Phoneme-to-viseme mappings: the good, the bad, and the ugly

Helen L Bear, Richard Harvey

PII: S0167-6393(17)30028-6
DOI: 10.1016/j.specom.2017.07.001
Reference: SPECOM 2472

To appear in: *Speech Communication*

Received date: 15 January 2017
Revised date: 4 June 2017
Accepted date: 28 July 2017

Please cite this article as: Helen L Bear, Richard Harvey, Phoneme-to-viseme mappings: the good, the bad, and the ugly, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.07.001

Phoneme-to-viseme mappings: the good, the bad, and the ugly

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

[a] *University of East London, London E16 2RD UK*
[b] *University of East Anglia, Norwich, Norfolk, NR4 7TJ UK*

Abstract

Visemes are the visual equivalent of phonemes. Although not precisely defined, a working definition of a viseme is "a set of phonemes which have identical appearance on the lips". Therefore a phoneme falls into one viseme class but a viseme may represent many phonemes: a many to one mapping. This mapping introduces ambiguity between phonemes when using viseme classifiers. Not only is this ambiguity damaging to the performance of audio-visual classifiers operating on real expressive speech, there is also considerable choice between possible mappings.

In this paper we explore the issue of this choice of viseme-to-phoneme map. We show that there is definite difference in performance between viseme-to-phoneme mappings and explore why some maps appear to work better than others. We also devise a new algorithm for constructing phoneme-to-viseme mappings from labeled speech data. These new visemes, 'Bear' visemes, are shown to perform better than previously known units.

Keywords: lipreading, speaker-dependent, viseme, phoneme, resolution, speech recognition, classification, visual speech, visual units.

Email address: h.l.bear@uel.ac.uk (Helen L Bear)
URL: https://www.uea.ac.uk/computing/people/profile/r-w-harvey (Richard Harvey)

1. Introduction

Recognition and synthesis of expressive audio-visual speech has proven to be a most challenging problem. When comparing audio-visual speech with acoustic recognition, one can identify several sources of difficulty. Firstly, the visual component of speech brings new problems such as pose, lighting, frame rate, resolution, and so on. Secondly, old problems in acoustic recognition, such as person specificity or the optimal recognition units, appear in new ways in the visual domain. While some of these aspects have been partially studied, progress has been hampered by very small datasets. Furthermore, reliable tracking has eluded many researchers which in turn has led to sub-optimal feature extraction, consequent poor performance and hence, incorrect conclusions about the parts of the problem that are tractable or intractable. A further challenge is the lack of consensus on the recognition units and it is commonplace to need to compare, say, word error rates with viseme error rates computed from a different set of visemes. Our contention is that progress in expressive audio-visual speech will remain stunted while this fundamental uncertainty remains. In this paper we review the choice of visual recognition units and provide a comprehensive set of evaluations of the competing phoneme-to-viseme mappings. We give guidance on what works well and provide explanations for the differences in performance. We also devise new algorithms for selecting optimal visual units should this be desired.

We should note that while this paper tends to focus on visual-only recognition, or lipreading, this aspect is by far the most challenging so progress on lipreading can be used to provide more useful audio-visual systems.

The rest of this paper is structured as follows: we discuss the current restrictions on a conventional lipreading system and identify the limitation of each upon the system. We then study the current sets of published visemes, before presenting a new speaker-dependent clustering algorithm for creating sets of visemes for individual speakers. We show that creating these speaker-dependent visemes follows from simple clustering and merge algorithms. These

new visemes are tested on both isolated words and continuous speech datasets before we evaluate the efficacy of the improved performance against the extra investment into a new lipreading system. Since it is computationally simple to develop these speaker-dependent visemes we contend they are also a useful step in the analysis of speaker variablity which itself is one of the more challenging problems in general lipreading.

2. Limitations in lipreading systems

It is often said that lipreading is difficult because not all sounds appear on the lips[1]. This is true but in reality there are a number of problems that can corrupt the lipreading signal even before one reaches the problem of trying to decode the visual signal. Table 1 provides a taxonomy of the challenges in lipreading. Some of them relate to the problems of extracting useful information from the visual signal whereas some appear later in the signal processing chain and relate to the coding and classification of the visual signal.

Motion is an important part of almost all realistic settings. It is therefore essential to have either some form of tracking or to devise features that are invariant to non-informational motions. An early dataset which captured speaker motion (not camera motion) is CUAVE [37]. Lipreading experiments on this dataset such as [38] examine two different features, one based on the Discrete Cosine Transform (DCT) and another on the Active Appearance Model (AAM). The AAM (which can be shape-only, appearance-only or shape and appearance models) [4] sometimes preceded by Linear Predictors (LP) [2]. An AAM [4] is a model trained on a combination of shape and/or appearance information from a subset of video frames. The model is usually built from video frames manually labeled with landmarks which are chosen to cover the full range of motion throughout the video. In [38] they prefer the DCT but note that there were implementation difficulties with the AAM which meant it was improperly

[1] [1] compares the performance of a system that measures, via electromagnetic articulography, the hidden and visual parts of the mouth so the extent of this statement can be quantified.

Table 1: Challenges to successful machine lipreading. Each challenge has some references.

Evaluation	Previously studied?
Motion	Yes, [2, 3, 4]
Pose	Yes, [5, 6, 7, 8, 9, 10, 11]
Expression	Yes, [6, 7]
Frame rate	Yes, [12, 13]
Video quality	Yes [14, 15, 16]
Color	Yes, [9]
Unit choice	Yes, [17, 18, 19, 20, 21]
Feature	Yes, [22, 23, 24, 4, 3]
Classifier technology	Yes, [25, 26, 27, 17, 28]
Multiple persons	Yes, [29, 30, 31, 32]
Speaker identity	Yes, [33, 34, 35]
Rate of speech	Yes, [36, 21]

tracked. Further lip-reading experiments on CUAVE [39] clarifies how challenging comparing results is, because there is no agreed evaluation protocol which could account for the motion challenge/face alignment. This is attributed to their partial success with particular speakers.

The majority of automatic lipreading systems use a frontal **pose** in which the speaker's facial place is normal to the principal ray of the camera. However in [7] for example, an improvement in expression recognition is seen by both computers and humans when the pose is rotated to 45°. Other work [8, 9], looks more specifically at visual speech recognition and suggests that a profile view of a speaker may not lead to catastrophically low accuracies. This observation is consistent with [10] which measures human sentence perception from three viewing angles: full-frontal view (0°), angled view (45°), and side view (90°). In this single-subject study a post-lingual deaf woman was tested to measure accuracy at the three angles independently. The three angles were randomly presented in every lipreading session. The results indicated that the side-view

angle is most effective. A model for pose-mismatched lipreading is presented in [11] in which it is shown that without training data at the correct pose, the recognition accuracy falls dramatically. However, the authors also show that this can be mitigated by projecting the features back to a canonical pose. This transformation principle is also used in [5] which presents a view-independent lipreading system. This investigation uses a continuous speech corpus compared to the small vocabulary dataset in [11]. This later study acknowledges a human lipreaders preference for a non-frontal view and suggests it could be attributed to lip protrusion. They show that the 45° angle is preferable. In short, when it comes to pose, there is evidence that it can be accounted for and need not be insurmountable. Therefore, for this work we stick to frontal pose.

Expression can be difficult to disentangle with the spoken word when lipreading natural speech. Smiling (a happy expression) has an known effect on lip motions during speech [40]. Effects on the inner, outer lips and lip protrusions have been measured in [41] who shows that smiling during speech (particularly vowels) places a restriction on lip motion with greater demand placed on the inner lips as variation in outer lips and lip protrusion is reduced. This in turn creates a greater challenge when lipreading non-neutral speech as gestures become less distinct. Furthermore, expression also effects the temporal property of speech [42, 43]. When a particular phoneme is uttered, its duration can be shortened (for example when angry and vowels particularly become shorter) or elongated, for example when a speaker is sad.

To the best of our knowledge there is no systematic study which specifically investigates lipreading expressive speech. Rather, tasks focus on either, synthesizing expression in faces [44, 45, 46] or expression recognition during speech [47, 48, 49].

Studies such as [12] on the effect of low video **frame-rate** on human speech intelligibility during video communications, suggest that lower frame rates, if they are visible to the speaker, encourage humans to over-articulate to compensate for the reduced visual information available, akin to a visual Lombard effect. Accuracy is maximized when the same frame rate is used for both train-

ing and testing [13]. They further recommend that when the training data cannot be recorded at the same frame rate as the test data, then it is best if the training data has a higher frame rate (for feature extraction) than the test data. A further observation is that word classification rates vary in a non-linear fashion as the frame rate is reduced.

When it comes to dependence of lipreading on **video quality**, an investigation into the effects of compression artifacts, visual noise (simulated with white noise) and localization errors in training is presented in [15], and in [16]. The authors undertake two experiments, of which the first includes some attention to spatial resolution (the number of pixels). However, here, resolution varies along with other parameters. Neither of these papers consider the simple removal of information from a smaller image compared to a larger one. A more systematic study of resolution can be found in [14] in which video of varying resolution is parameterized using AAMs [50]. This work shows that machines can lipread continuous speech with as little as two pixels per lip.

With regard to **color**, it has been surprisingly under used. In [9] algorithms are derived which contain three key components: shape models, motion models, and focused color feature detectors. In early works it was common to use colored lip-stick or markers to help track the lips (tracking remains challenging) but many authors convert the image to grayscale and use grayscale features.

Unit choice refers to the question of whether to use phonemes, visemes, words or something else. Classifiers built on phonemes [18], visemes [19], and words [20] have all been previously presented. Sometimes the unit choice is linked to the problem: word classifiers often use word units, whereas continuous speech has to use phonemes or visemes. It is essentially a trade-off since using phonemes means accepting that there will be units that do not appear on the lips (the words “bad”, “pad”, and “mad” are usually said to be visually indistinguishable) whereas using visemes leads to better unit accuracy but there is then the problem of homopheny (words that have identical visemic transcriptions but different spellings). One study has reviewed how the unit selection affects recognition in relation to the unit selection of the supporting language

model [21] and have shown that phoneme networks work best for both phoneme and viseme classifiers. However the practical reality is that many systems use visemes and there is need to resolve which choice of visemes works best. Comparative studies such as [17] have attempted to compare some previous viseme sets but, these often only consider a few different sets rather than the gulf available.

Lan *et al.* present in [24] a comparison of different **features** first presented in [4]. Revisited in [3], AAM features are produced as either model-based (using shape information) or pixel-based (using appearance information). In [24] 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 [24] classified words with the an Audio-Visual dataset known as RMAV but recommended in future creating classifiers with viseme labels for lipreading, and advises that most information is from the inner of the mouth. Some works have attempted to adapt features to address different problems, such as motion described above. For example, in [51] the authors suggest altering HMM modeling to permit either frozen or occluded frames, and demonstrate that even low level jitter will significantly affect the quality of lip reading features.

When it comes to the choice of **classifier technology** it is the norm that machine lipreading systems adapt methods from acoustic recognition. This not only follows from the observation that visual and acoustic speech have the same origins but also from the practical observation that language models are expensive to create and it makes sense to re-use the models across the two modalities. The conventional classifier process is 1) data preparation (an acoustic example is creating MFCC's [27], whereas a visual example might be [17]), 2) build Hidden Markov Model classifiers, and 3) feed the classification outputs through a language network to produce a transcript. Like feature selection, the choice of classifier is affected by the problem in hand. An optimal audio recognizer will

not guarantee optimal performance in an audio-visual, or visual only domain. In [52], for example, it is noted that their audio-visual results should not be "read across" to lipreading.

More modern deep learning techniques for lipreading are an alternative approach which require much more training data [28]. A key disadvantage of these methods is a lack of understanding about what exactly a neural network is learning in order for it to classify unseen gestures. So often the results from deep learning are good but the scientific insight can be poor. Thus recent work has begun to demonstrate performance of different deep learning approaches with a variety of neural network architectures. Convolution neural networks (CNN) have been particularly prevalent for image classification ([53, 54]) and Long Short Term Memory networks (LSTM) are performing well on temporal problems (e.g. language modeling [55] or, scene labeling [56]). For lipreading, we have evidence that both of these achieve good recognition rates in end-to-end systems, in [57] a CNN achieves 61.1% top 1 accuracy and in [58] an LSTM achieves 79.6% top 1 accuracy on a small dataset. However, our lipreading is a combination of these challenges, that is a temporal-visual classification problem.

For lipreading **multiple persons**, [30, 31] detailed human lipreading of multiple people, [30] recognizes consonants, and [31] visual vowels. [32] 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. In summary there is no reason to think that multi-person lipreading is any less viable than single-person lipreading, although the challenge of variability due to speaker identity is real.

Speaker identity is a major challenge in machine lipreading because Visual speech is not consistent across individuals. Sometimes this can be advantageous as in [33] where they use lipreading to identify speakers. With known speakers - lipreading recognition rates can be high, but with unknown speakers (referred to as speaker-independent lipreading) this is as yet not at the same standard as speaker dependent lipreading. In [34] results show that classifiers trained and tested on distinct speakers compared to those trained and tested on the

same speakers are statistically significantly different. This is supported in [35] where the authors strive to discriminate languages from visual speech and they conclude that in order to improve performance would be to move away from speaker-dependent features.

For acoustic speech it is acknowledged that people have different speaking styles, accents and **rates of speech**. For visual speech there is the additional confusion of what we call a "visual accent" in which very similar sounds can be made by persons with very different mouth shapes – examples of visual accent effects include people who talk out of the side of their mouths; ventriloquists and mimics. The rate of speech alters both an utterance duration and articulator positions. Therefore, both the sounds produced, but particularly, visible appearance are altered. In [36], 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. Anecdotal evidence suggests that speaker visual style can evolve as speakers age due to co-articulation reduction as a person travels/interacts with other adults [21].

In summary, while audio-visual speech processing has a great number of challenges, one of the pivotal ones is the question of the visual units and how they should be derived. Since all language models are defined in terms of phonemes, the practical question is the choice of the mapping from phonemes to visemes. The literature has presented a great number of these phoneme-to-viseme (P2V) mappings and few consistent comparisons between them so this is the topic for the next section.

3. Comparison of phoneme-to-viseme mappings

A summary of published P2V maps is provided in [59] Tables 2.3 and 2.4. This list is not exhaustive and these mappings motivated by: a focus on just consonants [60, 61, 62, 63]; being speaker-dependent [64], prioritizing particular visemes [65]; or a focus on vowels [66, 67]. 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 maps for the vowels and one map for the consonants. A full list of the mappings used is in Tables 2 and 3. Of these mappings , the most common are 'the Disney 12' [66], the 'lipreading 18' by Nichie [68], and Fisher's [61].

Table 2: Vowel phoneme-to-viseme maps previously presented in literature.

Classification	Viseme phoneme sets
Bozkurt [69]	{/ei/ /ʌ/} {/ei/ /e/ /æ/} {/ɜ/} {/i/ /ɪ/ /ə/ /y/} {/ɑʊ/} {/ɔ/ /ɑ/ /ɔɪ/ /əʊ/} {/u/ /ʊ/ /w/}
Disney [66]	{/ʊ/ /h/} {/ɛə/ /i/ /ai/ /e/ /ʌ/} {/u/} {/ʊə/ /ɔ/ /ɔə/}
Hazen [19]	{/ɑʊ/ /ʊ/ /u/ /əʊ/ /ɔ/ /w/ /ɔɪ/} {/ʌ/ /ɑ/} {/æ/ /e/ /ai/ /ei/} {/ə/ /ɪ/ /i/}
Jeffers [70]	{/ɑ/ /æ/ /ʌ/ /ai/ /e/ /ei/ /ɪ/ /i/ /ɔ/ /ə/ /ɪ/} {/ɔɪ/ /ɔ/} {/ɑʊ/} {/ɜ/ /əʊ/ /ʊ/ /u/}
Lee [71]	{/i/ /ɪ/} {/e/ /ei/ /æ/} {/ɑ/ /ɑʊ/ /ai/ /ʌ/} {/ɔ/ /ɔɪ/ /əʊ/} {/ʊ/ /u/}
Montgomery [67]	{/i/ /ɪ/} {/e/ /æ/ /ei/ /ai/} {/ɑ/ /ɔ/ /ʌ/} {/ʊ/ /ɜ/ /ə/}{/ɔɪ/} {/i/ /hh/} {/ɑʊ/ /əʊ/} {/u/ /u/}
Neti [72]	{/ɔ/ /ʌ/ /ɑ/ /ɜ/ /ɔɪ/ /ɑʊ/ /fi/} {/u/ /ʊ/ /əʊ/} {/æ/ /e/ /ei/ /ai/} {/ɪ/ /i/ /ə/}
Nichie [68]	{/uw/} {/ʊ/ /əʊ/} {/ɑʊ/} {/i/ /ʌ/ /ay/} {/ʌ/} {/iy/ /æ/} {/e/ /ɪə/} {/u/} {/ə/ /ei/}

In total, eight vowel- and fifteen consonant-maps are identified here and all of these are paired with each other to provide 120 P2V maps to test.

Recent comparisons between maps include [17] and as part of [59]. In [59] the following list of reasons are given for discrepancies between classifier sets.

- Variation between speakers - i.e. speaker identity.
- Variation between viewers - indicating lipreading ability varies by individ-

Table 3: Consonant phoneme-to-viseme maps previously presented in literature.

