This implies to lip-read a younger person we need more support from the language model, than for an older speaker. Our own informal observation is young people have more co-articulation than older people, but this is something for further investigation.

## 9.8 Decoding visemes

This chapter has described a viseme derivation method which allows us to construct any number of visual units. The reader is reminded this is not a proposal of a new method for the best visemes, the priority objective in this case was a method for enabling comparison of viseme sets in a controlled manner.

The presence of an optimum number of visemes within a set of classes is the result of two competing effects. In the first, as the number of visemes shrinks the number of homophones rises and it becomes more difficult to recognise words (correctness drops). In the second, as the number of visemes rises sufficient training data is no longer available in order to learn the subtle differences in lip-shapes (if they exist), so again, correctness drops. Thus, in theory the optimum number of visual units lies beween 1 and 45. In practice we see this optimum is between the number of phonemes and twelve (the size of one of the smaller viseme sets).

The choice of visual units in lip-reading has caused some debate. Some researchers use visemes as adduced by, for example Fisher [48] (in which visemes are a theoretical construct representing phonemes should look identical on the lips [57]). Others have noted lip-reading using phonemes can give superior performance to visemes [67].

Here, we supply further evidence to the more nuanced hypothesis there are intermediary units, which for convenience we call visemes, that can provide superior performances provided they are derived by an analysis of the data.

Furthermore, we have presented a novel learning algorithm which shows improved performance for these new data-driven visemes when used in hierarchical classifier training. The essence of our method is to re-train the viseme models in a fashion similar to weak learning in order they become better phoneme-labelled classifiers. This produces significantly better classification and is our second augmentation to the lip-reading system. This is shown in Figure 9.31, the extra steps are the dash-edged boxes on the right hand side.

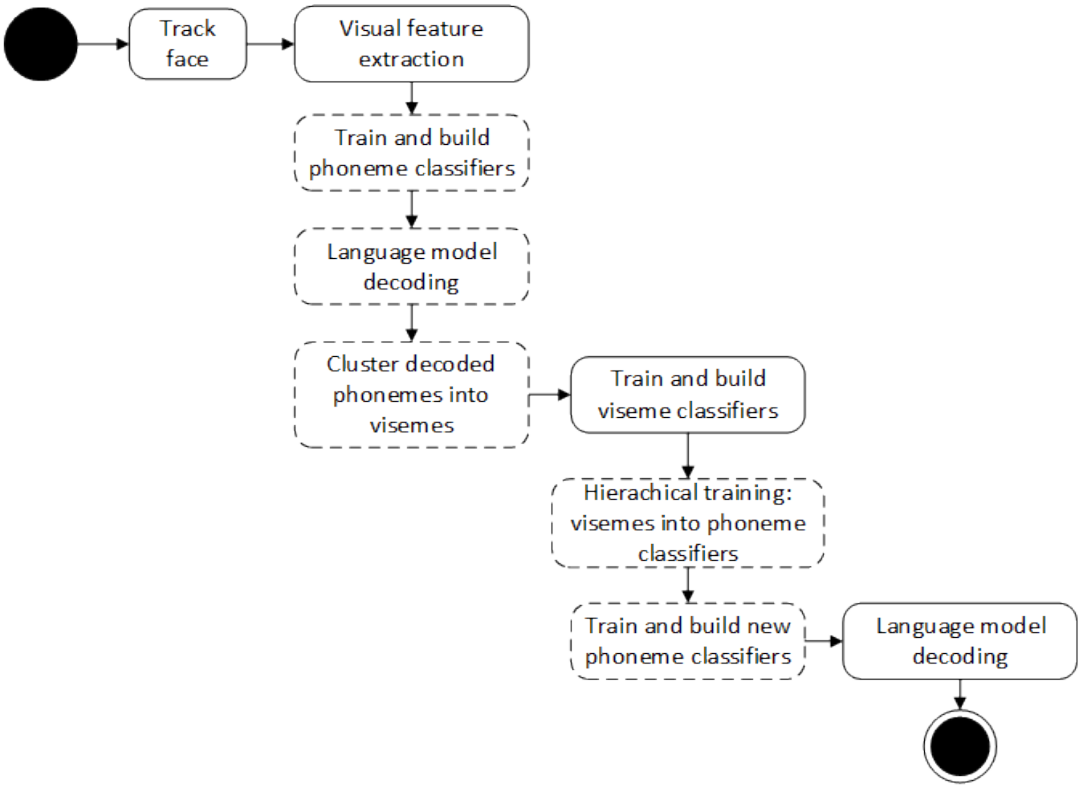



Figure 9.31: Second augmentation to the conventional lip-reading system to include hierarchical training of phoneme-labelled classifiers with visemes.

# Chapter 10

# Summary of research outputs

In this final chapter we summarise all we have learned throughout this thesis about decoding visemes.

## 10.1 Conclusions of research

Our original research question was how can we further understand visemes in order to augment or replace the current HMM classifiers in conventional automatic lip-reading systems? We have learnt through our experiments that:

There is a lower limit to the resolution at which a machine can lip-read, which is at least two pixels per lip. As long as videos of speakers have at least this then we can achieve some lip-reading. This is important because a high resolution video where a person's face is so far away the pixels per lip are less than two would be worse than a close up low resolution video [14].

There is also a limitation on how useful all speaker-independent (or multi-speaker) visemes within a set are towards the overall recognition. A badly trained viseme is worse than no viseme to represent certain phonemes [17]. When training visemes it is not enough to say we need more data, having bad training data is more detrimental to classification than having less.

In our comparison of many of the phoneme to viseme maps in literature we have seen there is little difference between each of them but Lee's marginally outperforms all others [16]. The majority of previous presented P2V maps have been designed from the observations of human lip-readers which are biased towards the individual perception of the human participating. The higher performing maps are more recent presentations and are data driven and/or machine trained.

When clustering phonemes into visemes we can say with confidence that vowel and consonant phonemes should be isolated. This was shown by our two methods for devising speaker-dependent visemes whereby one permitted mixing of all phonemes, and the second method restricted this clustering. The second method significantly outperformed the former. Speaker individuality is important in visual speech and should be recognised when devising viseme sets due to the variability with which different people use visual gestures whilst talking [16].

Viseme sets which are too small, (less than 11), are negatively affected by homophone confusions. The sets which are too large are not able to be trained sufficiently to achieve good classification. This means with the wrong speaker and training volume combination the size of the viseme set is fragile. We have shown a range of optimum sizes from 11 to 35 [12], and demonstrated how this varies by speaker and is higher than the phoneme-to-viseme maps previously presented in literature. We show for speaker dependent recognition there is a range of choices when selecting a set of visual units containing fewer members than the phoneme set, yet these sets outperform phoneme labelled classifiers. It is considered however, for speaker independent recognition, it is still most likely that phonemes are the desirable choice for classifier units as these are consistent across speakers.

Thus, in speaker-dependent recognition, the right visemes can not just outperform phoneme-labelled classifiers, but also when used to help train phoneme classifiers, they classify visual speech significantly better [13].

To support good classifiers we have seen the effect of different unit labels in the supporting language network. Best results are achieved when the unit labels are

the same for both classifiers and the network, but classification is not significantly affected if not. Therefore, for the purposes of decoding phonemes back to the words spoken, the preferred network unit is words [13].

In conclusion, and to answer our research question, we have improved machine lip-reading by adopting the current HMM classification system to use speaker-specific phoneme confusions within our new clustering algorithm to produce speaker-dependent viseme sets which, in turn, make good prototype HMM classifiers to train phoneme labelled classifiers. These, together with a word labelled language network, mean we can decode visemes to improve machine lip-reading classification.

## 10.2 Future work

Machine lip-reading is a large and complicated problem. There are many sub-problems which need to be solved within this challenge to achieve high, consistent classification. Remaining problems can be grouped into three classes using the Van Trees categories of detection, classification and estimation [141]:

- Some detection problems remaining are: automatically finding a face in an image and re-identifying same speakers between cameras, when is a person speaking/not speaking? Is the face occluded?
- Classification problems which remain include: Classification rates still need further improvement to be considered robust and speaker independence between the classifier training and test data is yet to produce good results.
- Finally estimation problems such as: how fast is a person speaking? What shapes do the lips form? And what is the the velocity of the lip movement? still need to be addressed.

Speech recognition is a maturing field of research, but if we refer to our introduction (Section 1) we remember our motivation to improve machine lip-reading

classification is for two major reasons. Firstly the use of such a system would be applicable in a range of areas from entertainment (e.g. sports events) to criminal detection (e.g. CCTV recordings). Secondly, and this is the main expectation of a lip-reading system, is the integration of such a system into AVSR. A robust lip-reading system could both improve the robustness and accuracy of an AVSR system by better use of the visual channel alone (channel independence) and as a fallback during times when the audio signal drops or is deteriorated by noise. This goal raises a number of significant questions which extend beyond the demands of achieving visual-only speech recognition. Audio-visual signal fusion, environmental noise, camera/microphone movement are three examples of further challenges in AVSR [155].

It is a difficult problem to classify acoustic utterances from a signal of sparse visual cues, so whilst acoustic recognition is achieving ubiquity with commercial applications (in 2015 Google's Translate app added speech recognition as a novel feature to assist travellers to communicate abroad), machine lip-reading is yet to achieve the same level of robustness. Independence to speaker identity, camera view, occlusions and language all still need to be robustly accomplished before we see such technology as a reality.

# Bibliography


[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transfom. *IEEE transactions on Computers*, pages 90–93, 1974.

[2] Ibrahim Almajai, Ben Milner, and Jonathan Darch. Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise. In *INTERSPEECH*, pages 2470–2473. ISCA, 2006.

[3] Anonymous. Zidane head butt story. `http://www.football-bible.com/soccer-info/zidane-head-butt-story.html`, 2006. Accessed: September 2015.

[4] S. Antar, A. Sagheer, S. Aly, and M. F. Tolba. Avas: Speech database for multimodal recognition applications. In *Hybrid Intelligent Systems (HIS), 2013 13th International Conference on*, pages 123–128, Dec 2013.

[5] International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[6] Jan Baetens. The silent movie revisited. *Online magazine of the visual narrative*, 6, Feb 2003.

[7] Simon Baker. Inverse compositional algorithm. In Katsushi Ikeuchi, editor, *Computer Vision*, pages 426–428. Springer US, 2014.

[8] J Andrew Bangham, Richard Harvey, Paul D Ling, and Richard V Aldridge. Nonlinear scale-space from n-dimensional sieves. In *Computer Vision, ECCV'96*, pages 187–198. Springer, 1996.

[9] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato. Sift features tracking for video stabilization. In *14th International Conference on Image Analysis and Processing (ICIAP)*, pages 825–830, Sept 2007.

[10] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, pages 164–171, 1970.

[11] Sara L. Bauman and Georgia Hambrecht. Analysis of view angle used in speechreading training of sentences. *American Journal of Audiology*, 4(3):67–70, 1995.

[12] Helen L Bear, Stephen J Cox, and Richard Harvey. Speaker independent machine lip reading with speaker dependent viseme classifiers. In *1st Joint International Conference on Facial Analysis, Animation and Audio-Visual Speech Processing (FAAVSP)*, pages 115–120. ISCA, 2015.

[13] Helen L Bear and Richard Harvey. Decoding visemes: improving machine lip-reading. In *41st International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[14] Helen L Bear, Richard Harvey, Barry-John Theobald, and Yuxuan Lan. Resolution limits on visual speech recognition. In *Image Processing (ICIP), IEEE International Conference on*, pages 1371–1375. IEEE, 2014.

[15] Helen L Bear, Richard Harvey, Barry-John Theobald, and Yuxuan Lan. Finding phonemes: improving machine lip-reading. In *1st Joint International Conference on Facial Analysis, Animation and Audio-Visual Speech Processing (FAAVSP)*, pages 190–195. ISCA, 2015.

[16] Helen L Bear, Richard W Harvey, Barry-John Theobald, and Yuxuan Lan. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In *Advances in Visual Computing*, pages 230–239. Springer, 2014.

[17] Helen L Bear, Gari Owen, Richard Harvey, and Barry-John Theobald. Some observations on computer lip-reading: moving from the dream to the reality. In *SPIE Security+ Defence*, pages 92530G–92530G. International Society for Optics and Photonics, 2014.

[18] Joseph C Beaver. A Grammar of Prosody. *College English*, pages 310–321, 1968.

[19] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *The Journal of Machine Learning Research*, 5:1089–1105, 2004.

[20] Christian Benoit, Thierry Guiard-Marigny, B Le Goff, and Ali Adjoudani. *Which components of the face do humans and machines best speechread?*, volume 150, pages 315–328. Berlin: NATO-ASI Series 150 Springer, 1996.

[21] Carl A Binnie, Pamela L Jackson, and Allen A Montgomery. Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders*, 41(4):530–539, 1976.

[22] Art Blokland and Anne H Anderson. Effect of low frame-rate video on intelligibility of speech. *Speech Communication*, 26(12):97–103, 1998.

[23] Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in knowledge discovery and data mining*, pages 3–12. Springer, 2004.

[24] Richard Bowden, Stephen Cox, Richard Harvey, Yuxuan Lan, Eng-Jon Ong, Gari Owen, and Barry-John Theobald. Recent developments in automated lip-reading. In *Symposium of Photonics and Intelligent Engineering (SPIE) Security+ Defence*, pages 89010J–89010J. International Society for Optics and Photonics, 2013.

[25] Elif Bozkurt, CE Erdem, Engin Erzin, Tanju Erdem, and M Ozkan. Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. *Proceedings of Signal Processing and Communications Applications*, pages 1–4, 2007.

[26] Cambridge University, UK. The BEEP Pronunciation Dictionary - British English. `ftp://svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/`, 1997. Accessed: January 2013.

[27] L. Cappelletta and N. Harte. Viseme definitions comparison for visual-only speech recognition. In *Signal Processing Conference, 2011 19th European*, pages 2109–2113, Aug 2011.

[28] Luca Cappelletta and Naomi Harte. Phoneme-to-viseme mapping for visual speech recognition. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 322–329, 2012.

[29] Carnegie Mellon University. CMU pronounciation dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict#lookup`, 2008. Accessed: April 2012.

[30] Tsuhan Chen and Ram R Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[31] Reuben Conrad. Lip-reading by deaf and hearing children. *British Journal of Educational Psychology*, 47(1):60–65, 1977.

[32] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[33] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681 –685, June 2001.

[34] Claire Cowie. The accents of outsourcing: The meanings of neutral in the Indian call centre industry. *World Englishes*, 26(3):316–330, 2007.

[35] Stephen Cox, Richard Harvey, Yuxuan Lan, Jacob Newman, and Barry Theobald. The challenge of multispeaker lip-reading. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 179–184, 2008.

[36] James L Crowley and Francois Berard. Multi-modal tracking of faces for video communications. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 640–645. IEEE, 1997.

[37] R. Cutler and L. Davis. Look who's talking: speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 1589–1592, 2000.

[38] KH Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

[39] Janez Demar. Statistical comparisons of classifiers over multiple datasets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[40] Barbara Dodd, Geoff Plant, and Mark Gregory. Teaching lip-reading: The efficacy of lessons on video. *British journal of audiology*, 23(3):229–238, 1989.

[41] Harris Drucker, Corinna Cortes, Lawrence D Jackel, Yann LeCun, and Vladimir Vapnik. Boosting and other ensemble methods. *Neural Computation*, 6(6):1289–1301, 1994.

[42] Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37:36–48, 1983.

[43] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247 – 261, 1989.

[44] Norman P Erber. Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40(4):481, 1975.

[45] Nicolas Eveno, Alice Caplier, and Pierre-Yves Coulon. Accurate and quasi-automatic lip tracking. *Circuits and Systems for Video technology, IEEE Transactions on*, 14(5):706–715, 2004.

[46] Irene Rosetta Ewing. *Lipreading and hearing aids*. Manchester University Press, 1944.

[47] Kathleen E Finn and Allen A Montgomery. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.

[48] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of Speech, Language and Hearing Research*, 11(4):796, 1968.

[49] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall. The DARPA speech recognition research database: specifications and status. In *Proceedings of the DARPA Workshop on speech recognition*, pages 93–99, 1986.

[50] J Richard Franks and Joan Kimble. The confusion of english consonant clusters in lipreading. *Journal of Speech, Language and Hearing Research*, 15(3):474, 1972.

[51] Georgios Galatas, Gerasimos Potamianos, Alexandros Papangelis, and Fillia Makedon. Audio visual speech recognition in noisy visual environments. In *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*, page 19. ACM, 2011.

[52] Consuelo Gonzalez. Consuelo gonzalez, professional lip reader. `http://www.lipreadingtranslation.com/translator.html`, 2016. Accessed: March 2016.

[53] John N Gowdy, Amarnag Subramanya, Chris Bartels, and Jeff Bilmes. Dbn based multi-stream models for audio-visual speech recognition. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–993. IEEE, 2004.

[54] J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[55] C.J. Hall. *An introduction to language and linguistics: breaking the language spell*. Open linguistics series. Continuum, 2005.

[56] N. Harte and E. Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, May 2015.

[57] Timothy J Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):1082–1089, 2006.

[58] Timothy J. Hazen, Kate Saenko, Chia-Hao La, and James R. Glass. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 235–242, New York, NY, USA, 2004. ACM.

[59] Martin Heckmann, Frédéric Berthommier, Christophe Savariaux, and Kristian Kroschel. Effects of image distortions on audio-visual speech recognition. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

[60] Martin Heckmann, Kristian Kroschel, Christophe Savariaux, Frédéric Berthommier, et al. Dct-based video features for audio-visual speech recognition. In *INTERSPEECH*, 2002.

[61] F Heider and Grace M Heider. An experimental investigation of lipreading. *Psychological Monographs*, 52(232):124–153, 1940.

[62] Jim L Hieronymus, D McKelvie, and FR McInnes. Use of acoustic sentence level and lexical stress in hsmm speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 225–227. IEEE, 1992.

[63] Sarah Hilder, Richard Harvey, and Barry-John Theobald. Comparison of human and machine lip-reading. *Proceedings of the International Conference on Audio-Visual Speech Processing (AVSP)*, pages 86–89, 2009.

[64] Wendy Holmes. *Speech synthesis and recognition*. CRC press, 2001.

[65] Xiaopeng Hong, Hongxun Yao, Yuqi Wan, and Rong Chen. A pca based visual dct feature extraction method for lip-reading. In *Intelligent Information Hiding and Multimedia Signal Processing, 2006. IIH-MSP'06. International Conference on*, pages 321–326. IEEE, 2006.

[66] Dominic Howell, Barry-John Theobald, and Stephen Cox. Confusion modelling for automated lip-reading using weighted finite-state transducers. In *Auditory-Visual Speech Processing (AVSP)*, pages 197–202, 2013.

[67] Dominic Liam Howell. *Confusion Modelling for Lip-Reading. PhD thesis*. University of East Anglia, 2014.

[68] Fu Jie Huang and Tsuhan Chen. Tracking of multiple faces for human-computer interfaces and virtual environments. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 1563–1566, 2000.

[69] Janet Jeffers and Margaret Barley. *Speechreading (lipreading)*. Thomas Springfield, IL:, 1971.

[70] Jintao Jiang, Abeer Alwan, Lyme E. Bernstein, Edward T. Auer, and Patricia A. Keating. Similarity structure in perceptual and physical measures for visual consonants across talkers. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–441 –I–444, May 2002.

[71] Robert Kaucic and Andrew Blake. Accurate, real-time, unadorned lip tracking. In *Sixth International Conference on Computer Vision.*, pages 370–375. IEEE, 1998.

[72] Tristan Kleinschmidt, David Dean, Sridha Sridharan, and Michael Mason. A continuous speech recognition evaluation protocol for the avicar database. 2008.

[73] Patricia B Kricos and Sharon A Lesner. Differences in visual intelligibility across talkers. *The Volta Review*, 84:219–225, 1982.

[74] K. Kumar, Tsuhan Chen, and R.M. Stern. Profile view lip reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–429–IV–432, 2007.

[75] Kush Kumar, Tsuhan Chen, and Richard M Stern. Profile view lip reading. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–429. IEEE, 2007.

[76] Yuxuan Lan, Richard Harvey, and Theobald Barry-John. Insights into machine lip reading. In *38th International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[77] Yuxuan Lan, Richard Harvey, B Theobald, Eng-Jon Ong, and Richard Bowden. Comparing visual features for lipreading. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 102–106, 2009.

[78] Yuxuan Lan, B.-J. Theobald, and R. Harvey. View independent computer lipreading. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 432–437, 2012.

[79] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden. Improving visual features for lip-reading. *Proceedings of the International Conference on Audio-Visual Speech Processing (AVSP)*, 7(3):42–48, 2010.

[80] Jeff Lander. Read my lips: Facial animation techniques. `http://www.gamasutra.com/view/feature/131587/read_my_lips_facial_animation_.php`. Accessed: 2014-01-28.

[81] Ryan Layne, Tim Hospedales, and Shaogang Gong. Re-id: Hunting attributes in the wild. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[82] Soonkyu Lee and DongSuk Yook. Audio-to-visual conversion using hidden markov models. In *Proceedings of Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 563–570. Springer, 2002.

[83] Charay Lerdsudwichai and Mohamed Abdel-Mottaleb. Algorithm for multiple faces tracking. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–777. IEEE, 2003.

[84] S.A. Lesner and P.B Kricos. Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology*, 14:252–258, 1981.

[85] Luhong Liangi, Yu Luo, Feiyue Huang, and A. V. Nefian. A multi-stream audio-video large-vocabulary mandarin chinese speech database. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 3, pages 1787–1790 Vol.3, June 2004.

[86] Bernice Eisman Lott and Joel Levy. The influence of certain communicator characteristics on lip reading efficiency. *The Journal of Social Psychology*, 51(2):419–425, 1960.

[87] P. Lucey and G. Potamianos. Lipreading using profile versus frontal views. In *2006 IEEE Workshop on Multimedia Signal Processing*, pages 24–28, Oct 2006.

[88] Patrick Lucey, Gerasimos Potamianos, and Sridha Sridharan. Visual speech recognition across multiple views. *Visual Speech Reognition: Lip Segmentation and Mapping*, 2009.

[89] J. Luettin, N.A. Thacker, and S.W. Beet. Speaker identification by lipreading. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)*, volume 1, pages 62–65, 1996.

[90] Juergen Luettin and Neil A Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[91] Björn Lyxell. *Beyond lips: Components of speechreading skill. PhD Thesis.* Umeå universitet, 1989.

[92] Alison MacLeod and Quentin Summerfield. Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, 21(2):131–141, 1987.

[93] Kenji Mase and Alex Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991.

[94] D. Massaro. *Perceiving Talking Faces.* The MIT Press, 1998.

[95] Jiří Matas, Karel Zimmermann, Tomáš Svoboda, and Adrian Hilton. Learning efficient linear predictors for motion estimation. In *Computer Vision, Graphics and Image Processing*, pages 445–456. Springer, 2006.

[96] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(2):198 –213, Feb 2002.

[97] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[98] Iain Matthews, J Bangham, Richard Harvey, and Stephen Cox. Non-linear scale decomposition based features for visual speech recognition. *Proceedings of the IX European Signal Processing Conference (EUSIPCO)*, pages 303–305, 1998.

[99] Iain Matthews, Gerasimos Potamianos, Chalapathy Neti, and Juergen Luettin. A comparison of model and transform-based visual features for audio-visual lvcsr. In *null*, page 210. IEEE, 2001.

[100] Matthew McGrath. *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces.* PhD thesis, University of Nottingham, 1985.

[101] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[102] Stephen McKenna and Shaogang Gong. Tracking faces. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 271–276. IEEE, 1996.

[103] Neil Midgley. New technology catches hitler off guard. `http://www.telegraph.co.uk/news/uknews/1534830/New-technology-catches-Hitler-off-guard.html`, 2006. Accessed: September 2015.

[104] Ben Milner and Danny Websdale. Analysing the importance of different visual feature coefficients. *FAAVSP 2015*, 2015.

[105] Allen A Montgomery and Pamela L Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, 73:2134, 1983.

[106] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541 – 558, 2011.

[107] Ara V Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1–15, 2002.

[108] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, Azad Mashari, and Jie Zhou. Audio-visual speech recognition. In *Final Workshop 2000 Report*, volume 764, 2000.

[109] EB Nichie. Lipreading principles and practice, 1912.

[110] E B Nitchie. *Lip-Reading, principles and practise: A handbook for teaching and self-practise*. Frederick A Stokes Co, New York, 1912.

[111] E.J. Ong and R. Bowden. Robust lip-tracking using rigid flocks of selected linear predictors. In *8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[112] E.J. Ong and R. Bowden. Robust facial feature tracking using shape-constrained multi-resolution selected linear predictors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1844–1859, 2011.

[113] Eng-Jon Ong, Yuxuan Lan, Barry Theobald, R. Harvey, and R. Bowden. Robust facial feature tracking using selected multi-resolution linear predictors. In *IEEE 12th International Conference on Computer Vision*, pages 1483–1490, Sept 2009.

[114] Elmer Owens and Barbara Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28(3):381, 1985.

[115] Alasdair Palmer. Lip reader saw fraser's incriminating conversations. `http://www.telegraph.co.uk/news/uknews/1420816/Lip-reader-saw-{F}rasers-incriminating-conversations.html`, 2003. Accessed: September 2015.

[116] Adrian Pass, Jianguo Zhang, and Darryl Stewart. An investigation into features for multi-view lipreading. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2417–2420. IEEE, 2010.

[117] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2017. IEEE, 2002.

[118] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha. Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 129–136, 2013.

[119] Gerasimos Potamianos, Hans Peter Graf, and Eric Cosatto. An image transform approach for hmm based automatic lipreading. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, pages 173–177. IEEE, 1998.

[120] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing*, 22:23, 2004.

[121] Patrick F Quinn. The critical mind of Edgar Poe: Claude Richard. Edgar Allan Poe: Journaliste et critique. *Poe Studies-Old Series*, 13(2):37–40, 1980.

[122] Jessica Rees. Jessica rees, forensic and professional lip reader. `http://www.lipreadingtranslation.com/translator.html`, 2016. Accessed: March 2016.

[123] Jerker Rönnberg, Stefan Samuelsson, and Björn Lyxell. Conceptual constraints in sentence-based lipreading in the hearing-impaired. *Hearing by eye: II. The psychology of speechreading and auditory–visual speech*, pages 143–153, 1998.

[124] Takeshi Saitoh and Ryosuke Konishi. A study of influence of word lip reading by change of frame rate. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, 2010.

[125] Stefan Samuelsson and Jerker Rönnberg. Script activation in lipreading. *Scandinavian journal of psychology*, 32(2):124–143, 1991.

[126] Stefan Samuelsson and Jerker Ronnberg. Implicit and explicit use of scripted constraints in lip-reading. *European Journal of Cognitive Psychology*, 5(2):201–233, 1993.

[127] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[128] Karl Schwerdt and James L Crowley. Robust face tracking using color. In *Proceedings of the fourth IEEE International Conference on Automatic Face and Gesture Recognition.*, pages 90–95. IEEE, 2000.

[129] AYAZ A. SHAIKH, DINESH K. KUMAR, and JAYAVARDHANA GUBBI. Visual speech recognition using optical flow and support vector machines. *International Journal of Computational Intelligence and Applications*, 10(02):167–187, 2011.

[130] Cambridge University Speech. The BEEP Pronunciation Dictionary - British English. `ftp://svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/`, 2007. Accessed: March 2013.

[131] M. Stommel, M. Langer, O. Herzog, and K.-D. Kuhnert. A fast, robust and low bit-rate representation for sift and surf features. In *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 278–283, Nov 2011.

[132] David G Stork and Marcus E Hennecke. *Speechreading by humans and machines: models, systems, and applications*, volume 150. Springer, 1996.

[133] Quentin Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 335(1273):71–78, 1992.

[134] S. Taylor, B.-J. Theobald, and I. Matthews. The effect of speaking rate on audio and visual speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3037–3041, May 2014.

[135] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284. Eurographics Association, 2012.

[136] Herrie ten Cate. Battle of the somme. the true story., 2006. Video Documentary.

[137] Barry J. Theobald, Richard Harvey, Stephen J. Cox, Colin Lewis, and Gari P. Owen. Lip-reading enhancement for law enforcement. In *Optics and Photonics for Counterterrorism and Crime Fighting II*, volume 6402, pages 640205–640214. SPIE, 2006.

[138] Barry-John Theobald. *Visual Speech Synthesis Using Shape and Appearance Models. PhD thesis*. University of East Anglia, 2003.

[139] Tadafusa Tomitaka. Human face tracking system, 1995. US Patent 5,430,809.

[140] unknown. ffmpeg, a complete, cross-platform solution to record, convert and stream audio and video. `https://www.ffmpeg.org/`, 2012-2015. Accessed: January 2012.

[141] Harry L Van Trees. *Detection, estimation, and modulation theory*. Wiley. com, 2004.

[142] Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.

[143] Melanie Vitkovitch and Paul Barber. Visible speech as a function of image quality: Effects of display parameters on lipreading ability. *Applied Cognitive Psychology*, 10(2):121–140, 1996.

[144] Brian E Walden, Robert A Prosek, Allen A Montgomery, Charlene K Scherr, and Carla J Jones. Effects of training on the visual recognition of consonants. *Journal of Speech, Language and Hearing Research*, 20(1):130, 1977.

[145] Alan Wee-Chung Liew and Shilin Wang. *Visual speech recognition: lip segmentation and mapping*. Information Science Reference, 2009.

[146] Xiaozhou Wei, Lijun Yin, Zhiwei Zhu, and Qiang Ji. Avatar-mediated face tracking and lip reading for human computer interaction. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 500–503, New York, NY, USA, 2004. ACM.

[147] Yee Wan Wong, Sue Inn Chng, Kah Phooi Seng, Li-Minn Ang, Siew Wen Chin, Wei Jen Chew, and King Hann Lim. A new multi-purpose audio-visual unmc-vier database with multiple variabilities. *Pattern Recognition Letters*, 32(13):1503 – 1510, 2011.

