Finding phonemes: improving machine lip-reading

Helen L. Bear[1]*, Richard W. Harvey*[1]*, Yuxuan Lan*[1]

[1]University of East Anglia, UK

{helen.bear,r.w.harvey,y.lan}@uea.ac.uk

Abstract

In machine lip-reading there is continued debate and research around the correct classes to be used for recognition.

In this paper we use a structured approach for devising speaker-dependent viseme classes, which enables the creation of a set of phoneme-to-viseme maps where each has a different quantity of visemes ranging from two to 45. Viseme classes are based upon the mapping of articulated phonemes, which have been confused during phoneme recognition, into viseme groups.

Using these maps, with the LiLIR dataset, we show the effect of changing the viseme map size in speaker-dependent machine lip-reading, measured by word recognition correctness and so demonstrate that word recognition with phoneme classifiers is not just possible, but often better than word recognition with viseme classifiers. Furthermore, there are intermediate units between visemes and phonemes which are better still.
Index Terms: visual-only speech recognition, computer lip-reading, visemes, classification, pattern recognition

1. Introduction

Although visemes are yet to be formally defined, many possibilities can be found across literature [1, 2, 3, 4]. Here we use the definition "a viseme is a visual cue representative of a subset of phonemes on the lips". Therefore, a set of viseme classifiers is inherently smaller than a set of phoneme classifiers. Whilst this means that there are more training samples per class (addressing the limitation of currently available dataset sizes), this also introduces generalisation between articulated sounds. So, to find optimal viseme classes, we need to minimise this generalisation in order to maximise recognition of correct utterances, but also maximise the use of the data available.

The relationship between phonemes (the units of acoustic speech) and visemes (the units of visual speech) can be described with Phoneme-to-Viseme (P2V) maps. In [1] it is shown how these maps can be derived automatically from phoneme confusions. A by-product of the method is that we can control how many visemes we need. This allows considerable precision when answering questions about the optimal number and nature of visemes.

2. Data

Our selected dataset is LiLIR [5]. This data consists of 12 British speakers (seven male and five female), 200 utterances per speaker of resource management context independent sentences from [6] which totals around 1000 words. The original videos were recorded in high definition and in a full-frontal position. Individual speakers are tracked using Active Appearance Models [7] and we extract features of concatenated shape and appearance information.

The pronunciation dictionary used throughout these experiments is British English [8] which we take to be represented by 46 phonemes.

3. Method

A high level overview of our method is shown in Figure 1 and is described in [1]. We begin by performing word recognition using classifiers based upon phoneme labels. This provides us with both a baseline to benchmark against and, crucially, a set of confusion matrices for each speaker which are used to cluster together potential monophones.

However, we undertake a different clustering process (section 3.2) during which we make a new P2V mapping each time a phoneme is re-classified to a new viseme grouping, thereby deriving up to 45 (subject to the number of phonemes recognised during the phoneme recognition stage) P2V maps per speaker. These new classifiers (visemes) are then used to repeat our word recognition task.

We use the word recognition as our performance measure as this normalises for variance in training samples for each set of classifiers. We note that it is not the performance itself which is relevant here, rather it is any improvement a variance in classes can provide. The reader should also note that we are not suggesting our clustering process will deliver the optimum visemes but rather address our need in this case for a method to enable a controlled comparison of the visemes.

3.1. Step one: phoneme recognition

We implement 10-fold cross-validation with replacement [9], of 200 sentences per speaker, 20 are randomly selected as test samples and these are not included in the training folds. Using the HTK toolkit [10] to use Hidden Markov Model (HMM) classes, we flat-start the HMMs, re-estimate them 11 times with forced alignment between seventh and eighth estimates. Our prototype is based upon a Gaussian mixture of five components and three state HMMs. We use a single-state tied short-pause, or 'sp' HMM for short silences between words in the sentence utterances. We also use a bigram word network to support recognition. There are a maximum of 46 phonemes within our phoneme recognition results, but not all speakers used all phonemes within their speech utterances.