Classification	Viseme phoneme sets
Binnie [60]	{/p/ /b/ /m/} {/f/ /v/} {/θ/ /ð/} {/ʃ/ /ʒ/} {/k/ /g/} {/w/} {/r/} {/l/ /n/} {/t/ /d/ /s/ /z/}
Bozkurt [69]	{/g/ /ɦ/ /k/ /ŋ/} {/l/ /d/ /n/ /t/} {/s/ /z/} {/tʃ/ /ʃ/ /dʒ/ /ʒ/} {/θ/ /ð/} {/r/} {/f/ /v/} {/p/ /b/ /m/}
Disney [66]	{/p/ /b/ /m/} {/w/} {/f/ /v/} {/θ/} {/l/} {/d/ /t/ /z/ /s/ /r/ /n/} {/ʃ/ /tʃ/ /j/} {/y/ /g/ /k/ /ŋ/}
Finn [73]	{/p/ /b/ /m/} {/θ/ /ð/} {/w/ /s/} {/k/ /h/ /g/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/y/} {/z/} {/f/} {/v/} {/t/ /d/ /n/ /l/ /r/}
Fisher [61]	{/k/ /g/ /ŋ/ /m/} {/p/ /b/} {/f/ /v/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/t/ /d/ /n/ /θ/ /ð/ /z/ /s/ /r/ /l/}
Franks [62]	{/p/ /b/ /m/} {/f/} {/r/ /w/} {/ʃ/ /dʒ/ /tʃ/}
Hazen [19]	{/l/} {/r/} {/y/} {/b/ /p/} {m} {/s/ /z/ /h/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/t/ /d/ /θ/ /ð/ /g/ /k/} {/ŋ/} {/f/ /v/}
Heider [74]	{/p/ /b/ /m/} {/f/ /v/} {/k/ /g/} {/ʃ/ /tʃ/ /dʒ/} {/θ/} {/n/ /t/ /d/} {/l/} {/r/}
Jeffers [70]	{/f/ /v/} {/r/ /ɋ/ /w/} {/p/ /b/ /m/} {/θ/ /ð/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/s/ /z/} {/d/ /l/ /n/ /t/} {/g/ /k/ /ŋ/}
Kricos [64]	{/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 [71]	{/d/ /t/ /s/ /z/ /θ/ /ð/} {/g/ /k/ /n/ /ŋ/ /l/ /y/ /ɦ/} {/dʒ/ /tʃ/ /ʃ/ /ʒ/} {/r/ /w/} {/f/ /v/} {/p/ /b/ /m/}
Neti [72]	{/l/ /r/ /y/} {/s/ /z/} {/t/ /d/ /n/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/p/ /b/ /m/} {/ŋ/ /k/ /g/ /w/} {/f/ /v/} {/θ/ /ð/}
Nichie [68]	{/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 [63]	{/p/ /b/ /m/} {/f/ /v/} {/θ /ð/} {/ʃ/ /ʒ/} {/w/} {/s/ /z/} {/r/} {/l/} {/t/ /d/ /n/ /k/ /g/ /j/}
Woodward [75]	{/p/ /b/ /m/} {/f/ /v/} {/w /r/ /ɯ/} {/t/ /d/ /n/ /l/ /θ/ /ð/ /s/ /z/ /tʃ/ /dʒ/ /ʃ/ /ʒ/ /j/ /k/ /g/ /h/}

uals, those with more practice 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.'[59].

These last two points are reinforced by [17] who achieved highest accuracy with the phoneme-to-viseme map of Jeffers in an HMM-based lipreading system. They attribute this to the use of continuous speech which encapsulates the same viseme in more contexts within the training data, and suggest that the Jeffers map has better clustering of consonant visemes for those contexts.

In Table 4 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 testing with few test subjects, for example Finn [73] used only one lipreader, and Kricos [64] twelve. Data-driven methods are most recent, e.g. Lee's [71] visemes were presented in 2002 and Hazen's [19] in 2004. The remaining visemes are based around linguistic/phonemic rules.

As an example, the clustering method of Hazen [19] involved bottom-up clustering using maximum Bhattacharyya distances [76] to measure similarity between the phoneme-labeled Gaussian models. Before clustering, some phonemes were manually merged, /*em*/ with /*m*/, /*en*/ with /*n*/, and /*Z*/ with /*S*/.

Table 4: 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	2004	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
Montgomery	1983	Human perception	Confusion patterns	10
Neti	2000	Linguistics	Decision tree clusters	26
Nichie	1912	Human observations	Human observation of lip movements	unknown
Walden	1977	Human testing	Hierarchical clustering	31
Woodward	1960	Linguistics	Language rules and context	unknown

A P2V map may be summarized as a ratio we call "compression factor," CF_s

$$CF_s = \frac{NV}{NP} \tag{1}$$

which is the ratio of number output visemes, NV to input phonemes NP. The compression factors for the P2V maps are listed in Table 5. Silence and garbage visemes are not included in Compression Factors.

Because we have a British English dataset and some works were formu-

Table 5: 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	-	-	-

lated using American English diacritics [77] we omit the following phonemes from some mappings: */si/* (Disney [66]), */axr/ /en/ /el/ /em/* (Bozkirt [69]), */axr/ /em/ /epi/ /tcl/ /dcl/ /en/ /gcl/ kcl/*(Hazen [19]), and */axr/ /em/ /el/ /nx/ /en/ /dx/ /eng/ /ux/* (Jeffers [70]). Moreover, Kricos provides speaker-dependent visemes [64]. These have been generalized for our tests using the most common mixtures of phonemes. Where a viseme map does not include phonemes present in the ground truth transcript these are grouped into one viseme denoted (*/gar/*). Note that all phonemes in each P2V map are in the dataset but no mapping includes all 29 phonemes in the AVL2 vocabulary.

3.1. Data preparation

The AVLetters2 (AVL2) dataset [78] is used to train and test HMM classifiers based upon our 120 P2V mappings with HTK [26]. AAM features (concatenated as in (4)) are used as they are known to outperform other feature methods in machine lipreading [17]. AVL2 [78] is an HD version of the AVLetters dataset [22]. It is a single word dataset of five male British English speakers reciting the alphabet seven times. We use four of these speakers at the fifth tracked too poorly to have confidence in lipreading accuracy. The speakers in this dataset are illustrated in [79]. 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 varying co-articulation.

Table 6: 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

Table 6 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 [3]. All features retain 95% variance of facial shape and appearance information.

The RMAV dataset consists of 20 British English speakers (we use 12 speak-

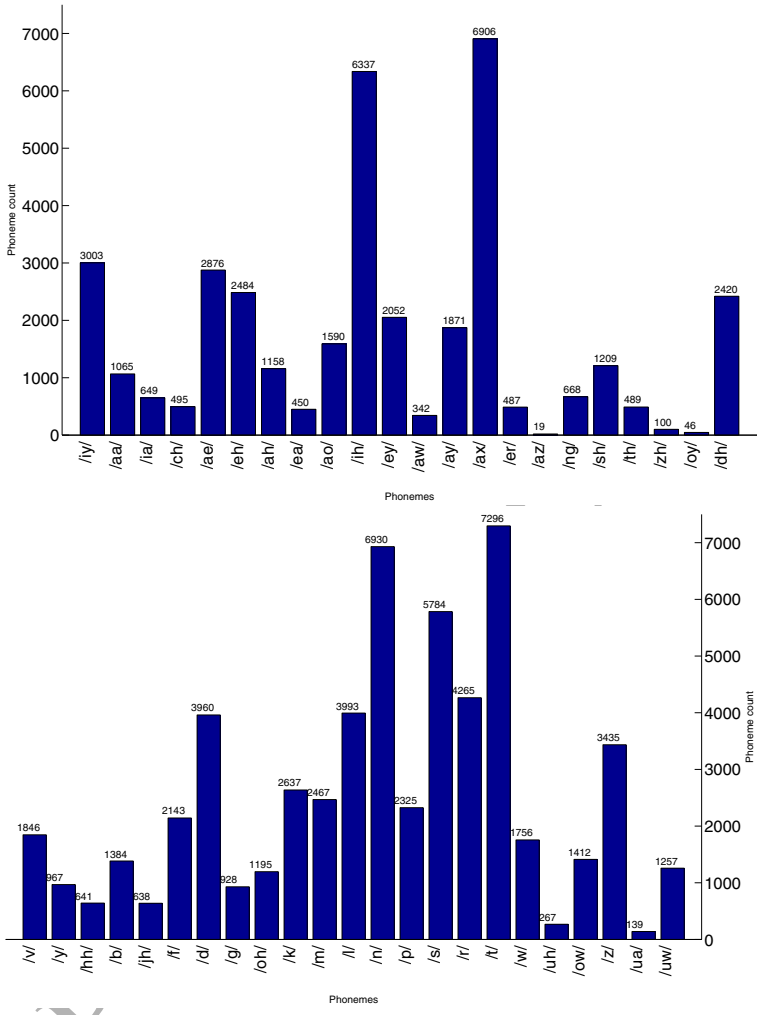

Figure 1: Occurrence frequency of phonemes in the RMAV dataset.

ers,seven male and five female, who have been tracked to maintain comparability with earlier work), 200 utterances per speaker of a subset of the Resource Management (RM) context independent sentences from [80] which totals around 1000 words each. The sentences are selected to maintain a good coverage all phonemes [81] and to represent the coverage of phonemes in spoken speech. The original videos were recorded in high definition and in a full-frontal posi-

tion. Individual speakers are tracked using Active Appearance Models [3] and AAM features of concatenated shape and appearance information have been extracted.

Figure 1 plots the frequency of all phonemes within the RMAV dataset over 200 sentences and Table 7 lists the number of parameters of shape, appearance, and combined shape and appearance AAM features where the features retain 95% variance of facial information.

Table 7: 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

3.2. Classification method

The method for these speaker-dependent classification tests on our combined shape and appearance features uses HMM classifiers built with HTK [26]. The features selected are from the AVL2 and RMAV datasets. The videos are tracked with a full-face AAM (Figure 2 (left)) and the features extracted consist of only the lip information (Figure 2 (right)). This means that we obtain

a robust tracking from the full-face model, then using this fit information, we apply a sub-active appearance model of only the lips. The HMM 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 [82] and a word transcription. The phonetic transcript is converted to a viseme transcript assuming the visemes in the mapping being tested (Tables 3 and 2). 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, this is increased to 10-fold cross-validation for RMAV speakers. The HMMs are initialized 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.

3.3. Active appearance models

An example full-face shape model example is in Figure 2 where there are 76 landmarks, 34 of which are modeling the inner and outer lip contours.

Figure 2: Example Active Appearance Model shape mesh (left), a lips only model is on the right.

The shape s of an AAM is the collection of coordinates of the v vertices (landmarks) which make up a mesh,

$$s = (x_1, y_1, x_2, y_2, ..., x_v, y_v)^T \tag{2}$$

These landmarks are aligned and normalized via Procrustes analysis [83] and then analyzed via a Principal Component Analysis (PCA) to

$$s = s_0 + \sum_{i=1}^{n} p_i s_i \tag{3}$$

where s_0 is the mean shape, p_i are coefficient shape parameters, and s_i are the eigenvectors of the co-variance matrix of the n largest eigenvalues [3].

Having built an Active Shape Model, the next step is to augment it with appearance data and hence compute an Active Appearance Model (AAM). Each shape model is used to warp the image data back to the mean shape. The appearance of those warped images is now modeled again using PCA [4],

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \quad \forall x \in s_0 \tag{4}$$

where λ_i are the appearance parameters, A_0 is the shape-free-mean appearance, and $A_i(x)$ are the appearance image eigenvectors of the co-variance matrix.

Usually the best results are obtained using both shape and appearance information combined within a single AAM [25, 4]. Therefore, unless explicitly stated otherwise, we use these. Once an AAM is built and trained, we fit the model using the Inverse Compositional algorithm [84] to all frames in the video sequence [3].

3.4. Comparison of current phoneme-to-viseme maps

Recognition performance of the HMMs can be measured by both correctness, C, and accuracy, A,

$$C = \frac{N - D - S}{N} \tag{5}$$

$$A = \frac{N - D - S - I}{N} \tag{6}$$

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 [26]. An insertion error (which are notoriously common

in lip reading [85]) occurs when the recognizer output has extra words/visemes missing from the original transcript [26]. 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 has deleted one[2].

In this experiment, classification performance of the HMMs is measured by correctness, C (5), as there are no insertion errors to consider [26]. It is acknowledged that 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, words, are used so we can compare different viseme sets. It is the difference between each set, rather than the individual performance, which is of interest in this investigation.

Figure 3 shows the correctness of each pair of viseme sets. On the top is the isolated word case (the AVL2 data) and on the bottom the continuous data (RMAV). Each diagram is ordered by the mean correctness over all speakers. For the isolated words the Lee vowel and consonant sets [71] are the best with the Montgomery vowels [67] and Hazen consonants [19] close behind. The worst performers are Disney vowels [66] and the Franks [62] and Woodward consonants [75]. For continuous speech the Disney vowels are the best performer [66] as are the Woodward consonants [75]. It is notable that for continuous speech the high compression factor visemes sets work better than those with larger numbers of visemes. The most likely explanation is that continuous speech has additional variability due to co-articulation so a few coarsely defined visemes are better than a greater number of finely defined ones.

Figure 4 shows the mean word correctness, C, over all speakers, $\pm 1 s.e$ for

[2]Once this utterance has been translated to one of viseme labels rather than words, as an example using Montgomery's visemes, this sentence becomes "v09 v12 v04 v05 - v12 v01 v12 v04 - v12 - v01 v10 v04 v11 v04 - v04 v07 v16 v07 v16" (hyphens are included to show breaks between words). In this case, the same insertion errors would create predicted outputs of "v09 v12 v04 v05 - v12 v01 v12 v04 - v12 v01 v12 v04 - v01 v10 v04 v11 v04 - v04 v07 v16 v07 v16 - v04 v07 v16 v07 v16."

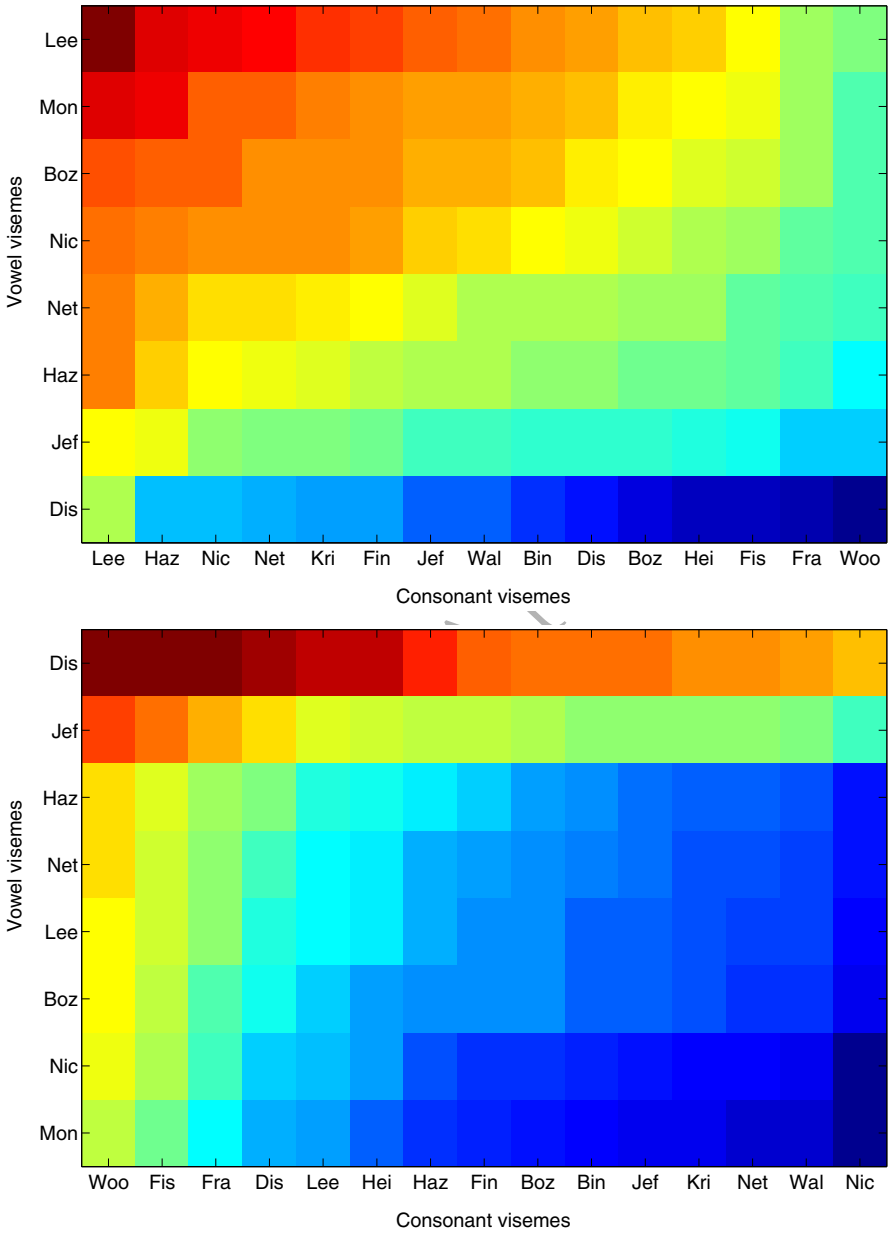

Figure 3: Speaker-dependent all-speaker mean word classification, C, comparing viseme classes on isolated word speech (top) and continuous speech (bottom)