[148] Mary F Woodward and Carroll G Barber. Phoneme perception in lipreading. *Journal of Speech, Language and Hearing Research*, 3(3):212, 1960.

[149] Jie Yang and Alex Waibel. A real-time face tracker. In *Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on*, pages 142–147. IEEE, 1996.

[150] S Young, G Evermann, M Gales, T Hain, D Kershaw, X A Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchec, and P. Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[151] S. J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.2*. Cambridge University Engineering Department, Cambridge, UK, 2002.

[152] Steve J Young, Gunnar Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev, and PC Woodland. The HTK book version 3.4, 2006.

[153] Jerrold H Zar. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580, 1972.

[154] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265, Nov 2009.

[155] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen. A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9):590–605, 2014.

# Appendices

# Appendix A

# Edgar Allen Poe's, The Raven

Once upon a midnight dreary, while I pondered, weak and weary,
Over many a quaint and curious volume of forgotten lore-
While I nodded, nearly napping, suddenly there came a tapping,
As of some one gently rapping, rapping at my chamber door.
'Tis some visitor, I muttered, tapping at my chamber door-
Only this and nothing more.

Ah, distinctly I remember it was in the bleak December;
And each separate dying ember wrought its ghost upon the floor.
Eagerly I wished the morrow;-vainly I had sought to borrow
From my books surcease of sorrow-sorrow for the lost Lenore-
For the rare and radiant maiden whom the angels name Lenore-
Nameless here for evermore.

And the silken, sad, uncertain rustling of each purple curtain
Thrilled me-filled me with fantastic terrors never felt before;
So that now, to still the beating of my heart, I stood repeating
'Tis some visitor entreating entrance at my chamber door-
Some late visitor entreating entrance at my chamber door;-
This it is and nothing more.

Presently my soul grew stronger; hesitating then no longer,
Sir, said I, or Madam, truly your forgiveness I implore;
But the fact is I was napping, and so gently you came rapping,
And so faintly you came tapping, tapping at my chamber door,
That I scarce was sure I heard you-here I opened wide the door;-
Darkness there and nothing more.

Deep into that darkness peering, long I stood there wondering, fearing,
Doubting, dreaming dreams no mortal ever dared to dream before;
But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word, Lenore?
This I whispered, and an echo murmured back the word, Lenore!-
Merely this and nothing more.

Back into the chamber turning, all my soul within me burning,
Soon again I heard a tapping somewhat louder than before.
Surely, said I, surely that is something at my window lattice;
Let me see, then, what thereat is, and this mystery explore-
Let my heart be still a moment and this mystery explore;-
'Tis the wind and nothing more!

Open here I flung the shutter, when, with many a flirt and flutter,
In there stepped a stately Raven of the saintly days of yore;
Not the least obeisance made he; not a minute stopped or stayed he;
But, with mien of lord or lady, perched above my chamber door-
Perched upon a bust of Pallas just above my chamber door-
Perched, and sat, and nothing more.

Then this ebony bird beguiling my sad fancy into smiling,
By the grave and stern decorum of the countenance it wore,
Though thy crest be shorn and shaven, thou, I said, art sure no craven,
Ghastly grim and ancient Raven wandering from the Nightly shore-

Tell me what thy lordly name is on the Night's Plutonian shore!
Quoth the Raven Nevermore.

Much I marvelled this ungainly fowl to hear discourse so plainly,
Though its answer little meaning-little relevancy bore;
For we cannot help agreeing that no living human being
Ever yet was blessed with seeing bird above his chamber door-
Bird or beast upon the sculptured bust above his chamber door,
With such name as Nevermore.

But the Raven, sitting lonely on the placid bust, spoke only
That one word, as if his soul in that one word he did outpour.
Nothing farther then he uttered-not a feather then he fluttered-
Till I scarcely more than muttered Other friends have flown before-
On the morrow he will leave me, as my Hopes have flown before.
Then the bird said Nevermore.

Startled at the stillness broken by reply so aptly spoken,
Doubtless, said I, what it utters is its only stock and store
Caught from some unhappy master whom unmerciful Disaster
Followed fast and followed faster till his songs one burden bore-
Till the dirges of his Hope that melancholy burden bore
Of Never-nevermore'.

But the Raven still beguiling all my fancy into smiling,
Straight I wheeled a cushioned seat in front of bird, and bust and door;
Then, upon the velvet sinking, I betook myself to linking
Fancy unto fancy, thinking what this ominous bird of yore-
What this grim, ungainly, ghastly, gaunt, and ominous bird of yore
Meant in croaking Nevermore.

This I sat engaged in guessing, but no syllable expressing

To the fowl whose fiery eyes now burned into my bosom's core;
This and more I sat divining, with my head at ease reclining
On the cushion's velvet lining that the lamp-light gloated o'er,
But whose velvet-violet lining with the lamp-light gloating o'er,
She shall press, ah, nevermore!

Then, methought, the air grew denser, perfumed from an unseen censer
Swung by Seraphim whose foot-falls tinkled on the tufted floor.
Wretch, I cried, thy God hath lent thee-by these angels he hath sent thee
Respite-respite and nepenthe from thy memories of Lenore;
Quaff, oh quaff this kind nepenthe and forget this lost Lenore!
Quoth the Raven Nevermore.

Prophet! said I, thing of evil!-prophet still, if bird or devil!-
Whether Tempter sent, or whether tempest tossed thee here ashore,
Desolate yet all undaunted, on this desert land enchanted-
On this home by Horror haunted-tell me truly, I implore-
Is there-is there balm in Gilead?-tell me-tell me, I implore!
Quoth the Raven Nevermore.

Prophet! said I, thing of evil!-prophet still, if bird or devil!
By that Heaven that bends above us-by that God we both adore-
Tell this soul with sorrow laden if, within the distant Aidenn,
It shall clasp a sainted maiden whom the angels name Lenore-
Clasp a rare and radiant maiden whom the angels name Lenore.
Quoth the Raven Nevermore.

Be that word our sign of parting, bird or fiend! I shrieked, upstarting
Get thee back into the tempest and the Night's Plutonian shore!
Leave no black plume as a token of that lie thy soul hath spoken!
Leave my loneliness unbroken!-quit the bust above my door!
Take thy beak from out my heart, and take thy form from off my door!

Quoth the Raven Nevermore.

And the Raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted -nevermore!

# Appendix B

# Phonetic notation

Table B.1: For translating vowel phonemes from phonetic symbols to their respective alphabet character representations

| Phonetic Symbol | Character Symbol | Latexipa | Example |
|---|---|---|---|
| /ai/ | /ay/ | /ai/ | bo**uy** |
| /ʌ/ | /ah/ | textturnv | **hu**t |
| /æ/ | /ae/ | ae | p**a**n |
| /ə/ | /ax/ | textschwa | **a**lbeit |
| /ɑʊ/ | /aw/ | textscripta textupsilon | cl**ou**d |
| /ɔ/ | /ao/ | textopeno | s**ou**r |
| /ɑ/ | /aa/ | textscripta | c**ar**d |
| /ei/ | /ey/ | /ei/ | st**ay** |
| /e/ | /eh/ | /e/ | dw**e**ll |
| /ɜ/ | /er/ | textrevepsilon | c**ur**t |
| /ɛ/ | /ea/ | {E} | cha**ir** |
| /i/ | /iy/ | /i/ | cr**ee**d |
| /ɪ/ | /ih/ | textsci | k**i**d |
| /ɪə/ | /ia/ | textsci textschwa | l**ea**r |
| /ɪ/ | /ix/ | textsci | **i**ll |
| /ɔɪ/ | /oy/ | textopeno textsci | c**oy** |
| /əʊ/ | /ow/ | textschwa textupsilon | c**o**de |
| /ʊə/ | /oo/ | textupsilon textschwa | p**ru**de |
| /ɔ/ | /oa/ | textopeno | g**oa**t |
| /ɔə/ | /ou/ | textopeno textschwa | p**ou**r |
| /ɒ/ | /oh/ | textturnscripta | t**o**t |
| /u/ | /uw/ | /u/ | c**ue** |
| /ʊ/ | /uh/ | textupsilon | f**oo**d |
| /ɔə/ | /ua/ | textopeno textschwa | c**or**e |

Table B.2: For translating consonant phonemes from phonetic symbols to their respective alphabet character representations

| Phonetic Symbol | Character Symbol | Latex textipa | Example |
|---|---|---|---|
| /θ/ | /th/ | {T} | **th**in |
| /ð/ | /dh/ | {D} | **th**ere |
| /ʃ/ | /sh/ | {S} | **sh**eer |
| /ʒ/ | /zh/ | {Z} | vi**s**ual |
| /dʒ/ | /jh/ | d{Z} | **j**udge |
| /tʃ/ | /ch/ | t{S} | **ch**runch |
| /ɦ/ | /H/ (or /hh/) | {H} | **h**unt |
| /ŋ/ | /ng/ | {N} | ki**ng** |
| /ɯ/ | /W/ | {W} | **wh**isky |
| /b/ | /b/ | /b/ | **b**ar |
| /d/ | /d/ | /d/ | **d**art |
| /f/ | /f/ | /f/ | **f**ete |
| /g/ | /g/ | /g/ | **g**reat |
| /h/ | /hh/ | /h/ | **h**unt |
| /k/ | /k/ | /k/ | **c**ane |
| /l/ | /l/ | /l/ | **l**ake |
| /m/ | /m/ | /m/ | **m**other |
| /n/ | /n/ | /n/ | **n**one |
| /p/ | /p/ | /p/ | **p**ot |
| /r/ | /r/ | /r/ | g**r**ate |
| /s/ | /s/ | /s/ | **s**ilk |
| /t/ | /t/ | /t/ | **t**ack |
| /v/ | /v/ | /v/ | **v**erge |
| /w/ | /w/ | /w/ | **w**eed |
| /y/ | /y/ | /y/ | **y**aught |
| /z/ | /z/ | /z/ | **z**ulu |

# Appendix C

# Example confusion matrices

| | /v01/ | /v02/ | /v03/ | /v04/ | /v05/ | /v06/ | /v07/ | /v08/ | /v09/ | /v10/ | /sil/ | /gar/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| /v01/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /v02/ | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /v03/ | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /v04/ | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /v05/ | 0 | 1 | 1 | 0 | 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| /v06/ | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| /v07/ | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 1 |
| /v08/ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 |
| /v09/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| /v10/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 |
| /sil/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 52 | 0 |
| /gar/ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 4 |

Figure C.1: An example viseme confusion matrix from AVL2 Speaker 3, fold 2 classification output with Fishers viseme set)

| | /aa/ | /ah/ | /ax/ | /ay/ | /b/ | /ey/ | /eh/ | /f/ | /iy/ | /jh/ | /l/ | /m/ | /ow/ | /s/ | /sil/ | /t/ | /v/ | /w/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| /aa/ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ah/ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ax/ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ay/ | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| /b/ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ch/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /d/ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| /eh/ | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ey/ | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /f/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /iy/ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| /jh/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /k/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /l/ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /m/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| /n/ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ow/ | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /p/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /s/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| /sil/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 52 | 0 | 0 | 0 |
| /t/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| /uw/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| /v/ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /w/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /y/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /z/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure C.2: An example phoneme confusion matrix from AVL2 Speaker 2, fold 2 phoneme classification output)

# Appendix D

# RMAV P2V maps

Table D.1: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 1 and 2

| Speaker 1 $M_1$ | | Speaker 2 $M_2$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ae/ /ax/ /eh/ /ɜ/ /ey/ /ɪ/ /iy/ | /v01/ | /əʊ/ |
| /v02/ | /ɔ/ /ɒ/ /əʊ/ | /v02/ | /ax/ /ay/ /eh/ /ɜ/ /ɪ/ /iy/ /oh/ |
| /v03/ | /ɪə/ /ɔə/ | /v03/ | /ʌ/ /ɛ/ /ey/ |
| /v04/ | /ʊ/ | /v04/ | /ɑʊ/ /ɔə/ |
| /v05/ | /ʌ/ /ɛ/ | /v05/ | /ɪə/ |
| /v06/ | /ɑ/ /ay/ | /v06/ | /ɑ/ /ae/ /ɔ/ |
| /v07/ | /ua/ | /v07/ | /ʊ/ |
| /v08/ | /ɔɪ/ | /v08/ | /ua/ |
| /v09/ | /ə/ | /v09/ | /ɔɪ/ |
| /v10/ | /ɑʊ/ | /v10/ | /b/ /l/ /m/ /n/ /p/ /r/ /s/ /ʃ/ /t/ /v/ /w/ /z/ |
| /v11/ | /d/ /ð/ /f/ /dʒ/ /k/ /l/ /m/ /n/ /p/ /s/ | /v11/ | /d/ /ð/ /f/ /g/ /dʒ/ /k/ /ng/ |
| /v12/ | /ŋ/ /t/ /θ/ /v/ /z/ | /v12/ | /hh/ /y/ |
| /v13/ | /ʃ/ | /v13/ | /tʃ/ /θ/ |
| /v14/ | /r/ /w/ /y/ | /v14/ | /ʒ/ |
| /v15/ | /b/ /g/ /hh/ | /sil/ | /sil/ |
| /v16/ | /tʃ/ | /sp/ | /sp/ |
| /sil/ | /sil/ | /gar/ | /ə/ /c/ |
| /sp/ | /sp/ | | |
| /gar/ | /ʒ/ /c/ | | |

Table D.2: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 3 and 4

| Speaker 3 $M_3$ | | Speaker 4 $M_4$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɔ/ /ax/ /eh/ /ɜ/ /ey/ /ɪ/ /iy/ /oh/ /ow/ | /v01/ | /əʊ/ |
| /v02/ | /ʊ/ | /v02/ | /ae/ /ɔ/ /ax/ /eh/ /ɜ/ /ey/ /ih/ /iy/ /oh/ |
| /v03/ | /ay/ /ɛ/ /ɔə/ | /v03/ | /ay/ /ɪə/ /ɔə/ |
| /v04/ | /ɪə/ | /v04/ | /ʌ/ /ɑʊ/ |
| /v05/ | /ae/ /ʌ/ | /v05/ | /ɑ/ /ɛ/ |
| /v06/ | /ɪə/ | /v06/ | /ʊ/ |
| /v07/ | /ɑ/ | /v07/ | /ua/ |
| /v08/ | /ua/ | /v08/ | /k/ /l/ /m/ /n/ /p/ /r/ /s/ /t/ /v/ /z/ |
| /v09/ | /ə/ | /v09/ | /d/ /ŋ/ |
| /v10/ | /ɑʊ/ | /v10/ | /ð/ /f/ /g/ /w/ |
| /v11/ | /k/ /l/ /m/ /n/ /ŋ/ /p/ | /v11/ | /dʒ/ /ʃ/ |
| /v12/ | /f/ /r/ /s/ /ʃ/ /t/ /θ/ /w/ /y/ /z/ | /v12/ | /hh/ |
| /v13/ | /tʃ/ /d/ /ð/ /g/ | /v13/ | /tʃ/ /y/ |
| /v14/ | /hh/ /dʒ/ /v/ | /v14/ | /b/ /θ/ |
| /v15/ | /ʒ/ | /v15/ | /ʒ/ |
| /v16/ | /b/ | /sil/ | /sil/ |
| /sil/ | /sil/ | /sp/ | /sp/ |
| /sp/ | /sp/ | /gar/ | /ɔɪ/ /ə/ /c/ |
| /gar/ | /ɔɪ/ /c/ | | |

Table D.3: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 5 and 6

| Speaker 5 $M_5$ | | Speaker 6 $M_6$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /əʊ/ | /v01/ | /ɑ/ /ae/ /ɔ/ /ax/ /ay/ /ey/ /ih/ /uw/ |
| /v02/ | /ɔ/ /ax/ /ay/ /eh/ /ɜ/ /ey/ /ih/ /iy/ | /v02/ | /iy/ /ɒ/ /əʊ/ |
| /v03/ | /ae/ /ɔə/ | /v03/ | /ɜ/ |
| /v04/ | /ʊ/ | /v04/ | /eh/ |
| /v05/ | /ɑʊ/ /ua/ | /v05/ | /ɛ/ |
| /v06/ | /ʌ/ /ɒ/ | /v06/ | /ɑʊ/ |
| /v07/ | /ɛ/ | /v07/ | /ʌ/ |
| /v08/ | /ɑ/ /ɪə/ | /v08/ | /ʊ/ |
| /v09/ | /ə/ | /v09/ | /ɪə/ |
| /v10/ | /w/ | /v10/ | /ə/ |
| /v11/ | /y/ | /v11/ | /ð/ /f/ /hh/ /l/ /m/ /ŋ/ /p/ /r/ /s/ /t/ |
| /v12/ | /t/ /θ/ /z/ | /v12/ | /ʃ/ /v/ /y/ |
| /v13/ | /l/ /m/ /n/ /p/ /r/ /s/ /ʃ/ /v/ | /v13/ | /g/ /dʒ/ /k/ /z/ |
| /v14/ | /b/ /dʒ/ | /v14/ | /b/ /d/ /w/ |
| /v15/ | /g/ /hh/ | /v15/ | /tʃ/ /n/ |
| /v16/ | /ð/ /f/ /ŋ/ | /v16/ | /θ/ /ʒ/ |
| /v17/ | /tʃ/ /d/ /k/ | /sil/ | /sil/ |
| /v18/ | /ʒ/ | /sp/ | /sp/ |
| /sil/ | /sil/ | /gar/ | /ɔɪ/ /c/ /ua/ |
| /sp/ | /sp/ | | |
| /gar/ | /ɔɪ/ /c/ | | |

Table D.4: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 7 and 8

| Speaker 7 $M_7$ | | Speaker 8 $M_8$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɛ/ /eh/ /ɜ/ | /v01/ | /ae/ /ʌ/ /ay/ /eh/ /ɪ/ /iy/ /ow/ /uh/ |
| /v02/ | /ae/ /ɔ/ /ax/ /ay/ /ey/ /ɪ/ /iy/ /oh/ | /v02/ | /ɔ/ /ax/ /ɛ/ /ɪə/ |
| /v03/ | /ɑ/ /əʊ/ /ɔə/ | /v03/ | /ɑ/ /ɑʊ/ /ey/ |
| /v04/ | /ʊ/ | /v04/ | /ua/ /ɔə/ |
| /v05/ | /ua/ | /v05/ | /ɜ/ /ɒ/ |
| /v06/ | /ɪə/ | /v06/ | /ə/ |
| /v07/ | /ɑʊ/ | /v07/ | /ɔɪ/ |
| /v08/ | /ʌ/ | /v08/ | /b/ /d/ /ð/ /f/ /k/ /l/ /m/ /n/ /p/ /r/ /s/ /t/ |
| /v09/ | /tʃ/ /d/ /ð/ /g/ /k/ /l/ /m/ /n/ /p/ /r/ /t/ | /v09/ | /ʃ/ /v/ /z/ |
| /v10/ | /ʃ/ | /v10/ | /dʒ/ /w/ /y/ |
| /v11/ | /s/ /v/ /w/ /y/ /z/ | /v11/ | /g/ |
| /v12/ | /b/ /ŋ/ /dʒ/ | /v12/ | /hh/ /θ/ |
| /v13/ | /f/ /θ/ | /v13/ | /tʃ/ /ŋ/ |
| /v14/ | /ŋ/ | /v14/ | /ʒ/ |
| /v15/ | /ʒ/ | /sil/ | /sil/ |
| /v16/ | /hh/ | /sp/ | /sp/ |
| /sil/ | /sil/ | /gar/ | /c/ |
| /sp/ | /sp/ | | |
| /gar/ | /ɔɪ/ /c/ /ə/ | | |

Table D.5: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 9 and 10

| Speaker 9 $M_9$ | | Speaker 10 $M_{10}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ae/ /ʌ/ /ey/ | /v01/ | /ax/ /ay/ /eh/ /ey/ /ɪ/ /iy/ /oh/ /ow/ |
| /v02/ | /ɔə/ | /v02/ | /ɑʊ/ /ɔə/ |
| /v03/ | /ɑ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ih/ /iy/ | /v03/ | /ʊ/ |
| /v04/ | /ɪə/ /əʊ/ | /v04/ | /ae/ /ʌ/ /ɔ/ /ɛ/ |
| /v05/ | /ɔ/ /ɒ/ | /v05/ | /ɜ/ /ɪə/ |
| /v06/ | /ɑʊ/ | /v06/ | /ɑ/ |
| /v07/ | /ʊ/ | /v07/ | /ua/ |
| /v08/ | /ua/ | /v08/ | /ə/ |
| /v09/ | /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /t/ /w/ | /v09/ | /k/ /l/ /m/ /n/ /p/ /r/ /s/ /ʃ/ /t/ /w/ /y/ /z/ |
| /v10/ | /tʃ/ | /v10/ | /g/ /θ/ /v/ |
| /v11/ | /d/ /ð/ /f/ /v/ | /v11/ | /tʃ/ /d/ /ð/ /f/ /hh/ |
| /v12/ | /ʃ/ | /v12/ | /b/ |
| /v13/ | /b/ /z/ | /v13/ | /ŋ/ |
| /v14/ | /ʃ/ | /v14/ | /ʒ/ |
| /v15/ | /hh/ | /v15/ | /dʒ/ |
| /v16/ | /y/ | /sil/ | /sil/ |
| /v17/ | /g/ /dʒ/ | /sp/ | /sp/ |
| /v18/ | /ʒ/ | /gar/ | /ɔɪ/ /c/ |
| /v19/ | /θ/ | | |
| /sil/ | /sil/ | | |
| /sp/ | /sp/ | | |
| /gar/ | /ɔɪ/ /ə/ /c/ | | |

Table D.6: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 11 and 12

| Speaker 11 $M_{11}$ | | Speaker 12 $M_{12}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɛ/ /ʊ/ | /v01/ | /ax/ /ay/ /eh/ /ey/ /ɪ/ /iy/ /ow/ /uw/ |
| /v02/ | /ae/ /eh/ /ɜ/ /ey/ /ɪ/ /iy/ /oh/ | /v02/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ɒ/ |
| /v03/ | /ʌ/ /ɔ/ /ax/ | /v03/ | /ɛ/ /ɪə/ /ɔɪ/ |
| /v04/ | /ay/ /ɔə/ | /v04/ | /ua/ |
| /v05/ | /ɪə/ /əʊ/ | /v05/ | /ʊ/ |
| /v06/ | /ɑʊ/ | /v06/ | /ɜ/ |
| /v07/ | /ɑ/ | /v07/ | /ɑʊ/ |
| /v08/ | /k/ /l/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /z/ | /v08/ | /w/ |
| /v09/ | /m/ /t/ /θ/ /v/ | /v09/ | /k/ /l/ /m/ /n/ /p/ /r/ /s/ /ʃ/ /t/ /th/ |
| /v10/ | /g/ | /v10/ | /v/ /ʒ/ /tʃ/ |
| /v11/ | /w/ | /v11/ | /y/ /b/ |
| /v12/ | /tʃ/ /dʒ/ | /v12/ | /d/ /ð/ /f/ /g/ /n/ /ŋ/ |
| /v13/ | /b/ /d/ /ð/ /f/ | /v13/ | /hh/ /dʒ/ /z/ |
| /v14/ | /hh/ /y/ | /sil/ | /sil/ |
| /v15/ | /ʒ/ | /sp/ | /sp/ |
| /sil/ | /sil/ | /gar/ | /c/ /ə/ |
| /sp/ | /sp/ | | |
| /gar/ | /ɔɪ/ /ua/ /c/ /ə/ | | |

Table D.7: A speaker-independent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 1 and 2

| Speaker 1 $M_{!1}$ | | Speaker 2 $M_{!2}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ | /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ |
| /v02/ | /ua/ /ʊ/ /ɔə/ | /v02/ | /ua/ /ʊ/ /ɔə/ |
| /v03/ | /ɑʊ/ | /v03/ | /ɑʊ/ |
| /v04/ | /ɔɪ/ | /v04/ | /ɔɪ/ |
| /v05/ | /ə/ | /v05/ | /ə/ |
| /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ | /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /v07/ | /ʒ/ | /v07/ | /ʒ/ |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /sp/ | /sp/ | /sp/ | /sp/ |
| /gar/ | /c/ | /gar/ | /c/ |

Table D.8: A speaker-independent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 3 and 4

| Speaker 3 $M_{!3}$ | | Speaker 4 $M_{!4}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ | /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ |
| /v02/ | /ua/ /ʊ/ /ɔə/ | /v02/ | /ua/ /ʊ/ /ɔə/ |
| /v03/ | /ɑʊ/ /ɔɪ/ | /v03/ | /ɑʊ/ |
| /v04/ | /ə/ | /v04/ | /ɔɪ/ |
| /v05/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ | /v05/ | /ə/ |
| /v06/ | /ʒ/ | /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /sil/ | /sil/ | /v07/ | /ʒ/ |
| /sp/ | /sp/ | /sil/ | /sil/ |
| /gar/ | /c/ | /sp/ | /sp/ |
| | | /gar/ | /c/ |

Table D.9: A speaker-independent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 5 and 6

| Speaker 5 $M_{!5}$ | | Speaker 6 $M_{!6}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ | /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ |
| /v02/ | /ua/ /ʊ/ /ɔə/ | /v02/ | /ua/ /ʊ/ /ɔə/ |
| /v03/ | /ɑʊ/ /ɔɪ/ | /v03/ | /ɑʊ/ |
| /v04/ | /ə/ | /v04/ | /ɔɪ/ |
| /v05/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ | /v05/ | /ə/ |
| /v06/ | /ʒ/ | /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /sil/ | /sil/ | /v07/ | /ʒ/ |
| /sp/ | /sp/ | /sil/ | /sil/ |
| /gar/ | /c/ | /sp/ | /sp/ |
| | | /gar/ | /c/ |

Table D.10: A speaker-independent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 7 and 8

| Speaker 7 $M_{!7}$ | | Speaker 8 $M_{!8}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ | /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ |
| /v02/ | /ua/ /ʊ/ /ɔə/ | /v02/ | /ʊ/ /ɔə/ |
| /v03/ | /ɑʊ/ | /v03/ | /ua/ |
| /v04/ | /ɔɪ/ | /v04/ | /ə/ |
| /v05/ | /ə/ | /v05/ | /ɑʊ/ /ɔɪ/ |
| /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ | /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /v07/ | /ʒ/ | /v07/ | /ʒ/ |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /sp/ | /sp/ | /sp/ | /sp/ |
| /gar/ | /c/ | /gar/ | /c/ |

Table D.11: A speaker-independent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 9 and 10

| Speaker 9 $M_{!9}$ | | Speaker 10 $M_{!10}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ | /v01/ | /ʊ/ |
| /v02/ | /ua/ /ʊ/ /ɔə/ | /v02/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ |
| /v03/ | /ɑʊ/ | /v03/ | /ɑʊ/ /ua/ /ɔə/ |
| /v04/ | /ɔɪ/ | /v04/ | /ɔɪ/ |
| /v05/ | /ə/ | /v05/ | /ə/ |
| /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ | /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /v07/ | /ʒ/ | /v07/ | /ʒ/ |
| /sil/ | /sil/ | /sil/ | /sil/ |
| /sp/ | /sp/ | /sp/ | /sp/ |
| /gar/ | /c/ | /gar/ | /c/ |

Table D.12: A speaker-independent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speakers 11 and 12

| Speaker 11 $M_{!11}$ | | Speaker 12 $M_{!12}$ | |
|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ | /v01/ | /ɑ/ /ae/ /ʌ/ /ɔ/ /ax/ /ay/ /ɛ/ /eh/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /ɒ/ /əʊ/ /ɔə/ |
| /v02/ | /ua/ /ʊ/ /ɔə/ | /v02/ | /ua/ /ʊ/ |
| /v03/ | /ɑʊ/ | /v03/ | /ə/ |
| /v04/ | /ɔɪ/ | /v04/ | /ɑʊ/ /ɔɪ/ |
| /v05/ | /ə/ | /v05/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ |
| /v06/ | /b/ /tʃ/ /d/ /ð/ /f/ /g/ /hh/ /dʒ/ /k/ /l/ /m/ /n/ /ŋ/ /p/ /r/ /s/ /ʃ/ /t/ /θ/ /v/ /w/ /y/ /z/ | /v06/ | /ʒ/ |
| /v07/ | /ʒ/ | /sil/ | /sil/ |
| /sil/ | /sil/ | /sp/ | /sp/ |
| /sp/ | /sp/ | /gar/ | /c/ |
| /gar/ | /c/ | | |

# Appendix E

# RMAV DSD&D Experiments

Table E.1: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 2

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p, q)$ |
| Sp1 | Sp1 | Sp2 | $M_1(1, 2)$ |
| Sp3 | Sp3 | Sp2 | $M_3(3, 2)$ |
| Sp4 | Sp4 | Sp2 | $M_4(4, 2)$ |
| Sp5 | Sp5 | Sp2 | $M_5(4, 2)$ |
| Sp6 | Sp6 | Sp2 | $M_6(4, 2)$ |
| Sp7 | Sp7 | Sp2 | $M_7(4, 2)$ |
| Sp8 | Sp8 | Sp2 | $M_8(4, 2)$ |
| Sp9 | Sp9 | Sp2 | $M_9(4, 2)$ |
| Sp10 | Sp10 | Sp2 | $M_{10}(10, 2)$ |
| Sp11 | Sp11 | Sp2 | $M_{11}(11, 2)$ |
| Sp12 | Sp12 | Sp2 | $M_{12}(12, 2)$ |

Table E.2: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 3

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp3 | $M_1(1,3)$ |
| Sp2 | Sp2 | Sp3 | $M_2(2,3)$ |
| Sp4 | Sp4 | Sp3 | $M_4(4,3)$ |
| Sp5 | Sp5 | Sp3 | $M_5(4,3)$ |
| Sp6 | Sp6 | Sp3 | $M_6(4,3)$ |
| Sp7 | Sp7 | Sp3 | $M_7(4,3)$ |
| Sp8 | Sp8 | Sp3 | $M_8(4,3)$ |
| Sp9 | Sp9 | Sp3 | $M_9(4,3)$ |
| Sp10 | Sp10 | Sp3 | $M_{10}(10,3)$ |
| Sp11 | Sp11 | Sp3 | $M_{11}(11,3)$ |
| Sp12 | Sp12 | Sp3 | $M_{12}(12,3)$ |