3.2. Step two: speaker-dependent phoneme clustering

We cluster the phonemes into new visemes classes as follows; we have 10 confusion matrices for each speaker (one from each fold), these are summed together to form one confusion matrix representing all confusions for that speaker. We start with this phoneme confusion matrix:

$$[K_m]_{ij} = N(\hat{p}_j|p_i) \quad (1)$$

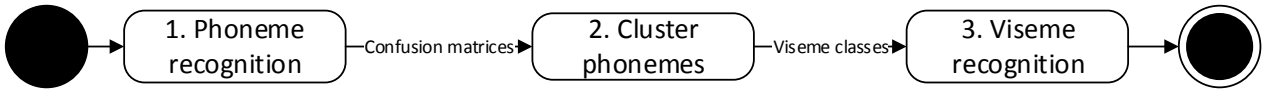


Figure 1: *Three step process for word recognition from visemes.*

Viseme	Phonemes
V01	/ax/
V02	/v/
V03	/oy/
V04	/f/ /zh/ /w/
V05	/k/ /b/ /d/ /th/ /p/
V06	/l/ /jh/
V07	/g/ /m/ /z/ /y/ /ch/ /dh/ /s/ /r/ /t/ /sh/
V08	/n/ /hh/ /ng/
V09	/ea/ /ae/ /ao/ /uw/ /oh/ /ia/ /ey/ /ua/ /er/
V10	/ay/ /aa/ /ah/ /aw/ /uh/ /ow/ /ih/ /iy/ /az/ /eh/

Table 1: An example P2V map, this is the P2V for Speaker 01 with ten visemes

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

$$[P_m]_{ij} = Pr\{p_i|\hat{p}_j\} \tag{2}$$

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

Specifically for each possible merged pair, Pr, Ps, we calculate a score:

$$q = [P_m]_{rs} + [P_m]_{sr} = Pr\{\hat{Pr}|Ps\} + Pr\{Pr|\hat{Ps}\} \tag{3}$$

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

This is a more formal approach than used in [1] and incorporates their conclusions that vowel and consonant phonemes should not be clustered together when devising phoneme-to-viseme mappings. An example P2V mapping is shown in Table 1.

3.3. Step three: viseme recognition

Similar to Step one, we implement 10-fold cross-validation with replacement [9], of 200 sentences per speaker, 20 are randomly selected as test samples and these are not included in the training folds. Using the HTK toolkit [10] to use Hidden Markov Model (HMM) classes, we flat-start the HMMs, re-estimate them 16 times over with forced alignment between seventh and eighth estimates.

Our prototype is based upon a gaussian mixture of five components and three state HMMs. We use a single-state tied short-pause, or 'sp' HMM for short silences between words in the sentence utterances. We also use a bigram word network to support recognition, apply a grammar scale factor of 1.0 (shown to be optimum in Howell's thesis [11]) and apply a transition penalty of 0.5.

This time around we have viseme classes to use as recognizers. By using these sets of classes which have shown in step one are confusing on the lips, we perform recognition for each class set. In total this is 45, where the smallest set is of two classes (one with all the vowel phonemes and the other all the consonant phonemes), and the largest set is of 45 classes with one phoneme in each - a repeat of the phoneme recognition task but using only phonemes which we know to have been identifiable.

4. Discussion

We note that word recognition performance of the HMMs can be measured by both correctness, C, and accuracy, A, of the recognition classes,

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

$$A = \frac{C - I}{N}, \tag{5}$$

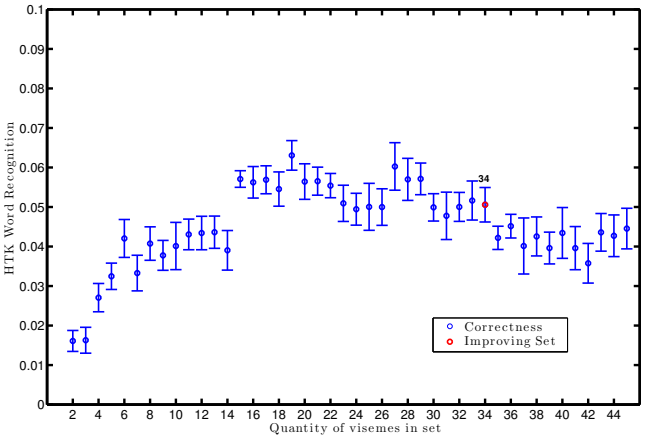
where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors and N the total number of labels in the reference transcriptions [10].