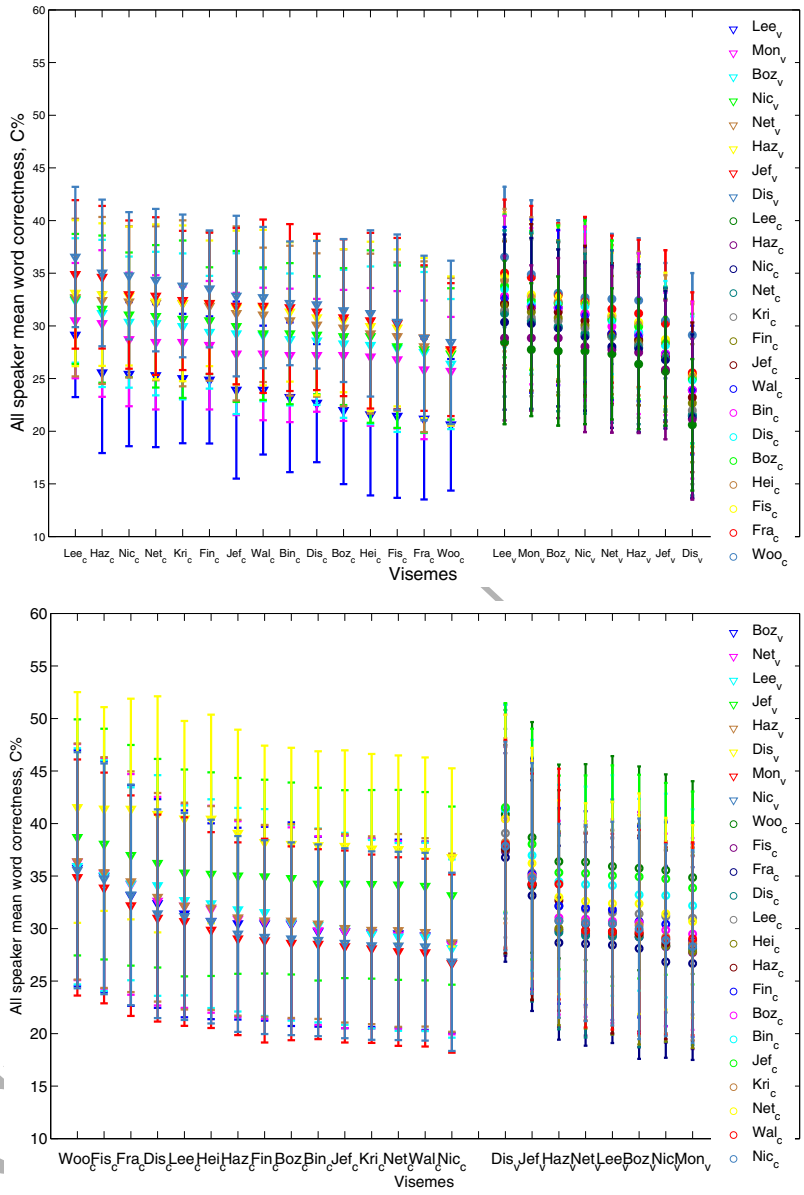

Figure 4: Speaker-independent all-speaker mean word classification, $C \pm 1s.e.$ For a given mapping ($x-$axis) the performance is measured after pairing with all vowel mappings (left) and vice versa on the right on AVL2 isolated words (top) and RMAV continuous (bottom)

pairings of vowel and consonant maps ordered by correctness from left to right. Again, isolated word results (the AVL2 data) at the top and continuous (RMAV) on the bottom. As previously, for isolated words, the Disney vowels are significantly worse than all others when paired with all consonant difference over the whole group. The Lee [71], Montgomery [67] and Bozkurt [69] vowels are consistently above the mean and above the upper error bar for Disney [66], Jeffers [70] and Hazen [19] vowels. In comparing the consonants, Lee [71] and Hazen [19] are the best whereas Woodward [75] and Franks [62] are the bottom performers. There is a significant difference between the 'best' visemes for individual speakers which arises from the unique way in which everyone articulates their speech.

The continuous speech experiment results in Figure 4 (bottom) show that, for vowel visemes, the Disney set surpasses all others, whereas Woodward's consonants are now a better fit. This is interesting as neither viseme set are data-derived. We recall that Disney's [66] are designed from human perception for synthesis of characters, and Woodward's [75] are from a pilot investigation into phoneme perception in lipreading using linguistic rules. As we move to more realistic data , continuous speech, many of the data-driven approaches degrade which implies that they data used to derive these visemes was unrealistic. For example the Lee visemes [71] were derived without any use of video data at all so it is hardly surprising that they are fragile when presented with more realistic data.

The idea that vowel and consonant visemes should be treated differently is no surprise. The suggestion that vowel visemes are essentially mouth shapes and the consonants govern how we move in and out of them was first presented by Nichie in 1912 from human observations by a profoundly deaf educator [68] and is supported by results in [86] which show we should not mix vowel and consonant visemes for best results. Therefore, it is reassuring to see that the better speaker-independent phoneme-to-viseme mapping for continuous speech is a combination of two previous maps, where the two maps have differing derivation methods; perception and language rules.

Generally speaking the continuous case (bottom of Figure 4) gives improved accuracies compared to the isolated word case (top of Figure 4. The first response to explain this is to suggest the increase is caused by better training of classifiers with the greater volume of training samples in RMAV than in AVL2. However, we should note that this effect is marginally countered by the co-articulation effects in continuous speech, so a set of classifiers trained on a larger isolated word dataset and compared to AVL2 would provide a greater increase in recognition.

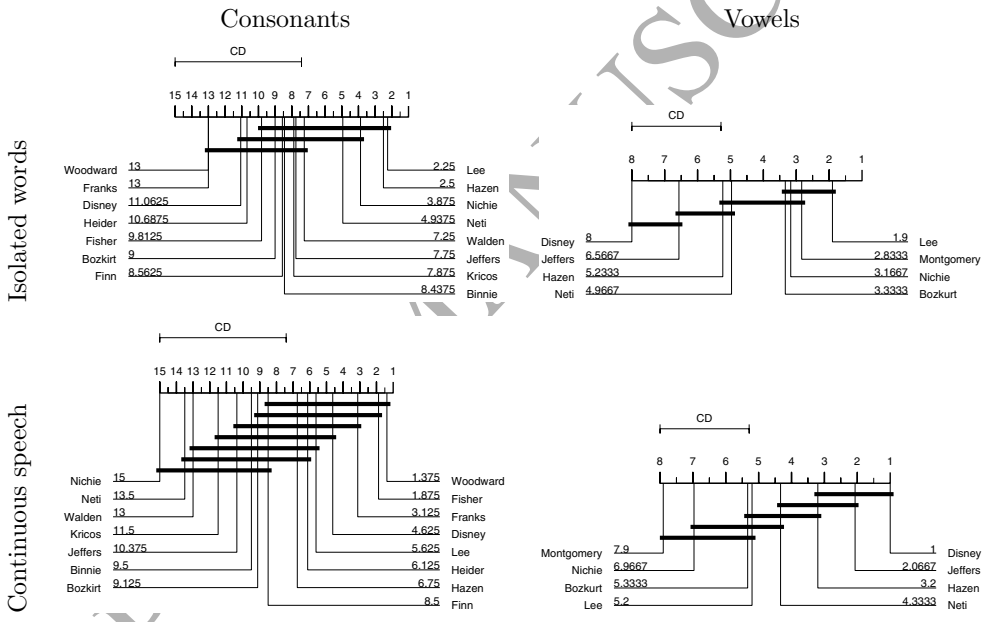

Figure 5: Critical difference of all phoneme-to-viseme maps independent of phoneme-to-viseme pair partner. Vowel maps are on the left side, consonants on the right. Isolated words are in the top row, and continuous speech along the bottom row.

Figure 5 are critical difference plots between the viseme class sets based upon their classification performance [87] with isolated word training. Critical difference is a measure of the confidence intervals between different machine learning algorithms derived from Wilcoxon tests on the ranked scores (here $p = 0.05$). Two assumptions within critical difference are: all measured results

are 'reliable', and all algorithms are evaluated using the same random samples [87]. As we use the HTK standard metrics [88], 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 on different datasets, whereas such as [89, 90], often require paired data or identical datasets.

Figure 5 shows a significant difference between some sub-sets of visemes. This is shown by the horizontal bars which do not overlap all viseme sets. Where the horizontal bars do overlap, this shows the viseme sets are indistinguishable at a 95% confidence. When comparing isolated words with continuous speech we see fewer significant differences with continuous speech despite there being more test data.

Table 8 summarises the best-performing visemes (consonant and vowels) for the isolated and continuous word data. The first column shows that the Lee consonants are the best performing for isolated words. But also that Hazan, Nichie, Neti etc are indistinguishable from Lee (they within Lee's critical difference). For continuous speech, the Woodward consonant visemes are the best but Fisher, Franks Disney etc are indistinguisable. In bold are the viseme sets that are common to both isolated words and continuous speech: Lee, Hazen, Finn and Fisher. For the vowels (second column) there are no common sets. However if we look at best and second-best (the third column of Table 8) then Hazen and Neti emerge as common. Looking across all sets the common method that performs near the top is that due to Hazen [19]. Interestingly these visemes were derived using the most realistic data (an audio-visual corpus based on TIMIT) and formed by a tree-based clustering of phoneme-trained HMMs. Note that the Hazan visemes were derived from American English data whereas here we use British English speakers.

The effectiveness of each mapping as a function of compression factor is presented in Figure 6. The two plots representing continuous speech (bottom of Figure 6) show improving performance with decreasing compression factor – we speculated earlier that the coarser visemes were better able to handle co-

Table 8: Critically different viseme sets changes with isolated word and continuous speech data. Sets are listed in the order they appear in Figure 5.

First Position Consonants	First Position Vowels	Second Position Vowels
Lee	Lee	Montgomery
Hazen	Montgomery	Nichie
Nichie	Nichie	Bozkurt
Neti	Bozkurt	**Hazen**
Walden		**Neti**
Jeffers		
Kricos		
Binnie		
Finn		
Bozkurt		
Fisher		
Woodward	Disney	Jeffers
Fisher	Jeffers	**Hazen**
Franks	Hazen	**Neti**
Disney		
Lee		
Heider		
Hazen		
Finn		

articulation. For the isolated word case (top) there is little difference. Very roughly, the best performing methods appear to have around 2 to 4 phonemes per viseme.

So far we have seen that there are noticeable differences between classification performances associated with a variety of viseme sets in the literature. Given that quite a few of the viseme sets are incremental improvements on previous sets, it is good to see confirmation that these sets are have rather similar

performance. We have identified the best sets for the various conditions and have used critical difference plots to explain the similarity between methods. We have identified that the most robust methods seem to be based on clustering large amounts of data but a questions arises when it comes to individual speakers – is it viable to create viseme sets per speaker and, if so, how similar are they? This is the topic of the next section.

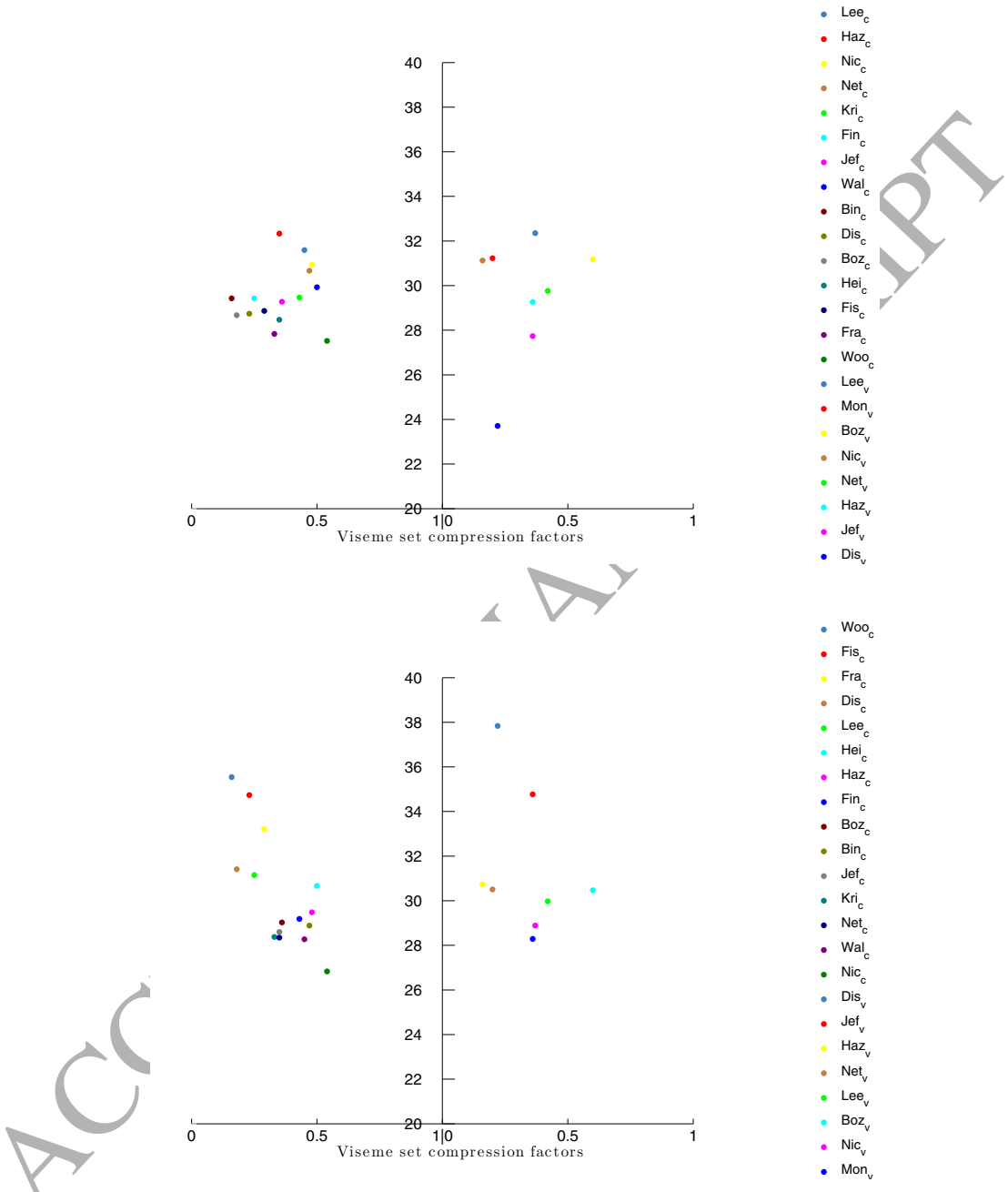

Figure 6: Scatter plot showing the relationship between compression factors, CF_s (x-axes), and word correctness, C, classification (y-axes) with consonant phoneme-to-viseme maps (left) and vowel phoneme-to-viseme maps (right), isolated word results are at the top, and continuous speech along the bottom.

4. Encoding speaker-dependent visemes

In the second part of our phoneme-to-viseme mapping study, two approaches are used to find a better method of mapping phonemes to visemes. These 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 be allocated to only one viseme class. These new P2V maps are tested on the AVL2 dataset using the same classification method as described in Section 3.2. The results from the best performing P2V map from our comparison study (Lee [71] or Woodward [75] and Disney [66]) is the benchmark to measure improvements with respect to the training data.

4.1. Viseme classes with strictly confusable phonemes

Our approaches for identifying visemes are speaker-dependent, data-driven and based on phoneme confusions within the classifier. The idea of speaker-dependent visemes is not new [31, 34] but our algorithm is, and in conjunction with the fixed outputs available from HTK enables easy reuse. The first undertaking in this work is to complete classification using phoneme labeled HHM classifiers. The classifiers are built in HTK with flat-start 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 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. For both data-driven speaker-dependent approaches, this is the first step of completing phoneme classification is essential to create the data to derive the

Table 9: Mean per speaker Correctness, C, of phoneme-labeled HMM classifiers.

	Speaker 1	Speaker 2	Speaker 3	Speaker 4
Phoneme C	24.72	23.63	57.69	43.41

P2V maps from. This is completed for each speaker in both AVL2 and RMAV datasets. Now, let us use a smaller seven-unit confusion matrix example, as in Table 10, to explain our clustering method.

Table 10: Demonstration confusion matrix showing confusions between phoneme-labeled 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.

	/$p1$/	/$p2$/	/$p3$/	/$p4$/	/$p5$/	/$p6$/	/$p7$/
/$p1$/	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

For the 'strictly-confused' viseme set (remembering 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 recognized and have no false positive/negative classifications are permitted to be single phoneme visemes. In Table 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 (Table 11).

Table 11: List of all possible subgroups of phonemes with an example set of seven phonemes

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

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 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 Table 10) with all others.

The first phoneme possibility in our list where this is true is $\{/p1/, /p3/, /p7/\}$.

This is confirmed by the Table 10 values:

$$N\{/p1/|/p3/\} + N\{/p3/|/p1/\} = 0 + 1 = 1 > 0$$

also,

$$N\{/p1/|/p7/\} + N\{/p7/|/p1/\} = 4 + 1 = 5 > 0$$

and,

$$N\{/p3/|/p7/\} + N\{/p7/|/p3/\} = 1 + 3 = 4 > 0.$$

This becomes our second viseme and thus our current viseme list looks like Table 12.