Table E.3: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 4

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp4 | $M_1(1,4)$ |
| Sp2 | Sp2 | Sp4 | $M_2(2,4)$ |
| Sp3 | Sp3 | Sp4 | $M_3(3,4)$ |
| Sp5 | Sp5 | Sp4 | $M_5(4,4)$ |
| Sp6 | Sp6 | Sp4 | $M_6(4,4)$ |
| Sp7 | Sp7 | Sp4 | $M_7(4,4)$ |
| Sp8 | Sp8 | Sp4 | $M_8(4,4)$ |
| Sp9 | Sp9 | Sp4 | $M_9(4,4)$ |
| Sp10 | Sp10 | Sp4 | $M_{10}(10,4)$ |
| Sp11 | Sp11 | Sp4 | $M_{11}(11,4)$ |
| Sp12 | Sp12 | Sp4 | $M_{12}(12,4)$ |

Table E.4: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 5

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp5 | $M_1(1,5)$ |
| Sp2 | Sp2 | Sp5 | $M_2(2,5)$ |
| Sp3 | Sp3 | Sp5 | $M_3(3,5)$ |
| Sp4 | Sp4 | Sp5 | $M_4(4,5)$ |
| Sp6 | Sp6 | Sp5 | $M_6(6,5)$ |
| Sp7 | Sp7 | Sp5 | $M_7(7,5)$ |
| Sp8 | Sp8 | Sp5 | $M_8(7,5)$ |
| Sp9 | Sp9 | Sp5 | $M_9(8,5)$ |
| Sp10 | Sp10 | Sp5 | $M_{10}(10,5)$ |
| Sp11 | Sp11 | Sp5 | $M_{11}(11,5)$ |
| Sp12 | Sp12 | Sp5 | $M_{12}(12,5)$ |

Table E.5: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 6

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp6 | $M_1(1,6)$ |
| Sp2 | Sp2 | Sp6 | $M_2(2,6)$ |
| Sp3 | Sp3 | Sp6 | $M_3(3,6)$ |
| Sp4 | Sp4 | Sp6 | $M_4(4,6)$ |
| Sp5 | Sp5 | Sp6 | $M_5(5,6)$ |
| Sp7 | Sp7 | Sp6 | $M_7(6,6)$ |
| Sp8 | Sp8 | Sp6 | $M_8(7,6)$ |
| Sp9 | Sp9 | Sp6 | $M_9(9,6)$ |
| Sp10 | Sp10 | Sp6 | $M_{10}(10,6)$ |
| Sp11 | Sp11 | Sp6 | $M_{11}(11,6)$ |
| Sp12 | Sp12 | Sp6 | $M_{12}(12,6)$ |

Table E.6: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 7

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp7 | $M_1(1,7)$ |
| Sp2 | Sp2 | Sp7 | $M_2(2,7)$ |
| Sp3 | Sp3 | Sp7 | $M_3(3,7)$ |
| Sp4 | Sp4 | Sp7 | $M_4(4,7)$ |
| Sp5 | Sp5 | Sp7 | $M_5(5,7)$ |
| Sp6 | Sp6 | Sp7 | $M_6(6,7)$ |
| Sp8 | Sp8 | Sp7 | $M_8(7,7)$ |
| Sp9 | Sp9 | Sp7 | $M_9(9,7)$ |
| Sp10 | Sp10 | Sp7 | $M_{10}(10,7)$ |
| Sp11 | Sp11 | Sp7 | $M_{11}(11,7)$ |
| Sp12 | Sp12 | Sp7 | $M_{12}(12,7)$ |

Table E.7: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 8

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp8 | $M_1(1,8)$ |
| Sp2 | Sp2 | Sp8 | $M_2(2,8)$ |
| Sp3 | Sp3 | Sp8 | $M_3(3,8)$ |
| Sp4 | Sp4 | Sp8 | $M_4(4,8)$ |
| Sp5 | Sp5 | Sp8 | $M_5(5,8)$ |
| Sp6 | Sp6 | Sp8 | $M_6(6,8)$ |
| Sp7 | Sp7 | Sp8 | $M_7(7,8)$ |
| Sp9 | Sp9 | Sp8 | $M_9(9,8)$ |
| Sp10 | Sp10 | Sp8 | $M_{10}(10,8)$ |
| Sp11 | Sp11 | Sp8 | $M_{11}(11,8)$ |
| Sp12 | Sp12 | Sp8 | $M_{12}(12,8)$ |

Table E.8: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 9

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp9 | $M_1(1,9)$ |
| Sp2 | Sp2 | Sp9 | $M_2(2,9)$ |
| Sp3 | Sp3 | Sp9 | $M_3(3,9)$ |
| Sp4 | Sp4 | Sp9 | $M_4(4,9)$ |
| Sp5 | Sp5 | Sp9 | $M_5(5,9)$ |
| Sp6 | Sp6 | Sp9 | $M_6(6,9)$ |
| Sp7 | Sp7 | Sp9 | $M_7(7,9)$ |
| Sp8 | Sp8 | Sp9 | $M_8(8,9)$ |
| Sp10 | Sp10 | Sp9 | $M_{10}(10,9)$ |
| Sp11 | Sp11 | Sp9 | $M_{11}(11,9)$ |
| Sp12 | Sp12 | Sp9 | $M_{12}(12,9)$ |

Table E.9: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 10

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp10 | $M_1(1,10)$ |
| Sp2 | Sp2 | Sp10 | $M_2(2,10)$ |
| Sp3 | Sp3 | Sp10 | $M_3(3,10)$ |
| Sp4 | Sp4 | Sp10 | $M_4(4,10)$ |
| Sp5 | Sp5 | Sp10 | $M_5(5,10)$ |
| Sp6 | Sp6 | Sp10 | $M_6(6,10)$ |
| Sp7 | Sp7 | Sp10 | $M_7(7,10)$ |
| Sp8 | Sp8 | Sp10 | $M_8(8,10)$ |
| Sp9 | Sp9 | Sp10 | $M_9(9,10)$ |
| Sp11 | Sp11 | Sp10 | $M_{11}(11,10)$ |
| Sp12 | Sp12 | Sp10 | $M_{12}(12,10)$ |

Table E.10: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 11

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp11 | $M_1(1,11)$ |
| Sp2 | Sp2 | Sp11 | $M_2(2,11)$ |
| Sp3 | Sp3 | Sp11 | $M_3(3,11)$ |
| Sp4 | Sp4 | Sp11 | $M_4(4,11)$ |
| Sp5 | Sp5 | Sp11 | $M_5(5,11)$ |
| Sp6 | Sp6 | Sp11 | $M_6(6,11)$ |
| Sp7 | Sp7 | Sp11 | $M_7(7,11)$ |
| Sp8 | Sp8 | Sp11 | $M_8(8,11)$ |
| Sp9 | Sp9 | Sp11 | $M_9(9,11)$ |
| Sp10 | Sp10 | Sp11 | $M_{10}(10,11)$ |
| Sp12 | Sp12 | Sp11 | $M_{12}(12,11)$ |

Table E.11: Different Speaker-Dependent maps and Data (DSD&D) experiments for RMAV speaker 12

| Different Speaker-Dependent maps and Data (DSD&D) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp1 | Sp12 | $M_1(1,12)$ |
| Sp2 | Sp2 | Sp12 | $M_2(2,12)$ |
| Sp3 | Sp3 | Sp12 | $M_3(3,12)$ |
| Sp4 | Sp4 | Sp12 | $M_4(4,12)$ |
| Sp5 | Sp5 | Sp12 | $M_5(5,12)$ |
| Sp6 | Sp6 | Sp12 | $M_6(6,12)$ |
| Sp7 | Sp7 | Sp12 | $M_7(7,12)$ |
| Sp8 | Sp8 | Sp12 | $M_8(8,12)$ |
| Sp9 | Sp9 | Sp12 | $M_9(9,12)$ |
| Sp10 | Sp10 | Sp12 | $M_{10}(10,12)$ |
| Sp11 | Sp11 | Sp12 | $M_{11}(11,12)$ |

# Appendix F

# RMAV DSD Experiments

Table F.1: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 2

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp2 | Sp2 | $M_1(2,2)$ |
| Sp3 | Sp2 | Sp2 | $M_3(2,2)$ |
| Sp4 | Sp2 | Sp2 | $M_4(2,2)$ |
| Sp5 | Sp2 | Sp2 | $M_5(2,2)$ |
| Sp6 | Sp2 | Sp2 | $M_6(2,2)$ |
| Sp7 | Sp2 | Sp2 | $M_7(2,2)$ |
| Sp8 | Sp2 | Sp2 | $M_8(2,2)$ |
| Sp9 | Sp2 | Sp2 | $M_9(2,2)$ |
| Sp10 | Sp2 | Sp2 | $M_{10}(2,2)$ |
| Sp11 | Sp2 | Sp2 | $M_{11}(2,2)$ |
| Sp12 | Sp2 | Sp2 | $M_{12}(2,2)$ |

Table F.2: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 3

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp3 | Sp3 | $M_1(3,3)$ |
| Sp2 | Sp3 | Sp3 | $M_3(3,3)$ |
| Sp4 | Sp3 | Sp3 | $M_4(3,3)$ |
| Sp5 | Sp3 | Sp3 | $M_5(3,3)$ |
| Sp6 | Sp3 | Sp3 | $M_6(3,3)$ |
| Sp7 | Sp3 | Sp3 | $M_7(3,3)$ |
| Sp8 | Sp3 | Sp3 | $M_8(3,3)$ |
| Sp9 | Sp3 | Sp3 | $M_9(3,3)$ |
| Sp10 | Sp3 | Sp3 | $M_{10}(3,3)$ |
| Sp11 | Sp3 | Sp3 | $M_{11}(3,3)$ |
| Sp12 | Sp3 | Sp3 | $M_{12}(3,3)$ |

Table F.3: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 4

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp4 | Sp4 | $M_1(4,4)$ |
| Sp2 | Sp4 | Sp4 | $M_3(4,4)$ |
| Sp3 | Sp4 | Sp4 | $M_4(4,4)$ |
| Sp5 | Sp4 | Sp4 | $M_5(4,4)$ |
| Sp6 | Sp4 | Sp4 | $M_6(4,4)$ |
| Sp7 | Sp4 | Sp4 | $M_7(4,4)$ |
| Sp8 | Sp4 | Sp4 | $M_8(4,4)$ |
| Sp9 | Sp4 | Sp4 | $M_9(4,4)$ |
| Sp10 | Sp4 | Sp4 | $M_{10}(4,4)$ |
| Sp11 | Sp4 | Sp4 | $M_{11}(4,4)$ |
| Sp12 | Sp4 | Sp4 | $M_{12}(4,4)$ |

Table F.4: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 5

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp5 | Sp5 | $M_1(5,5)$ |
| Sp2 | Sp5 | Sp5 | $M_3(5,5)$ |
| Sp3 | Sp5 | Sp5 | $M_4(5,5)$ |
| Sp4 | Sp5 | Sp5 | $M_5(5,5)$ |
| Sp6 | Sp5 | Sp5 | $M_6(5,5)$ |
| Sp7 | Sp5 | Sp5 | $M_7(5,5)$ |
| Sp8 | Sp5 | Sp5 | $M_8(5,5)$ |
| Sp9 | Sp5 | Sp5 | $M_9(5,5)$ |
| Sp10 | Sp5 | Sp5 | $M_{10}(5,5)$ |
| Sp11 | Sp5 | Sp5 | $M_{11}(5,5)$ |
| Sp12 | Sp5 | Sp5 | $M_{12}(5,5)$ |

Table F.5: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 6

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp6 | Sp6 | $M_1(6,6)$ |
| Sp2 | Sp6 | Sp6 | $M_3(6,6)$ |
| Sp3 | Sp6 | Sp6 | $M_4(6,6)$ |
| Sp4 | Sp6 | Sp6 | $M_5(6,6)$ |
| Sp5 | Sp6 | Sp6 | $M_6(6,6)$ |
| Sp7 | Sp6 | Sp6 | $M_7(6,6)$ |
| Sp8 | Sp6 | Sp6 | $M_8(6,6)$ |
| Sp9 | Sp6 | Sp6 | $M_9(6,6)$ |
| Sp10 | Sp6 | Sp6 | $M_{10}(6,6)$ |
| Sp11 | Sp6 | Sp6 | $M_{11}(6,6)$ |
| Sp12 | Sp6 | Sp6 | $M_{12}(6,6)$ |

Table F.6: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 7

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp7 | Sp7 | $M_1(7,7)$ |
| Sp2 | Sp7 | Sp7 | $M_3(7,7)$ |
| Sp3 | Sp7 | Sp7 | $M_4(7,7)$ |
| Sp4 | Sp7 | Sp7 | $M_5(7,7)$ |
| Sp5 | Sp7 | Sp7 | $M_6(7,7)$ |
| Sp6 | Sp7 | Sp7 | $M_7(7,7)$ |
| Sp8 | Sp7 | Sp7 | $M_8(7,7)$ |
| Sp9 | Sp7 | Sp7 | $M_9(7,7)$ |
| Sp10 | Sp7 | Sp7 | $M_{10}(7,7)$ |
| Sp11 | Sp7 | Sp7 | $M_{11}(7,7)$ |
| Sp12 | Sp7 | Sp7 | $M_{12}(7,7)$ |

Table F.7: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 8

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp8 | Sp8 | $M_1(8,8)$ |
| Sp2 | Sp8 | Sp8 | $M_3(8,8)$ |
| Sp3 | Sp8 | Sp8 | $M_4(8,8)$ |
| Sp4 | Sp8 | Sp8 | $M_5(8,8)$ |
| Sp5 | Sp8 | Sp8 | $M_6(8,8)$ |
| Sp6 | Sp8 | Sp8 | $M_7(8,8)$ |
| Sp7 | Sp8 | Sp8 | $M_8(8,8)$ |
| Sp9 | Sp8 | Sp8 | $M_9(8,8)$ |
| Sp10 | Sp8 | Sp8 | $M_{10}(8,8)$ |
| Sp11 | Sp8 | Sp8 | $M_{11}(8,8)$ |
| Sp12 | Sp8 | Sp8 | $M_{12}(8,8)$ |

Table F.8: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 9

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp9 | Sp9 | $M_1(9,9)$ |
| Sp2 | Sp9 | Sp9 | $M_3(9,9)$ |
| Sp3 | Sp9 | Sp9 | $M_4(9,9)$ |
| Sp4 | Sp9 | Sp9 | $M_5(9,9)$ |
| Sp5 | Sp9 | Sp9 | $M_6(9,9)$ |
| Sp6 | Sp9 | Sp9 | $M_7(9,9)$ |
| Sp7 | Sp9 | Sp9 | $M_8(9,9)$ |
| Sp8 | Sp9 | Sp9 | $M_9(9,9)$ |
| Sp10 | Sp9 | Sp9 | $M_{10}(9,9)$ |
| Sp11 | Sp9 | Sp9 | $M_{11}(9,9)$ |
| Sp12 | Sp9 | Sp9 | $M_{12}(9,9)$ |

Table F.9: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 10

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp10 | Sp10 | $M_1(10,10)$ |
| Sp2 | Sp10 | Sp10 | $M_3(10,10)$ |
| Sp3 | Sp10 | Sp10 | $M_4(10,10)$ |
| Sp4 | Sp10 | Sp10 | $M_5(10,10)$ |
| Sp5 | Sp10 | Sp10 | $M_6(10,10)$ |
| Sp6 | Sp10 | Sp10 | $M_7(10,10)$ |
| Sp7 | Sp10 | Sp10 | $M_8(10,10)$ |
| Sp8 | Sp10 | Sp10 | $M_9(10,10)$ |
| Sp9 | Sp10 | Sp10 | $M_{10}(10,10)$ |
| Sp11 | Sp10 | Sp10 | $M_{11}(10,10)$ |
| Sp12 | Sp10 | Sp10 | $M_{12}(10,10)$ |

Table F.10: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 11

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp11 | Sp11 | $M_1(11,11)$ |
| Sp2 | Sp11 | Sp11 | $M_3(11,11)$ |
| Sp3 | Sp11 | Sp11 | $M_4(11,11)$ |
| Sp4 | Sp11 | Sp11 | $M_5(11,11)$ |
| Sp5 | Sp11 | Sp11 | $M_6(11,11)$ |
| Sp6 | Sp11 | Sp11 | $M_7(11,11)$ |
| Sp7 | Sp11 | Sp11 | $M_8(11,11)$ |
| Sp8 | Sp11 | Sp11 | $M_9(11,11)$ |
| Sp9 | Sp11 | Sp11 | $M_{10}(11,11)$ |
| Sp10 | Sp11 | Sp11 | $M_{11}(11,11)$ |
| Sp12 | Sp11 | Sp11 | $M_{12}(11,11)$ |

Table F.11: Different Speaker-Dependent maps (DSD) RMAV experiments for RMAV speaker 12

| Different Speaker-Dependent maps (DSD) | | | |
|---|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) | $M_n(p,q)$ |
| Sp1 | Sp12 | Sp12 | $M_1(12,12)$ |
| Sp2 | Sp12 | Sp12 | $M_3(12,12)$ |
| Sp3 | Sp12 | Sp12 | $M_4(12,12)$ |
| Sp4 | Sp12 | Sp12 | $M_5(12,12)$ |
| Sp5 | Sp12 | Sp12 | $M_6(12,12)$ |
| Sp6 | Sp12 | Sp12 | $M_7(12,12)$ |
| Sp7 | Sp12 | Sp12 | $M_8(12,12)$ |
| Sp8 | Sp12 | Sp12 | $M_9(12,12)$ |
| Sp9 | Sp12 | Sp12 | $M_{10}(12,12)$ |
| Sp10 | Sp12 | Sp12 | $M_{11}(12,12)$ |
| Sp11 | Sp12 | Sp12 | $M_{12}(12,12)$ |

# Appendix G

# Publications

# RESOLUTION LIMITS ON VISUAL SPEECH RECOGNITION

Helen L. Bear, Richard Harvey, Barry-John Theobald, Yuxuan Lan

School of Computing Sciences , University of East Anglia, Norwich, NR4 7TJ, UK.

helen.bear@uea.ac.uk, r.w.harvey@uea.ac.uk, b.theobald@uea.ac.uk, y.lan@uea.ac.uk

## ABSTRACT

Visual-only speech recognition is dependent upon a number of factors that can be difficult to control, such as: lighting; identity; motion; emotion and expression. But some factors, such as video resolution are controllable, so it is surprising that there is not yet a systematic study of the effect of resolution on lip-reading. Here we use a new data set, the Rosetta Raven data, to train and test recognizers so we can measure the affect of video resolution on recognition accuracy. We conclude that, contrary to common practice, resolution need not be that great for automatic lip-reading. However it is highly unlikely that automatic lip-reading can work reliably when the distance between the bottom of the lower lip and the top of the upper lip is less than four pixels at rest.

## 1. INTRODUCTION

A typical lip-reading system has a number of stages: first, the data are pre-processed and normalised; second, the face and lips are tracked; third, visual features are extracted and classified. In practice many systems find tracking challenging, which affects the overall recognition performance. However, the tracking problem is not insurmountable and it is now realistic to track talking heads in outdoor scenes filmed with shaky hand-held cameras [2], so we focus on feature extraction using Active Appearance Models (AAMs) [4]. We select AAMs since they have been shown to have robust performance on a number of datasets ( [8, 9, 10, 11] for example) and out perform other feature types [6].

## 2. DATASET AND FEATURE EXTRACTION

An AAM is a combined model of shape and appearance trained to fit to a whole video sequence [4]. Training creates a mean model and a set of modes, which may be varied to create shape and appearance changes. In training, a small number of frames are identified and manually landmarked. These models are Procrustes-aligned and the mean and covariance of the shape are computed. The eigenvectors of the covariance matrix give a set of modes of variation, which are used to deform the mean shape. For appearance a mesh shape-normalizes the images via a piecewise affine transform so the pixels of all images are aligned. We then compute the mean and the eigenvectors of their covariance. Concatenating the shape and appearance features forms the feature vector for training and testing. Having built a model on a few frames, it is fitted to unseen data using inverse compositional fitting [8].

The Rosetta Raven data are four videos of two North American talkers (each talker in two videos), reciting Edgar Allen Poe's 'The Raven'. The poem was published in 1845 and, recited properly, the poem has trochaic octameter [13], but this does not appear to have been followed by the talkers in this dataset. Figure 3(a) shows example frames from the high-definition video of the two talkers. The database summarised in Table 1 was recorded at $1440 \times 1080$ non-interlaced resolution at 60 frames per second. The talkers wore no make-up.

| Video | Train Images | Fit Images | Duration |
|---|---|---|---|
| Talker1 - 1 | 10 | 21,648 | 00:06:01 |
| Talker1 - 2 | 10 | 21,703 | 00:06:02 |
| Talker2 - 1 | 11 | 31,858 | 00:08:52 |
| Talker2 - 2 | 11 | 33,328 | 00:09:17 |

**Table 1**: Frame images from each video

All four videos were converted into a set of images (one per frame) with ffmpeg using image2 encoding at full high-definition resolution ($1440 \times 1080$).

To build an initial model we select the first frame and nine or ten others randomly. These *key frames* are hand-labelled with a model of a face and lips. This preliminary model is then fitted, via inverse compositional fitting [8] to the remaining frames (Table 1 lists total frames for each video). At this stage therefore we have tracked and fitted full face talker dependent AAMs on full resolution lossless PNG frame images as in Figure 1.

These models are then decomposed into sub-models for the eyes, eyebrows, nose, face outline and lips (this allows us to obtain a robust fit from the full face model but process only the lips). Figure 2 shows both talker's lips sub-model. Next, the video frames used in the high-resolution fitting were down-sampled to each of the required resolutions (Table 2) by nearest neighbor sampling and then up-sampled via bilinear sampling (Figure 3) to provide us with 18 sets of frames. These new frames are the same physical size as the origi-

**Fig. 1**: Showing full face mesh for talker T1 (left) and T2 (right)

nal ($1440 \times 1080$) but contain far less information due to the downsampling.

| | | | |
|---|---|---|---|
| $1440 \times 1080$ | $960 \times 720$ | $720 \times 540$ | $360 \times 270$ |
| $240 \times 180$ | $180 \times 135$ | $144 \times 108$ | $120 \times 90$ |
| $90 \times 67$ | $80 \times 60$ | $72 \times 54$ | $65 \times 49$ |
| $69 \times 45$ | $55 \times 42$ | $51 \times 39$ | $48 \times 36$ |
| $45 \times 34$ | $42 \times 32$ | | |

**Table 2**: Resolutions

We are most interested in the affect of low resolution on the loss of lip-reading information rather than, say the affect it would also have on the tracker (many AAM trackers lose track quite easily at low resolutions and we do not wish to be overwhelmed with catastrophic errors due to tracking problems which can often be solved in other ways [12]). Consequently the shape features in this experiment are unaffected by the downsample whereas as the appearance features vary (a useful benchmark as it will turn out).

**Fig. 2**: Showing lip-only mesh for talker T1 (left) and talker T2 (right)

For talker1 (T1), we retain 6 shape and 14 appearance parameters and for talker2 (T2), 7 shape and 14 appearance parameters. The number of parameters was chosen to retain 95% of the variance in the usual way [4].

## 3. RECOGNITION METHOD

| vID | Phones | vID | Phones |
|---|---|---|---|
| v01 | /p/ /b/ /m/ | v10 | /i/ /ih/ |
| v02 | /f/ /v/ | v11 | /eh/ /ae/ /ey/ /ay/ |
| v03 | /th /dh/ | v12 | /aa/ /ao/ /ah/ |
| v04 | /t/ /d/ /n/ /k/ /g/ /h/ /j/ /ng/ /y/ | v13 | /uh/ /er/ /ax/ |
| v05 | /s/ /z/ | v14 | /u/ /uw/ |
| v06 | /l/ | v15 | /oy/ |
| v07 | /r/ | v16 | /iy/ /hh/ |
| v08 | /sh/ /zh/ /ch/ /jh/ | v17 | /aw/ /ow/ |
| v09 | /w/ | v18 | silence |

**Table 3**: Phone to viseme mapping

To produce the ground truth we listen to each recitation of the poem and produced a ground truth text (some recitations of the poem were not word-perfect). This word transcript is converted to an American English phone level transcript using the CMU pronunciation dictionary [3]. However not all phones are visible on the lips, so we select a mapping from phones to *visemes* (which are the visual equivalent of phonemes). Here, the viseme mapping is based upon Walden's trained consonants [14] and Montgomery et al's vowel [7] classifications as illustrated in Table 3. Viseme recognition is selected over phoneme recognition as, on a small data set, it has the benefits of reducing the number of classes needed (the model for each class forms a single recogniser) and increasing the training data available for each viseme classifier. Note that not all visemes are equally represented in the data as is shown by the viseme counts in Figures 4 and 5.

For each talker, a test fold is randomly selected as 42 of the 108 lines in the poem. The remaining lines are used as training folds. Repeating this five times gives five-fold cross-validation. Visemes cannot be equally represented in all folds. For recognition we use Hidden Markov Models (HMMs) implemented in the Hidden Markov Toolkit (HTK) [15]. An HMM is initialised using the 'flat start' method using a prototype of five states and five mixture components and the information in the training samples. We choose five states and five mixtures via [9]. We define an HMM for each viseme plus silence and short-pause labels (Table 3) and re-estimate the parameters four times with no pruning. We use the HTK tool `HHEd` to tie together the short-pause and silence models between states two and three before re-estimating the HMMs a further two times. Then `HVite` is used to force-align the data using the word transcript [1].

[1] We use the `-m` flag with `HVite` with the manual creation of a viseme version of the CMU dictionary for word to viseme mapping so that the force-alignment produced uses the break points of the words.

(a) (b) (c)

**Fig. 3**: (a) $1440 \times 1080$-Original resolution image for T1 & T2, (b) $60 \times 45$-T1 downsampled, and (c) $1440 \times 1080$-T1 restored

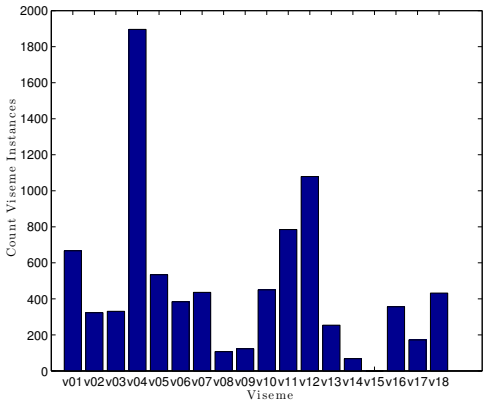



**Fig. 4**: Visemes present in both T1 videos

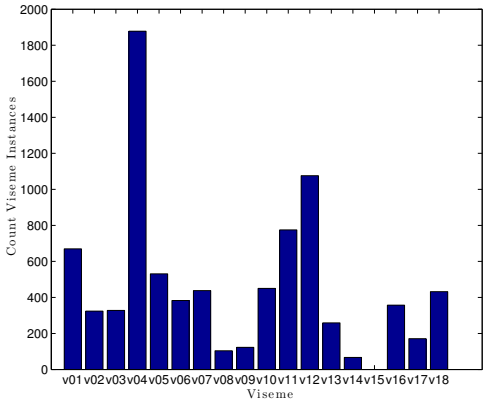



**Fig. 5**: Visemes present in both T2 videos

The HMMs are now re-estimated twice more, however now we use the force-aligned viseme transcript rather than the original viseme transcript used in the previous HMM re-estimations. To complete recognition using our HMMs we require a word network. We use `HLStats` and `HBuild` to make both a Unigram Word-level Network (UWN) and a Bi-gram Word-level Network (BWN). Finally `HVite` is used with the different network support for the recognition task and `HResults` gives us the correctness and accuracy values.

## 4. RESULTS

Recognition performance of the HMMs can be measured by both correctness, $C$, and accuracy, $A$,

$$C = \frac{N - D - S}{N}, \quad A = \frac{N - D - S - I}{N},$$

where $S$ is the number of substitution errors, $D$ is the number of deletion errors, $I$ is the number of insertion errors and $N$ the total number of labels in the reference transcriptions [15]. We use accuracy as a measure rather than correctness since it accounts for all errors including insertion errors which are notoriously common in lip reading. An insertion error occurs when the recognizer output has extra words/visemes missing from the original transcript [15]. As an example one could say "Once upon a midnight dreary", but the recognizer outputs "Once upon upon midnight dreary dreary". Here the recognizer has inserted two words which were never present and it has deleted one.

Figure 6 shows the acurracy, $A$, versus resolution for an UWN. The $x$-axis is calibrated by the vertical height of the lips of each talker in their rest position. For example, at the maximum resolution of $1440 \times 1080$ talker T1 has a lip-height of approximately 26 pixels in the rest position whereas T2 has a lip-height of approximately 17 pixels. The worst performance is from talker T2 using shape-only features. Note that the shape features do not vary with resolution so any variation in this curve is due to the cross-fold validation error (all folds do not contain all visemes equally). Nevertheless the variation is within an error bar. The poor performance is, as usual with lip-reading, a standard error dominated by insertion errors (hence the negative $A$ values). The usual explanation for this effect is that shape data contains a few characteristic shapes (which are easily recognised) in a sea of indistinct shapes - it is easier for a recogniser to insert garbage symbols than it is to learn the duration of a symbol which has indistinct start and end shapes due to co-articulation. Talker T1 has more distinctive shapes so scores better on the shape feature.

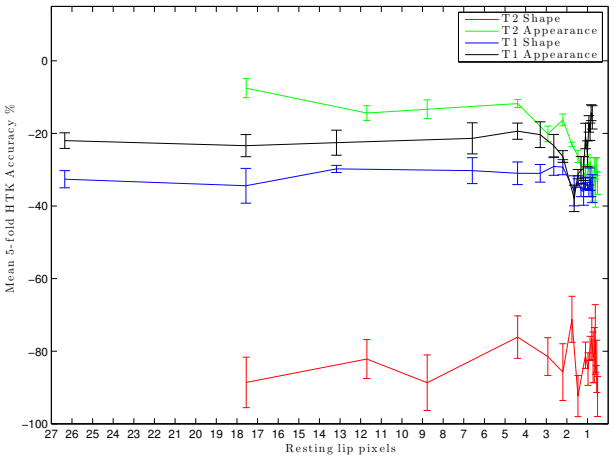



**Fig. 6**: Mean viseme recognition accuracy supported by UWN at 18 degraded resolutions shown by vertical resting lip height in pixels. Error bars show ± one standard error.