Figure 2 (subfigures a-l), show the correctness for all 12 speakers. Viseme sets containing fewer visemes produce more viseme strings that represent more than one word: homophones. An example of a homophone in these data are the words 'port' and 'bass'. Using Speaker 1's 10-viseme P2V map these both become 'v5 v9 v7' i.e. a single identifier for identifying two words. Thus distinguishing between 'port' and 'bass' becomes impossible. The effect of these can be seen on the left side of the graphs in Figure 2.

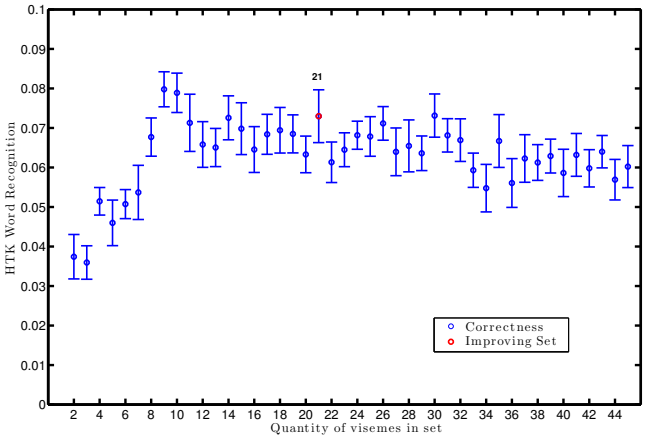
Although the correctness scores are low they are all significantly above chance. The results for each speaker vary but the overall trend is very clear. Superior performances are to be found with larger numbers of visemes. Note that, had we reported viseme error (as is commonplace) then this effect is not visible and the imperative for large numbers of visemes would be missed.

Also in Figure 2 (subfigures a-l), class sets are highlighted in red and labelled which show where a particular combination of two previous viseme classes delivers a significant improvement in recognition. These combinations are listed in Table 2. Whilst there is no apparent pattern through these pairings, this does further reinforce our knowledge that all speakers are visually unique and how difficult finding a set of cross-talker viseme sets will be when different phonemes require alternative grouping arrangements for each individual.

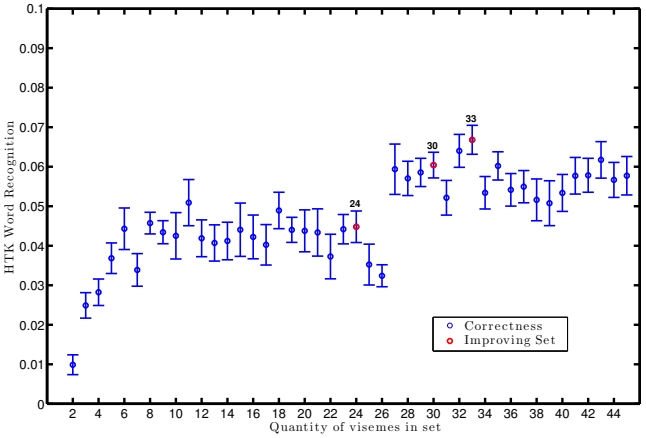
As has been noted before [12] the conventional wisdom which is that visemes are needed for lip-reading is not bourne out by these experiments. However it is an over simplification