Table 12: Demonstration example 1: first-iteration of clustering, a phoneme-to-viseme map for strictly-confused phonemes.

Viseme	Phonemes
$/v1/$	$\{/p6/\}$
$/v2/$	$\{/p1/, /p3/, /p7/\}$

We now only have three remaining phonemes to cluster, $/p2/, /p4/$ and $/p5/$. This reduces our list of possible combinations substantially, see Table 13.

Table 13: List of all possible subgroups of phonemes with an example set of seven phonemes after the first viseme is formed.

$\{/p2/, /p4/, /p5/\}$
$\{/p2/, /p4/\}$
$\{/p2/, /p5/\}$
$\{/p4/, /p5/\}$

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 14.

Table 14: Demonstration example 2: final phoneme-to-viseme map for strictly-confused phonemes.

Viseme	Phonemes
/*v*1/	{/*p*6/}
/*v*2/	{/*p*1/, /*p*3/, /*p*7/}
/*v*3/	{/*p*2/, /*p*4/}
/*v*4/	{/*p*5/}

Our original phoneme classification has produced confusion matrices which permit confusions between vowel and consonant phonemes. We can see in Section 3.1 (Tables 2 and 3), 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 recognized 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 15.

Table 15: Strictly-confused phoneme speaker-dependent visemes. The score in brackets is the compression factor.$B1$ is listed on top, $B2$ visemes are listed at the bottom.

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

4.2. 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

Table 16: Demonstration example 3: final phoneme-to-viseme map for relaxed-confused phonemes.

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

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 14. 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 Table 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.

$$N\{/p1/|/p5/\} + N\{/p5/|/p1/\} = 0 + 3 = 3$$
$$N\{/p3/|/p5/\} + N\{/p5/|/p3/\} = 0 + 1 = 1$$
$$N\{/p4/|/p5/\} + N\{/p5/|/p4/\} = 2 + 1 = 3$$

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 14 to form a second set shown in Table 16. This has the added benefit that we have also increased the number of training samples for each classifier.

Remember, as we have two versions of Table 14 - one with mixed vowel and consonant phonemes and a second with divided vowels and consonant phonemes

Table 17: The four variations on speaker-dependent phoneme-to-viseme maps derived from phoneme confusion in phoneme classification.

Bear1, $B1$: Mixed vowels and consonants + Strict-confusion of phonemes	Bear2, $B2$: Split vowels and consonants + Strict-confusion of phonemes
Bear3, $B3$: Mixed vowels and consonants + Relaxed-confusion of phonemes	Bear4, $B4$ Split vowels and consonants + Relaxed-confusion of 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 17. For our strictly-confused P2V maps in Table 15, these become the relaxed P2V maps in Table 18. In Table 17 we have labeled each of the four variations $B1$, $B2$, $B3$ and $B4$ for ease of reference.

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.

Table 18: Relaxed-confused phoneme speaker-dependent visemes. The score in brackets is the ratio of visemes to phonemes. $B3$ visemes are on top, and $B4$ listed below.

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/}

4.3. Results analysis

Figure 7 (top) compares the new speaker-dependent viseme method with the Lee visemes which are the benchmark from the isolated word study. 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.

The second set of our experiments with continuous speech training data (RMAV) is to repeat our investigation with speaker-dependent visemes. These have been derived with the same methods described in Section 4.1 & 4.2 and are listed in full for each speaker in Appendix A. Our classification method is identical to that used previously with HMMs. In the previous work of [86], we see limited improvement in word classification with viseme classes due to the size of the dataset.

In Figure 7 (bottom) we have plotted the word correctness achieved for each RMAV speaker using all four variants of the speaker-dependent visemes. Our first observation is that on this figure, the correctness scores achieved range from 26.67% to 41.53%, whereas in Figure 7 (top) the values range from 20.60% to 36.53%. As before, this overall increase is attributed to the larger volume of training samples in RMAV compared to AVLetters2.

Compared to the benchmark of the Disney vowels and Montgomery consonant visemes which has been plotted in black on Figure 7 (bottom) we see that the comparison between speaker-dependent visemes and the best speaker-independent visemes is subject to the speaker. For three out of 12 speakers (sp01, sp03, sp05), the speaker-dependent visemes are all worse than our benchmark. For another three of our 12 speakers (sp02, sp09, sp14) all of the speaker-dependent visemes out-perform the benchmark. For all six remaining speakers, the results are mixed. This suggests that it is possible that speaker-dependent visemes could improve on speaker-independent ones, but that it is essential that they are exactly right for the individual otherwise they become at worse, detrimental, or a lot of effort for no significant improvement.

Careful observation of Figure 7 (top) shows that when considering the performance of mixed or split visemes, split visemes signfificantly ($>$ 1se) outperform mixed. When considering relaxed versus split the split has a marginal advantage but it is not significant ($<$1se).

The comparison of strict and split visemes for continuous speech (Figure 7 (bottom) is consistent with the isolated word observations. The strictly-confused visemes perform better than those with a relaxed confusion, but not statistically significantly ($<$1se). Again, we see that mixing vowel and consonants phonemes within individual viseme classes reduces the classification performance but not significantly.

In Figure 8 we have plotted accuracy, A, and correctness, C, for our best performing speaker-dependent visemes ($B1$) on continuous speech. We also plot, the accuracy scores of our benchmark from Woodward and Disney's visemes. These are compared with the correctness scores as a baseline to show the improvement. Whilst the improvement of speaker-dependent visemes is not significant when measured by Correctness, by plotting the accuracy of the viseme classifiers we can see that they do have a positive influence in reducing insertion errors which are a bugbear of lipreading.

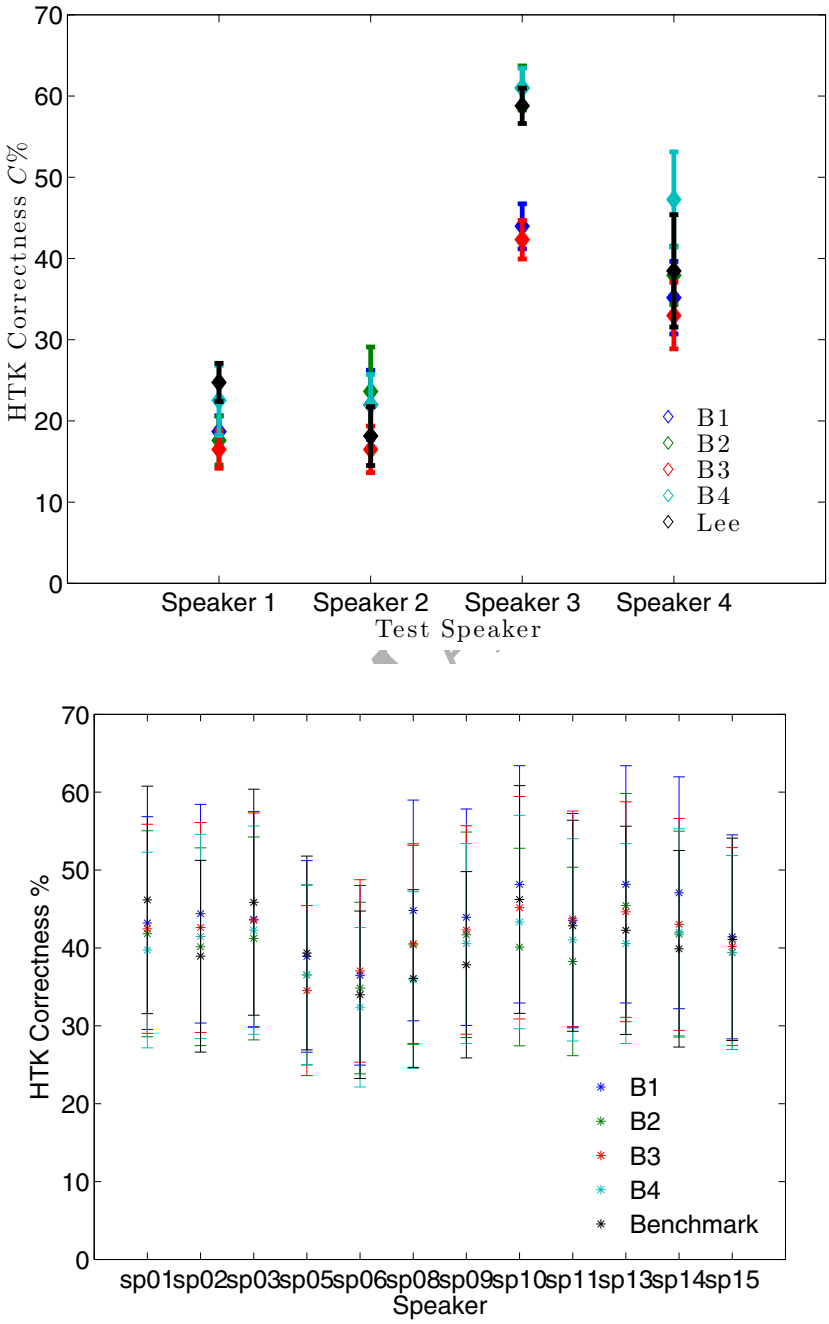

Figure 7: Word classification correctness $C \pm 1$se, using all four new methods of deriving speaker dependent visemes. AVL2 (top) and RMAV (bottom) speakers against Lee (top) and Woodward and Disney (bottom) benchmarks in black.

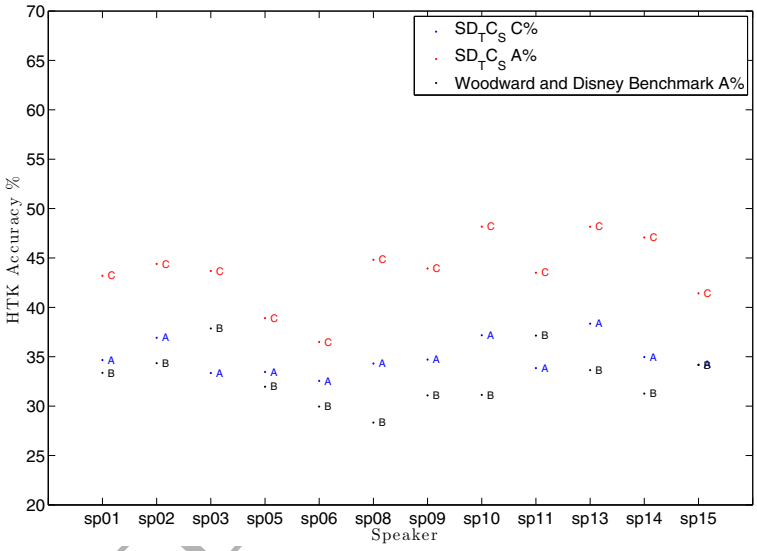

Figure 8: Comparing the accuracy change between strict and relaxed visemes to show the improvement in accuracy/reduction in insertion errors for all 12 speakers in continuous speech. The baseline is the correctness classification which ignores insertion error penalties.

5. Performance of individual visemes

In Figures 9 and 10, 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 probability of each class, $\Pr\{v|\hat{v}\}$. These values have been calculated from the `HResults` confusion matrices.

This analysis of visemes within a set is also used in [91], which proposes a threshold subject to the information in the features.

The same viseme comparison analysis has been repeated for our continuous speech recognition experiments and the results are shown in Figures 11 and 12.

In the isolated word data (Figures 9 and 10) the difference between a high-performing speaker map and a poor one is striking. Speaker 3 for example has at least five visemes in which $\Pr\{v|\hat{v}\} = 1$ (more in some configurations) whereas Speaker 1 has only one good viseme. Referring to Tables 15 and 18 there is no consistency on the best viseme although generally visual silence appears to be easy to spot. This variation is to be expected – speaker variablity is a very serious problem in lipreading.

Figures 11 and 12 show the same thing for the continuous speech data. Now there is a shallower drop-off to the curve and there are certainly no visemes for which $\Pr\{v|\hat{v}\} = 1$. Although there appears to be less variablity among speakers this is an illusion caused by the poorly-performing visemes to be similar among speakers – within the top five visemes there are significant differences among speakers.

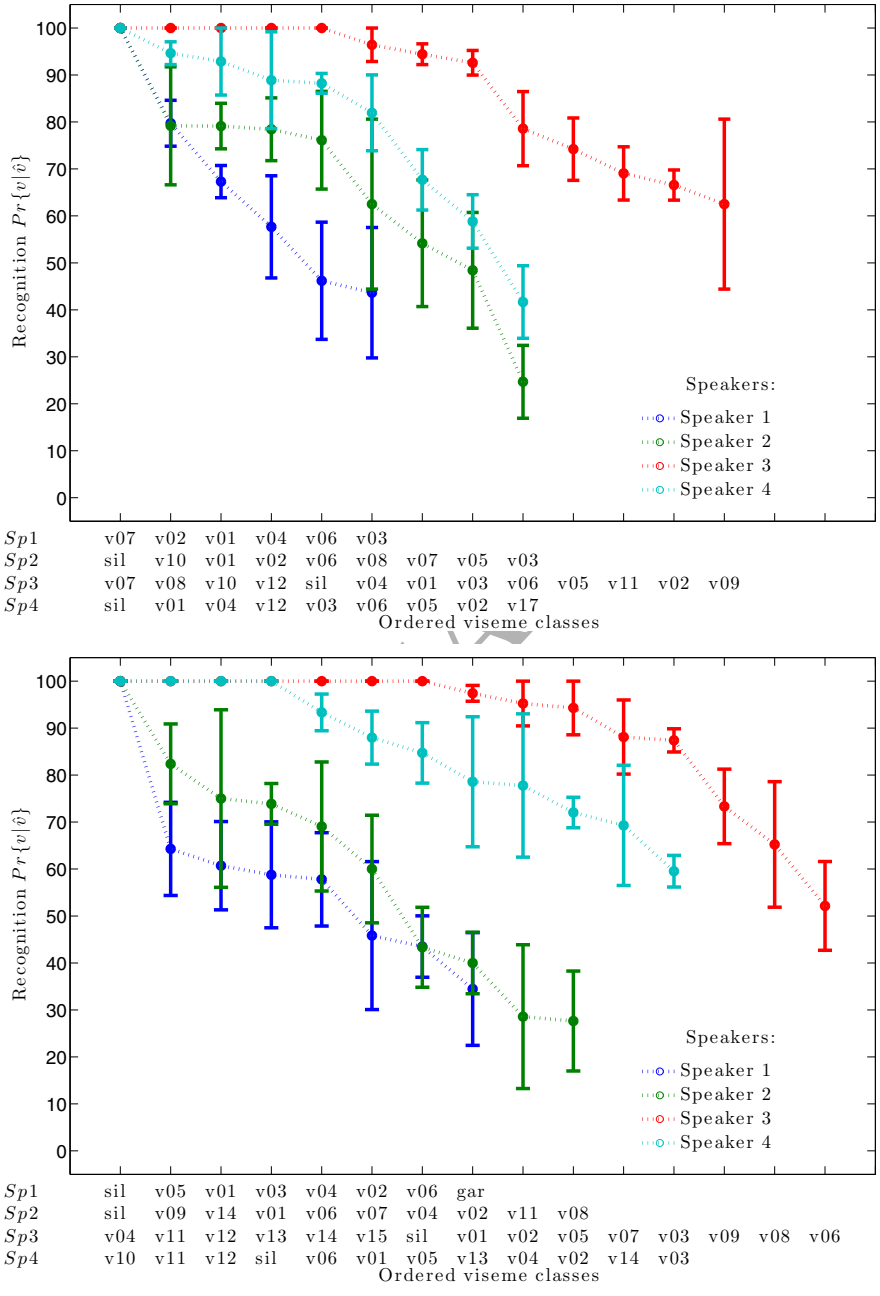

Figure 9: Individual viseme classification, Pr$\{v|\hat{v}\}$ with speaker-dependent visemes for four speakers with isolated word training of classifiers B1 visemes (top) and B2 visemes (bottom).