However it is the appearance that is of more interest since this varies as we downsample. At resolutions lower than four pixels it is difficult to be confident that the shape information is effective. However the basic problem is a very low error rate (shown in Figure 6) therefore we adopt a more supportive word model.

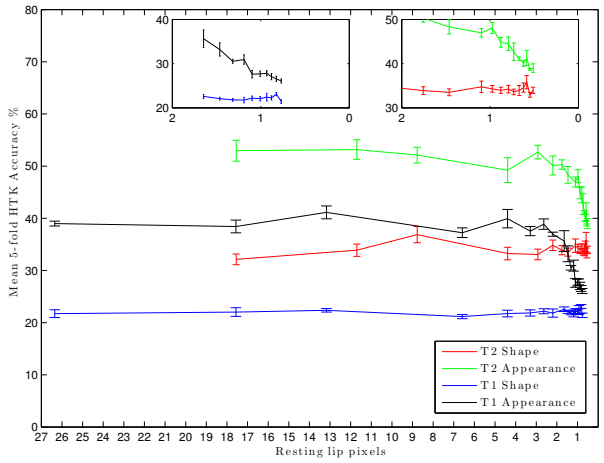



**Fig. 7**: Mean viseme recognition accuracy supported by BWN at 18 degraded resolutions shown by vertical resting lip height in pixels. Error bars show ± one standard error.

Figure 7 shows the recognition accuracy versus resolution (represented by the same $x$-axis calibration in Figure 6) for a BWN. It also includes two sub-plots which zoom the right-most part of the graph. Again the shape models perform worse than the appearance models but, looking at the zoomed plots, appearance never becomes as poor as shape performance even at very low resolutions. As with the UWN accuracies, there is clear inflection point at around four pixels (two pixels per lip) and by two pixels the performance has declined noticeably.

| Rest | Talker 1 | | | Talker 2 | | |
|---|---|---|---|---|---|---|
| Pixels | Ins | Del | Sub | ins | Del | Sub |
| $> 4$ | 69.8 | 667.0 | 259.6 | 114.2 | 467.8 | 284.6 |
| $< 4$ | 61.0 | 729.2 | 271.0 | 106.0 | 464.4 | 300.0 |

**Table 4**: Error rates for insertions, deletions and substitutions where the pixels are more than four covering the lips at rest (where recognition is still reliable), and less than four pixels where recognition performance falls. Values are averaged over all five folds.

Table 4 shows the deletion, insertion and substitution error rates for the recognition performance of resolutions which are just above and below the four pixels at rest. We see that the insertion errors are significantly lower than both deletions and substitutions so we are confident that our accuracy scores are accurate insertions despite negative accuracy scores being achieved with the Unigram Word Network support in Figure 6.

## 5. CONCLUSIONS

We have shown that the performance of simple visual speech recognizers has a threshold effect with resolution. For successful lip-reading one needs a minimum four pixels across the closed lips. However the surprising result is the remarkable resilience that computer lip-reading shows to resolution. Given that modern experiments in lip-reading usually take place with high-resolution video ([16] and [1] for example) the disparity between measured performance (shown here) and assumed performance is very striking.

Of course higher resolution may be beneficial for tracking but, in previous work we have been able to show other factors believed to be highly detrimental to lip-reading such as off-axis views [5] actually have the ability to improve performance rather than degrade it. We have also noted that previous shibboleths of outdoor video, poor lighting and agile motion affecting performance can all be overcome [1]. It seems that in lip-reading it is better to trust the data than conventional wisdom.

## 6. ACKNOWLEDGEMENTS

The authors wish to thank Professor Eamonn Keogh UCLA for providing the Rosetta Raven videos used for this work.

## 7. REFERENCES


[1] R. Bowden, S. Cox, R. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B.-J. Theobald. Recent developments in automated lip-reading. In *SPIE Security+ Defence*, pages 89010J–89010J. International Society for Optics and Photonics, 2013.

[2] R. Bowden, S. J. Cox, R. W. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B. Theobald. Is automated conversion of video to text a reality? In C. Lewis and D. Burgess, editors, *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*, volume SPIE 8546, pages 85460U–85460U–9. SPIE, 2012.

[3] Carnegie Mellon University. CMU pronounciation dictionary, 2008.

[4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681 –685, Jun 2001.

[5] Y. Lan, B.-J. Theobald, and R. Harvey. View independent computer lip-reading. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 432–437. IEEE, 2012.

[6] Y. Lan, B.-J. 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*, volume 201, 2010.

[7] D. Massaro. *Perceiving Talking Faces*. The MIT Press, 1998.

[8] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[9] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(2):198 –213, feb 2002.

[10] E. Ong and R. Bowden. Robust lip-tracking using rigid flocks of selected linear predictors. In *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG2008)*, pages 247–254, 2008.

[11] E. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden. Robust facial feature tracking using selected multi-resolution linear predictors. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1483–1490. IEEE, 2009.

[12] E.-J. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden. Robust facial feature tracking using selected multi-resolution linear predictors. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1483–1490, Sept 2009.

[13] P. F. Quinn. The critical mind of Edgar Poe: Claude Richard. Edgar Allan Poe: Journaliste et critique. *Poe Studies-Old Series*, 13(2):37–40, 1980.

[14] B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones. Effects of training on the visual recognition of consonants. *Journal of Speech, Language and Hearing Research*, 20(1):130, 1977.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchec, and P. Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[16] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen. A compact representation of visual speech data using latent variables. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):181–187, Jan 2014.

# Some observations on computer lip-reading: moving from the dream to the reality

Helen L. Bear [a], Gari Owen[b], Richard Harvey[a], and Barry-John Theobald[a]

[a]University of East Anglia, Norwich, NR4 7TJ, UK.
[b]Annwvyn Solutions, Bromley, Kent, BR1 3DW, UK.

## ABSTRACT

In the quest for greater computer lip-reading performance there are a number of tacit assumptions which are either present in the datasets (high resolution for example) or in the methods (recognition of spoken visual units called "visemes" for example). Here we review these and other assumptions and show the surprising result that computer lip-reading is not heavily constrained by video resolution, pose, lighting and other practical factors. However, the working assumption that visemes, which are the visual equivalent of phonemes, are the best unit for recognition does need further examination. We conclude that visemes, which were defined over a century ago, are unlikely to be optimal for a modern computer lip-reading system.

**Keywords:** Lip-reading, speech recognition, pattern recognition

## 1. BACKGROUND

There has been consistent and sustained interest in building computer systems that can understand what humans are saying without hearing the audio channel.[1–4] There are obvious applications for such systems in security but also in noisy environments such as cockpits, battlefields and crowds where audio recognition is likely to be impossible or highly degraded. Early work consisted of very small vocabularies (often fewer than 10 words),[5] single speakers, high-definition video (often the camera would be zoomed into the lip region or the frame rate would be greater than 60 fields per second)[6] and, often, the talker would wear special lipstick to allow easy segmentation and analysis of the lips.[7] Subsequently, our understanding of the problem has improved such that lip-reading in outdoor conditions (which requires very robust lip-tracking) and with 1000-voclabularies (which requires good machine learning) looks feasible. The problem of speaker dependence is still only partially solved.[8] One surprising recent result was a characterisation of the effect of resolution on lip-reading. An informal understanding was that relatively high resolution was required (at least a couple of hundred pixels to span the lips). In practice, it was reported in[9] that, provided the tracking was perfect, then fewer than 10 pixels can give acceptable results. A further observation[10] was that off-axis lip-reading gave slightly better performance than full frontal (which is the default for most experiments). It seems, when it comes to lip-reading, one's intuition might often be wrong – indeed experimenters in the field are often confounded by one of the most counter-intuitive illusions in the field – the McGurk effect.[11]

Experimental recognition systems for audio are almost always built using phonemes. There appears to be good agreement as to which phonemes appear in the major languages and what their expected frequency might be. Once these phonetic units have been recognised then the sequence (together with their probabilities and next-most probable sequence and so on) is fed into a language model which generates hypotheses for words and sentences. In modern speech recognition language models are powerful and important and have been the subject of decades of work. There is clearly a huge advantage in a lip-reading system re-using the language model so many lip-reading systems recognise using the visual units, visemes, and then feed the sequence into an acoustic language model modified to cope with visemes. If visemes exist in the form postulated by linguists e.g.,[12,13] then there are many choices of visemes. However there has been surprisingly few examinations of which visemes give the best performance or how fragile that performance is compared to phonetic recognition.

Further author information: (Send correspondence to RWH)
RWH: E-mail: r.w.harvey@uea.ac.uk

## 2. INTRODUCTION

A phoneme is generally regarded as the smallest sound which can be uttered.[14] A viseme, which is often said to be the visual equivalent of a phoneme, is not so precisely defined[15–17] so we use the working definition: 'a viseme is a set of phonemes that have identical appearance on the lips'. Therefore any phoneme falls into one viseme class but a viseme may represent many phonemes: a many to one mapping.

A typical lip-reading system is a sequence of tasks as in Figure 1 and our work is focused within the recognition step.

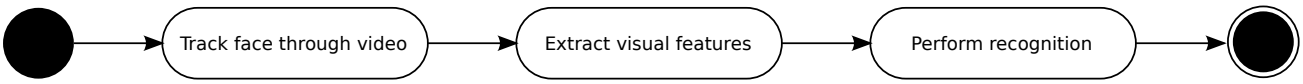



Figure 1. Steps in a typical lip-reading system

Similar to a simplified version of audio recognition whereby we seek to identify a string of unique phonemes, each recognizer is based upon training data of the correctly labelled phoneme. In visual-only recognition we use the same concept of building recognizers based upon visual-only training samples correctly labelled according to a viseme mapping. There is still debate over what the *correct* phoneme-to-viseme mapping is and many have been suggested, e.g[16,18–20] but our interest is in the contribution of each viseme to the recognition performance. We look for any particular visemes (or combinations of phonemes) that contribute more to the recognition accuracy.

We aim to measure the reduction of each unique visemic recogniser in contribution value to the whole task of accurate recognition in continuous speech. To demonstrate the influence of reduced recogniser classes in visual speech recognition we compare the outputs with those of audio recognition of the same data. For a fair comparison we use the same groupings of phonemes into faux 'audio-viseme' recognisers on the audio data. Audio recognition has a higher quantity of classifiers (phonemes) than proposed viseme classes, therefore we hypothesise visual classes have bigger variance in use/purpose towards the whole recognition task. We anticipate, fewer visemes will be used in visual speech recognition than 'audio-visemes' in audio recognition.

## 3. DATASET AND FEATURE EXTRACTION

For the first two steps in Figure 1 we use full face Active Appearance Models (AAMs)[21] to track the faces through the videos, and lip-only AAMs (one for shape and another for appearance) and using the methods of[22] we produce two sets of talker-dependent features; shape-only visual features and appearance-only visual features.

Table 1. Frame images from each video.

| Video | Num. of AAM train images | Video Length (frames) | Duration |
|---|---|---|---|
| Talker1 - 1 | 10 | 21,658 | 00:06:01 |
| Talker1 - 2 | 10 | 21,713 | 00:06:02 |
| Talker2 - 1 | 11 | 31,868 | 00:08:52 |
| Talker2 - 2 | 11 | 33,338 | 00:09:17 |

Shape features (1) are based solely upon the lip shape and positioning during the duration of the talker speaking e.g. the landmarks in Figure 2. 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:

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

(a) (b)

Figure 2. Showing full face shape landmarks for talker T1 (a) and a lip shape landmarks for talker T1 (b).

The Rosetta Raven data is four videos of recitations of Edgar Allen Poe's poem 'The Raven'. There are two talkers, one male, one female. Neither are trained actors and they do not recite the poem with the intended trochaic octameter.[23] The videos were recorded at 1440 × 1080 resolution (non-interlaced) at 60 frames per second. Table 1 summarises the video data.

A set of images are extracted from each video (one image per frame) via ffmpeg using image2 encoding at full high-definition resolution (1440 × 1080). To construct an initial AAM we select the first frame and nine or ten others randomly. These *training frames* are hand-labelled with a shape model of a face and lips to build a preliminary model for each talker. These models are then fitted, via inverse compositional fitting[22] to the remaining frames (Table 1). Thus we get tracked and fitted full-face talker-dependent AAMs (Figure 2 left) on full resolution lossless PNG frame images (Figure 1 step 1).

Next we create a sub-model of only the lips for each talker by decomposing the two full face models (Figure 2 right). From the fitted landmarks, the shape and appearance parameters for each frame are extracted. For talker1 (T1), we retain 6 shape and 14 appearance parameters and for talker2 (T2), 7 shape and 14 appearance parameters. We restrict the feature parameters to retain 95% of variation from the mean AAM model produced using the whole tracked video data.[21] (Figure 1 step 2.)

We did not implement $\Delta\Delta's$ into our extracted features to address co-articulation because we used a phonetic-alignment in the production of our ground-truth benchmark and forced-alignment within the training process of our HMM recognizers.

Table 2. Phone to viseme mapping.

| vID | Phones | vID | Phones |
|---|---|---|---|
| v01 | /p/ /b/ /m/ | v10 | /i/ /ih/ |
| v02 | /f/ /v/ | v11 | /eh/ /ae/ /ey/ /ay/ |
| v03 | /th/ /dh/ | v12 | /aa/ /ao/ /ah/ |
| v04 | /t/ /d/ /n/ /k/ /g/ /h/ /j/ /ng/ /y/ | v13 | /uh/ /er/ /ax/ |
| | | v14 | /u/ /uw/ |
| v05 | /s/ /z/ | v15 | /oy/ |
| v06 | /l/ | v16 | /iy/ /hh/ |
| v07 | /r/ | v17 | /aw/ /ow/ |
| v08 | /sh/ /zh/ /ch/ /jh/ | v18 | silence |
| v09 | /w/ | | |

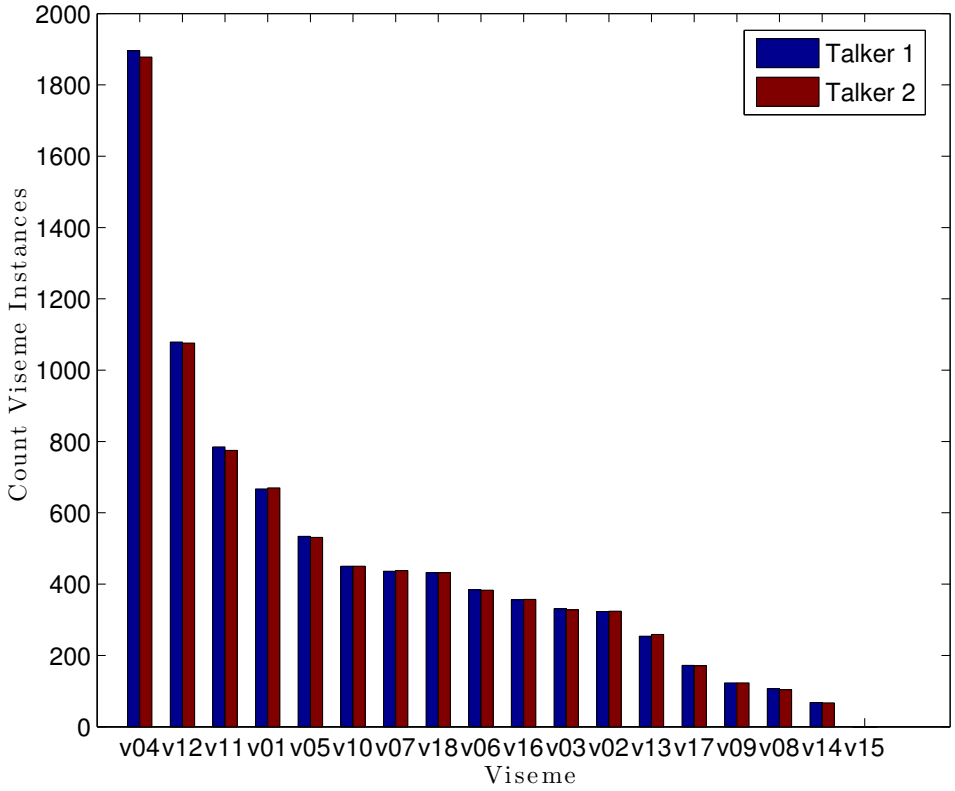



Figure 3. Viseme counts for both talker transcripts

To have a benchmark for measuring our recognition outputs we produce a ground-truth viseme transcription using the Carnegie Mellon University (CMU) North American pronunciation dictionary,[24] and a word transcription. We convert a phonetic transcript to a viseme transcript assuming 15 visemes, listed in Table 2 which is a combination of Montgomery *et al's* vowel mapping and Walden's consonant mapping.[25,26]

The limited availability of large datasets is documented[2] so we work within the restrictions of short datasets. Here we note these may not provide adequate training examples of all visemes. Where this happens, we group the untrainable visemes into a single garbage viseme. In this case we select a 150 sample threshold so visemes /v08/, /v09/, /v14/ and /v15/ are grouped. Figure 3 shows the occurrence of visemes listed in Table 3 in our data and Table 4 shows our revised viseme mapping.

Table 3. Phone to viseme mapping modified to accomodate restrictions in dataset.

| vID | Phones | vID | Phones |
|---|---|---|---|
| v01 | /p/ /b/ /m/ | v11 | /eh/ /ae/ /ey/ /ay/ |
| v02 | /f/ /v/ | v12 | /aa/ /ao/ /ah/ |
| v03 | /th /dh/ | v13 | /uh/ /er/ /ax/ |
| v04 | /t/ /d/ /n/ /k/ /g/ /h/ /j/ /ng/ /y/ | v16 | /iy/ /hh/ |
| | | v17 | /aw/ /ow/ |
| v05 | /s/ /z/ | v18 | silence |
| v06 | /l/ | garb | /u/ /uw/ /oy/ /w/ /sh/ /zh/ /ch/ /jh/ |
| v07 | /r/ | | |
| v10 | /i/ /ih/ | | |

For each talker, a test fold is randomly selected as 42 of the 108 lines in the poem with replacement. The remaining lines are used as training folds. Repeating this five times gives five-fold cross-validation. Note that visemes cannot be equally represented in all folds.

For recognition we use Hidden Markov Models (HMMs) implemented in the Hidden Markov Toolkit (HTK).[27] An HMM is initialised using the 'flat start' method using a prototype of five states and five mixture components and the information in the training samples. We choose five states and five mixture components based upon.[28] We define an HMM for each viseme plus silence and short-pause labels (Table 3) and re-estimate the parameters four times with no pruning.

Next, we use the HTK tool `HHEd` to tie together the short-pause and silence models between states two and three before re-estimating the HMMs a further two times. Then `HVite` is used to force-align the data using the word transcript*. The HMMs are now re-estimated twice more, however now we use the force-aligned viseme transcript rather than the original viseme transcript used in the previous HMM re-estimations.

To complete recognition using our HMMs we require a word network as we have a continuous speech dataset. We use `HLStats` and `HBuild` to make a Bi-gram Word-level Network (BWN). Finally `HVite` is used with the network support for the recognition task and `HResults` gives us both correctness and accuracy viseme recognition values and a viseme confusion matrix for all folds. We have provided the reader with technical details to enable repeatability of our experiments. Please contact the author for original videos.

## 4. RESULTS

We have extracted figures from the `HResults` confusion matrices for analysis. For each viseme we have calculated the inverse probability of its recognition $\Pr\{v|\hat{v}\}$.

Figure 4 shows the probability of correct recognition using shape-only features (mean and $\pm 1$ standard error) plotted against the probability of correct recognition using appearance-only features for each viseme. As usual some talkers are better recognised with shape and some with appearance[1†]. Note that the top right-hand point is the visual silence phoneme. In general, visual silence can be quite variable compared to audio silence because talkers breathe and show emotion. However here, because the source text is a poem, there are well-defined visual silence periods at the start of each line.

Table 4. Ranked mean viseme recognition for Shape, Appearance, Talker 1 and Talker 2.

| Feature | Viseme order |
|---|---|
| Shape | /v18/ {/v04/, /v12} /v11/ /v01/ /v07/ /v05/ {/v02/ /v06/ /garb/} /v10/ {/v03/ /v13/} /v16/ /v17/ |
| Appearance | /v18/ /v04/ /v12/ /v11/ /v01/ /v07/ {/v02/, /v05/} /v06/ /garb/ /v10/ /v03/ {/v13, /v16/} /v17/ |
| Talker 1 | /v18/ {/v04/, /v12/} /v11/ /v01/ /v07/ {/v02/, /v05/} /v06/ {v10/, /garb/} /v03/ /v13/ /v16/ /v17/ |
| Talker 2 | /v18/ {/v04/, /v12/} /v11/ /v01/ /v07/ {/v02/, /v05/} {/v06/,/garb/} /v10/ /v03/ /v13/ /v16/ /v17/ |
| Overall | /v18/ /v04/ /v12/ /v01/ /v11/ /v07/ {/v02/, /v05/} /v19/ /v06/ /v10/ /v13/ /v03/ /v16/ /v17/ |

Figures 5 and 6 show, for the T1 and T2 shape and appearance models, the probability of correctly recognising the top ten visemes, $\Pr\{v|\hat{v}\}$. They also show, the audio performance measured on visemes. The $x$-axis varies by performance; the best performing viseme is on the left hand side which for visual shape and appearance features is silence for all features.

*We use the `-m` flag with `HVite` with the manual creation of a viseme version of the CMU dictionary for word to viseme mapping so that the force-alignment produced uses the break points of the words.

†The conventional wisdom is that appearance features give the best results but only in studio-type conditions with good tracking.

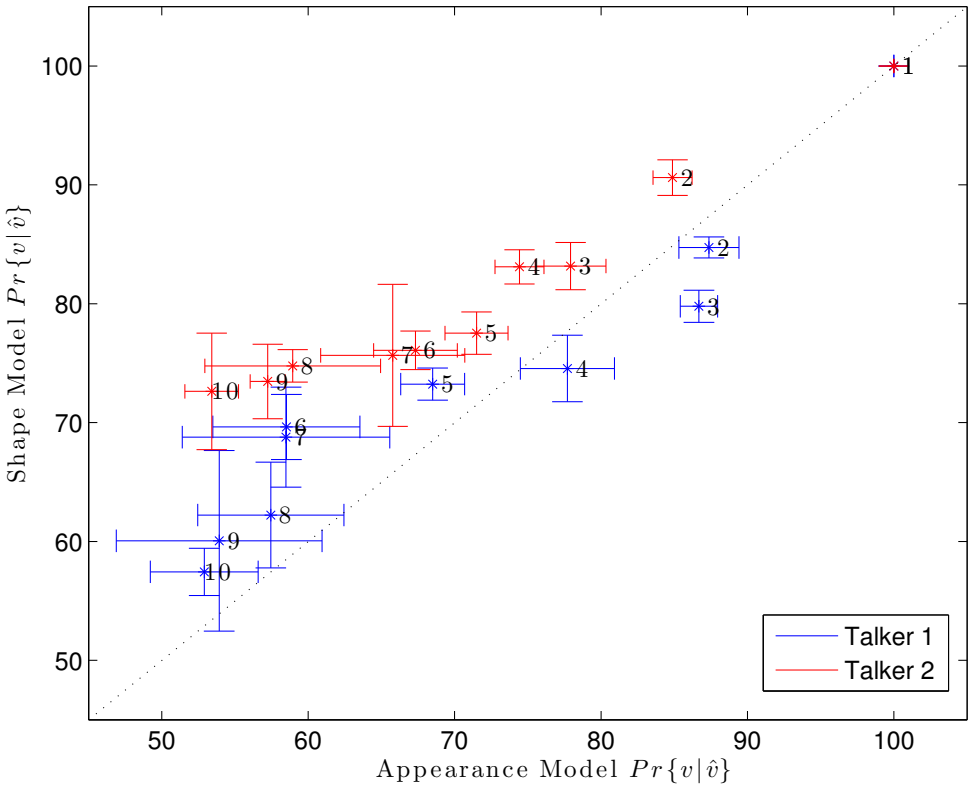



Figure 4. Relationship between Shape and Appearance model features for both talkers.

It has been observed in human lip reading there are few visual cues that are reliable and humans use these combined with rich contextual information to interpret or 'fill in the gaps' of what a talker is saying.[29,30] Therefore our hypothesis is that robust audio recognition is based upon a large spread of recognised phones and the resilience in recognition is due to the number of phones contributing to the accuracy. Visually, as with human lip-readers, it is anticipated that fewer visemes would perform the equivalent recognition and, as such, the graph would demonstrate a steeper performance decline over the top performing visemes.

In Figure 5 we do see a greater decline from left to right over the top ten visemes for visual features than for audio for both talkers. We also note that the error bars after the $5^{th}$ position viseme increase, which is consistent with our hypothesis that audio recognition is spread over more visemes to be correct. The top visemes (after silence /v18/) are /v04/, /v12/, /v11/ and /v01/. These are vowels (/v12/, /v11/) and front-of-mouth consonant visemes (/v04/, /v01/).

Figure 6 demonstrates a shallower decline from left to right than the shape graph in Figure 5 but still there is a greater decline for visual features than for audio. The error bars here increase after the $7^{th}$ position viseme[‡]. The shape of the graph in Figure 6 is similar between audio and video which implies that appearance-based recognition is similar to noisy acoustic recognition for both talkers and hence is less fragile. The top visemes in Figure 6 (not including silence /v18/) are: /v04/ /v12/ /v11/ /v01/ /v7/ i.e. identical for shape-only in the first six positions.

Where the error bars increase, we consider this may be due to the small data available, which makes recognition more unreliable due to less well trained HMM classifiers. We have reduced the impact of this with the

[‡]Note that the order of the audio viseme ordering is identical in both Figures 5 and 6 as this is the same experiment.

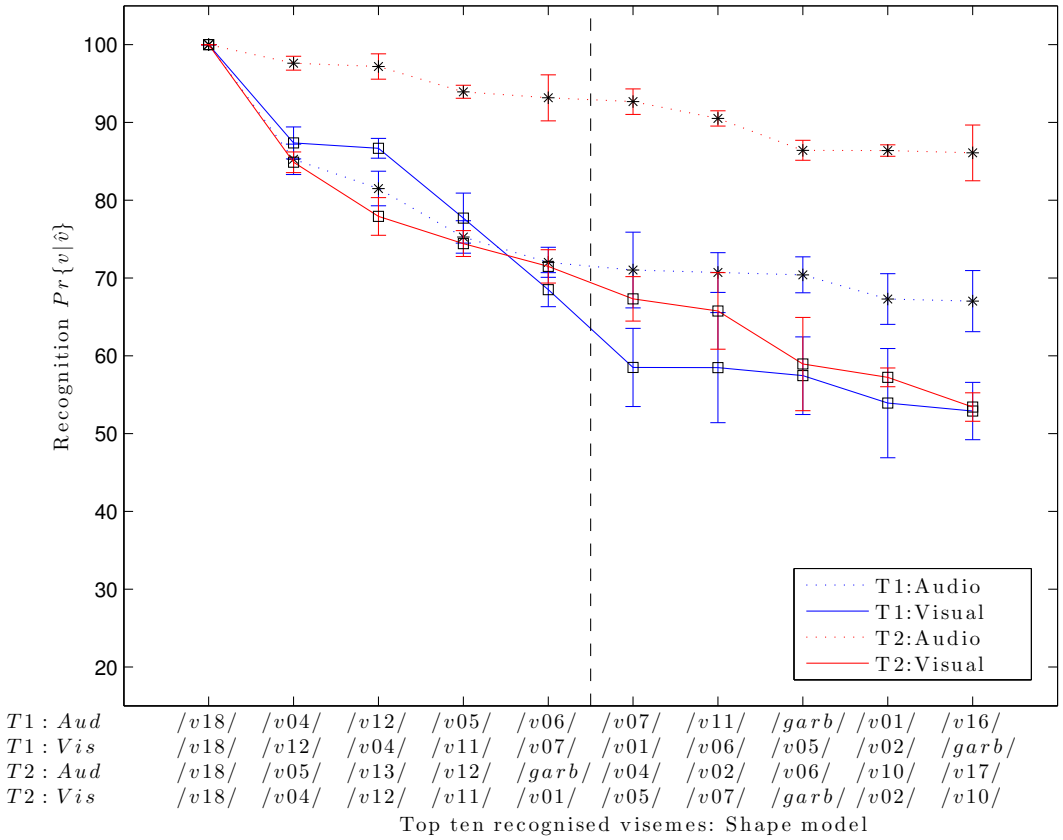



Figure 5. Top ten viseme recognition probability in descending order with a shape model.

/garb/ viseme but note with Figure 3 there are similarities between our top performing visemes and those with the most training samples.

Table 4 is the visemes ordered by correctness showing, for example that viseme 18 /v18/ is the best performing viseme overall. It is natural to ask if the differences in ranking are significant. To compare the viseme ordering we compute the Spearman rank correlation coefficient, $r$. The results are shown in Tables 5 and 6. Also shown is the $p$-value for the null that $r = 0$ (randomly ordered). Those that are significant at the 5% threshold are underlined. Talker 2 has poor audio performance which tends to degrade the audio correlation. Lip-reading does not depend on audio though so these results confirm the strong relation between shape-only and viseme-only classification. Also note for T1 (Figure 6) the audio ranking is similar to the video ranking although as we have previously noticed there is a more rapid drop-off for video.