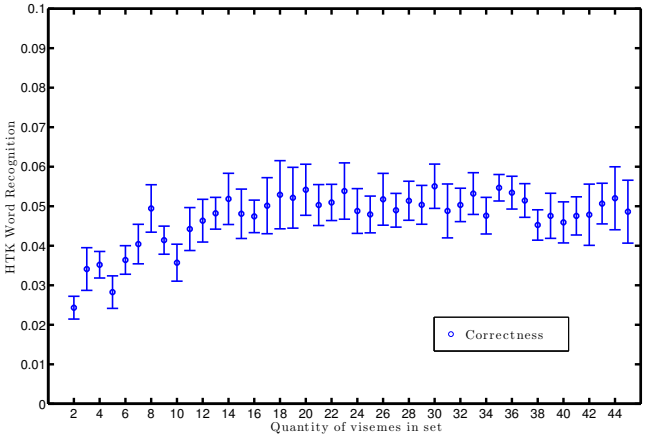
(a) Speaker 1



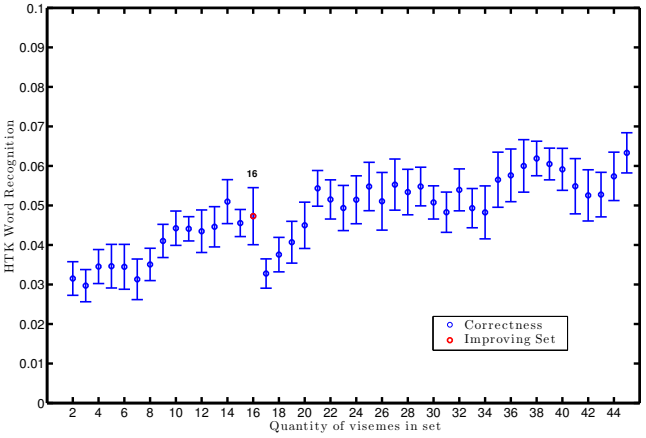
(b) Speaker 2



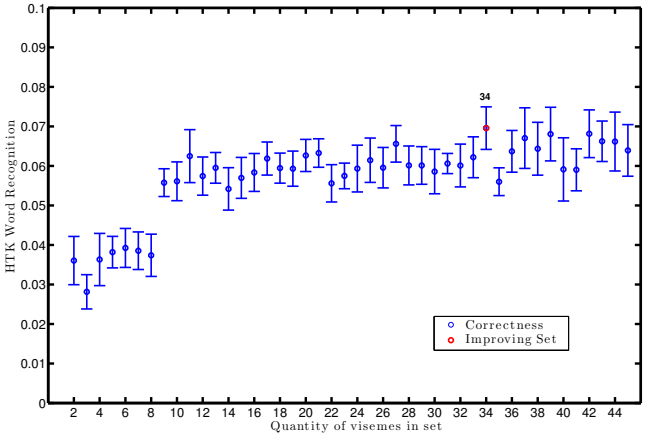
(c) Speaker 3



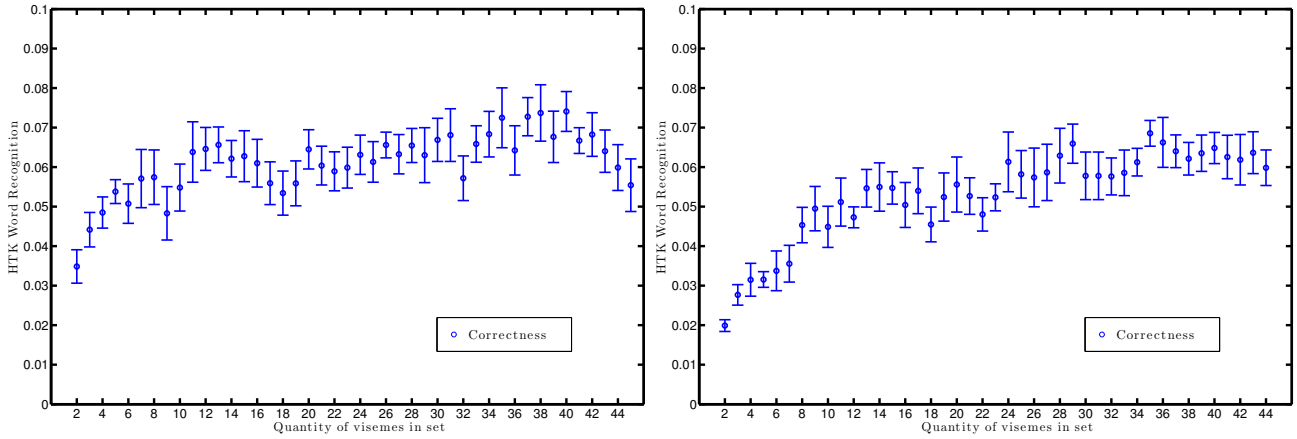
(d) Speaker 4



(e) Speaker 5

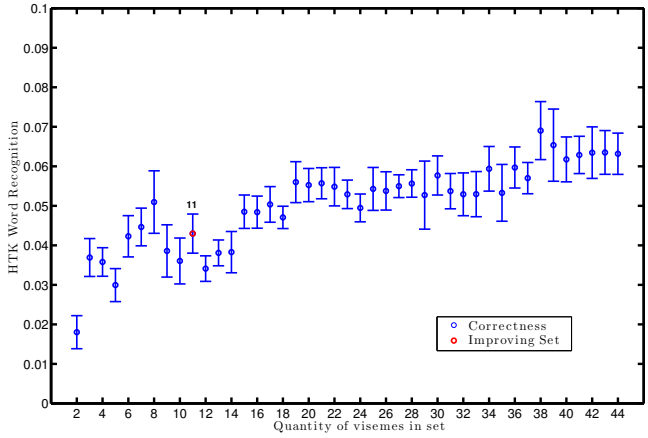


(f) Speaker 6

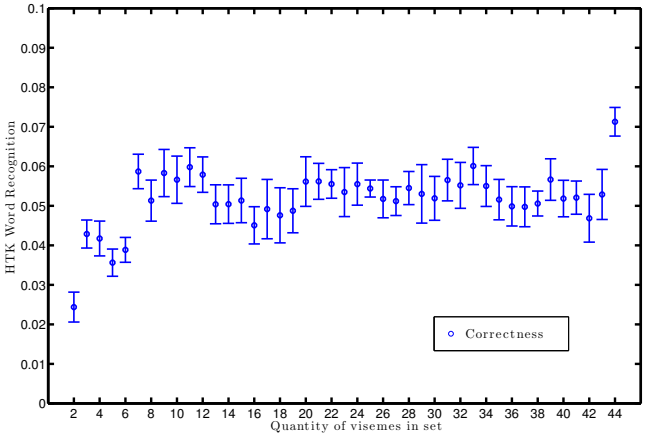


(g) Speaker 7

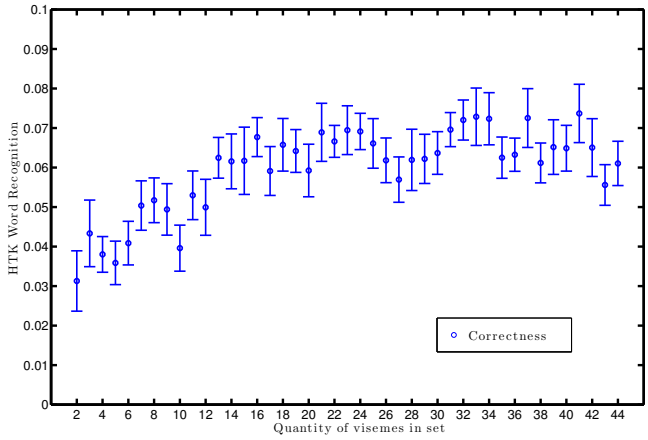