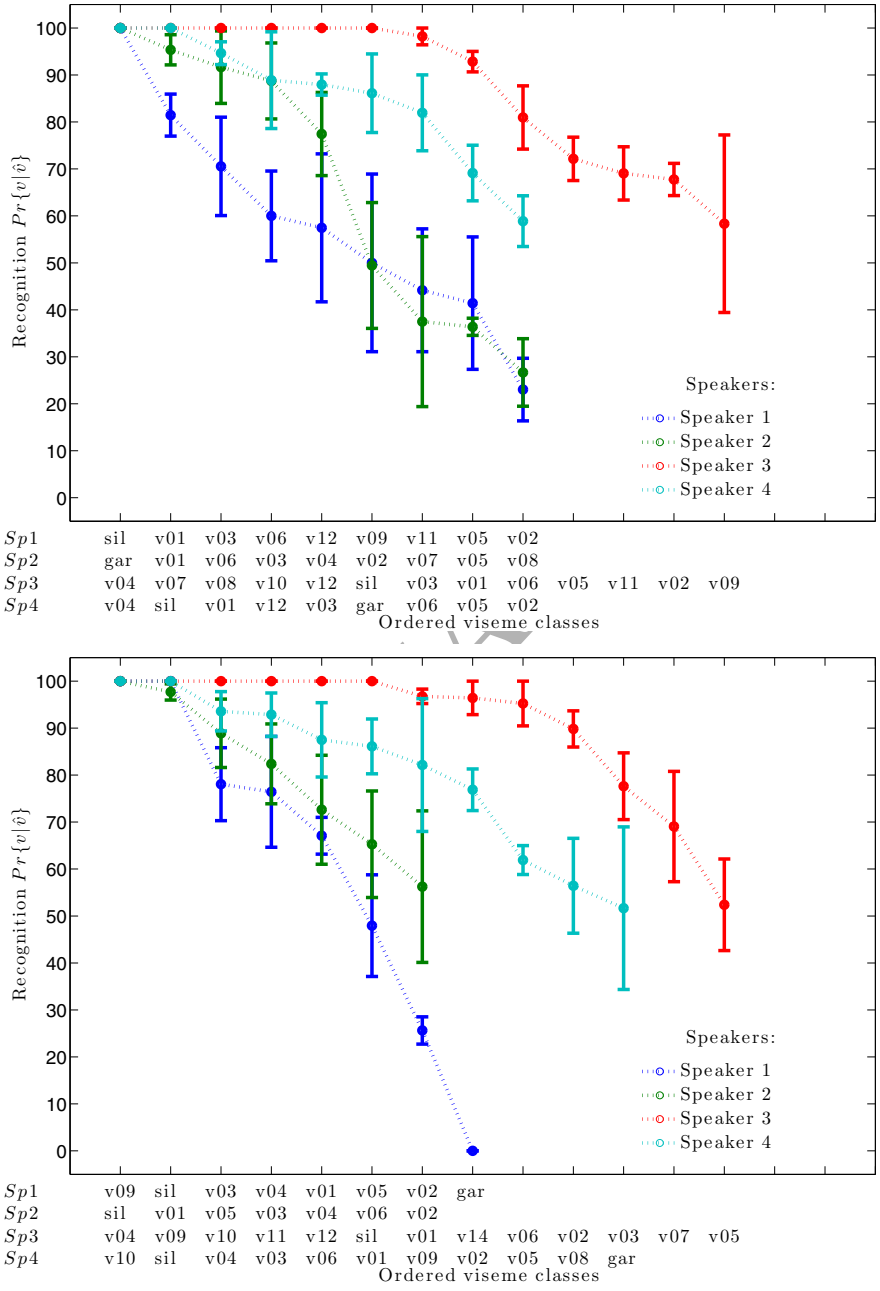

Figure 10: Individual viseme classification, Pr$\{v|\hat{v}\}$ with speaker-dependent visemes for four speakers with isolated word training of classifiers. B3 visemes (top) and B4 visemes (bottom).

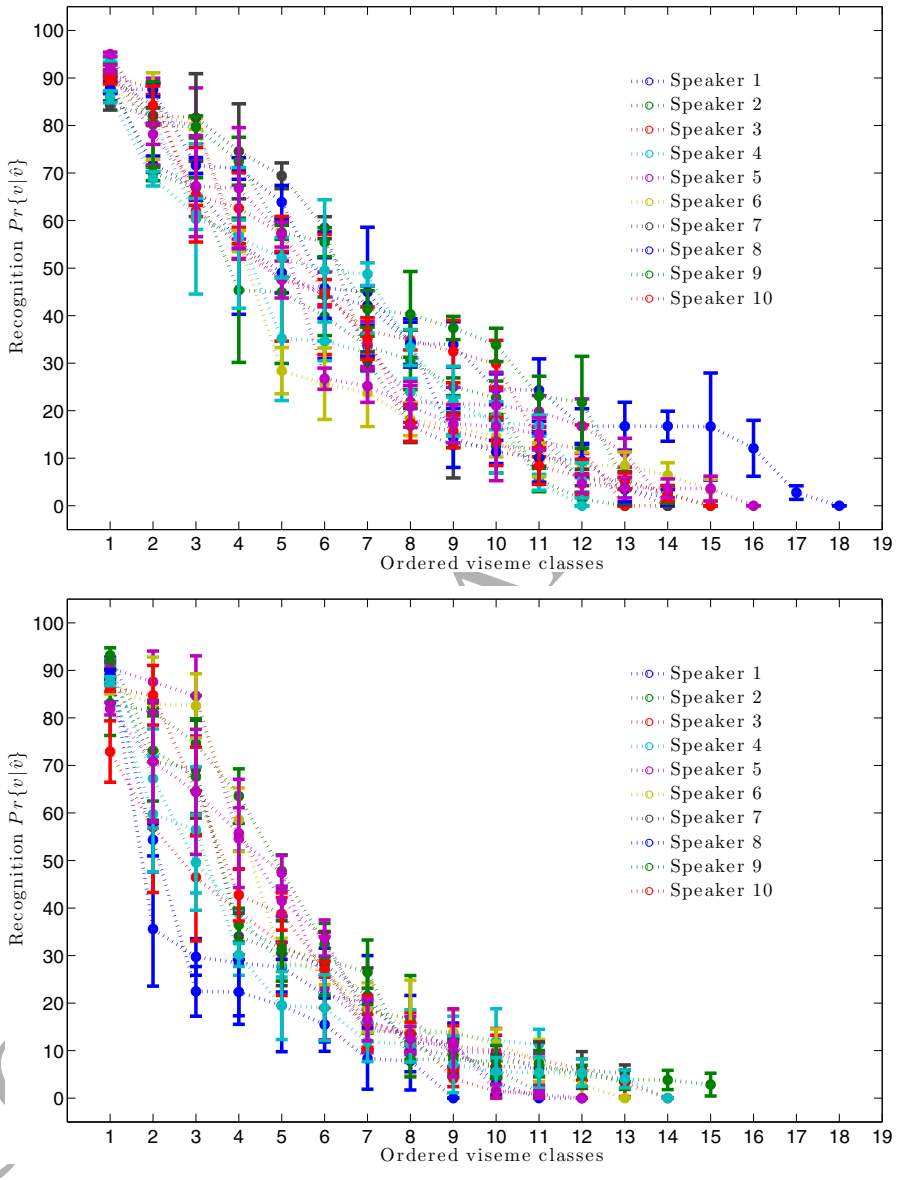

Figure 11: Individual viseme classification, $\Pr\{v|\hat{v}\}$ with speaker-dependent visemes for twelve speakers with continuous speech training of classifiers. B1 visemes (top) and B2 visemes (bottom).

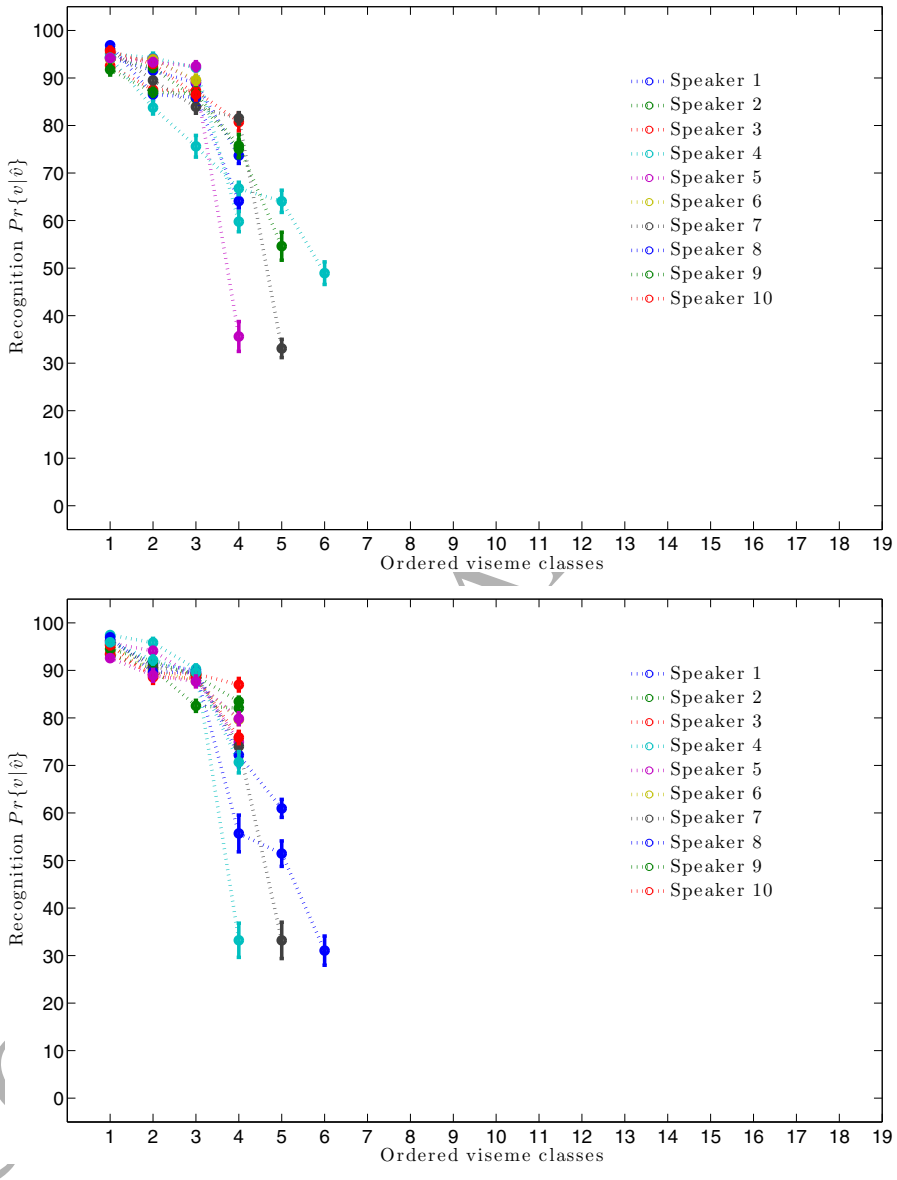

Figure 12: Individual viseme classification, $\Pr\{v|\hat{v}\}$ with speaker-dependent visemes for twelve speakers with continuous speech training of classifiers. B3 visemes (top) and B4 visemes (bottom).

6. Conclusions

While lipreading and hence expressive audio-visual speech recognition face a number of challenges, one the persistent difficulties has been the multiplicity of mappings between phonemes and visemes. This paper has described a study of previously suggested Phoneme-to-Viseme (P2V) maps. For isolated word classification, Lee's [71] is the best of the previously published maps. For continuous speech a combination of Woodward's and Disney's visemes are better. The best performing viseme sets have on average, between two and four phonemes per viseme.

When looking at speaker-independent visemes, whilst most viseme sets do not experience any difference in correctness between isolated and continuous speech, it is interesting to note that Woodward consonant visemes are better for continuous speech and are linguistically derived, whereas Lee visemes are better for isolated words and are data-derived. This suggests that an optimal set of visemes for all speakers would need to consider both the visual speech gestures of the individual *and* the rules of language. Which in essence is the dilemma for visemes: does one choose units that make sense in terms of likely visual gestures or in terms of the linguistic problem that is trying to be solved.

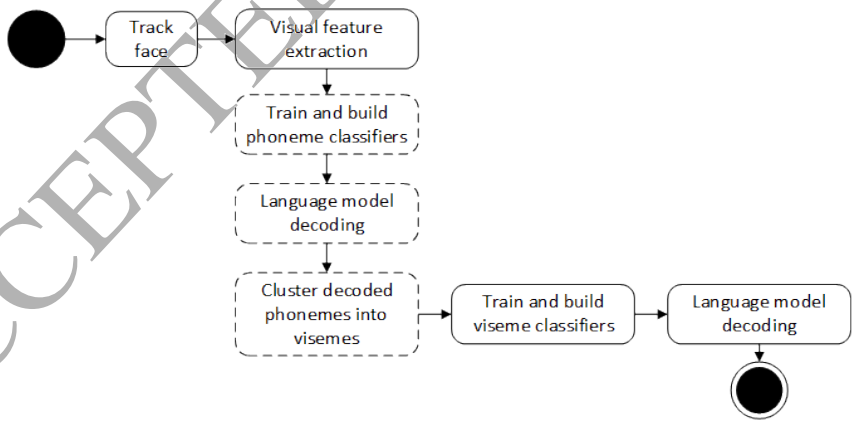

Figure 13: A simple augment to the conventional lip-reading system to include speaker-dependent visemes.

We have also derived some new visemes, the 'Bear' visemes. These new data-driven visemes respect speaker individuality in speech and uses this property to demonstrate that our second data-driven method tested, a strictly-confused viseme derivation with split vowel and consonant phonemes, can improve word classification. The best of Bear visemes is the strict confused phonemes with split vowels and consonants ($B2$) for both isolated and continuous speech.

Furthermore, a review of these speaker-dependent visemes (listed in Tables 15, 18, and Appendix A) shows that formally 'accepted' visemes such as { /p/ /b/ /m/ } and {/ʃ/ /ʒ/ /dʒ/ /tʃ/} are no longer present. Similarly with our previous vowel based visemes, six of our eight prior viseme sets pair /ʌ/ with /ɑ/ (albeit not as a complete viseme, others are also present) but with our best speaker-dependent visemes these two phonemes are not paired. This is an interesting insight because it suggests that formerly 'accepted' strong visemes might not be so useful for all speakers, and some adaptability, or further investigation into understanding viseme variation is still needed. Our suggestion at this time, is that linguistics or co-articulation in continuous speech, are a strong influence causing this variation.

In practical terms, our new viseme derivation method is simple and can be included within a conventional lipreading system easily. This is demonstrated in Figure 13 where our clustering method is shown in dashed boxes. We recommend this approach for viseme classification since speaker-independent visemes are unlikely to perform well.

In general, for cases, Speaker-dependent visemes reduce insertion errors when classifying continuous speech. This is thought to be because the phoneme confusions in speaker-dependent visemes are affected by speaker specific visual co-articulation. For all viseme sets, not mixing vowel and consonant phonemes significantly improves classification.

7. Acknowledgments

We gratefully acknowledge the assistance of Dr Yuxuan Lan and Dr Barry-John Theobald, formerly of the University of East Anglia for their help with HTK and general advice and guidance.

This work was conducted while Helen L. Bear was in receipt of a studentship from the UK Engineering and Physical Sciences Research Council (EPSRC).

References

[1] Limitations of visual speech recognition, in: Proceegins of the International Conference on Audio-Visual Speech Processing, 2010.

[2] E. Ong, R. Bowden, Robust lip-tracking using rigid flocks of selected linear predictors, in: 8th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG2008), 2008, pp. 247–254.

[3] I. Matthews, S. Baker, Active appearance models revisited, International Journal of Computer Vision 60 (2) (2004) 135–164.
URL http://www.springerlink.com/openurl.asp?id=doi:10.1023/B:VISI.0000029666.37597.d3

[4] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685. doi:10.1109/34.927467.

[5] Y. Lan, B.-J. Theobald, R. Harvey, View independent computer lip-reading, in: IEEE International Conference on Multimedia and Expo (ICME), 2012, pp. 432–437. doi:10.1109/ICME.2012.192.

[6] A. Pass, J. Zhang, D. Stewart, An investigation into features for multi-view lipreading, in: Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE, 2010, pp. 2417–2420.

[7] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Computer Vision and Image Understanding 115 (4) (2011) 541 – 558. doi:10.1016/j.cviu.2010.12.001.

[8] K. Kumar, T. Chen, R. Stern, Profile view lip reading, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, 2007, pp. IV–429–IV–432. doi:10.1109/ICASSP.2007.366941.

[9] R. Kaucic, A. Blake, Accurate, real-time, unadorned lip tracking, in: Computer Vision, 1998. Sixth International Conference on, IEEE, 1998, pp. 370–375.

[10] S. L. Bauman, G. Hambrecht, Analysis of view angle used in speechreading training of sentences, American Journal of Audiology 4 (3) (1995) 67–70. URL http://aja.asha.org/cgi/content/abstract/4/3/67

[11] P. Lucey, G. Potamianos, S. Sridharan, Visual speech recognition across multiple views, in: A. W.-C. Liew, S. Wang (Eds.), Visual Speech Reognition: Lip Segmentation and Mapping, 2009. doi:10.4018/978-1-60566-186-5.

[12] A. Blokland, A. H. Anderson, Effect of low frame-rate video on intelligibility of speech, Speech Communication 26 (1-2) (1998) 97–103. doi:http://dx.doi.org/10.1016/S0167-6393(98)00053-3.

[13] T. Saitoh, R. 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.

[14] H. Bear, R. W. Harvey, B.-J. Theobald, Y. Lan, Resolution limits on visual speech recognition, in: IEEE International Conference on Image Processing, 2014, pp. 2009–2013. doi:10.1109/ICIP.2014.7025274.

[15] M. Heckmann, F. Berthommier, C. Savariaux, K. Kroschel, Effects of image distortions on audio-visual speech recognition, in: AVSP 2003-International Conference on Audio-Visual Speech Processing, 2003, pp. 163–168.

[16] M. Vitkovitch, P. Barber, Visible speech as a function of image quality: Effects of display parameters on lipreading ability, Applied Cognitive Psychology 10 (2) (1996) 121–140. doi:10.1002/(SICI)1099-0720(199604)10:2<121::AID-ACP371>3.0.CO;2-V.
URL http://dx.doi.org/10.1002/(SICI)1099-0720(199604)10:2<121::AID-ACP371>3.0.CO;2-V

[17] L. Cappelletta, N. Harte, Phoneme-to-viseme mapping for visual speech recognition., in: ICPRAM (2), 2012, pp. 322–329.