In Table 7 we have provided the overall mean and Standard Error Accuracy score for the whole viseme set recognition performance over all five folds. Talker 2 outperforms Talker 1 with all features but for visual features also has a larger degree of error. Appearance features outperform shape for both talkers and audio outperforms appearance for both talkers. As we have seen in Figures 5 and 6 this recognition is based upon a larger spread of visemes than the shape models with the audio having the largest spread of visemes and hence being the most

Table 5. Spearman rank correlation, $r$ and $p$-value for visemes ranked by performance for Talker 1 and Talker 2

| Talker 1 | Talker 2 | r | p |
|---|---|---|---|
| Audio | Audio | <u>0.43</u> | $1.63 \times 10^{-2}$ |
| Shape | Shape | <u>0.92</u> | 0.00 |
| Appearance | Appearance | <u>0.93</u> | 0.00 |

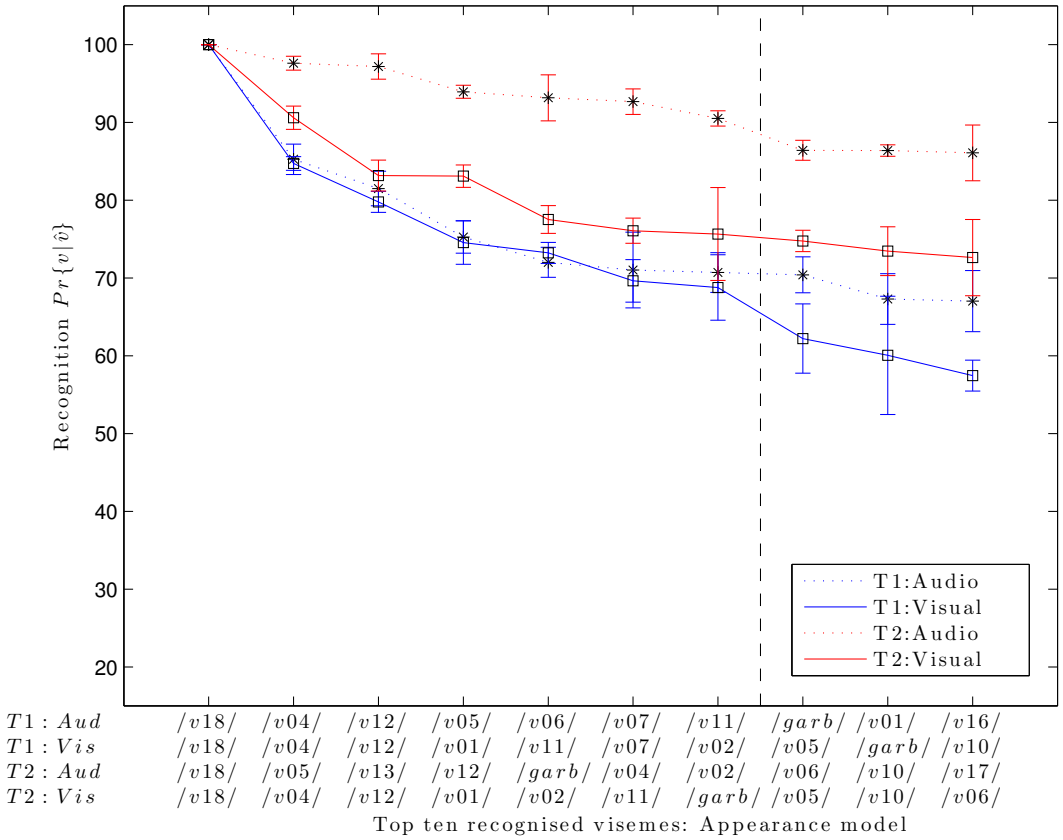



Figure 6. Top ten viseme recognition probability in descending order with an appearance model.

Table 6. Spearman rank correlation, $r$ and $p$-values for visemes ordered by feature for Talker 1 (left) and Talker 2 (right)

| Talker 1 | Talker 1 | $r$ | $p$ |
|---|---|---|---|
| Shape | Appearance | 0.90 | 0.00 |
| Audio | Shape | 0.85 | 0.00 |
| Audio | Appearance | 0.74 | 0.00 |

| Talker 2 | Talker 2 | $r$ | $p$ |
|---|---|---|---|
| Shape | Appearance | 0.92 | 0.00 |
| Audio | Shape | 0.42 | 0.12 |
| Audio | Appearance | 0.48 | 0.07 |

robust recognition mode.

Table 7. Mean accuracy scores of each feature type by talker

| Feature type | Mean | Standard error |
|---|---|---|
| T1 Audio | 45.6380 | 2.0086 |
| T2 Audio | 75.8780 | 1.7839 |
| T1 Shape | 21.7360 | 0.7501 |
| T1 Appearance | 38.9860 | 0.4637 |
| T2 Shape | 32.1360 | 1.0339 |
| T2 Appearance | 52.9540 | 1.9996 |

## 5. CONCLUSIONS

Our principal observations are:

- Assuming there is enough data to properly train classifiers, then the performance ordering of the visemes is relatively stable across modes of recognition (audio, shape and appearance).

- That said, the visual classifiers are far more dependent on the good performance of a few visemes than the audio.
- Of the video classifiers, shape is the most dependent on a few visemes.

These are important results because they illuminate the often made observation that lip-reading is fragile. In other words if one cannot build classifiers for a few critical visemes then lip-reading is impossible. In a human lip-reading context, humans are often trained to recognise a small number of critical gestures which are then processed via a very sophisticated language and context model to create a transcript.

In audio is it surprisingly rare to see this effect measured even though a good acoustic unit will have accuracies that are at least 10% higher than an average unit (the mean audio viseme performance on T2 is 76% for the whole viseme set).

It is important to acknowledge that most work in this field focuses on improving mean accuracies over the set of all visemes which can conceal the real source of overall performance. A system that achieves a mean viseme accuracy of, say 53% maybe one that contains one supremely accurate viseme classifier or it maybe a system that has a set of classifiers of much more modest performance.

This paper therefore raises two different tactics for improving lip-reading systems. Either one makes the best viseme classifiers better or, one focuses upon improving the worst. At this stage we do not know which tactic is likely to be more successful but we hope this methodology allows future work to focus attention where it is likely to do the most good.

## REFERENCES


[1] Bowden, R., Cox, S., Harvey, R., Lan, Y., Ong, E.-J., Owen, G., and Theobald, B.-J., "Recent developments in automated lip-reading," in [*SPIE Security+ Defence*], 89010J–89010J, International Society for Optics and Photonics (2013).

[2] Cappelletta, L. and Harte, N., "Phoneme-to-viseme mapping for visual speech recognition.," in [*ICPRAM (2)*], 322–329 (2012).

[3] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on* **28**(4), 357–366 (1980).

[4] Bowden, R., Cox, S. J., Harvey, R. W., Lan, Y., Ong, E.-J., Owen, G., and Theobald, B., "Is automated conversion of video to text a reality?," in [*Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*], Lewis, C. and Burgess, D., eds., **SPIE 8546**, 85460U–85460U–9, SPIE (2012).

[5] Petajan, E. D., *Automatic Lipreading to Enhance Speech Recognition*, PhD thesis, University of Illinois, Urbana-Champaign (1984).

[6] Brooke, N. M. and Summerfield, Q., "Analysis, synthesis and perception of visible articulatory movements," *Journal of Phonetics* **11**, 63–76 (1983).

[7] Kaucic, R. and Blake, A., "Accurate, real-time, unadorned lip tracking," in [*Computer Vision, 1998. Sixth International Conference on*], 370–375, IEEE (1998).

[8] Z., Z., X., H., and M., Z. G. . P., "A compact representation of visual speech data using latent variables.," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(1), 181–187 (2014).

[9] Bear, H., Harvey, R. W., Theobald, B.-J., and Lan, Y., "Resolution limits on computer lip-reading," in [*IEEE International Conference on Image Processing*], (2014).

[10] Lan, Y., Theobald, B.-J., and Harvey, R., "View independent computer lip-reading," in [*Multimedia and Expo (ICME), 2012 IEEE International Conference on*], 432–437, IEEE (2012).

[11] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices," *Nature* **264**, 746–748 (1976).

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

[13] Bozkurt, E., Erdem, C., Erzin, E., Erdem, T., and Ozkan, M., "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," in [*3DTV Conference*], 1–4, IEEE (May 2007).

[14] Association, I. P., [*Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*], Cambridge University Press (1999).

[15] Chen, T. and Rao, R. R., "Audio-visual integration in multimodal communication," *Proceedings of the IEEE* **86**(5), 837–852 (1998).

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

[17] Hazen, T. J., Saenko, K., La, C.-H., and Glass, J. R., "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in [*Proceedings of the 6th International Conference on Multimodal Interfaces*], *ICMI '04*, 235–242, ACM, New York, NY, USA (2004).

[18] Binnie, C. A., Jackson, P. L., and Montgomery, A. A., "Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation," *Journal of Speech and Hearing Disorders* **41**(4), 530 (1976).

[19] Kricos, P. B. and Lesner, S. A., "Differences in visual intelligibility across talkers.," *The Volta Review* **84**, 219–226 (1982).

[20] Nitchie, E. B., [*Lip-Reading, principles and practise: A handbook for teaching and self-practise*], Frederick A Stokes Co, New York (1912).

[21] Cootes, T., Edwards, G., and Taylor, C., "Active appearance models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **23**, 681 –685 (Jun 2001).

[22] Matthews, I. and Baker, S., "Active appearance models revisited," *International Journal of Computer Vision* **60**(2), 135–164 (2004).

[23] Quinn, P. F., "The critical mind of Edgar Poe: Claude Richard. Edgar Allan Poe: Journaliste et critique.," *Poe Studies-Old Series* **13**(2), 37–40 (1980).

[24] Carnegie Mellon University, "CMU pronounciation dictionary," (2008).

[25] Massaro, D., [*Perceiving Talking Faces*], The MIT Press (1998).

[26] Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J., "Effects of training on the visual recognition of consonants," *Journal of Speech, Language and Hearing Research* **20**(1), 130 (1977).

[27] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchec, V., and Woodland, P., [*The HTK Book (for HTK Version 3.4)*], Cambridge University Engineering Department (2006).

[28] Matthews, I., Cootes, T., Bangham, J., Cox, S., and Harvey, R., "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**, 198 –213 (feb 2002).

[29] Erber, N. P., "Auditory-visual perception of speech," *Journal of Speech and Hearing Disorders* **40**(4), 481 (1975).

[30] Stork, D. G. and Hennecke, M. E., [*Speechreading by humans and machines: models, systems, and applications*], vol. 150, Springer (1996).

# Which phoneme-to-viseme maps best improve visual-only computer lip-reading?

No Author Given

No Institute Given

**Abstract.** A critical assumption of all current visual speech recognition systems is that there are visual speech units called visemes which can be mapped to units of acoustic speech, the phonemes. Despite there being a number of published maps it is infrequent to see the effectiveness of these tested, particularly on visual-only lip-reading (many works use audio-visual speech). Here we examine 120 mappings and consider if any are stable across talkers. We show a method for devising maps based on phoneme confusions from an automated lip-reading system, and we present new mappings that show improvements for individual talkers.

## 1 Introduction

Phonemes are the discriminate sounds of a language [1] and the visual equivalent, although not precisely defined, are the visemes; [2–4]. A working definition of a viseme is a set of phonemes that have identical appearance on the lips. Therefore a phoneme falls into one viseme class but a viseme may map to many phonemes: a many-to-one mapping. In computer lip-reading there are several possibilities for Phoneme-to-Viseme (P2V) mappings and some are listed in, for example, [5] Tables 2.3 and 2.4. Such mappings are often consonant-only mappings [6, 3, 7, 8]; or devised from single-talker data (so are talker-dependent [9]) or devised from highly stylised vocabularies ([10] for example). These are useful starting points but a P2V mapping should cover all phonemes. So here we consider the possibility of using combinations of the various known mappings which cover the consonants (listed in Table 2) and which cover vowels (Table 1). In total we use 15 consonant maps and eight vowel maps, all of these are paired with each other to produce 120 P2V maps to test.

## 2 Dataset and Data Preparation

We use the AVLetters2 (AVL2) dataset [11], to train and test recognisers based upon the 120 P2V mappings. This dataset is British-English talkers reciting the alphabet seven times. We use four talkers for training which involves tracking their faces with Active Appearance Models (AAMs) [12] and extracting combined shape and appearance features. We select AAM features because they are known to out-perform other feature methods in machine visual-only lip-reading [13].

**Table 1.** Vowel Viseme:Phoneme maps

| Classification | Viseme phoneme sets |
|---|---|
| Bozkurt [14] | {/ei/ /ʌ/} {/ei/ /e/ /æ/} {/ɜ/} {/i/ /ɪ/ /ə/ /y/} {/u/ /ʊ/ /w/} {/ɑʊ/} {/ɔ/ /ɑ/ /ɔɪ/ /əʊ/} |
| Disney [15] | {/ʊ/ /h/} {/ɛə/ /i/ /ai/ /e/ /a/} {/u/} {/ʊə/ /ɔ/ /ɔə/} |
| Hazen [4] | {/ɑʊ/ /ʊ/ /u/ /əʊ/ /ɔ/ /w/ /ɔɪ/} {/ʌ/ /ɑ/} {/æ/ /e/ /ai/ /ei/} {/ə/ /ɪ/ /i/} |
| Jeffers [16] | {/ɑ/ /æ/ /ʌ/ /ai/ /e/ /ei/ /ɪ/ /i/ /ɔ/ /ə/ /ɪ/} {/ɔɪ/ /ɔ/} {/ɑʊ/} {/ɜ/ /əʊ/ /ʊ/ /u/} |
| Lee [17] | {/i/ /ɪ/} {/e/ /ei/ /æ/} {/ɑ/ /ɑʊ/ /ai/ /ʌ/} {/ɔ/ /ɔɪ/ /əʊ/} {/ʊ/ /u/} |
| Montgomery [18] | {/i/ /ɪ/} {/e/ /æ/ /ei/ /ai/} {/ɑ/ /ɔ/ /ʌ/} {/ʊ/ /ɜ/ /ə/}{/ɔɪ/} {/i/ /hh/} {/ɑʊ/ /əʊ/} {/u/ /u/} |
| Neti [19] | {/u/ /ʊ/ /əʊ/} {/æ/ /e/ /ei/ /ai/} {/ɪ/ /i/ /ə/} {/ɔ/ /ʌ/ /ɑ/ /ɜ/ /ɔɪ/ /ɑʊ/ /ɦ/} |
| Nichie [20] | {/u/} {/ʊ/ /əʊ/} {/ɑʊ/} {/i/ /ʌ/ /ɪ/} {/ʌ/} {/i/ /æ/} {/e/ /ɪə/} {/u/} {/ə/ /ei/} |

Figure 1 shows the count of the 29 phonemes in training component of AVL2 with the silence phoneme omitted. As is often the case, the rare phonemes in British English are not represented [13]. The division of these phoneme across viseme classes will vary with each different map. P2V mappings are contractive which is illustrated in Table 3 which lists the ratio of phonemes to visemes (excluding silence and phonemes not handled by that mapping). Thus, in Table 3, the Woodward map covers 24 consonant phonemes to four viesmes and has a confusion factor (CF) of $4/24 = 0.167$, whereas Jeffers vowels maps cover 23 phonemes which are mapped to eight visemes.

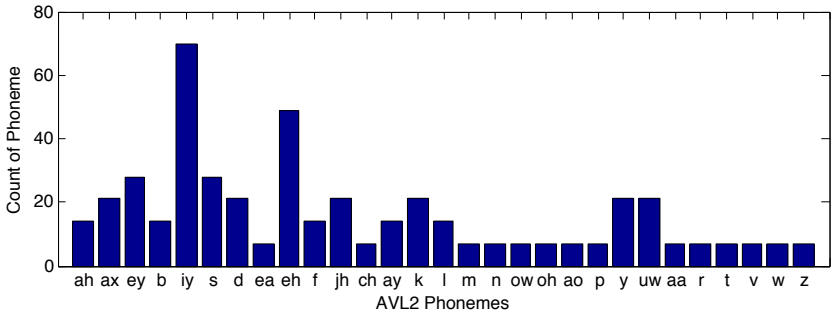



**Fig. 1.** Phoneme histogram of AVLetters-2 dataset

**Table 2.** Consonant Viseme:Phoneme maps

| Classification | Viseme phoneme sets |
|---|---|
| Binnie [6] | {/p/ /b/ /m/} {/f/ /v/} {/θ/ /ð/} {/ʃ/ /ʒ/} {/k/ /g/} {/w/} {/r/} {/l/ /n/} {/t/ /d/ /s/ /z/} |
| Bozkurt [14] | {/g/ /ɦ/ /k/ /ŋ/} {/l/ /d/ /n/ /t/} {/s/ /z/} {/tʃ/ /ʃ/ /dʒ/ /ʒ/} {/r/} {/θ/ /ð/} {/f/ /v/} {/p/ /b/ /m/} |
| Disney [15] | {/p/ /b/ /m/} {/w/} {/f/ /v/} {/θ/} {/l/} {/d/ /t/ /z/ /s/ /r/ /n/} {/ʃ/ /tʃ/ /j/} {/y/ /g/ /k/ /ŋ/} |
| Finn [21] | {/p/ /b/ /m/} {/θ/ /ð/} {/w/ /s/} {/k/ /h/ /g/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/y/} {/z/} {/f/} {/v/} {/t/ /d/ /n/ /l/ /r/} |
| Fisher [3] | {/k/ /g/ /ŋ/ /m/} {/p/ /b/} {/f/ /v/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/t/ /d/ /n/ /θ/ /ð/ /z/ /s/ /r/ /l/} |
| Franks [7] | {/p/ /b/ /m/} {/f/} {/r/ /w/} {/ʃ/ /dʒ/ /tʃ/} |
| Hazen [4] | {/l/} {/r/} {/y/} {/b/ /p/} {m} {/s/ /z/ /h/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/ŋ/} {/f/ /v/} {/t/ /d/ /θ/ /ð/ /g/ /k/} |
| Heider [22] | {/p/ /b/ /m/} {/f/ /v/} {/k/ /g/} {/ʃ/ /tʃ/ /dʒ/} {/n/ /t/ /d/} {/l/} {/r/} {/θ/} |
| Jeffers [16] | {/f/ /v/} {/r/ /q/ /w/} {/p/ /b/ /m/} {/θ/ /ð/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/g/ /k/ /ŋ/} {/s/ /z/} {/d/ /l/ /n/ /t/} |
| Kricos [9] | {/p/ /b/ /m/} {/f/ /v/} {/w/ /r/} {/t/ /d/ /s/ /z/} {/l/} {/θ/ /ð/} {/ʃ/ /ʒ/ /tʃ/ /dʒ/} {/k/ /n/ /j/ /h/ /ŋ/ /g/} |
| Lee [17] | {/d/ /t/ /s/ /z/ /θ/ /ð/} {/g/ /k/ /n/ /ŋ/ /l/ /y/ /ɦ/} {/f/ /v/} {/r/ /w/} {/dʒ/ /tʃ/ /ʃ/ /ʒ/} {/p/ /b/ /m/} |
| Neti [19] | {/l/ /r/ /y/} {/s/ /z/} {/t/ /d/ /n/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/f/ /v/} {/ŋ/ /k/ /g/ /w/} {/p/ /b/ /m/} {/θ/ /ð/} |
| Nichie [20] | {/p/ /b/ /m/} {/f/ /v/} {/ɯ/ /w/} {/s/ /z/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/t/ /d/ /n/} {/y/} {/θ/} {/l/} {/k/ /g/ /ŋ/} {/ɦ/} {/r/} |
| Walden [8] | {/p/ /b/ /m/} {/f/ /v/} {/θ /ð/} {/ʃ/ /ʒ/} {/w/} {/s/ /z/} {/r/} {/t/ /d/ /n/ /k/ /g/ /j/} {/l/} |
| Woodward [23] | {/t/ /d/ /n/ /l/ /θ/ /ð/ /s/ /z/ /tʃ/ /dʒ/ /ʃ/ /ʒ/ /j/ /k/ /g/ /h/} {/p/ /b/ /m/} {/f/ /v/} {/w /r/ /ɯ/} |

We deliberately omit the following phonemes from some mappings; /si/ (Disney), /axr/ /en/ /el/ /em/ (Bozkirt), /axr/ /em/ /epi/ /tcl/ /dcl/ /en/ /gcl/ kcl/ (Hazan), and /axr/ /em/ /el/ /nx/ /en/ /dx/ /eng/ /ux/ (Jeffers) because these are American diacritics which are not appropriate for a British English phonetic dataset. Note that all 29 phonemes in AVL2 appear across the existing P2V maps, but no mapping uses all of these phonemes. Missing phonemes from a viseme map are grouped into a garbage viseme (/gar/) to ensures we measure only the performance of the previously described viseme sets. That is, we are not creating a new map by defining new visemes within an existing map.

## 3 Recognition Method

Our ground truth for measuring correct recognition is a viseme transcription produced by converting a phonetic transcript of the training data to viseme

labels assuming the mapping being tested (Tables 1 & 2). Using HTK [24], we build viseme-level Hidden Markov Model (HMM) recognisers with five states and five mixture components per state. We implement a leave-one-out seven-fold cross validation. Seven folds are selected as we have seven utterances of the alphabet per talker in AVL2. The HMMs are initialised using 'flat start' training and re-estimated eight times and then force-aligned using HTK's `HVite`. Training is completed by re-estimating the HMMs three more times.

## 4 Comparison of current P2V maps results

We measure recognition performance of the HMMs by correctness, $C$, as there are no insertion errors to consider at the word level (AVL2 contains isolated words). Correctness is measured using:

$$C = \frac{N - D - S}{N}, \tag{1}$$

where $S$ is the number of substitution errors, $D$ is the number of deletion errors and $N$ the total number of labels in the reference transcriptions.

Word recognition is less accurate than viseme recognition. However, viseme recognition performance is not a fair test since each viseme set has a different number of visemes. Instead, words are a common comparator that can be cross-referenced from each viseme set, and ultimately it is the difference between sets that we are interested in rather than the absolute level of performance.

Figure 2 shows mean word correctness ± one standard error over all talkers for each consonant map along the $x$-axis paired with each vowel map. Figure 3 shows the same but for each vowel map along the $x$-axis paired with each consonant map. The black line is the mean word correctness. Both $x$-axes are ordered by the mean correctness. This means we can see clearly that the 'best' performing map for both consonants and vowels are from Lee (as this is left-most on the $x$-axis) for all talkers.

Comparing the consonant P2V maps in Figure 2 we see that the Disney vowels are significantly worse than all others when paired with all consonant maps. Over the other vowels there is overlap with the majority of error bars suggesting little significant difference over the whole group but Lee [17] and Bozkurt [14] vowels are consistently above the mean and above the upper error bar for Disney [15], Jeffers [16] and Hazen [4] vowels. In comparing the vowel P2V maps in Figure 3, Lee[17] and Hazen [4] are the best consonants by a margin above the mean whereas Woodward [23] and Franks [7] vie for bottom performance. The best performance in terms of correctness is of a combination of vowels from Lee and consonants from Jeffers but close second best is a combination of Lee's consonants and vowels and this has a much smaller error bar.

In Table 3 we present data to suggest the best performing vowel P2Vs have a ratio of phonemes-to-visemes around 0.44 (top four CF mean = 0.44), and the better performing consonant maps have a CF of approximately 0.41 (top four CF mean = 0.41) so the better P2V is $<\sim 2$ phonemes per viseme.

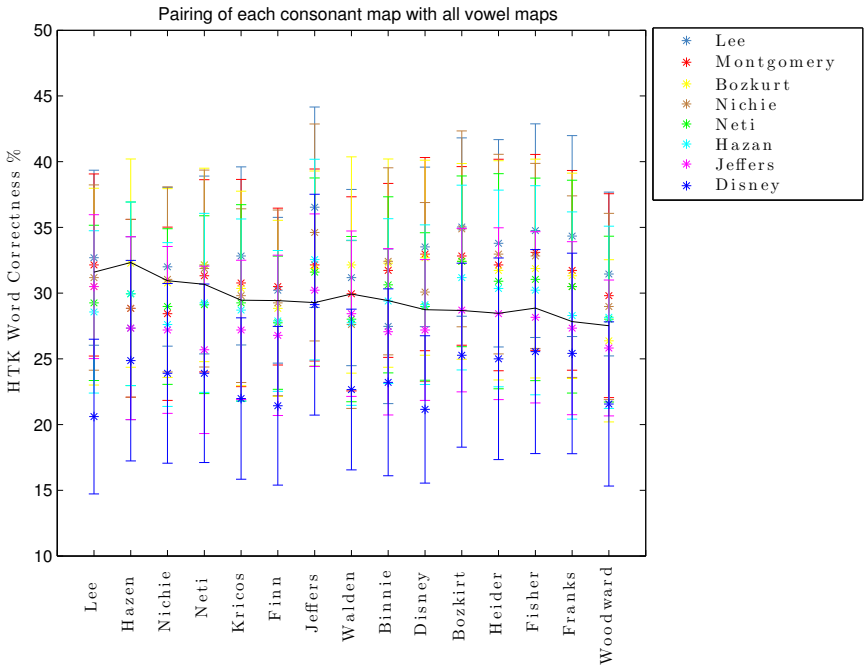



**Fig. 2.** Talker-dependent mean word recognition ± one standard error over all four talkers comparing consonant P2V maps paired with all vowel mappings

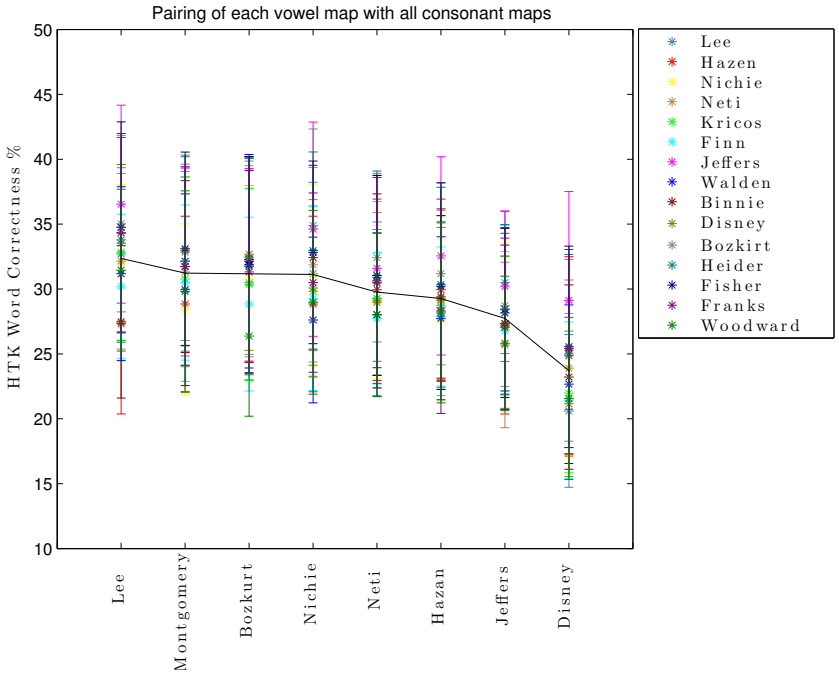



**Fig. 3.** Talker-dependent mean word recognition ± one standard error over all four talkers comparing vowel P2V maps paired with all consonant mappings

**Table 3.** Confusion Factors for each viseme map tested

| Consonant Map | V:P | CF | Mean C | Vowel Map | V:P | CF | Mean C |
|---|---|---|---|---|---|---|---|
| Woodward | 4:24 | 0.16 | 27.52 | Jeffers | 3:19 | 0.16 | 27.74 |
| Disney | 6:22 | 0.18 | 28.74 | Neti | 4:20 | 0.20 | 29.76 |
| Fisher | 5:21 | 0.23 | 28.86 | Hazen | 4:18 | 0.22 | 29.27 |
| Lee | 6:24 | 0.25 | 31.55 | Disney | 4:11 | 0.36 | 23.71 |
| Franks | 5:17 | 0.29 | 27.83 | Lee | 5:14 | 0.36 | 32.35 |
| Kricos | 8:24 | 0.33 | 29.46 | Bozkurt | 7:19 | 0.37 | 31.17 |
| Jeffers | 8:23 | 0.35 | 29.28 | Montgomery | 8:19 | 0.42 | 31.23 |
| Neti | 8:23 | 0.35 | 30.67 | Nichie | 9:15 | 0.60 | 31.13 |
| Bozkurt | 8:22 | 0.36 | 28.67 | - | - | - | - |
| Finn | 10:23 | 0.43 | 29.43 | - | - | - | - |
| Walden | 9:20 | 0.45 | 29.93 | - | - | - | - |
| Binnie | 9:19 | 0.47 | 29.43 | - | - | - | - |
| Hazen | 10:21 | 0.48 | 32.33 | - | - | - | - |
| Heider | 8:16 | 0.50 | 28.47 | - | - | - | - |
| Nichie | 18:33 | 0.54 | 30.94 | - | - | - | - |

## 5 New viseme mappings

Given that Lee [17] provides the best pairing of the existing phoneme to viseme maps, we now ask if there are alternatives that can perform better? Our first approach is to find talker-dependent P2V maps based upon phoneme confusion matrices generated by a visual-only automated recognition system using phoneme HMM classifiers. Where a phoneme is only ever correctly identified as itself (true positives on the confusion matrix diagonal), this is quickly allocated to be a viseme of that single phoneme.

Now we address the remaining phonemes which have been confused. The first candidate for viseme class 1 is a subset of Phonemes: $V_1 = \{\phi_1, \phi_2, \phi_{M_1}\}$ such that every pair, $(\phi_i, \phi_j)$ in $V_1$ has $N_{ij} > 0$. $V_1$ is chosen as the largest such set. $V_2$, which is the second viseme set, is determined in the same way from the remaining phonemes until all phonemes are accounted for. Within this process phonemes are grouped into a viseme class only if *all* of the phonemes within the candidate group are mutually confused. Once a phoneme has been assigned to a viseme class, it is no longer considered for grouping and so any possible other viseme combinations that include this phoneme are discarded.

Our phoneme recognition produces confusions between consonant and vowel phonemes so we make two types of map, one that permits vowel and consonant phonemes to be mixed within the same viseme and a second which restricts visemes to be vowel or consonant phonemes only. These P2V maps for each talker are in Table 4. These are the "tightly confused" maps because all phonemes within each viseme have been confused with each other in the phoneme recognition.