[18] D. Howell, B.-J. Theobald, S. J. Cox, Confusion modelling for automated lip-reading using weighted finite-state transducers., in: AVSP, 2013, pp. 197–202.

[19] T. J. Hazen, K. Saenko, C.-H. La, J. R. Glass, A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments, in: Proceedings of the 6th International Conference on Multimodal Interfaces, ICMI '04, ACM, New York, NY, USA, 2004, pp. 235–242. doi:10.1145/1027933.1027972.
URL http://doi.acm.org/10.1145/1027933.1027972

[20] J. Shin, J. Lee, D. Kim, Real-time lip reading system for isolated korean word recognition, Pattern Recognition 44 (3) (2011) 559–571.

[21] H. L. Bear, R. Harvey, Decoding visemes: Improving machine lip-reading, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2009–2013.

[22] I. Matthews, J. Bangham, R. Harvey, S. Cox, Non-linear scale decomposition based features for visual speech recognition, Proceedings of the IX European Signal Processing Conference (EUSIPCO) (1998) 303–305.

[23] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, R. Bowden, Comparing visual features for lipreading, in: International Conference on Auditory-Visual Speech Processing 2009, 2009, pp. 102–106.

[24] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, R. Bowden, Improving visual features for lip-reading, Proceedings of the International Conference on Audio-Visual Speech Processing (AVSP) 7 (3) (2010) 42–48.

[25] I. Matthews, T. Cootes, J. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (2) (2002) 198 –213. doi:10.1109/34.982900.

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

[27] Q. Zhu, A. Alwan, On the use of variable frame rate analysis in speech recognition, in: Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, Vol. 3, IEEE, 2000, pp. 1783–1786.

[28] K. Thangthai, R. Harvey, S. Cox, B.-J. Theobald, Improving lip-reading performance for robust audiovisual speech recognition using dnns, in: Proc. FAAVSP, 1St Joint Conference on Facial Analysis, Animation and Audio–Visual Speech Processing, 2015.

[29] F. J. Huang, T. Chen, Tracking of multiple faces for human-computer interfaces and virtual environments, in: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Vol. 3, 2000, pp. 1563–1566. doi:10.1109/ICME.2000.871067.

[30] J. Jiang, A. Alwan, L. E. Bernstein, E. T. Auer, P. 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), Vol. 1, 2002, pp. I–441 –I–444. doi:10.1109/ICASSP.2002.5743749.

[31] S. Lesner, P. Kricos, Visual vowel and diphthong perception across speakers, Journal of the Academy of Rehabilitative Audiology 14 (1981) 252–258.

[32] R. Cutler, L. Davis, Look who's talking: speaker detection using video and audio correlation, in: IEEE International Conference on Multimedia and Expo (ICME), Vol. 3, 2000, pp. 1589–1592. doi:10.1109/ICME.2000.871073.

[33] J. Luettin, N. Thacker, S. Beet, Speaker identification by lipreading, in: Proceedings of the Fourth International Conference on Spoken Language (ICSLP), Vol. 1, 1996, pp. 62–65. doi:10.1109/ICSLP.1996.607030.

[34] H. L. Bear, S. J. Cox, R. W. Harvey, Speaker-independent machine lip-reading with speaker-dependent viseme classifiers, Facial Animation and Audio-Visual Speech Processing (FAAVSP) 2015 (2015) 190–195. URL http://www.isca-speech.org/archive/avsp15/papers/av15_190.pdf

[35] J. L. Newman, S. J. Cox, Speaker independent visual-only language identification, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010, pp. 5026–5029.

[36] S. Taylor, B.-J. Theobald, I. Matthews, The effect of speaking rate on audio and visual speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3037–3041. doi:10.1109/ICASSP.2014.6854158.

[37] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy, Cuave: A new audio-visual database for multimodal human-computer interface research, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp. II–2017–II–2020. doi:10.1109/ICASSP.2002.5745028.

[38] J. F. G. Perez, A. F. Frangi, E. L. Solano, K. Lukas, Lip reading for robust speech recognition on embedded devices, in: Proceedings. (ICASSP '05).

IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Vol. 1, 2005, pp. 473–476. doi:10.1109/ICASSP.2005.1415153.

[39] K. Paleek, Lipreading using spatiotemporal histogram of oriented gradients, in: 2016 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 1882–1885. doi:10.1109/EUSIPCO.2016.7760575.

[40] R. E. Shor, The production and judgment of smile magnitude, The Journal of General Psychology 98 (1) (1978) 79–96.

[41] S. Fagel, Effects of smiling on articulation: Lips, larynx and acoustics, in: Development of multimodal interfaces: active listening and synchrony, Springer, 2010, pp. 294–303.

[42] M. Kienast, A. Paeschke, W. Sendlmeier, Articulatory reduction in emotional speech, in: Sixth European Conference on Speech Communication and Technology, 1999.

[43] M. Kienast, W. F. Sendlmeier, Acoustical analysis of spectral and temporal changes in emotional speech, in: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000.

[44] F. Shaw, B.-J. Theobald, Expressive modulation of neutral visual speech, IEEE MultiMedia 23 (4) (2016) 68–78.

[45] W. Hamza, E. Eide, R. Bakis, M. Picheny, J. Pitrelli, The ibm expressive speech synthesis system, in: Eighth International Conference on Spoken Language Processing, 2004.

[46] N. N. Khatri, Z. H. Shah, S. A. Patel, Facial expression recognition: A survey, International Journal of Computer Science and Information Technologies (IJCSIT) 5 (1) (2014) 149–152.

[47] S. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, IEEE transactions on Affective Computing 6 (1) (2015) 1–12.

[48] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, B. Wang, Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech, IEEE Transactions on Multimedia 18 (7) (2016) 1319–1329.

[49] S. Zhang, X. Wang, G. Zhang, X. Zhao, Multimodal emotion recognition integrating affective speech with facial expression, WSEAS Transactions on Signal Processing 10 (2014) (2014) 526–537.

[50] T. Cootes, G. Edwards, C. Taylor, Active appearance models, Pattern Analysis and Machine Intelligence, IEEE Transactions on 23 (6) (2001) 681 –685. doi:10.1109/34.927467.

[51] R. Seymour, D. Stewart, J. Ming, Comparison of image transform-based features for visual speech recognition in clean and corrupted videos, Journal on Image and Video Processing 2008 (2008) 14.

[52] G. Potamianos, C. Neti, J. Luettin, I. Matthews, Audio-visual automatic speech recognition: An overview, Issues in Visual and Audio-Visual Speech Processing 22 (2004) 23.

[53] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.

[54] Z. Yan, V. Jagadeesh, D. DeCoste, W. Di, R. Piramuthu, Hd-cnn: Hierarchical deep convolutional neural network for image classification, in: International Conference on Computer Vision (ICCV), Vol. 2, 2015.

[55] M. Sundermeyer, R. Schlüter, H. Ney, Lstm neural networks for language modeling., in: Interspeech, 2012, pp. 194–197.

[56] W. Byeon, T. M. Breuel, F. Raue, M. Liwicki, Scene labeling with lstm recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3547–3555.

[57] J. S. Chung, A. Zisserman, Lip Reading in the Wild, Springer International Publishing, Cham, 2017, pp. 87–103. doi:10.1007/978-3-319-54184-6_6.
URL http://dx.doi.org/10.1007/978-3-319-54184-6_6

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

[59] B.-J. Theobald, Visual speech synthesis using shape and appearance models, Ph.D. thesis, University of East Anglia (2003).

[60] C. A. Binnie, P. L. Jackson, A. A. Montgomery, Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation, Journal of Speech and Hearing Disorders 41 (4) (1976) 530.

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

[62] J. R. Franks, J. Kimble, The confusion of english consonant clusters in lipreading, Journal of Speech, Language and Hearing Research 15 (3) (1972) 474.

[63] B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, C. J. Jones, Effects of training on the visual recognition of consonants, Journal of Speech, Language and Hearing Research 20 (1) (1977) 130.

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

[65] E. Owens, B. Blazek, Visemes observed by hearing-impaired and normal-hearing adult viewers, Journal of Speech and Hearing Research 28 (3) (1985) 381.

[66] J. Lander, Read my lips: Facial animation techniques, http://www.gamasutra.com/view/feature/131587/read_my_lips_facial_animation_.php, accessed: 2014-01-28 (2014).

[67] A. A. Montgomery, P. L. Jackson, Physical characteristics of the lips underlying vowel lipreading performance, The Journal of the Acoustical Society of America 73 (1983) 2134–2144.

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

[69] E. Bozkurt, C. Erdem, E. Erzin, T. Erdem, M. Ozkan, Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation, in: 3DTV Conference, IEEE, 2007, pp. 1–4.

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

[71] S. Lee, D. Yook, Audio-to-visual conversion using Hidden Markov Models, in: PRICAI 2002: Trends in Artificial Intelligence, Springer, 2002, pp. 563–570.

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

[73] K. E. Finn, A. A. Montgomery, Automatic optically-based recognition of speech, Pattern Recognition Letters 8 (3) (1988) 159–164.

[74] F. Heider, G. M. Heider, An experimental investigation of lipreading, Psychological Monographs 52 (232) (1940) 124–153.

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

[76] A. Bhattachayya, On a measure of divergence between two statistical population defined by their population distributions, Bulletin Calcutta Mathematical Society 35 (1943) 99–109.

[77] K. Wilson, The Columbia guide to standard American English, New York : Columbia University Press, 1993.

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

[79] H. L. Bear, Decoding visemes: improving machine lip-reading. PhD thesis, University of East Anglia, 2016.

[80] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, The DARPA speech recognition research database: specifications and status, in: Proceedings of the DARPA Workshop on speech recognition, 1986, pp. 93–99.

[81] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, R. Bowden, Improving visual features for lip-reading, in: Proceedings of International Conference on Auditory-Visual Speech Processing, Vol. 201, 2010.

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

[83] J. Gower, Generalized procrustes analysis, Psychometrika 40 (1) (1975) 33–51. doi:10.1007/BF02291478.
URL http://dx.doi.org/10.1007/BF02291478

[84] S. Baker, Inverse compositional algorithm, in: K. Ikeuchi (Ed.), Computer Vision, Springer US, 2014, pp. 426–428. doi:10.1007/978-0-387-31439-6_759.
URL http://dx.doi.org/10.1007/978-0-387-31439-6_759

[85] T. J. Hazen, Automatic alignment and error correction of human generated transcripts for long speech recordings., in: INTERSPEECH, Vol. 2006, 2006, pp. 1606–1609.

[86] H. L. Bear, R. W. Harvey, B.-J. Theobald, Y. Lan, Which phoneme-to-viseme maps best improve visual-only computer lip-reading?, in: Advances in Visual Computing, Springer, 2014, pp. 230–239. doi:10.1007/978-3-319-14364-4_22.

[87] J. Demsar, Statistical comparisons of classifiers over multiple datasets, Journal of Machine Learning Research 7 (2006) 1–30.

[88] S. J. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK book version 3.4 (2006).

[89] R. R. Bouckaert, E. Frank, Evaluating the replicability of significance tests for comparing learning algorithms, in: Advances in knowledge discovery and data mining, Springer, 2004, pp. 3–12.

[90] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, The Journal of Machine Learning Research 5 (2004) 1089–1105.

[91] H. L. Bear, G. Owen, R. Harvey, B.-J. Theobald, Some observations on computer lip-reading: moving from the dream to the reality, in: SPIE Security+ Defence, International Society for Optics and Photonics, 2014, pp. 92530G–92530G. doi:10.1117/12.2067464.

Appendix A. RMAV Speaker-dependent P2V maps

Table A.19: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp01

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp01	/v01/	/dʒ/ /m/	/v01/	/æ/ /ʌ/ /ə/ /ay/	/v01/	/ɪə/ /t/ /θ/ /uw/	/v01/	/æ/ /ʌ/ /ə/ /ay/
	/v02/	/ɜ/ /ɪ/ /iy/ /k/		/eh/ /ɪə/ /ɪ/ /iy/		/z/		/eh/ /ɪə/ /ɪ/ /iy/
		/n/ /ŋ/ /r/ /s/	/v02/	/ɒ/ /əʊ/	/v02/	/ɜ/ /ɪ/ /iy/ /k/	/v02/	/ʃ/ /θ/ /v/ /w/
	/v03/	/ey/	/v03/	/ɔ/ /ɜ/ /ey/		/n/ /ŋ/ /r/ /s/		/z/
	/v04/	/ə/ /ð/ /ɛ/ /eh/	/v04/	/ɑ/ /sil/	/sil/	/sil/ /sil/ /sp/	/v03/	/b/ /d/ /f/ /k/
		/ʊ/	/v05/	/uw/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/		/m/ /n/ /ŋ/ /p/ /r/
	/v05/	/ɑ/	/v06/	/ʊ/		/ə/ /ay/ /ə/ /b/ /tʃ/		/r/ /s/ /t/
	/v06/	/ɪə/ /t/ /θ/ /uw/	/v07/	/ɔə/		/tʃ/ /d/ /ð/ /ɛ/ /eh/	/sil/	/sil/ /sil/ /sp/
		/z/	/v08/	/ɔɪ/		/eh/ /ey/ /f/ /g/ /ɦ/	/gar/	/gar/ /ɑ/ /ɔ/ /ɑʊ/ /ə/
	/v07/	/ɒ/ /əʊ/ /p/ /w/	/v09/	/ə/		/ɦ/ /dʒ/ /m/ /ɒ/ /əʊ/		/ð/ /ɜ/ /ey/ /g/ /ɦ/
	/v08/	/ʃ/	/v10/	/ɑʊ/		/əʊ/ /ɔɪ/ /p/ /ʃ/ /ɔə/		/ɦ/ /dʒ/ /ɒ/ /əʊ/ /ɔɪ/
	/v09/	/ɔ/	/v11/	/b/ /d/ /f/ /k/		/ɔə/ /ʊ/ /w/ /y/ /ʒ/		/ɔɪ/ /ɔə/ /ʊ/ /uw/ /ʒ/
	/v10/	/æ/		/m/ /n/ /ŋ/ /p/ /r/		/ʒ/		/ʒ/
	/v11/	/d/ /g/ /ɦ/		/r/ /s/ /t/				
	/v12/	/b/	/v12/	/ð/ /dʒ/				
	/v13/	/y/	/v13/	/ʃ/ /θ/ /v/ /w/				
	/v14/	/ʌ/ /ay/		/z/				
	/v15/	/ʒ/	/v14/	/g/				
	/v16/	/ɔə/	/v15/	/tʃ/ /ɦ/				
	/v17/	/sil/	/v16/	/ʒ/				
	/v18/	/ɔɪ/	/sil/	/sil/ /sil/ /sp/				
	/v19/	/tʃ/						
	/v20/	/ə/						
	/v21/	/ɑʊ/						
	/gar/	/gar/ /sp/						

Table A.20: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp02

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp02	/v01/	/l/ /m/ /n/ /p/ /s/ /ʃ/ /t/ /v/ /w/	/v01/	/ə/ /ay/ /ɛ/ /eh/ /ey/ /ɪ/ /iy/	/v01/	/ə/ /ay/ /b/ /d/ /eh/ /ey/ /dʒ/	/v01/	/ə/ /ay/ /ɛ/ /eh/ /ey/ /ɪ/ /iy/
		/w/	/v02/	/ɔ/ /ɪə/ /ɒ/ /əʊ/	/v02/	/l/ /m/ /n/ /p/	/v02/	/b/ /m/ /n/ /ŋ/
	/v02/	/g/ /ɦ/ /ɪə/ /ɪ/ /k/	/v03/	/æ/ /ʌ/ /ɑʊ/ /ɔɪ/		/s/ /ʃ/ /t/ /v/ /w/		/ɾ/ /s/ /ʃ/ /t/ /v/
			/v04/	/ʊ/ /uw/		/w/		/v/ /w/ /y/ /z/
	/v03/	/ə/ /ay/ /b/ /d/ /eh/ /ey/ /dʒ/	/v05/	/əə/	/sil/	/sil/ /sil/ /sp/	/sil/	/sil/ /sil/ /sp/
			/v06/	/sil/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/ /ə/ /tʃ/ /ɛ/ /ʒ/ /ɾ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/ /ə/ /tʃ/ /d/ /ð/ /f/
	/v04/	/ɑ/ /ɔ/	/v07/	/ɑ/				
	/v05/	/ɜ/ /uw/ /y/ /z/	/v08/	/b/ /m/ /n/ /ŋ/ /ɾ/ /s/ /ʃ/ /t/ /v/ /w/ /y/ /z/		/f/ /g/ /ɦ/ /ɪə/ /ɪ/ /ɪ/ /iy/ /k/ /ŋ/ /ɒ/ /ɒ/ /əʊ/ /əə/ /ɔɪ/ /ɵ/ /ɔə/ /ʊ/ /uw/ /ʒ/		/f/ /g/ /ɦ/ /ɪə/ /dʒ/ /k/ /l/ /ɒ/ /əʊ/ /əʊ/ /ɔɪ/ /θ/ /ɔə/ /ʊ/ /uw/ /ʒ/
	/v06/	/ɒ/ /əʊ/						
	/v07/	/æ/ /ʌ/ /ɑʊ/ /ɔɪ/						
	/v08/	/f/ /ŋ/ /ɔə/	/v09/	/dʒ/				
	/v09/	/ɛ/	/v10/	/d/ /ð/ /f/ /g/ /k/ /l/		/z/ /ʒ/		
	/v10/	/tʃ/ /θ/						
	/v11/	/ʒ/	/v11/	/tʃ/ /θ/				
	/v12/	/ʊ/	/v12/	/ʒ/				
	/v13/	/sil/	/sil/	/sil/ /sil/ /sp/				
	/gar/	/gar/ /ə/ /sp/	/gar/	/gar/ /ə/				

Table A.21: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp03

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp03	/v01/	/ey/ /f/ /ɪ/ /iy/ /k/ /l/ /m/ /n/ /ʃ/ /ʃ/	/v01/	/ɛ/ /ʒ/ /sil/ /uw/	/v01/	/ey/ /f/ /ɪ/ /iy/ /k/ /l/ /m/ /n/ /ʃ/ /ʃ/	/v01/	/ay/ /eh/ /ey/ /ɪə/ /iy/ /ɒ/ /əʊ/
			/v02/	/ʊ/			/v02/	/g/ /k/ /l/ /m/ /p/ /r/ /s/ /t/ /θ/ /θ/
	/v02/	/ð/ /g/	/v03/	/ay/ /eh/ /ey/ /ɪə/ /iy/ /ɒ/ /əʊ/	/v02/	/ɛ/ /r/ /s/ /t/ /z/		
	/v03/	/ɛ/ /ɾ/ /s/ /sil/ /uw/ /z/	/v04/	/ɔ/		/z/		
			/v05/	/əə/	/sil/	/sil/ /sil/ /sp/	/sil/	/sil/ /sil/ /sp/
	/v04/	/d/ /θ/ /v/ /w/	/v06/	/æ/ /ʌ/ /ə/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/ /ə/ /ay/ /ə/ /b/ /tʃ/ /tʃ/ /d/ /ð/ /eh/ /ɜ/ /ɜ/ /g/ /ɦ/ /ɪə/ /ŋ/ /ŋ/ /ɒ/ /əʊ/ /ɔɪ/ /p/ /ɵ/ /ɔə/ /ʊ/ /v/ /v/ /w/ /y/ /ʒ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/ /ə/ /ə/ /b/ /tʃ/ /d/ /d/ /ð/ /ɛ/ /ɜ/ /f/ /f/ /ɦ/ /dʒ/ /ŋ/ /ɔɪ/ /ɔɪ/ /ʃ/ /əə/ /ʊ/ /uw/ /uw/ /v/ /w/ /y/ /z/ /z/ /ʒ/
	/v05/	/ɔ/ /əʊ/ /p/	/v07/	/ə/				
	/v06/	/æ/	/v08/	/ɑʊ/		/tʃ/ /d/ /ð/ /eh/ /ɜ/		
	/v07/	/ə/ /ay/ /b/ /tʃ/	/v09/	/ɑ/				
	/v08/	/ŋ/	/v10/	/g/ /k/ /l/ /m/ /p/ /r/ /s/ /t/ /θ/ /θ/				
	/v09/	/ɦ/						
	/v10/	/ɑ/ /eh/ /ɜ/						
	/v11/	/əə/ /ʊ/	/v11/	/tʃ/ /d/ /ð/ /f/				
	/v12/	/ʌ/ /ɪə/	/v12/	/dʒ/ /v/ /w/ /z/				
	/v13/	/ʒ/	/v13/	/b/				
	/v14/	/ə/	/v14/	/ʃ/ /ʒ/				
	/v15/	/ɑʊ/	/v15/	/ɦ/ /ŋ/				
	/gar/	/gar/ /ɔɪ/ /sp/	/sil/	/sil/ /sil/ /sp/				
			/gar/	/gar/ /ɔɪ/				

Table A.22: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp05

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp05	/v01/	/æ/ /ɔ/ /ə/ /ð/ /ɜ/ /ey/ /ɪ/ /iy/ /k/ /k/ /l/ /n/	/v01/	/æ/ /ɔ/ /ə/ /eh/ /ey/ /ɪ/ /iy/ /ɒ/ /əʊ/ /əʊ/	/v01/ /v02/	/aɣ/ /b/ /d/ /w/ /æ/ /ɔ/ /ə/ /ð/ /ɜ/ /ey/ /ɪ/ /iy/ /k/ /k/ /l/ /n/	/v01/ /v02/	/aɣ/ /uw/ /d/ /ð/ /f/ /dʒ/ /l/ /m/ /n/ /r/ /s/ /s/ /ʃ/
	/v02/	/p/ /r/ /s/ /t/ /z/	/v02/ /v03/	/ɛ/ /ʊ/ /aɣ/ /uw/	/sil/	/sil/ /sil/ /sp/	/sil/	/sil/ /sil/ /sp/
	/v03/	/ə/ /ŋ/ /uw/ /v/	/v04/	/ɔə/	/gar/	/gar/ /ɑ/ /ʌ/ /ɑʊ/ /ə/ /ɛ/ /f/ /g/ /ñ/ /ɪə/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/ /ə/ /ə/ /b/ /tʃ/ /ɛ/
	/v04/	/tʃ/ /ɒ/	/v05/	/ʌ/ /ɑʊ/				
	/v05/	/aɣ/ /b/ /d/ /w/	/v06/	/ɑ/ /ɪə/		/ɪə/ /dʒ/ /m/ /ŋ/ /ɒ/		/ɛ/ /eh/ /ɜ/ /ey/ /g/
	/v06/	/f/ /m/	/v07/	/g/ /ñ/ /t/ /v/		/ɒ/ /əʊ/ /ɔɪ/ /p/ /r/		/g/ /ñ/ /ɪə/ /ɪ/ /iy/
	/v07/	/ɑ/ /g/ /ñ/	/v08/	/p/ /w/ /y/		/r/ /s/ /ʃ/ /t/ /θ/		/iy/ /ŋ/ /ɒ/ /əʊ/ /ɔɪ/
	/v08/	/əʊ/ /ʃ/	/v09/	/d/ /ð/ /f/ /dʒ/ /l/ /m/ /n/ /r/ /s/ /s/ /ʃ/		/θ/ /ɔə/ /ʊ/ /uw/ /v/		/ɔɪ/ /p/ /t/ /θ/ /əə/
	/v09/	/dʒ/				/v/ /y/ /z/ /ʒ/		/əə/ /ʊ/ /v/ /w/ /y/
	/v10/	/dʒ/						/y/ /z/ /ʒ/
	/v11/	/ɛ/ /y/	/v10/	/ŋ/ /θ/				
	/v12/	/θ/	/v11/	/b/ /tʃ/				
	/v13/	/ʌ/ /ɑʊ/	/v12/	/ʒ/				
	/v14/	/ʒ/	/sil/	/sil/ /sil/ /sp/				
	/v15/	/ɔə/	/gar/	/gar/ /ə/ /ɔɪ/				
	/v16/	/j/ /h/						
	/gar/	/gar/ /ə/ /ɔɪ/ /sil/ /sp/						

Table A.23: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp06

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp06	/v01/	/ə/ /aɣ/ /d/ /ð/ /eh/ /ɪ/ /k/ /l/ /n/ /n/ /p/ /s/ /t/	/v01/	/ɑ/ /æ/ /ʌ/ /ə/ /ɜ/ /ɪə/ /ɪ/ /ɒ/ /əʊ/ /əʊ/	/v01/ /v02/	/ñ/ /ŋ/ /ɒ/ /əʊ/ /ə/ /aɣ/ /d/ /ð/ /eh/ /ɪ/ /k/ /l/ /n/	/v01/	/ɑ/ /æ/ /ʌ/ /ə/ /ɜ/ /ɪə/ /ɪ/ /ɒ/ /əʊ/ /əʊ/
	/v02/	/v/ /w/ /y/ /z/	/v02/	/sil/ /uw/		/n/ /p/ /s/ /t/	/v02/	/k/ /l/ /m/ /n/ /r/ /s/ /ʃ/ /t/ /v/ /v/ /w/ /y/ /z/
	/v03/	/m/	/v03/	/aɣ/ /ey/ /iy/ /ʊ/	/sil/ /gar/	/sil/ /sil/ /sp/ /gar/ /ɑ/ /æ/ /ʌ/ /ə/ /ə/ /b/ /tʃ/ /ɜ/ /ey/ /ey/ /f/ /g/ /ɪə/ /iy/ /iy/ /dʒ/ /m/ /ɔɪ/ /r/ /r/ /ʃ/ /θ/ /əə/ /ʊ/ /ʊ/ /uw/ /v/ /w/ /y/ /y/ /z/ /ʒ/		
	/v04/	/ñ/ /ŋ/ /ɒ/ /əʊ/	/v04/	/ɑʊ/ /ɔə/				
	/v05/	/ey/ /iy/ /r/ /ʃ/	/v05/	/ɛ/			/sil/ /gar/	/sil/ /sil/ /sp/ /gar/ /ɔ/ /ɑʊ/ /aɣ/ /ə/ /tʃ/ /d/ /ð/ /ɛ/ /ey/ /ey/ /f/ /g/ /ñ/ /iy/ /iy/ /dʒ/ /ŋ/ /ɔɪ/ /θ/ /θ/ /ɔə/ /ʊ/ /uw/ /ʒ/ /ʒ/
	/v06/	/ɪə/	/v06/	/ə/				
	/v07/	/ɑ/ /æ/ /ʌ/ /ɜ/	/v07/	/ɔ/				
	/v08/	/f/ /θ/ /əə/	/v08/	/k/ /l/ /m/ /n/ /r/ /s/ /ʃ/ /t/ /v/ /v/ /w/ /y/ /z/				
	/v09/	/uw/						
	/v10/	/uw/						
	/v11/	/b/ /tʃ/ /g/	/v09/	/b/ /tʃ/ /d/ /ð/ /g/ /dʒ/				
	/v12/	/ɔ/ /dʒ/						
	/v13/	/ʒ/	/v10/	/ʒ/				
	/v14/	/sil/	/v11/	/ñ/ /θ/				
	/v15/	/ə/	/v12/	/ŋ/				
	/v16/	/ɑʊ/	/sil/	/sil/ /sil/ /sp/				
	/v17/	/u/ /w/	/gar/	/gar/ /ɔɪ/				
	/gar/	/gar/ /ɔɪ/ /sp/						

Table A.24: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp08

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp08	/v01/	/eh/ /f/ /ɦ/ /ɪ/ /l/ /m/ /ŋ/ /p/ /r/ /ɾ/ /s/ /t/	/v01/	/ɑ/ /æ/ /ɔ/ /ə/ /eh/ /ey/ /ɪ/ /iy/ /uw/	/v01/	/eh/ /f/ /ɦ/ /ɪ/ /l/ /m/ /ŋ/ /p/ /r/ /ɾ/ /s/ /t/	/v01/	/ɑ/ /æ/ /ɔ/ /ə/ /eh/ /ey/ /ɪ/ /iy/ /uw/
	/v02/	/ɑ/ /æ/ /ɔ/ /ə/ /ey/ /n/ /ʊ/	/v02/	/ʊ/	/sil/	/sil/ /sil/ /sp/	/v02/	/k/ /l/ /n/ /p/ /s/ /t/ /θ/ /v/ /w/ /z/
	/v03/	/ay/ /b/ /uw/	/v03/	/ɒ/ /əʊ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/		
	/v04/	/g/	/v04/	/ɪə/		/ə/ /ay/ /ɔ/ /b/ /tʃ/		
	/v05/	/tʃ/	/v05/	/aʊ/ /ɛ/		/tʃ/ /d/ /ð/ /ɛ/ /ɜ/	/sil/	/sil/ /sil/ /sp/
	/v06/	/ʃ/ /y/	/v06/	/ʌ/ /ɜ/		/ɜ/ /ey/ /g/ /ɪə/ /dʒ/	/gar/	/gar/ /ɹ/ /ɑʊ/ /ə/ /b/
	/v07/	/ɒ/	/v07/	/ə/		/dʒ/ /k/ /n/ /ɒ/ /əʊ/		/d/ /ð/ /ɛ/ /ɜ/ /f/
	/v08/	/k/	/v08/	/k/ /l/ /n/ /p/ /s/ /t/ /θ/ /v/ /w/		/əʊ/ /ɔɪ/ /ʃ/ /θ/ /ɔə/ /ɔə/ /ʊ/ /ɔɪ/ /ʃ/ /ɔə/ /ʊ/ /ɪ/ /ʊ/ /ʃ/ /ʒ/		/f/ /g/ /ɦ/ /ɪə/ /dʒ/ /dʒ/ /m/ /ŋ/ /ɒ/ /əʊ/
	/v09/	/dʒ/		/w/ /z/		/w/ /y/ /z/ /ʒ/		/əʊ/ /ɔɪ/ /ʃ/ /ɔə/ /ʊ/
	/v10/	/ð/ /v/ /w/ /z/	/v09/	/d/ /ð/ /f/ /ɦ/				/ʊ/ /y/ /ʒ/
	/v11/	/θ/ /ʒ/		/ŋ/				
	/v12/	/ɜ/ /ɪə/	/v10/	/g/ /dʒ/				
	/v13/	/ɑʊ/ /əʊ/	/v11/	/b/ /tʃ/ /ʃ/				
	/v14/	/ʌ/ /ɛ/	/v12/	/ʒ/				
	/v15/	/ɔə/	/v13/	/y/				
	/v16/	/ə/	/sil/	/sil/ /sil/ /sp/				
	/gar/	/gar/ /ɔɪ/ /sil/ /sp/	/gar/	/gar/ /ɔɪ/ /ɔə/				

Table A.25: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp09

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp09	/v01/	/ð/ /ɛ/ /ɜ/ /g/ /k/ /l/ /m/ /n/ /p/	/v01/	/ɛ/ /ɜ/	/v01/	/ɔ/ /ŋ/ /ɒ/ /əʊ/	/v01/	/ʌ/ /əʊ/
			/v02/	/ɑ/ /æ/ /ɔ/ /ə/ /eh/ /ey/ /ɪ/ /iy/ /ɒ/ /ɒ/	/v02/	/ð/ /ɛ/ /ɜ/ /g/ /k/ /l/ /m/ /n/ /p/ /p/	/v02/	/ɑ/ /æ/ /ɔ/ /ə/ /eh/ /ey/ /ɪ/ /iy/ /ɒ/ /ɒ/
	/v02/	/ɪə/ /y/						
	/v03/	/ay/ /ɾ/ /s/ /ʃ/ /v/ /w/ /z/	/v03/	/ʊ/ /uw/	/v03/	/ay/ /r/ /s/ /ʃ/ /v/ /w/ /z/	/v03/	/k/ /l/ /m/ /n/ /p/ /r/ /s/ /ʃ/ /t/ /t/ /θ/ /z/
			/v04/	/ɔə/				
	/v04/	/æ/ /ʌ/ /ə/ /b/ /θ/	/v05/	/ɪə/	/sil/	/sil/ /sil/ /sp/		
			/v06/	/ɑʊ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɑʊ/ /ə/ /b/ /tʃ/ /d/ /eh/	/sil/	/sil/ /sil/ /sp/
	/v05/	/eh/ /ey/ /f/ /ɪ/	/v07/	/ʌ/ /əʊ/			/gar/	/gar/ /ɑʊ/ /ə/ /b/ /tʃ/
	/v06/	/ɔ/ /ŋ/ /ɒ/ /əʊ/	/v08/	/sil/		/eh/ /ey/ /f/ /ɦ/ /ɪə/		/ð/ /ɛ/ /ɜ/ /f/ /g/
	/v07/	/ɑ/	/v09/	/k/ /l/ /m/ /n/ /p/ /r/ /s/ /ʃ/ /t/		/ɪə/ /ɪ/ /dʒ/ /ɔɪ/ /θ/ /θ/ /ɔə/ /ʊ/ /uw/ /y/ /y/ /ʒ/		/g/ /ɦ/ /ɪə/ /dʒ/ /ɔɪ/ /ɔɪ/ /ɔə/ /ʊ/ /uw/ /v/ /v/ /w/ /y/ /ʒ/
	/v08/	/ɑ/						
	/v09/	/ʊ/ /uw/		/t/ /θ/ /z/				
	/v10/	/dʒ/	/v10/	/f/				
	/v11/	/tʃ/	/v11/	/d/ /ð/ /dʒ/				
	/v12/	/ʒ/	/v12/	/g/ /v/ /w/ /y/				
	/v13/	/ɔə/	/v13/	/b/				
	/v14/	/sil/	/v14/	/tʃ/ /ɦ/				
	/v15/	/ɦ/	/v15/	/ʒ/				
	/v16/	/ʊ/	/sil/	/sil/ /sil/ /sp/				
	/v17/	/a/ /a/	/gar/	/gar/ /ə/ /ɔɪ/				
	/gar/	/gar/ /ə/ /ɔɪ/ /sp/						