These viseme sets will contain spurious phonemes that cannot be grouped into a viseme because they are not confused with *all* of the phonemes of the viseme. This leaves some single-phoneme visemes (e.g. /u/ in Talker 1 with mixed

**Table 4.** Tightly confused phoneme talker-dependent visemes. The score in brackets is the ratio of phonemes to visemes

| Classification P2V mapping - permitting mixing of vowels and consonants | |
|---|---|
| Talker1 | {/ʌ/ /ai/ /i/ /n/ /əʊ/} {/b/ /e/ /ei/ /y/ } {/d/ /s/} {/tʃ/ /l/} |
| (CF:0.48) | {/t/} {/w/} {/f/} {/k/} {/ə/ /v/} {/dʒ/ /z/} {/ɑ/ /u/} |
| Talker2 | {/ə/ /ai/ /ei/ /i/ /s/} {/e/ /v/ /w/ /y/} {/l/ /m/ /n/} {/ʌ/ /f/} |
| (CF: 0.44) | {/z/} tʃ/} {/t/} {/ɑ/} {/əʊ/ /u/} {/dʒ/ /k/} {/b/ /d/ /p/} |
| Talker3 | {/ei/ /f/ /n/} {/d/ /t/ /p/} {/b/ /s/} {/l/ /m/} {/ə/ /e/} {/i/} |
| (CF: 0.68) | {/ɑ/} {/dʒ/} {/əʊ/} {/z/} {/y/} {/tʃ}/ {/ai/} {/ʌ/} {/ɑ/} {/dʒ/} {/k/ /w/} {/əʊ/} {/z/} {/v/} {/u/} |
| Talker4 | {/ʌ/ /ai/ /i/ /ei/ } {/m/ /n/} {/ə/ /e/ /p/} {/k/ /w/} {/d/ /s/} |
| (CF: 0.64) | {/f/} {/v/} {/ɑ/} {/z/} {/tʃ/} {/b/} {/əʊ/} {/dʒ/ /t/} {/b/} {/əʊ/} {/l/} {/u/} |
| Classification P2V mapping - restricting mixing of vowels and consonants | |
| Talker1 | {/ʌ/ /i/ /əʊ/ /u/} {/ɑ/ /ei/} {/ə/ /e/ /ei/} {/d/ /s/ /t/ } {/tʃ/ /l/} |
| (CF:0.50) | {/k/} {/z/} {/w/} {/f/} {/m/ /n/} {/dʒ/ /v/} {/b/ /y/} |
| Talker2 | {/ai/ /ei/ /i/ /u/} {/əʊ/} {/ə/} {/e/} {/ʌ/} {/ɑ/} {/v/ /w/} {/k/} |
| (CF: 0.58) | {/d/ /b/} {/t/} {/tʃ/} {/l/ /m/ /n/} {/dʒ/ /p/ /y/} {/f/ /s/} |
| Talker3 | {/ei/ /i/} {/ai/} {/ə/ /e/} {/ʌ/} {/d/ /p/ /t/} {/l/ /m/} {/k/ /w/} |
| (CF: 0.68) | {/tʃ/} {/əʊ/} {/y/} {/u/} {/ɑ/} {/z/} {/b/ /s/} {/v/} {/dʒ/} {/f/ /n/} |
| Talker4 | {/ʌ/ /ai/ /i/ /ei/} {/ə/ /e/} {/m/ /n/} {/k/ /l/} {/dʒ/ /t/} {/b/} |
| (CF: 0.65) | {/əʊ/} {/y/} {/u/} {/ɑ/} {/w/} {/f/} {/v/} {/tʃ/} {/d/ /s/} |

vowel and consonant phonemes), so our second approach relaxes the condition requiring confusion with all of the phonemes. We execute a second pass through the viseme sets. Any single-phoneme viseme classes are then permitted to merge with existing multi-phoneme classes if they share any confusions with that class. In the event that a phoneme has multiple class confusions it is merged with the class with the greatest confusion. We term these the "loosely confused" maps. Again we do two sets with vowel and consonant phonemes both mixed and separate. The final P2V maps are in Table 5 for four talkers.

Looking at Tables 4 and 5 there are no identical visemes with each map type between talkers, this confirms our variability of individual talker visual speech (excluding the true positive single phoneme visemes). We observe that none of the new visemes match the previously suggested visemes in the comparison study (Tables 1 and 2), e.g. the most common previous viseme was {/p/ /b/ /m/} and this is never created with our new method.

Figure 4 shows the word recognition performance using both the tightly confused map and the loosely confused map for each talker. Also shown is the performance using the Lee map as a benchmark. For Talker 1 no new viseme map significantly improves upon the benchmark performance, but we do see significant improvements for both Talker 2 and Talker 4 and a minor improvement within the error bars for Talker 3. For Talkers 2 and 3, both types of the split vowels and consonant maps demonstrate improvement on the benchmark, and for Talker 4 the tightly confused split vowels and consonants shows a significant

**Table 5.** Loosely confused phoneme talker-dependent visemes. The score in brackets is the ratio of phonemes to visemes

| Classification | P2V mapping - permitting mixing of vowels and consonants |
|---|---|
| Talker1 | {/b/ /e/ /ei/ /p/ /w/ /y/ /k/} {/ʌ/ /ai/ /f/ /i/ /m/ /n/ /əʊ/} |
| (CF:0.28) | {/dʒ/ /z/} {/ɑ/ /u/} {/d/ /s/ /t/} {/tʃ/ /l/} {/ə/ /v/} |
| Talker2 | {/ɑ/ /ə/ /ai/ /ei/ /i/ /s/ /tʃ/} {/e/ /t/ /v/ /w/ /y/} {/l/ /m/ /n/} |
| (CF: 0.32) | {/ʌ/ /f/} {/z/} {/b/ /d/ /p/} {/əʊ/ /u/} {/dʒ/ /k/} |
| Talker3 | {/ʌ/ /ai/ /ei/ /f/ /i/ /n/} {/ə/ /e/ /y/ /tʃ/} {/b/ /s/ /v/} |
| (CF: 0.40) | {/dʒ/} {/əʊ/} {/z/} {/l/ /m/ /u/} {/d/ /p/ /t/} {/k/ /w/} {/ɑ/} |
| Talker4 | {/ʌ/ /ai/ /tʃ/ /i/ /ei/ } {/ɑ/ /m/ /u/ /n/} {/ə/ /e/ /p/ /v/ /y/} |
| (CF: 0.32) | {/dʒ/ /t/} {/k/ /l/ /w/} {/əʊ/} {/d/ /f/ /s/} {/b/} |
| Classification | P2V mapping - restricting mixing of vowels and consonants |
| Talker1 | {/ʌ/ /i/ /əʊ/ /u/} {/ɑ/ /ai/} {/ə/ /e/ /ei/} {/b/ /w/ /y/} |
| (CF:0.47) | {/k/} {/z/} {/m/} {/l/} {/d/ /f/ /s/ /t/} {/tʃ/} {/dʒ/ /k/ /v/ /z/} |
| Talker2 | {/ɑ/ /ʌ/ /ə/ /ai/ /ei/ /i/ /əʊ/ /u/} {/k/ /t/ /v/ /w/} |
| (CF: 0.29) | {/f/ /s/} {/tʃ/ /l/ /m/ /n/} {/dʒ/ /p/ /y/} {/b/ /d/} {/z/} |
| Talker3 | {/ʌ/ /ai/ /i/ /ei/} {/ə/ /e/} {/b/ /s/ /v/} {/d/ /p/ /t/} |
| (CF: 0.56) | {/y/} {/dʒ/} {/əʊ/} {/z/} {/u/} {/ə/ /e/} {/l/ /m/} {/k/ /w/} {/f/ /n/} {/ɑ/} {/tʃ/} |
| Talker4 | {/ʌ/ /ai/ /i/ /ei/} {/tʃ/ /k/ /l/ /w/} {/d/ /f/ /s/ /v/} {/m/ /n/} |
| (CF: 0.50) | {/f/} {/ɑ/} {/dʒ/ /t/} {/əʊ/} {/u/} {/y/} {/b/} |

improvement. Comparing mixed consonant and vowel maps against split consonant and vowel maps, the split maps are always better than mixed maps for all talkers in this data. In comparing the loosely confused maps versus the tightly confused maps, the tight confusions are better for two out of our four talkers (Talkers 1 and 2) and equal for a third (Talker 1). These are talkers with highest confusion factor P2V maps (Tables 4 & 5). This is despite the tightly confused viseme set including single phoneme-viseme classes which can be confused with parts of the tightly confused classes.

## 6 Conclusions and Future work

We have completed a comprehensive experimental study of previously suggested P2V maps and shown that Lee [17] is the best of the previously published P2V maps. Puzzlingly the Lee mapping is not that popular among engineers of lip-reading systems so our finding should be of immediate use.

We have also outlined how it is possible to build phoneme-to-viseme maps in a systematic way using confusion matrices from real recognisers. We believe that this is a more principled approach than previous methods (including Lee's [17] whose method is bound by the Fisher [3] visemes) and also allows comparison between talkers using phonetic terminology. Further we have shown that the automatic method need do no worse than the Lee visemes and can exceed performance. We acknowledge that our dataset is still rather small and the sparsely represented phonemes are unlikely to be accurately modelled. In future we would

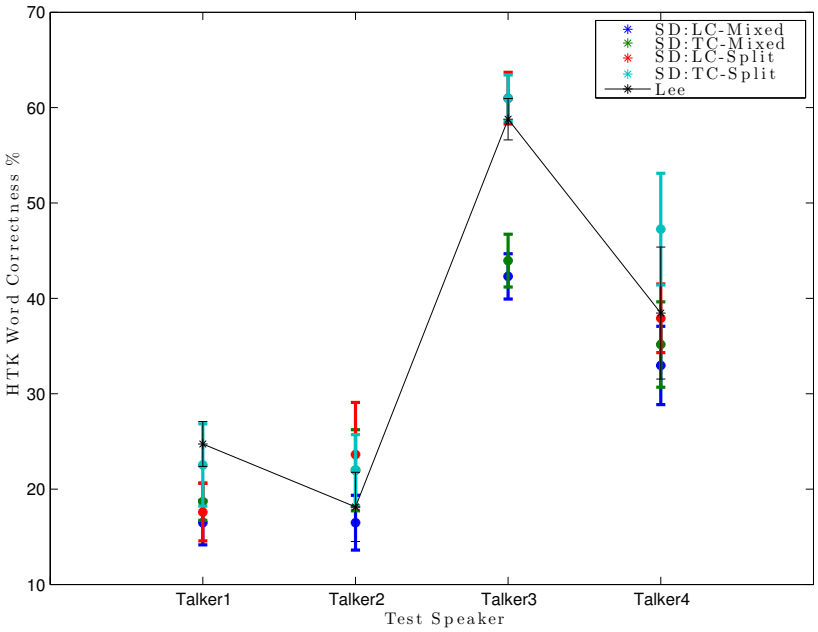



**Fig. 4.** HTK word Correctness using tightly confused and loosely confused viseme sets based on phoneme recognition confusions. SD = Speaker Dependent, LC = Loosely coupled, TC = Tightly Coupled, Mixed = Mixed vowels and consonant phonemes within viseme classes and Split = separated vowel and consonant visemes

like to extend this to full set of American and British phonemes but that will require a more extensive set of data.

## References

1. Association, I.P.: Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press (1999)
2. Chen, T., Rao, R.R.: Audio-visual integration in multimodal communication. Proceedings of the IEEE **86** (1998) 837–852
3. Fisher, C.G.: Confusions among visually perceived consonants. Journal of Speech, Language and Hearing Research **11** (1968) 796
4. Hazen, T.J., Saenko, K., La, C.H., Glass, J.R.: A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In: Proceedings of the 6th International Conference on Multimodal Interfaces. ICMI '04, New York, NY, USA, ACM (2004) 235–242
5. Theobald, B.J.: Visual speech synthesis using shape and appearance models. PhD thesis, University of East Anglia (2003)

6. Binnie, C.A., Jackson, P.L., Montgomery, A.A.: Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. Journal of Speech and Hearing Disorders **41** (1976) 530
7. Franks, J.R., Kimble, J.: The confusion of english consonant clusters in lipreading. Journal of Speech, Language and Hearing Research **15** (1972) 474
8. Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C.J.: Effects of training on the visual recognition of consonants. Journal of Speech, Language and Hearing Research **20** (1977) 130
9. Kricos, P.B., Lesner, S.A.: Differences in visual intelligibility across talkers. The Volta Review (1982)
10. Owens, E., Blazek, B.: Visemes observed by hearing-impaired and normal-hearing adult viewers. Journal of Speech and Hearing Research **28** (1985) 381
11. Cox, S., Harvey, R., Lan, Y., Newman, J., Theobald, B.J.: The challenge of multi-speaker lip-reading. In: International Conference on Auditory-Visual Speech Processing, Citeseer (2008) 179–184
12. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60** (2004) 135–164
13. Cappelletta, L., Harte, N.: Phoneme-to-viseme mapping for visual speech recognition. In: ICPRAM (2). (2012) 322–329
14. Bozkurt, E., Erdem, C., Erzin, E., Erdem, T., Ozkan, M.: Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. Proc. of Signal Proc. and Communications Applications (2007) 1–4
15. Lander, J.: Read my lips: Facial animation techniques. http://www.gamasutra.com/view/feature/131587/read_my_lips_facial_animation_.php (2014) Accessed: 2014-01-28.
16. Jeffers, J., Barley, M.: Speechreading (lipreading). Thomas Springfield, IL: (1971)
17. Lee, S., Yook, D.: Audio-to-visual conversion using hidden markov models. In: PRICAI 2002: Trends in Artificial Intelligence. Springer (2002) 563–570
18. Montgomery, A.A., Jackson, P.L.: Physical characteristics of the lips underlying vowel lipreading performance. The Journal of the Acoustical Society of America **73** (1983) 2134
19. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J.: Audio-visual speech recognition. In: Final Workshop 2000 Report. Volume 764. (2000)
20. Nitchie, E.B.: Lip-Reading, principles and practise: A handbook for teaching and self-practise. Frederick A Stokes Co, New York (1912)
21. Finn, K.E., Montgomery, A.A.: Automatic optically-based recognition of speech. Pattern Recognition Letters **8** (1988) 159–164
22. Heider, F., Heider, G.M.: An experimental investigation of lipreading. Psychological Monographs **52** (1940) 124–153
23. Woodward, M.F., Barber, C.G.: Phoneme perception in lipreading. Journal of Speech, Language and Hearing Research **3** (1960) 212
24. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchec, V., Woodland, P.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department (2006)

# Speaker-independent machine lip-reading with speaker-dependent viseme classifiers

*Helen L. Bear*[1]*, Stephen J. Cox*[1]*, Richard W. Harvey*[1]

[1]University of East Anglia, UK

{helen.bear,r.w.harvey,s.j.cox}@uea.ac.uk

## Abstract

In machine lip-reading, which is identification of speech from visual-only information, there is evidence to show that visual speech is highly dependent upon the speaker [1]. Here, we use a phoneme-clustering method to form new phoneme-to-viseme maps for both individual and multiple speakers. We use these maps to examine how similarly speakers talk visually. We conclude that broadly speaking, speakers have the same repertoire of mouth gestures, where they differ is in the use of the gestures.
**Index Terms**: visual-only speech recognition, computer lip-reading, visemes, classification, pattern recognition, speaker-independence

## 1. Introduction

Speaker identity is known to be important in the recognition of speech from visual-only information (lip-reading) [1], more so than in audio speech. One of the difficulties in dealing with visual speech is what the findamental units for recognition should be. The term *viseme* is loosely defined [2] 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 [3], [2] or [4]. In [5], a new idea of using speaker-dependent visemes is presented. The method can be summarised as follows:

1. Perform speaker-dependent phoneme recognition with recognisers that use phoneme units.
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. Phonemes are clustered into groups (visemes) based on the confusions identified in step two. 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 $V$. Consonant and vowel phonemes are not permitted to be mixed within a viseme class. The result of this process is a Phoneme-to-Viseme (P2V) map $M$ for each speaker—for further details, see [5].
4. These new speaker-dependent viseme sets are then used as units for visual speech recognition for a speaker.

This resulted in a small improvement in speaker-dependent recognition [5]. 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.

## 2. Dataset description

We use the AVLetters2 (AVL2) dataset [1], to train and test recognisers based upon the new P2V mappings. This dataset consists of four British-English speakers reciting the alphabet seven times. The full-faces of the speakers are tracked using Active Appearance Models (AAMs) [6] from which lip-only combined shape and appearance features are extracted. We select AAM features because they are known to out-perform other feature methods in machine visual-only lip-reading [7]. Figure 1 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 BEEP pronunciation dictionary used throughout these experiments is in British English [8].

## 3. Method overview

We use the clustering approach of [5] to produce a series of P2V maps. We construct

1. a speaker-dependent map for each speaker;
2. a multi-speaker map using *all* speakers' phoneme confusions;
3. a speaker-independent map for each speaker using confusions of all *other* speakers in the data.

Each P2V map is constructed using separate training and test data by using seven fold cross-validation [9]. In total each speaker utters 182 words (seven recitations of 26 words). Each one of seven recitations of the alphabet are selected as test folds in turn and are not included in the training folds.

We then use the HTK toolkit [10] to build Hidden Markov Model (HMM) classifiers whose models are the viseme classes in each P2V map. We flat-start the HMMs with `HCompV`, re-estimate them 11 times over (`HERest`) with forced alignment between seventh and eighth re-estimates. Finally we recognise using `HVite` and output our results with `HResults`. The models are three state HMMs each having an associated Gaussian mixture of five components. Our recognition network constrains the output to be one of the 26 letters of the alphabet. Therefore, our measure of accuracy is $\frac{\#letterscorrect}{\#lettersclassified}$.

## 4. Experimental setup

We designate the P2V maps formed in these experiments as

$$M_n(p, q) \tag{1}$$

This means that 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, $M_1(2, 3)$

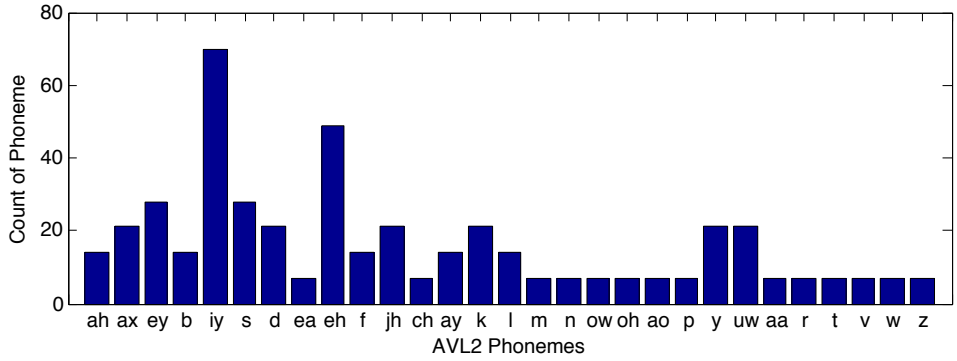



Figure 1: Phoneme histogram of AVLetters-2 dataset

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.

### 4.1. Baseline: Same Speaker Dependent maps (SSD)

We establish a baseline of performance using the speaker-dependent results: $M_1(1,1), M_2(2,2), M_3(3,3)$ and $M_4(4,4)$. They are same speaker dependent (SSD) because the map, the models and the testing data are all derived from the same speaker. Table 1 depicts how these maps are constructed. The resulting SSD P2V maps are listed in Table 3. The /garb/

| Same speaker-dependent (SD) | | |
|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) |
| Sp1 | Sp1 | Sp1 |
| Sp2 | Sp2 | Sp2 |
| Sp3 | Sp3 | Sp3 |
| Sp4 | Sp4 | Sp4 |

Table 1: Same Speaker-Dependent (SSD) experiments used as a baseline for comparison

viseme is made up of phonemes which did not appear in the output from the recogniser. Each viseme is listed with its associated mutually-confused phonemes e.g. for $M_1$, we see /v01/ is made up of phonemes {/ah/, /iy/, /ow/, /uw/}. These means in the phoneme recognition, all four phonemes {/ah/, /iy/, /ow/, /uw/} were confused with the other three in the viseme.

### 4.2. Different Speaker Dependent maps & Data (DSD&D)

In these tests, we use the HMM recognisers trained on each single speaker to recognise data from different speakers. This is done 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 $M_2(2,1), M_3(3,1)$ and $M_4(4,1)$ and so on for the other speakers—this is depicted in Table 2.

### 4.3. Different Speaker Dependent maps (DSD)

In our next experiment we train our models on speech from a single speaker but vary the speaker-dependent maps. This isolates the effects of the HMM recognition from the effect of different viseme classes. So for Speaker 1, we test the following

| Different Speaker Dependent maps & Data (DSD&D) | | |
|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) |
| Sp2,Sp3,Sp4 | Sp2,Sp3,Sp4 | Sp1 |
| Sp1,Sp3,Sp4 | Sp2,Sp3,Sp4 | Sp2 |
| Sp1,Sp2,Sp4 | Sp3,Sp2,Sp4 | Sp3 |
| Sp1,Sp2,Sp3 | Sp4,Sp2,Sp3 | Sp4 |

Table 2: Different Speaker Dependent maps & Data (DSD&D) experiments

Speaker-Independent Maps: $M_2(1,1), M_3(1,1)$ and $M_4(1,1)$. These are the same P2V maps in Table 3 but trained and tested differently. This is depicted in Table 4.

| Different Speaker Dependent maps (DSD) | | |
|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) |
| Sp2,Sp3,Sp4 | Sp1 | Sp1 |
| Sp1,Sp3,Sp4 | Sp2 | Sp2 |
| Sp1,Sp2,Sp4 | Sp3 | Sp3 |
| Sp1,Sp2,Sp3 | Sp4 | Sp4 |

Table 4: Different Speaker Dependent maps (DSD) experiments

### 4.4. Multi-speaker maps (MS)

In the third set of experiments, we use the multi-speaker (MS) P2V map to form the viseme classes. This map is constructed using phoneme confusions produced by *all* our speakers and is shown in Table 6. We test this map as follows:

| Multi-Speaker (MS) | | |
|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) |
| Sp1234 | Sp1 | Sp1 |
| Sp1234 | Sp2 | Sp2 |
| Sp1234 | Sp3 | Sp3 |
| Sp1234 | Sp4 | Sp4 |

Table 5: Multi-Speaker (MS) experiments used as a baseline for comparison

$M_{[1234]}(1,1), M_{[1234]}(2,2), M_{[1234]}(3,3)$ and $M_{[1234]}(4,4)$: this is explained in Table 5.

| Speaker 1 $M_1$ | | Speaker 2 $M_2$ | | Speaker 3 $M_3$ | | Speaker 4 $M_4$ | |
|---|---|---|---|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes | Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ah/ /iy/ /ow/ /uw/ | /v01/ | /ay/ /ey/ /iy/ /uw/ | /v01/ | /ey/ /iy/ | /v01/ | /ah/ /ay/ /ey/ /iy/ |
| /v02/ | /ax/ /eh/ /ey/ | /v02/ | /ow/ | /v02/ | /ax/ /eh/ | /v02/ | /ax/ /eh/ |
| /v03/ | /aa/ /ay/ | /v03/ | /ax/ | /v03/ | /ay/ | /v03/ | /aa/ |
| /v04/ | /d/ /s/ /t/ | /v04/ | /eh/ | /v04/ | /ah/ | /v04/ | /ow/ |
| /v05/ | /ch/ /l/ | /v05/ | /ah/ | /v05/ | /aa/ | /v05/ | /uw/ |
| /v06/ | /m/ /n/ | /v06/ | /aa/ | /v06/ | /ow/ | /v06/ | /m/ /n/ |
| /v07/ | /jh/ /v/ | /v07/ | /jh/ /p/ /y/ | /v07/ | /uw/ | /v07/ | /k/ /l/ |
| /v08/ | /b/ /y/ | /v08/ | /l/ /m/ /n/ | /v08/ | /d/ /p/ /t/ | /v08/ | /jh/ /t/ |
| /v09/ | /k/ | /v09/ | /v/ /w/ | /v09/ | /l/ /m/ | /v09/ | /d/ /s/ |
| /v10/ | /z/ | /v10/ | /d/ /b/ | /v10/ | /k/ /w/ | /v10/ | /w/ |
| /v11/ | /w/ | /v11/ | /f/ /s/ | /v11/ | /f/ /n/ | /v11/ | /f/ |
| /v12/ | /f/ | /v12/ | /t/ | /v12/ | /b/ /s/ | /v12/ | /v/ |
| | | /v13/ | /k/ | /v13/ | /v/ | /v13/ | /ch/ |
| | | /v14/ | /ch/ | /v14/ | /jh/ | /v14/ | /b/ |
| | | | | /v15/ | /ch/ | /v15/ | /y/ |
| | | | | /v16/ | /y/ | | |
| | | | | /v17/ | /z/ | | |
| /sil/ | /sil/ | /sil/ | /sil/ | /sil/ | /sil/ | /sil/ | /sil/ |
| /garb/ | /ea/ /oh/ /ao/ /r/ /p/ | /garb/ | /ea/ /oh/ /ao/ /r/ /z/ | /garb/ | /ea/ /oh/ /ao/ /r/ | /garb/ | /ea/ /oh/ /ao/ /r/ /p/ /z/ |

Table 3: Speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for each speaker in AVL2

| Multi-Speaker $M_{1234}$ | |
|---|---|
| Viseme | Phonemes |
| /v01/ | /ah/ /ay/ /ey/ /iy/ /ow/ /uw/ |
| /v02/ | /ax/ /eh/ |
| /v03/ | /aa/ |
| /v04/ | /d/ /s/ /t/ /v/ |
| /v05/ | /f/ /l/ /n/ |
| /v06/ | /b/ /w/ /y/ |
| /v07/ | /jh/ |
| /v08/ | /z/ |
| /v09/ | /p/ |
| /v10/ | /m/ |
| /v11/ | /k/ |
| /v12/ | /ch/ |
| /sil/ | /sil/ |
| /garb/ | /ea/ /oh/ /ao/ /r/ |

Table 6: Phoneme-to-viseme mapping derived from phoneme recognition confusions for all four speakers in AVL2

### 4.5. Speaker-Independent maps (SI)

Finally, we use our phoneme-clustering method to create a set of Speaker-Independent (SI) maps for each of the four speakers. These final P2V maps are shown in Table 8. We test these maps

| Speaker-Independent maps (SI) | | |
|---|---|---|
| Mapping ($M_n$) | Training data ($p$) | Test speaker ($q$) |
| Sp234 | Sp1 | Sp1 |
| Sp134 | Sp2 | Sp2 |
| Sp124 | Sp3 | Sp3 |
| Sp123 | Sp4 | Sp4 |

Table 7: Speaker-Independent (SI) maps experiments

as follows $M_{234}(1,1), M_{134}(2,2), M_{124}(3,3)$ and $M_{123}(4,4)$ as shown in Table 7.

### 4.6. Homophones

| Map | Unique words $T$ |
|---|---|
| $M_1$ | 19 |
| $M_2$ | 19 |
| $M_3$ | 24 |
| $M_4$ | 24 |
| $M_{1234}$ | 14 |
| $M_{234}$ | 17 |
| $M_{134}$ | 18 |
| $M_{124}$ | 20 |
| $M_{123}$ | 15 |

Table 9: Homophones created by each P2V mapping, allowing for variation in pronunciation

Because the P2V maps are a many-to-one mapping, there is the possibility of creating visual homophones. For example, the phonetic realisation of the word 'B' is *b iy* and of 'D' is *d iy*. Using map $M_2(2,2)$ they become B = *v08 v01* and D = *v08 v01* which are indistinguishable. The vocabulary of AVL2 is the 26 letters, A–Z. Permitting variations in pronunciation, we show the total unique words ($T$) for each map after each word (letter) has been translated from words, to phonemes, to visemes in Table 9. The higher the volume of homophones, the greater the chance of substitution errors.

## 5. Results

Figure 2 shows the word recognition of speaker-dependent viseme classes, measured by correctness. In this figure, our baseline is $n = p = q$ for all $M$. We compare these to: $M_2(2,1), M_3(3,1), M_4(4,1)$ 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. DSD HMM recognisers are significantly worse than SSD HMMs, as all results where $p$ is not the same speaker as $q$ are around the

| Speaker 1 $M_{234}$ | | Speaker 2 $M_{134}$ | | Speaker 3 $M_{124}$ | | Speaker 4 $M_{123}$ | |
|---|---|---|---|---|---|---|---|
| Viseme | Phonemes | Viseme | Phonemes | Viseme | Phonemes | Viseme | Phonemes |
| /v01/ | /ah/ /ax/ /ay/ /ey/ /iy/ | /v01/ | /ah/ /ay/ /ey/ /iy/ | /v01/ | /ah/ /ay/ /ey/ /iy/ /ow/ /uw/ | /v01/ | /ah/ /ay/ /ey/ /iy/ /ow/ /uw/ |
| v02 | ow uw | v02 | aa ow uw | v02 | aa | v02 | aa |
| v03 | eh | v03 | ax eh | v03 | ax eh | v03 | ax eh |
| v04 | aa | v04 | d s t | v04 | d s t v | v04 | jh s t v |
| v05 | d s t v | v05 | ch l | v05 | l m n | v05 | f l n |
| v06 | l m n | v06 | b jh | v06 | b w y | v06 | b d p |
| v07 | jh p y | v07 | v y | v07 | jh | v07 | w y |
| v08 | k w | v08 | k w | v08 | z | v08 | z |
| v09 | f | v09 | p | v09 | p | v09 | m |
| v10 | ch | v10 | z | v10 | k | v10 | k |
| v11 | b | v11 | m | v11 | f | v11 | ch |
| | | | | v12 | ch | | |
| sil | sil | sil | sil | sil | sil | sil | sil |
| garb | ea oh ao r z | garb | ea oh ao r f n | garb | ea oh ao r iy | garb | ea oh ao r |

Table 8: Phoneme-to-viseme mapping derived from phoneme recognition confusions of the three other speakers in AVL2

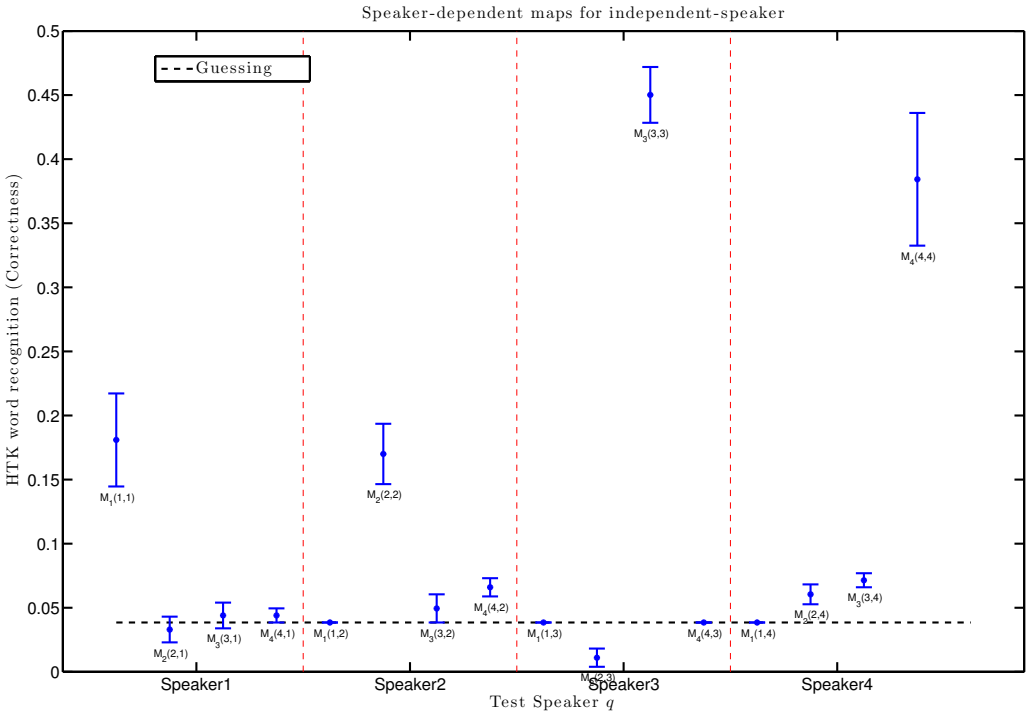



Figure 2: *Word recognition measured by correctness of the DSD&D trained HMM classifiers used on all three other speakers in AVL2. Baseline is the SSD maps and error bars show ± one standard error.*

equivalent performance of guessing. This correlates with similar tests of independent HMM's in [1]. We can attribute this gap to two possible effects, either - the visual units are incorrect, or they are trained on the incorrect speaker.

In Figure 3 we have repeated the same benchmark as in Figure 2($n = p = q$), but we have now allowed the HMM 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 that the previous poor performance was not due to the choice of visemes but rather, the badly trained HMMs.

We rank the performance of each viseme set on each speaker by weighting the effect of the DSD tests. We score each map as in Table 10. If a map increases on SSD performance within error bar range this scores $+1$ or outside error bar range scores $+2$. Likewise if a map decreases recognition on SSD performance, these values are negative.

So we see that $M-3$ is the best of the four SSD maps, followed by $M_4$, $M_2$ and finally $M_1$ is the most susceptible to speaker identity. We note that this order matches a decreasing order of quantity of visemes in the speaker-dependent viseme sets i.e. the more similar to phoneme classes visemes are, then the better the recognition performance. This ties in with Table 9,

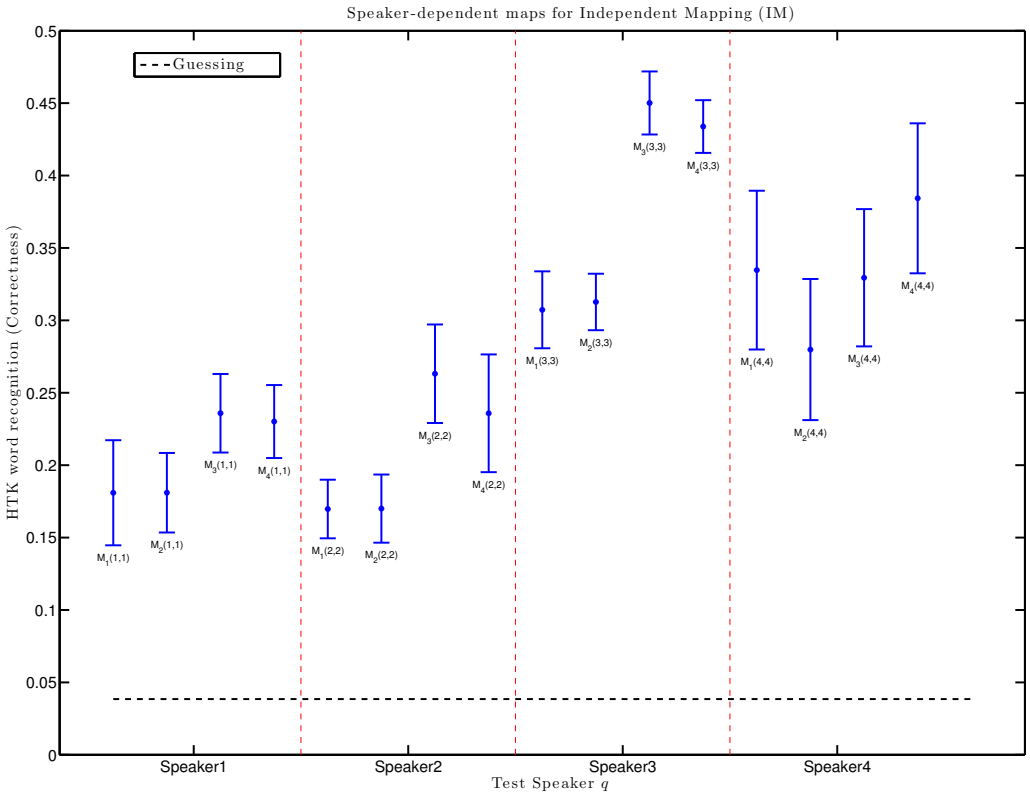



Figure 3: *Word recognition measured by correctness of the DSD classifiers constructed with single-speaker independent P2V maps for all four speakers in AVL2. Baseline is the SSD maps and error bars show* $\pm$ *one standard error.*

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| 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 10: Weighted scores from comparing the use of speaker-dependent maps for *other* speaker-dependent lip-reading

where the better P2V maps have less homorphous words.

In Table 3, phoneme pairs {/ax/, /eh/}, {/m/, /n/} and {/ey/, /iy/} are present for three speakers and {/ah/, /iy/} and {/l/, /m/} are pairs for two speakers. Of the single-phoneme visemes, /ch/ is present three times, /f/, /k/, /w/ & /z/ twice.

The important lesson from Figure 3, is that the selection of incorrect units, whilst detrimental, is not as devastating as training recognition classes on alternative speakers.

Figure 4 shows the correctness of both the MS viseme class set and the SI sets. For the multi-speaker classifiers, these are all built on the same map $M_{1234}$, and tested on the same speaker so, $p = q$. Therefore our tests are: $M_{1234}(1,1), M_{1234}(2,2), M_{1234}(3,3), M_{1234}(4,4)$. To test our SI maps, we plot $M_{234}(1,1), M_{134}(2,2), M_{124}(3,3)$ and $M_{123}(4,4)$. Again we repeat the same baseline where $n = p = q$ for reference.

There is no significant difference on Speaker 2, and while Speaker 3 word recognition is reduced, it is not eradicated. It is interesting that for Speaker 3, for whom their speaker-dependent recognition was the best of all speakers, the SIM map ($M_{124}$) out performs the multi-speaker viseme classes ($M_{1234}$) significantly. This maybe due to Speaker 3 having a unique visual talking style which reduces similarities with Speakers 1, 2 & 4.

If we compare all the P2V maps in Tables 6 & 8, there are similarities. Mostly because we know there is only one speaker at a time removed from within SIM P2V maps. However, if we compare these to the speaker-dependent maps in Table 3, we see a different picture. Speaker 4 is significantly affected by the introduction of /ow/ and /uw/ into viseme /v01/. Where Speaker 1 has these in $M_1(1,1)$, we see that his SD word recognition of $15.9\%$ is less than half of Speaker 4's $38.4\%$ (Figure 3).

## 6. Conclusions

Our principal conclusion can be seen by comparing Figures 3 & 4 with Figure 2. Figure 2 shows a very substantial reduction in performance when the system is truing on a speaker who is not the test speaker. 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 is it 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.

This is an important conclusion since it tells us that the repertoire of lip appearances does not vary significantly across speakers. This is comforting since the prospect of recognition using a symbol alphabet which varies by speaker is daunting.

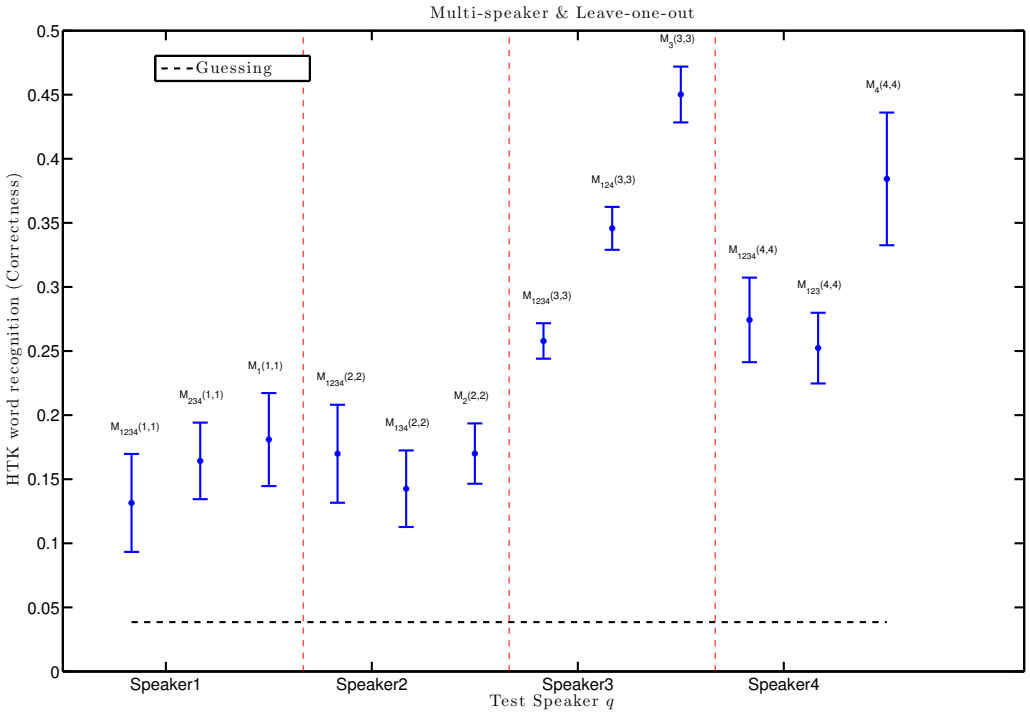



Figure 4: *Word recognition measured by correctness of the classifiers using MS and SI phoneme-to-viseme maps. Baseline is the SSD maps and error bars show ± one standard error.*

This is further reinforced by Tables 3, 6 & 8. There are differences between speakers, but not significant ones.

An analogy with acoustic speech would be the question of whether an accented Norfolk speaker requires a different set of phonemes to a standard British talker. No: they can be represented by the same set of phonemes; they just use these phonemes in a different way.

## 7. References


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

[2] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language and Hearing Research*, vol. 11, no. 4, p. 796, 1968.

[3] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

[4] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ser. ICMI '04. New York, NY, USA: ACM, 2004, pp. 235–242. [Online]. Available: http://doi.acm.org/10.1145/1027933.1027972

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

[6] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004. [Online]. Available: http://www.springerlink.com/openurl.asp?

[7] L. Cappelletta and N. Harte, "Phoneme-to-viseme mapping for visual speech recognition." in *ICPRAM (2)*, 2012, pp. 322–329.

[8] Cambridge University, UK. (1997) BEEP pronunciation dictionary. [Online]. Available: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

[9] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

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

[11] H. L. Bear, G. Owen, R. Harvey, and 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. 92 530G–92 530G.

# Finding phonemes: improving machine lip-reading

*Helen L. Bear*[1]*, Richard W. Harvey*[1]*, Yuxuan Lan*[1]

[1]University of East Anglia, UK

{helen.bear,r.w.harvey,y.lan}@uea.ac.uk

## Abstract

In machine lip-reading there is continued debate and research around the correct classes to be used for recognition.

In this paper we use a structured approach for devising speaker-dependent viseme classes, which enables the creation of a set of phoneme-to-viseme maps where each has a different quantity of visemes ranging from two to 45. Viseme classes are based upon the mapping of articulated phonemes, which have been confused during phoneme recognition, into viseme groups.

Using these maps, with the LiLIR dataset, we show the effect of changing the viseme map size in speaker-dependent machine lip-reading, measured by word recognition correctness and so demonstrate that word recognition with phoneme classifiers is not just possible, but often better than word recognition with viseme classifiers. Furthermore, there are intermediate units between visemes and phonemes which are better still.
**Index Terms**: visual-only speech recognition, computer lip-reading, visemes, classification, pattern recognition

## 1. Introduction

Although visemes are yet to be formally defined, many possibilities can be found across literature [1, 2, 3, 4]. Here we use the definition "a viseme is a visual cue representative of a subset of phonemes on the lips". Therefore, a set of viseme classifiers is inherently smaller than a set of phoneme classifiers. Whilst this means that there are more training samples per class (addressing the limitation of currently available dataset sizes), this also introduces generalisation between articulated sounds. So, to find optimal viseme classes, we need to minimise this generalisation in order to maximise recognition of correct utterances, but also maximise the use of the data available.

The relationship between phonemes (the units of acoustic speech) and visemes (the units of visual speech) can be described with Phoneme-to-Viseme (P2V) maps. In [1] it is shown how these maps can be derived automatically from phoneme confusions. A by-product of the method is that we can control how many visemes we need. This allows considerable precision when answering questions about the optimal number and nature of visemes.

## 2. Data

Our selected dataset is LiLIR [5]. This data consists of 12 British speakers (seven male and five female), 200 utterances per speaker of resource management context independent sentences from [6] which totals around 1000 words. The original videos were recorded in high definition and in a full-frontal position. Individual speakers are tracked using Active Appearance Models [7] and we extract features of concatenated shape and appearance information.

The pronunciation dictionary used throughout these experiments is British English [8] which we take to be represented by 46 phonemes.

## 3. Method

A high level overview of our method is shown in Figure 1 and is described in [1]. We begin by performing word recognition using classifiers based upon phoneme labels. This provides us with both a baseline to benchmark against and, crucially, a set of confusion matrices for each speaker which are used to cluster together potential monophones.

However, we undertake a different clustering process (section 3.2) during which we make a new P2V mapping each time a phoneme is re-classified to a new viseme grouping, thereby deriving up to 45 (subject to the number of phonemes recognised during the phoneme recognition stage) P2V maps per speaker. These new classifiers (visemes) are then used to repeat our word recognition task.

We use the word recognition as our performance measure as this normalises for variance in training samples for each set of classifiers. We note that it is not the performance itself which is relevant here, rather it is any improvement a variance in classes can provide. The reader should also note that we are not suggesting our clustering process will deliver the optimum visemes but rather address our need in this case for a method to enable a controlled comparison of the visemes.

### 3.1. Step one: phoneme recognition

We implement 10-fold cross-validation with replacement [9], of 200 sentences per speaker, 20 are randomly selected as test samples and these are not included in the training folds. Using the HTK toolkit [10] to use Hidden Markov Model (HMM) classes, we flat-start the HMMs, re-estimate them 11 times with forced alignment between seventh and eighth estimates. Our prototype is based upon a Gaussian mixture of five components and three state HMMs. We use a single-state tied short-pause, or 'sp' HMM for short silences between words in the sentence utterances. We also use a bigram word network to support recognition. There are a maximum of 46 phonemes within our phoneme recognition results, but not all speakers used all phonemes within their speech utterances.

### 3.2. Step two: speaker-dependent phoneme clustering

We cluster the phonemes into new visemes classes as follows; we have 10 confusion matrices for each speaker (one from each fold), these are summed together to form one confusion matrix representing all confusions for that speaker. We start with this phoneme confusion matrix:

$$[K_m]_{ij} = N(\hat{p}_j|p_i) \quad (1)$$

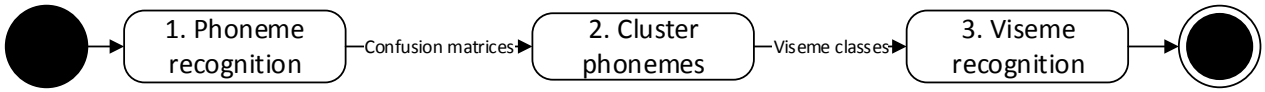



Figure 1: *Three step process for word recognition from visemes.*

| Viseme | Phonemes |
|---|---|
| V01 | /ax/ |
| V02 | /v/ |
| V03 | /oy/ |
| V04 | /f/ /zh/ /w/ |
| V05 | /k/ /b/ /d/ /th/ /p/ |
| V06 | /l/ /jh/ |
| V07 | /g/ /m/ /z/ /y/ /ch/ /dh/ /s/ /r/ /t/ /sh/ |
| V08 | /n/ /hh/ /ng/ |
| V09 | /ea/ /ae/ /ao/ /uw/ /oh/ /ia/ /ey/ /ua/ /er/ |
| V10 | /ay/ /aa/ /ah/ /aw/ /uh/ /ow/ /ih/ /iy/ /az/ /eh/ |

Table 1: An example P2V map, this is the P2V for Speaker 01 with ten visemes

where the $ij^{th}$ element is the count of the number of times phoneme $i$ is classified as phoneme $j$. Our algorithm works with the column normalised version,

$$[P_m]_{ij} = Pr\{p_i|\hat{p}_j\} \tag{2}$$

the probability that, given a classification of $p_j$ that the phoneme really was $p_i$. The subscript $m$ in $K_m$ and $P_m$ indicates that $K_m$ and $P_m$ have $m^2$ elements ($m$ phonemes). We merge phonemes by looking for the two most confused phonemes and hence create a new class with confusions $K_{m-1}, P_{m-1}$.

Specifically for each possible merged pair, $Pr, Ps$, we calculate a score:

$$q = [P_m]_{rs} + [P_m]_{sr} = Pr\{\hat{Pr}|Ps\} + Pr\{Pr|\hat{Ps}\} \tag{3}$$

Phonemes are assigned to one of two classes, $V\&C$, vowels and consonants. Vowels and consonants can not be mixed. The pair with the highest $q$ is merged. Equal scores are broken randomly. This process is repeated until $M = 2$. Each intermittent step, $M = 45, 44, 43...2$ forms a possible set of visual units.

This is a more formal approach than used in [1] and incorporates their conclusions that vowel and consonant phonemes should not be clustered together when devising phoneme-to-viseme mappings. An example P2V mapping is shown in Table 1.

### 3.3. Step three: viseme recognition

Similar to Step one, we implement 10-fold cross-validation with replacement [9], of 200 sentences per speaker, 20 are randomly selected as test samples and these are not included in the training folds. Using the HTK toolkit [10] to use Hidden Markov Model (HMM) classes, we flat-start the HMMs, re-estimate them 16 times over with forced alignment between seventh and eighth estimates.

Our prototype is based upon a gaussian mixture of five components and three state HMMs. We use a single-state tied short-pause, or 'sp' HMM for short silences between words in the sentence utterances. We also use a bigram word network to support recognition, apply a grammar scale factor of 1.0 (shown to be optimum in Howell's thesis [11]) and apply a transition penalty of 0.5.

This time around we have viseme classes to use as recognizers. By using these sets of classes which have shown in step one are confusing on the lips, we perform recognition for each class set. In total this is 45, where the smallest set is of two classes (one with all the vowel phonemes and the other all the consonant phonemes), and the largest set is of 45 classes with one phoneme in each - a repeat of the phoneme recognition task but using only phonemes which we know to have been identifiable.

## 4. Discussion

We note that word recognition performance of the HMMs can be measured by both correctness, $C$, and accuracy, $A$, of the recognition classes,

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

$$A = \frac{C - I}{N}, \tag{5}$$

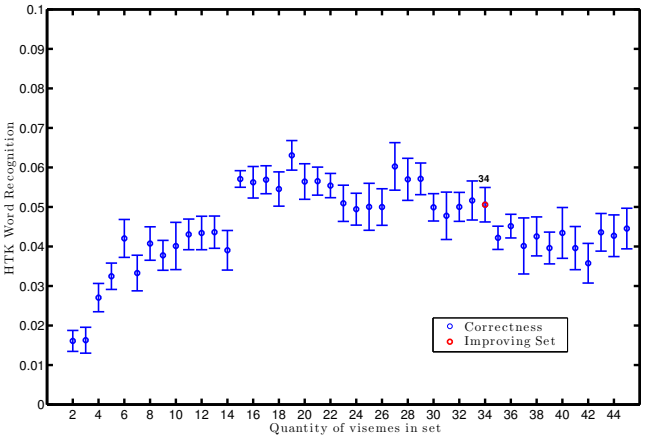
where $S$ is the number of substitution errors, $D$ is the number of deletion errors, $I$ is the number of insertion errors and $N$ the total number of labels in the reference transcriptions [10].

Figure 2 (subfigures a-l), show the correctness for all 12 speakers. Viseme sets containing fewer visemes produce more viseme strings that represent more than one word: homophones. An example of a homophone in these data are the words 'port' and 'bass'. Using Speaker 1's 10-viseme P2V map these both become 'v5 v9 v7' i.e. a single identifier for identifying two words. Thus distinguishing between 'port' and 'bass' becomes impossible. The effect of these can be seen on the left side of the graphs in Figure 2.

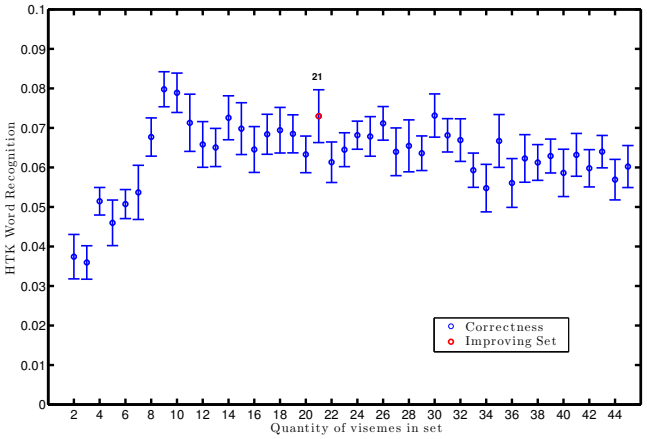
Although the correctness scores are low they are all significantly above chance. The results for each speaker vary but the overall trend is very clear. Superior performances are to be found with larger numbers of visemes. Note that, had we reported viseme error (as is commonplace) then this effect is not visible and the imperative for large numbers of visemes would be missed.

Also in Figure 2 (subfigures a-l), class sets are highlighted in red and labelled which show where a particular combination of two previous viseme classes delivers a significant improvement in recognition. These combinations are listed in Table 2. Whilst there is no apparent pattern through these pairings, this does further reinforce our knowledge that all speakers are visually unique and how difficult finding a set of cross-talker viseme sets will be when different phonemes require alternative grouping arrangements for each individual.

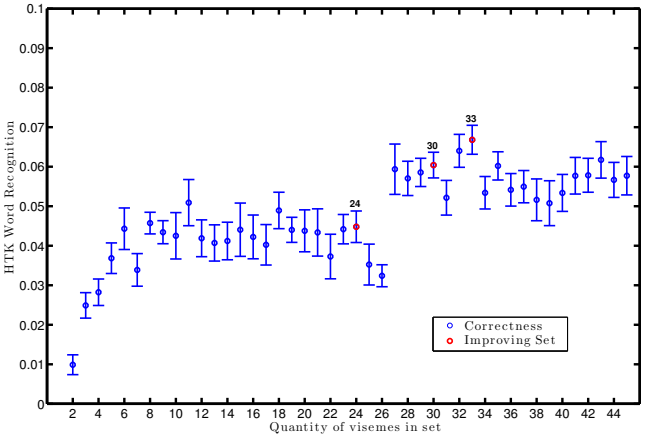
As has been noted before [12] the conventional wisdom which is that visemes are needed for lip-reading is not bourne out by these experiments. However it is an over simplification



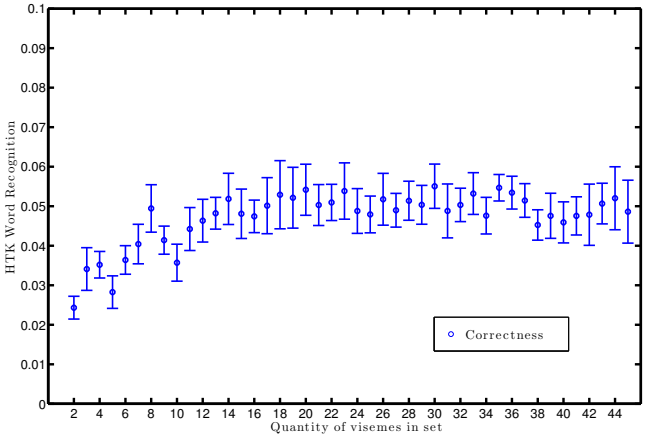

(a) Speaker 1



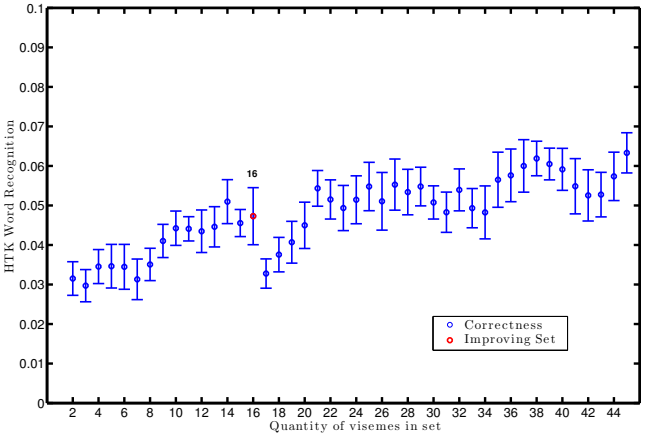

(b) Speaker 2



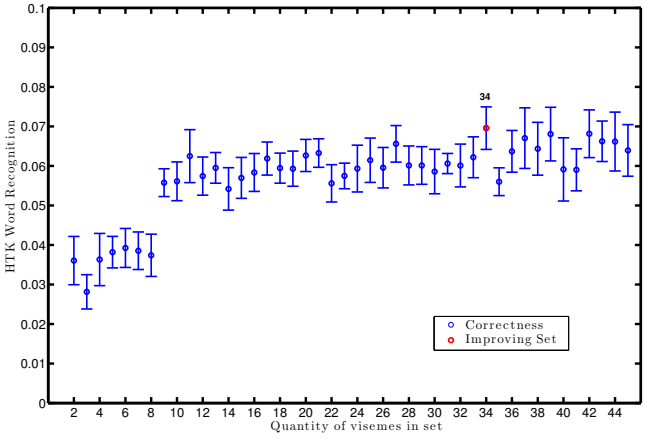

(c) Speaker 3



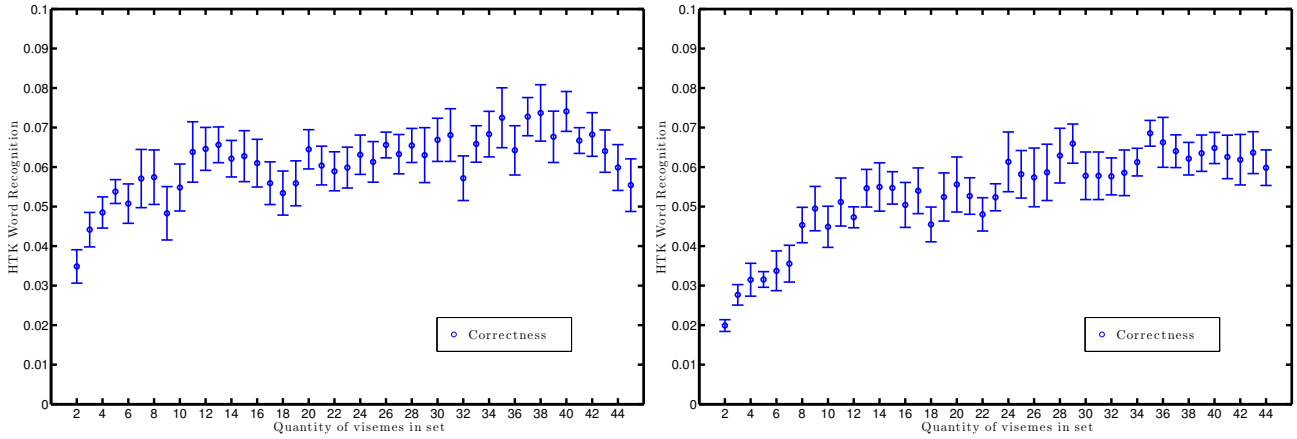

(d) Speaker 4




(e) Speaker 5



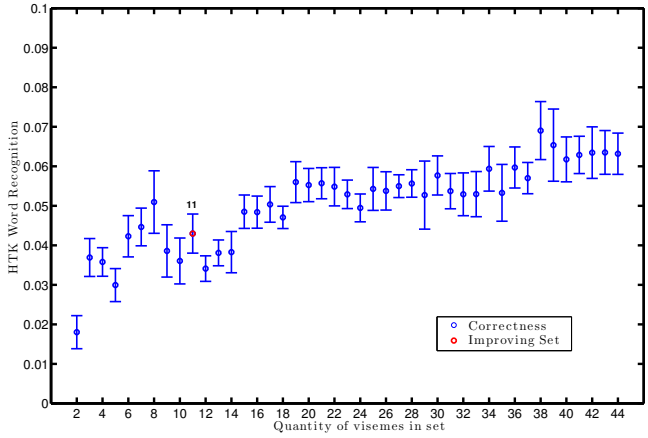

(f) Speaker 6



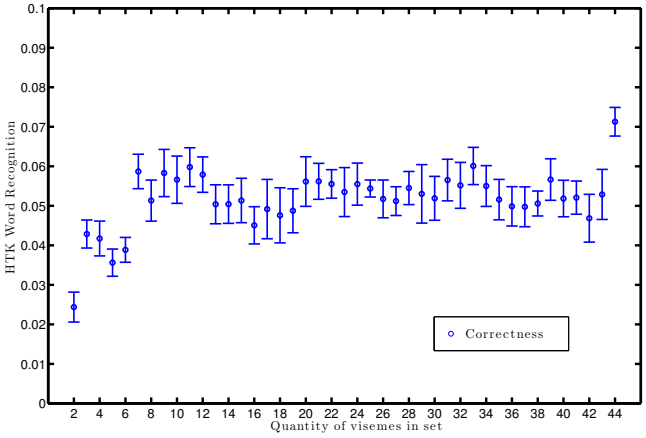

(g) Speaker 7



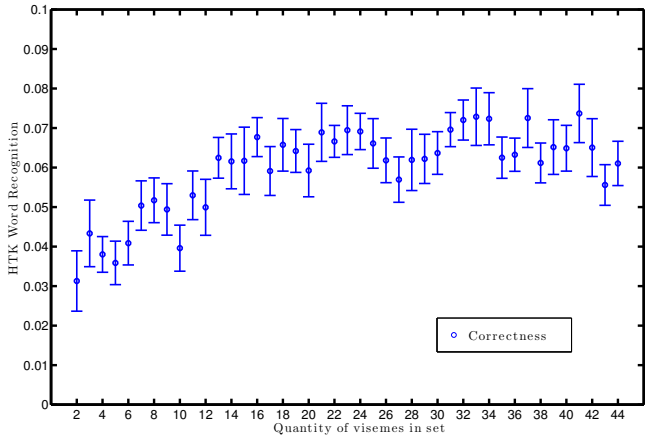

(h) Speaker 8



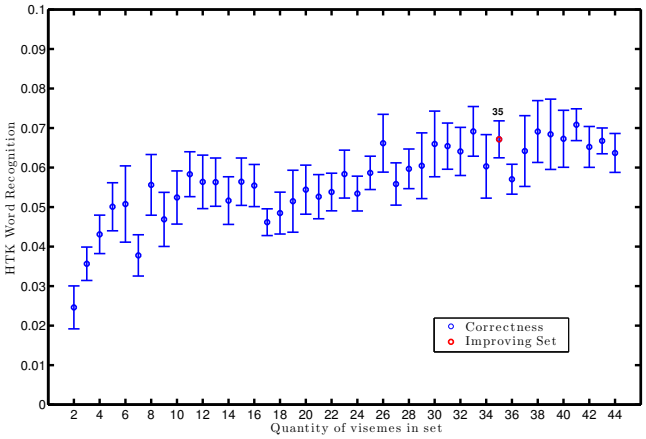

(i) Speaker 9




(j) Speaker 10




(k) Speaker 11




(l) Speaker 12

Figure 2: *Individual speaker word recognition in correctness C for all viseme map sizes*

| Speaker | Set No | $V_i$ | $V_j$ | Set No | $V_n$ |
|---|---|---|---|---|---|
| SP01 | 35 | /s/ /r/ | /dh/ | 34 | /s/ /r/ /dh/ |
| SP02 | 22 | /d/ | /z/ /y/ | 21 | /d/ /z/ /y/ |
| SP03 | 34 | /b/ /ch/ | /zh/ | 33 | /b/ /ch/ /zh/ |
| SP03 | 31 | /zh/ /b/ /ch/ | /z/ | 30 | /zh/ /b/ /ch/ /z/ |
| SP03 | 25 | /p/ /r/ | /ng/ | 24 | /p/ /r/ /ng/ |
| SP05 | 17 | /ae/ | /eh/ | 16 | /ae/ /eh/ |
| SP06 | 35 | /ae/ /ah/ | /iy/ | 34 | /ae/ /ah/ /iy/ |
| SP09 | 12 | /b/ /w/ /v/ | /jh/ /hh/ | 11 | /b/ /w/ /v/ /jh/ /hh/ |
| SP12 | 36 | /ah/ | /ao/ | 34 | /ah/ /ao/ |