Table A.26: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp10

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp10	/v01/	/ɪ/ /iy/ /dʒ/ /l/ /ŋ/	/v01/	/ə/ /ay/ /eh/ /ɜ/ /ɪ/ /iy/ /ɒ/ /əʊ/	/v01/ /v02/	/ə/ /uw/ /v/ /w/ /fi/ /n/ /ɒ/ /əʊ/ /r/ /s/ /t/ /θ/ /sil/ /sil/ /sil/ /sp/ /gar/ /gar/ /ɑ/ /æ/ /ʌ/ /ɔ/ /ay/ /ə/ /b/ /tʃ/ /d/ /d/ /ð/ /ɛ/ /eh/ /ɜ/ /ɜ/ /ey/ /f/ /g/ /ɪə/ /ə/ /iy/ /dʒ/ /k/ /k/ /l/ /m/ /ŋ/ /ɔɪ/ /ɔɪ/ /ʃ/ /əə/ /ʊ/ /z/ /z/ /ʒ/	/v01/ /v02/ /v03/ /v04/ /sil/ /gar/	/æ/ /ʌ/ /ə/ /ʊ/ /ə/ /ay/ /eh/ /ɜ/ /ɪ/ /iy/ /ɒ/ /əʊ/ /d/ /ð/ /f/ /fi/ /l/ /m/ /n/ /p/ /r/ /r/ /s/ /t/ /v/ /w/ /w/ /z/ /b/ /tʃ/ /y/ /sil/ /sil/ /sp/ /gar/ /d/ /ɑʊ/ /ə/ /ɛ/ /ɪə/ /dʒ/ /ŋ/ /ɔɪ/ /ʃ/ /ʃ/ /θ/ /əə/ /uw/ /ʒ/ /ʒ/
		/ŋ/		/ɪ/ /iy/ /ɒ/ /əʊ/		/fi/ /n/ /ɒ/ /əʊ/		/ə/ /ay/ /eh/ /ɜ/
	/v02/	/fi/ /n/ /ɒ/ /əʊ/	/v02/	/æ/ /ʌ/ /ɔ/ /ʊ/		/r/ /s/ /t/ /θ/		/ɪ/ /iy/ /ɒ/ /əʊ/
		/r/ /s/ /t/ /θ/	/v03/	/əə/	/sil/	/sil/ /sil/ /sp/	/v03/	/d/ /ð/ /f/ /fi/
	/v03/	/b/	/v04/	/ɛ/ /uw/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/		/l/ /m/ /n/ /p/ /r/
	/v04/	/æ/ /d/ /ð/ /ɛ/	/v05/	/ɑ/ /ɪə/		/ay/ /ə/ /b/ /tʃ/ /d/		/r/ /s/ /t/ /v/ /w/
		/ey/ /f/	/v06/	/ɑʊ/		/d/ /ð/ /ɛ/ /eh/ /ɜ/		/w/ /z/
	/v05/	/k/	/v07/	/sil/		/ɜ/ /ey/ /f/ /g/ /ɪə/	/v04/	/b/ /tʃ/ /y/
	/v06/	/ə/ /uw/ /v/ /w/	/v08/	/ɔɪ/		/ɪə/ /ɪ/ /iy/ /dʒ/ /k/	/sil/	/sil/ /sil/ /sp/
	/v07/	/ay/ /ʃ/ /sil/	/v09/	/ə/		/k/ /l/ /m/ /ŋ/ /ɔɪ/	/gar/	/gar/ /d/ /ɑʊ/ /ə/ /ɛ/
	/v08/	/ʊ/	/v10/	/d/ /ð/ /f/ /fi/		/ɔɪ/ /ʃ/ /əə/ /ʊ/ /z/		/ɪə/ /dʒ/ /ŋ/ /ɔɪ/ /ʃ/
	/v09/	/ʌ/ /ɔ/ /z/		/l/ /m/ /n/ /p/ /r/		/z/ /ʒ/		/ʃ/ /θ/ /əə/ /uw/ /ʒ/
	/v10/	/ɪə/		/r/ /s/ /t/ /v/ /w/				/ʒ/
	/v11/	/tʃ/ /g/		/w/ /z/				
	/v12/	/ə/ /ɜ/	/v11/	/ʃ/				
	/v13/	/ɑ/ /ɑʊ/	/v12/	/g/ /dʒ/ /ŋ/				
	/v14/	/ʒ/	/v13/	/b/ /tʃ/ /y/				
	/v15/	/əə/	/v14/	/ʒ/				
	/v16/	/ɔɪ/	/v15/	/θ/				
	/gar/	/gar/ /sp/	/sil/	/sil/ /sil/ /sp/				

Table A.27: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp11

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp11	/v01/	/iy/ /k/ /m/ /n/ /ɒ/ /p/ /r/ /s/ /t/ /t/	/v01/ /v02/	/uw/ /æ/ /ə/ /ay/ /ɛ/ /ɜ/ /ey/ /ɪ/ /iy/	/v01/ /v02/	/ɔ/ /ə/ /ay/ /tʃ/ /ey/ /d/ /ð/ /f/	/v01/ /v02/	/æ/ /ə/ /ay/ /ɛ/ /ɜ/ /ey/ /ɪ/ /iy/ /dʒ/ /k/ /l/ /m/
	/v02/	/v/	/v03/	/ɑ/	/v03/	/iy/ /k/ /m/ /n/ /ɒ/ /p/ /r/ /s/ /t/		/ŋ/ /p/ /r/ /s/ /t/ /t/ /w/
	/v03/	/ɔ/ /ə/ /ay/ /tʃ/ /ey/	/v04/ /v05/	/ʌ/ /ɔ/ /əʊ/ /ɒ/		/t/	/sil/	/sil/ /sil/ /sp/
	/v04/	/d/ /ð/ /f/	/v06/	/ʊ/	/sil/	/sil/ /sil/ /sp/	/gar/	/gar/ /ɑ/ /ʌ/ /ɔ/ /ɑʊ/
	/v05/	/w/	/v07/	/əə/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɑʊ/		/b/ /tʃ/ /d/ /ð/ /f/
	/v06/	/ʃ/	/v08/	/sil/		/b/ /ɛ/ /ɜ/ /g/ /fi/		/f/ /g/ /fi/ /ɪə/ /ɒ/
	/v07/	/d/ /æ/ /ʌ/ /b/	/v09/	/ɔɪ/		/fi/ /ɪə/ /ɪ/ /dʒ/ /l/		/ɒ/ /əʊ/ /ɔɪ/ /ʃ/ /θ/
	/v08/	/ɑʊ/ /ɛ/ /ɜ/ /ɪ/	/v10/	/ɪə/		/l/ /əʊ/ /ɔɪ/ /ʃ/ /θ/		/θ/ /əə/ /ʊ/ /uw/ /v/
	/v09/	/θ/ /əə/	/v11/	/ɑʊ/		/θ/ /əə/ /ʊ/ /uw/ /v/		/v/ /y/ /z/ /ʒ/
	/v10/	/əʊ/	/v12/	/dʒ/ /k/ /l/ /m/ /ŋ/ /p/ /r/ /s/ /t/		/v/ /w/ /y/ /z/ /ʒ/		
	/v11/	/g/ /y/ /z/				/ʒ/		
	/v12/	/fi/ /l/		/t/ /w/				
	/v13/	/ɪə/ /uw/	/v13/	/d/ /f/ /g/ /fi/				
	/v14/	/ʒ/	/v14/	/ʃ/				
	/v15/	/ʊ/	/v15/	/y/ /z/				
	/v16/	/sil/	/v16/	/ð/ /θ/ /v/				
	/v17/	/dʒ/	/v17/	/tʃ/				
	/v18/	/ə/	/v18/	/ʒ/				
	/gar/	/gar/ /ɔɪ/ /sp/	/v19/	/b/				
			/sil/	/sil/ /sil/ /sp/				
			/gar/	/gar/ /ə/				

Table A.28: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp13

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp13	/v01/	/ɔ/ /d/ /ɪ/ /k/ /n/ /p/ /s/ /uw/ /v/ /ɤ/ /z/ /ʒ/	/v01/	/æ/ /ɔ/ /ə/ /ay/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /iy/	/v01/	/ɔ/ /d/ /ɪ/ /k/ /n/ /p/ /s/ /uw/ /v/ /v/ /z/ /ʒ/	/v01/	/æ/ /ɔ/ /ə/ /ay/ /ɜ/ /ey/ /ɪə/ /ɪ/ /iy/ /iy/
	/v02/	/ɪə/	/v02/	/ɛ/ /ʊ/ /əʊ/ /uw/	/sil/	/sil/ /sil/ /sp/	/v02/	/d/ /f/ /g/ /k/
	/v03/	/ɜ/ /f/ /g/ /r/	/v03/	/ɑʊ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɑʊ/		/m/ /n/ /ŋ/ /p/ /s/
	/v04/	/b/ /ð/ /ɛ/ /eh/	/v04/	/ʌ/ /ʊ/		/ay/ /ə/ /b/ /tʃ/ /ð/		/ʃ/ /t/ /v/ /w/ /z/
	/v05/	/tʃ/	/v05/	/ɑ/ /ɔə/		/ð/ /ɛ/ /eh/ /ɜ/ /ey/		/z/
	/v06/	/ɑʊ/ /iy/ /ɒ/ /əʊ/	/v06/	/sil/		/ey/ /f/ /g/ /ɦ/ /ɪə/	/sil/	/sil/ /sil/ /sp/
	/v07/	/ə/ /ʊ/	/v07/	/ə/		/ɪə/ /iy/ /dʒ/ /m/ /ŋ/	/gar/	/gar/ /ɑ/ /ʌ/ /ɑʊ/ /ə/
	/v08/	/æ/ /ʌ/ /ə/ /ay/	/v08/	/dʒ/ /r/ /ʃ/ /ɣ/		/ŋ/ /ɒ/ /əʊ/ /ɔɪ/ /r/		/tʃ/ /ð/ /ɛ/ /ɦ/ /dʒ/
	/v09/	/ɑ/ /ɣ/	/v09/	/d/ /f/ /g/ /k/		/r/ /ʃ/ /t/ /θ/ /ɔə/		/dʒ/ /ɒ/ /əʊ/ /ɔɪ/ /r/
	/v10/	/m/ /sil/ /t/ /θ/		/m/ /n/ /ŋ/ /p/ /s/		/ɔə/ /ʊ/ /w/ /ɣ/		/r/ /ʃ/ /θ/ /ɔə/ /ʊ/
	/v11/	/ʃ/		/s/ /t/ /v/ /w/ /z/				/ʊ/ /uw/ /ɣ/ /ʒ/
	/v12/	/ey/		/z/				
	/v13/	/ɔə/ /w/	/v10/	/ɦ/				
	/v14/	/ŋ/	/v11/	/b/ /tʃ/ /ð/				
	/gar/	/gar/ /ɔɪ/ /sp/	/v12/	/ʒ/				
			/v13/	/θ/				
			/sil/	/sil/ /sil/ /sp/				
			/gar/	/gar/ /ɔɪ/				

Table A.29: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp14

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp14	/v01/	/tʃ/ /iy/ /dʒ/ /m/ /əʊ/ /p/ /r/ /s/ /t/ /t/ /θ/	/v01/	/æ/ /ɔ/ /ə/ /ay/ /eh/ /ɜ/ /ey/ /ɪ/ /iy/ /iy/	/v01/ /v02/ /v03/	/æ/ /ɜ/ /ey/ /f/ /ʃ/ /v/ /w/ /y/ /tʃ/ /iy/ /dʒ/ /m/ /əʊ/ /p/ /r/ /s/ /t/ /t/ /θ/	/v01/	/æ/ /ɔ/ /ə/ /ay/ /eh/ /ɜ/ /ey/ /ɪ/ /iy/ /iy/
	/v02/	/ə/ /ay/ /ŋ/	/v02/	/uw/			/v02/	/ð/ /f/ /ɦ/ /k/
	/v03/	/ɔ/ /b/ /d/ /ð/ /l/	/v03/	/ʊ/				/m/ /n/ /r/ /s/ /ʃ/
	/v04/	/ʃ/ /v/ /w/ /y/	/v04/	/ɪə/ /ɒ/ /əʊ/	/v04/	/ɔ/ /b/ /d/ /ð/ /l/		/ʃ/ /t/ /v/ /w/
	/v05/	/g/ /ɦ/ /k/	/v05/	/ʌ/ /sil/			/sil/	/sil/ /sil/ /sp/
	/v06/	/ɛ/ /ʊ/	/v06/	/ɑʊ/	/sil/	/sil/ /sil/ /sp/	/gar/	/gar/ /ɑ/ /ʌ/ /ɑʊ/ /ə/
	/v07/	/æ/ /ɜ/ /ey/ /f/	/v07/	/ɑ/	/gar/	/gar/ /ɑ/ /ʌ/ /ɑʊ/ /ə/		/tʃ/ /d/ /g/ /ɪə/ /dʒ/
	/v08/	/ɑ/ /uw/	/v08/	/ɑ/		/ə/ /ɛ/ /g/ /ɦ/ /ɪə/		/dʒ/ /ŋ/ /ɒ/ /əʊ/ /ɔɪ/
	/v09/	/ɪə/	/v09/	/ɔə/		/ɪə/ /k/ /ŋ/ /ɒ/ /ɔɪ/		/ɔɪ/ /p/ /θ/ /ɔə/ /ʊ/
	/v10/	/ʌ/ /ɒ/	/v10/	/a/ /a/		/ɔɪ/ /ɔə/ /ʊ/ /uw/ /ʒ/		/ʊ/ /uw/ /ɣ/ /z/ /ʒ/
	/v11/	/ɪə/	/v11/	/ð/ /f/ /ɦ/ /k/ /m/ /n/ /r/ /s/ /ʃ/		/ʒ/		/ʒ/
	/v12/	/ʒ/		/ʃ/ /t/ /v/ /w/				
	/v13/	/ɔə/	/v12/	/z/				
	/v14/	/sil/	/v13/	/ɣ/				
	/v15/	/ɑʊ/	/v14/	/b/ /tʃ/ /d/ /θ/				
	/v16/	/i/ /a/	/v15/	/p/				
	/gar/	/gar/ /ə/ /ɔɪ/ /sp/	/v16/	/g/				
			/v17/	/dʒ/ /ŋ/				
			/v18/	/ʒ/				
			/sil/	/sil/ /sil/ /sp/				
			/gar/	/gar/ /ə/ /ɔɪ/				

Table A.30: A speaker-dependent phoneme-to-viseme mapping derived from phoneme recognition confusions for RMAV speaker sp15

Speaker	Bear1		Bear2		Bear3		Bear4	
	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes	Viseme	Phonemes
sp15	/v01/	/ə/ /d/ /ð/ /ey/ /ɪ/ /iy/ /k/ /l/ /m/ /n/ /y/	/v01/	/ə/ /ay/ /eh/ /ey/ /iy/ /əʊ/ /uw/	/v01/	/ə/ /d/ /ð/ /ey/ /ɪ/ /iy/ /k/ /l/ /m/ /n/ /y/	/v01/	/ə/ /ay/ /eh/ /ey/ /iy/ /əʊ/ /uw/
	/v02/	/ɚ/ /p/ /r/ /s/ /t/ /θ/ /z/	/v02/	/ʌ/ /ɔ/ /ɑʊ/ /ɛ/ /ɔɪ/	/sil/	/sil/ /sil/ /sp/	/v02/	/b/ /d/ /ð/ /f/ /k/ /l/ /m/ /n/ /ŋ/
	/v03/	/eh/ /əʊ/	/v03/	/ɒ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/		/ŋ/ /p/ /v/
	/v04/	/ɑ/ /æ/ /ʌ/ /ɔ/	/v04/	/ɑ/ /æ/ /ɜ/		/ay/ /ə/ /b/ /tʃ/ /ɛ/	/sil/	/sil/ /sil/ /sp/
	/v05/	/ɒ/	/v05/	/sil/ /ɔə/		/ɛ/ /eh/ /ɜ/ /g/ /ɦ/	/gar/	/gar/ /ɑ/ /æ/ /ʌ/ /ɔ/
	/v06/	/ŋ/ /uw/ /v/	/v06/	/ʊ/		/ɦ/ /ɚ/ /dʒ/ /ŋ/ /ɒ/		/ə/ /tʃ/ /ɛ/ /ɜ/ /ɦ/
	/v07/	/ʊ/	/v07/	/ə/		/ɒ/ /əʊ/ /ɔɪ/ /p/ /r/		/ɦ/ /ɚ/ /dʒ/ /ɒ/ /ɔɪ/
	/v08/	/g/ /ɦ/ /dʒ/	/v08/	/b/ /d/ /ð/ /f/ /k/ /l/ /m/ /n/ /ŋ/		/r/ /s/ /ʃ/ /t/ /θ/ /θ/ /ɔə/ /ʊ/ /uw/ /v/		/ɔɪ/ /r/ /s/ /ʃ/ /t/ /t/ /θ/ /ɔə/ /ʊ/ /w/
	/v09/	/ɜ/		/ŋ/ /p/ /v/		/v/ /w/ /z/ /ʒ/		/w/ /y/ /z/ /ʒ/
	/v10/	/b/ /tʃ/	/v09/	/r/ /s/ /ʃ/ /t/ /æ/				
	/v11/	/ɜ/						
	/v12/	/ay/ /ɛ/	/v10/	/dʒ/				
	/v13/	/sil/ /ɔə/	/v11/	/ʒ/				
	/v14/	/ɑʊ/ /ɔɪ/	/v12/	/w/ /y/				
	/v15/	/ʒ/	/v13/	/ɦ/				
	/v16/	/e/ /r/	/v14/	/tʃ/				
	/gar/	/gar/ /ə/ /sp/	/sil/	/sil/ /sil/ /sp/				