Table 2: Viseme class merges which improve word recognition

| Speaker | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phoneme $C$ | 0.045 | 0.060 | 0.058 | 0.049 | 0.063 | 0.063 | 0.055 | 0.090 | 0.063 | 0.071 | 0.061 | 0.064 |

Table 3: Phoneme correctness values for each speaker, these are on the right hand side of each respective subfigure in Figure 2

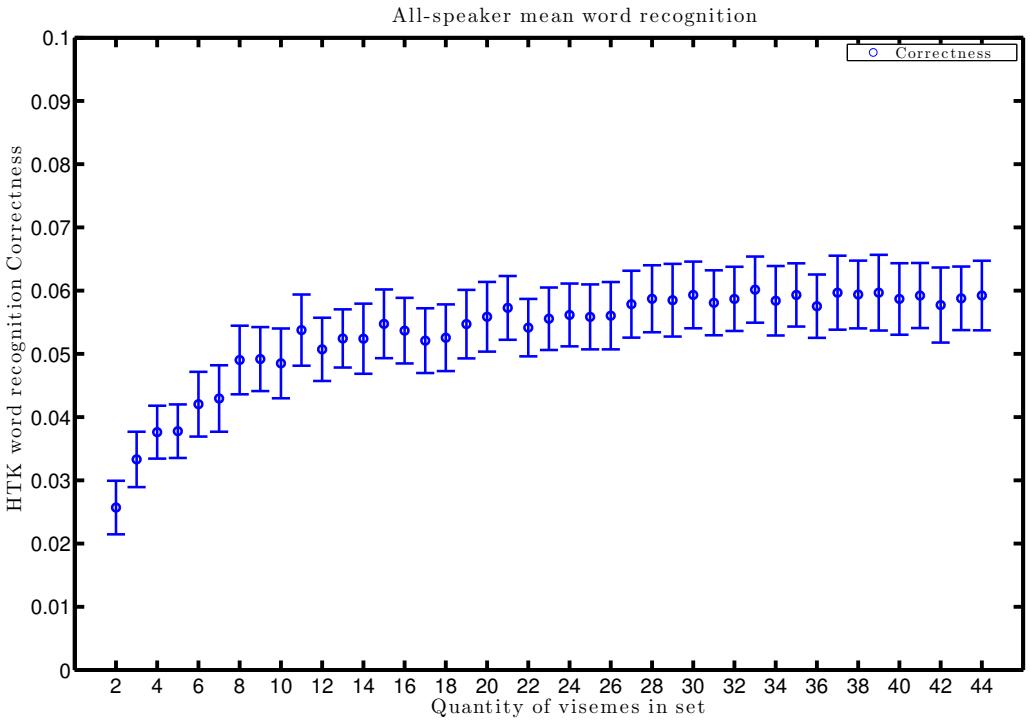



Figure 3: *Word recognition measured by correctness of the classifiers. Error bars show ± one standard error.*

to assert that better lip-reading can be achieved with phonemes than visemes. It is true that, generally speaking, larger numbers of visemes out-perform smaller numbers, but the curves in Figure 2 are far from monotonic. Even Figure 3, which is the mean performance over all speakers, is not monotonic.

There are a number of proposed phoneme-to-viseme maps in the literature, typically they generate between 10 and 20 visemes (see [1] for a summary) - the well known Lee set has six consonant visemes and five vowels [13]; Jeffers eight & three [14] and so on. Looking at Figures 2 & 3 there is certainly a rapid drop-off in performance for fewer than ten visemes but the region between ten and 20 contains the optimum viseme set for three out of the 12 speakers which is no more than chance. In other words, for each speaker there is an optimal number of visual units (shown by the best performing result in Figure 2) but that optimal number is not related to any of the conventional viseme definitions, nor is the number of phonemes. The correctness of the phoneme recognition for each speaker is shown in Table 3.

The two factors at play in these graphs are, the underlying accuracy with which the visual units represent the mouth shape and appearances versus the introduction of homophones. For large numbers of visemes we are close to phonetic recognition, (with fewer homophones) but we run the risk of visual units which are not visually very distinctive - several of the HMM models will "match" on a particular sub-sequence. This latter problem creates a decoding lattice in which there are several near equal probability paths which, in turn, implies that state-of-the-art language models would improve results still further.

## 5. Conclusions

We have described a method that allows us to construct any number of visual units. We remind the reader that we are not proposing that our visemes are the best, our priority in this case is a method for enabling comparison of viseme sets in a controlled manner.

The presence of an optimum is a result of two competing effects. In the first, as the number of visemes shrinks the number of homophones rises and it becomes more difficult to recognise words (correctness drops). In the second, as the number of visemes rises we run out of training data to learn the subtle differences in lip-shapes (if they exist), so again, correctness drops.

Thus, the optimum number of visual units lies beween one and 45. In practice we see this optimum is between the number of phonemes and eight (which is the size of one of the smaller viseme sets).

For future work we are interested to extend these methods to work across speakers with a view to identify combinations of phonemes which can improve more than an single speaker.

## 6. References


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

[2] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

[3] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language and Hearing Research*, vol. 11, no. 4, p. 796, 1968.

[4] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ser. ICMI '04. New York, NY, USA: ACM, 2004, pp. 235–242. [Online]. Available: http://doi.acm.org/10.1145/1027933.1027972

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

[6] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[7] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004. [Online]. Available: http://www.springerlink.com/openurl.asp?

[8] Cambridge University, UK. (1997) BEEP pronounciation dictionary. [Online]. Available: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

[9] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

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

[11] D. L. Howell, *Confusion Modelling for Lip-Reading. PhD thesis*. University of East Anglia, 2014.

[12] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1082–1089, 2006.

[13] S. Lee and D. Yook, "Audio-to-visual conversion using hidden markov models," in *PRICAI 2002: Trends in Artificial Intelligence*. Springer, 2002, pp. 563–570.

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

# DECODING VISEMES: IMPROVING MACHINE LIP-READING

*Helen L. Bear and Richard Harvey*

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

## ABSTRACT

To undertake machine lip-reading, we try to recognise speech from a visual signal. Current work often uses viseme classification supported by language models with varying degrees of success. A few recent works suggest phoneme classification, in the right circumstances, can outperform viseme classification. In this work we present a novel two-pass method of training phoneme classifiers which uses previously trained visemes in the first pass. With our new training algorithm, we show classification performance which significantly improves on previous lip-reading results.

***Index Terms***— visemes, weak learning, visual speech, lip-reading, recognition, classification

## 1. INTRODUCTION

In machine lip-reading, the classification of an utterance from a visual-only signal, there are many obstacles to overcome. Some, such as pose [1, 2], motion [3, 4] and resolution [5] have been studied and measured, including the selection of a phoneme-to-viseme mapping [6, 7]. However, visemes are not precisely defined. Many working definitions have been offered such as; "A set of phonemes that have identical appearance on the lips" [7] or "A visual equivalent of a phoneme" [8]. However, there are challenges with using viseme labelled classifiers including: the homophone effect, not enough training data per class, and the consequential lack of differentiation between classes when there are too many visemes within a set. More recently, there is evidence that viseme labels may not be needed at all because with enough data, classifiers based on phoneme labels can outperform viseme classification [9, 10]. As phonemes are well studied, this idea is attractive. However, others have tested small numbers of visual units: visemes and found they also give acceptable results [11, 12]. It would be very helpful to be able to systematically vary the number visual units and hence devise optimal strategies for learning.

The rest of this paper is as follows; a summary of the analysis into the effect of varying the quantity of visemes in a set on lip-reading performance presented in [13] is followed by a short test on unit selection effects between classifier and its supporting network, the results of these are used to introduce the hypothesis for applying weak learning during classifier training. A full description of the experimental setup to test the hypothesis is included before analysis of results and conclusions.

## 2. BACKGROUND

A systematic study into varying the number of visemes was conducted in [13] which generated viseme sets of varying size. HTK [14] was used to build Hidden Markov Model (HMM) classifiers for every viseme in each set. We initialised a set of HMMs (`HCompV`), that were trained (and retrained) using `HERest` during which there were options to tie any required model states together (e.g. for short pause models) (`HHed`) or to force align the HMMs to a time-aligned ground truth (`HVite`) before producing a classification output. The output of classification was supported by a word bigram model created with `HBuild` and `HLStats`. Finally, this classification output was compared to the ground truth to measure its efficacy (`HResults`) which we measured using Correctness, $C$.

$$C = \frac{N - D - S}{N}, \tag{1}$$

where $S$ is the number of substitution errors, $D$ is the number of deletion errors, $I$ is the number of insertion errors and $N$ the total number of labels in the reference transcriptions [14].

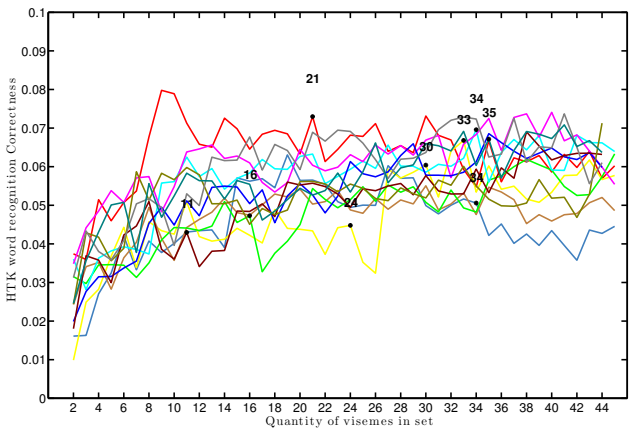



**Fig. 1**: Viseme correctness as the quantity of visemes changes in a set of classifiers for 12 LiLiR speakers. Results from [13].

Figure 1 shows our previous results [13], derived using the algorithm described in [7]. The algorithm works by merging visemes. For example, a label set with 44 visemes has been obtained from the label set of 45 visemes. At each merging stage we measure the difference in correctness compared to the previous set. Significant differences in Figure 1 are shown with black dots where the number represents the size of the significant set.

In Figure 1 the performance of classifiers with few visemes is poor due to the large number of homophones. An example of a homophone in the data are the words "port" and "bass". Using Speaker 1's 10-viseme P2V map these both become '$/v5/$ $/v9/$ $/v7/$' i.e. a single identifier for identifying two distinct words. Thus distinguishing between "port" and "bass" is impossible. Large numbers of visemes do not appear to further improve the correctness, probably because, as has been observed before, many phonemes look similar on the lips [15]. Looking at Figure 1 there appears to be a sweet spot where optimality might be found between visemes set sizes from 11 to 36.

## 3. DATA

For comparable experiments, we select the same 12 speakers from the dataset [16] presented in [13]. For the seven male and five female speakers, each utters 200 sentences from [15]. Individual speakers were tracked using Active Appearance Models (AAMs) [17] and the extracted features consist of concatenated shape and appearance information representing only the mouth area of the face.

## 4. METHOD

In previous work, we essentially examined two different algorithms. In the first, the data were labelled with phonemes, we use `HCompV` to initialise the phoneme classifiers, and 11 repetitions of `HERest` to train the classifiers. This system had the advantage that the output was a sequence of phonemes, but the disadvantage that phoneme models are hard to train. The alternative was to use a smaller number of visemes. The data were labelled with the visemes, and we learned the viseme classifiers in the same way, `HCompV` followed by `HERest`. Our new method is a hybrid. We initially learn the visemes, these trained visemes then become the starting point phoneme classifiers (we know the mapping from the visemes to the phonemes for all sets of visemes). We now train the the phoneme models via repeated applications of `HERest`, thus we have obtained phoneme models but with a new initialisation based upon what was learned for the visemes. This process is illustrated in Figure 2. In this example $p1$, $p2$ and $p4$ are associated with $v1$, so are initialised as replicas of HMM $v1$. Likewise $p3$ and $p5$ are initialised as replicas of $v2$. We now retrain the phoneme models using the same training data.

In full; we initialise *viseme* HMMs with `HCompV`. Our prototype HMM is based upon a Gaussian mixture of five components and three states [18]. These are re-estimated 11 times over with `HERest`, including both short pause model state tying (between re-estimates 3 & 4 with `HHed`), and forced alignment between re-estimates 7 & 8 with `HVite`. This is steps 1 & 2 in Figure 2. But before classification, these viseme HMM definitions are used as initialised definitions for phoneme labelled HMMs (Figure 2 step 3). The respective viseme HMM definition is used for all the phonemes in its relative phoneme-to-viseme mapping. These phoneme HMMs are retrained and used for classification. This amendment to training is analogous with weak learning. We complete classification twice. First with a phoneme bigram network, second with a word bigram network. For both we apply a grammar scale factor of 1.0 and a transition penalty of 0.5 (based on [9]) with `HVite`. This is implemented using 10-fold cross-validation with replacement [19].

The advantage of our new training approach is that the phoneme classifiers have seen only positive cases therefore have good mode matching, the disadvantage is they are not exposed to negative cases to the same degree as the visemes.

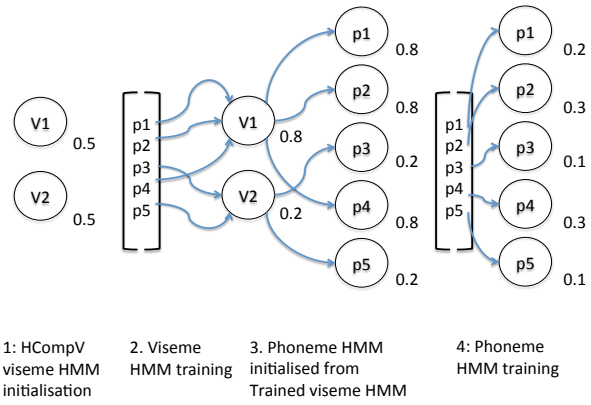



**Fig. 2**: Weak learning of visemes to initialise phoneme labelled classifiers.

### 4.1. Language network units

The systems under study in this paper have two components. The first component, the classifier takes the raw data and attempts to estimate a probable string of units. The second component, the language model, modifies that string on the basis of knowledge of how the units are co-located in the training data. In practice of course, these two components work together and there is no intermediate uncorrected string.

Here we are considering the problem of what the classification unit should be: a viseme? A phoneme? Or a word? But we also must consider how the language model should work. Should we use $n$-grams of phonemes? Visemes? Or words?

The further confusion is the unit on which we measure correctness. It is possible, for example, to build a word classifier followed by a bigram word network measured in terms of its viseme correctness. Such a system would be bizarre but is none-the-less possible. Table 1 shows some of the more sensible possibilities. The first row of Table 1 is a viseme classifier followed by a viseme bigram network with a viseme correctness of 0.0231. In Table 1 correctness is always measured by the units of the classifier. The dashed lines group different correctness units. The top group show viseme correctness which can be compared against each other, the second group show phoneme correctness and the bottom, word correctness.

In our data we have a large vocabulary (approximately 1000 words), so we eliminate word level classifiers as impractical. This leaves us with viseme classifiers for which the viseme word network is the worst performing so we do not consider this option either. For convenience the same data are plotted in Figure 3 with error bars of one standard error.

**Table 1**: Unit selection pairs for HMMs & language networks.

| Classifier units | Network units | Classifier unit, $C$ |
|---|---|---|
| Viseme | Viseme | 0.0231 |
| Viseme | Phoneme | 0.1914 |
| Viseme | Word | 0.0851 |
| Phoneme | Phoneme | 0.1980 |
| Phoneme | Word | 0.1980 |
| Word | Word | 0.1874 |

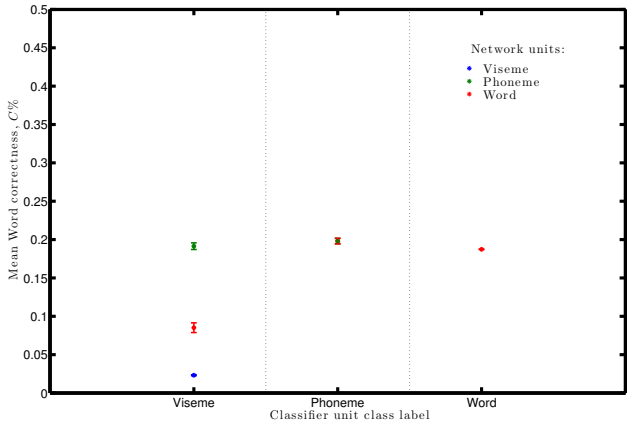



**Fig. 3**: Effects of support network unit choice with varying HMM classifier units measured in all speaker mean correctness, $C$.

## 5. RESULTS

Figure 4 shows the mean speaker-dependent correctness. We examine two configurations, one is phoneme classification where we measure phoneme correctness. These are the top two data series in Figure 4 (in green and pink), and the other is word classification where we measure word correctness. These are the lower two data series in Figure 4 in blue and red. Word correctness guessing is duplicated from [13] and is plotted in orange.

In the top two series, both have bigram phoneme networks, the lower of these two series uses a viseme classifier as in [13], and the upper our new phonemes denoted WLT. The lower pair of series use bigram word networks and again show the difference between visemes and our new method of training phoneme classifiers.

The situation in Figure 4 is summarised in Table 2. For hard to classify speakers, the new model training method gives a significant improvement.

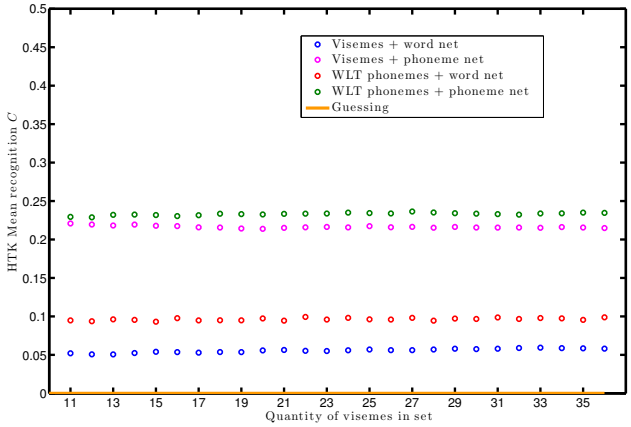
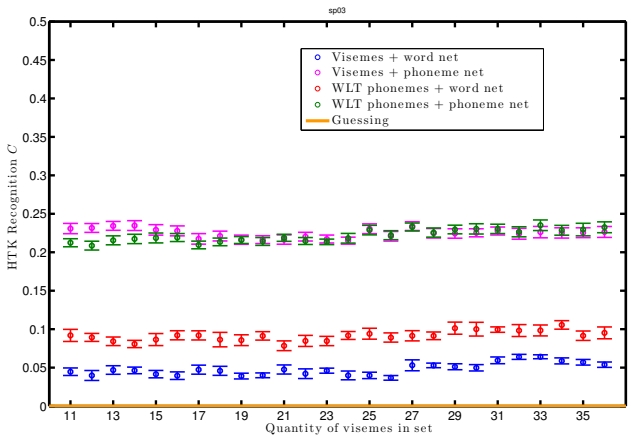


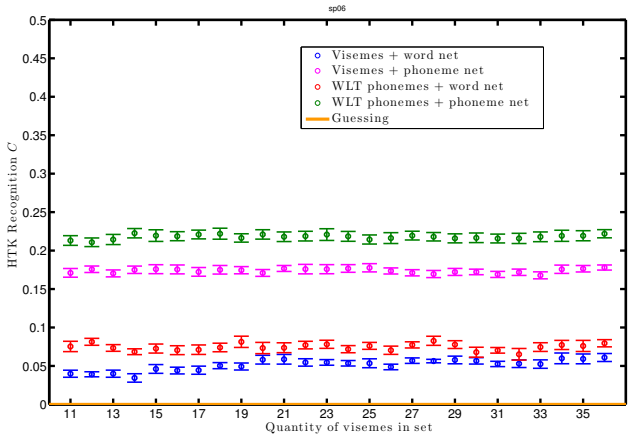

**Fig. 4**: HTK Correctness $C$ for both types of classifier with either a phoneme or a word language model averaged over all 12 speakers.

**Table 2**: Minimum, maximum, and range of mean correctness measured over all speakers for the various methods. Top of table shows word correctness, bottom of table phoneme correctness.

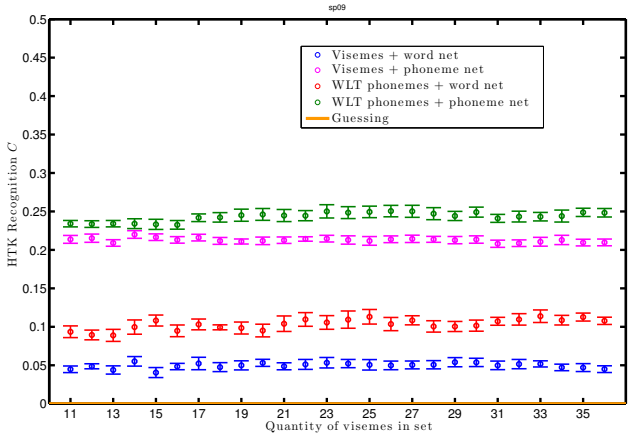
| | Min | Max | Range |
|---|---|---|---|
| WLT phonemes + phoneme net | 0.2253 | 0.2367 | 0.0114 |
| Visemes + phoneme net | 0.2036 | 0.2214 | 0.0179 |
| Effect of WLT | 0.0217 | 0.0153 | – |
| WLT phonemes + word net | 0.0905 | 0.0995 | 0.0090 |
| Visemes + word net | 0.0274 | 0.0601 | 0.0327 |
| Effect of WLT | 0.0631 | 0.0394 | – |




(a) Speaker 3




(b) Speaker 6




(c) Speaker 9

**Fig. 5**: HTK Correctness $C$ for a variety of classifiers with either phoneme or word language models for three speakers.

Figures 5a, b & c show example performances for three speakers. Whilst not monotonic, these graphs are much smoother than the speaker-dependent graphs shown in [13]. Which is encouraging because it implies that our new algorithm is optimising its learning for each speaker-dependent phoneme-to-viseme mapping.

Figure 5 shows that, for certain numbers of visemes, and for certain speakers, the weak learning method gives improvement. However, with the right number of visemes for a particular speaker, the new method will always give a significant improvement.

Looking at Figure 5 there appeard to be a few regions where the new training method gives only marginal improvement. Not all speakers have these regions. We think the presence of these regions is associated with speakers that have more co-articulation than others. If this is true, then the phonemes are blurred together, the learning is more difficult and performance declines. We do not have enough speakers to make this anything other than speculation at this stage. Our own observation is that young people have more co-articulation than old people, but this is something for further investigation.

## 6. CONCLUSIONS

The choice of visual units in lip-reading has caused some debate. Some workers use visemes as adduced by for example Fisher [20] (in which visemes are a theoretical construct representing phonemes should look identical on the lips [10]). Others have noted that lip-reading using phonemes gives superior performance to visemes such as in [9].

Here, we supply further evidence to the more nuanced hypothesis first presented in [13], which is that there are intermediary units, which for convenience we call visemes, that can provide superior performances provided they are derived by an analysis of the data. A good number of visemes in a set is higher than previously thought.

In this paper we have presented a novel learning algorithm which shows improved performance for these new data-driven visemes by using them as an intermediate step in training phoneme classifiers. The essence of our method is to re-train the viseme models in a fashion similar to weak learning. This two-pass approach on the same training data has improved the training of phoneme labelled classifiers and increased the classification performance.

## 7. REFERENCES

[1] K. Kumar, Tsuhan Chen, and R.M. Stern, "Profile view lip reading," in *IEEE International Conference on Acoustics, Speech and Signal Processing. (ICASSP)*, 2007, vol. 4, pp. IV–429–IV–432.

[2] Yuxuan Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 432–437.

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

[4] E.J. Ong and R. Bowden, "Robust facial feature tracking using shape-constrained multi-resolution selected linear predictors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1844–1859, 2011.

[5] Helen L Bear, Richard Harvey, Barry-John Theobald, and Yuxuan Lan, "Resolution limits on visual speech recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1371–1375.

[6] Luca Cappelletta and Naomi Harte, "Phoneme-to-viseme mapping for visual speech recognition.," in *ICPRAM (2)*, 2012, pp. 322–329.

[7] Helen L Bear, Richard W Harvey, Barry-John Theobald, and Yuxuan Lan, "Which phoneme-to-viseme maps best improve visual-only computer lip-reading?," in *Advances in Visual Computing*, pp. 230–239. Springer, 2014.

[8] Helen L Bear, Gari Owen, Richard Harvey, and Barry-John 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.

[9] Dominic Liam Howell, *Confusion Modelling for Lip-Reading. PhD thesis*, University of East Anglia, 2014.

[10] Timothy J Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1082–1089, 2006.

[11] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," in *Signal Processing Conference, 2011 19th European*, Aug 2011, pp. 2109–2113.

[12] Elif Bozkurt, CE Erdem, Engin Erzin, Tanju Erdem, and M Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," *Proceedings of Signal Processing and Communications Applications*, pp. 1–4, 2007.

[13] Helen L Bear, Richard Harvey, Barry-John Theobald, and Yuxuan Lan, "Finding phonemes: improving machine lip-reading," in *1st Joint International Conference on Facial Analysis, Animation and Audio-Visual Speech Processing (FAAVSP)*. ISCA, 2015, pp. 190–195.

[14] Steve J Young, Gunnar Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev, and PC Woodland, "The HTK book version 3.4," 2006.

[15] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[16] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden, "Improving visual features for lip-reading," *International Conference on Audio-Visual Speech Processing (AVSP)*, vol. 7, no. 3, 2010.

[17] Iain Matthews and Simon Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[18] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 198 –213, Feb 2002.

[19] Bradley Efron and Gail Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[20] Cletus G Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language and Hearing Research*, vol. 11, no. 4, pp. 796, 1968.