(h) Speaker 8



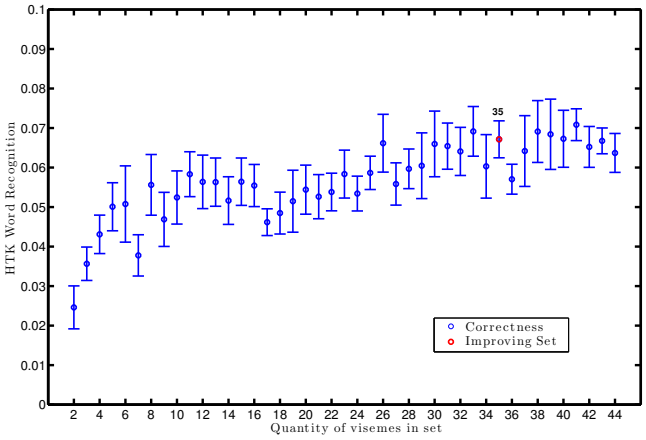
(i) Speaker 9



(j) Speaker 10



(k) Speaker 11



(l) Speaker 12

Figure 2: *Individual speaker word recognition in correctness C for all viseme map sizes*

Speaker	Set No	V_i	V_j	Set No	V_n
SP01	35	/s/ /r/	/dh/	34	/s/ /r/ /dh/
SP02	22	/d/	/z/ /y/	21	/d/ /z/ /y/
SP03	34	/b/ /ch/	/zh/	33	/b/ /ch/ /zh/
SP03	31	/zh/ /b/ /ch/	/z/	30	/zh/ /b/ /ch/ /z/
SP03	25	/p/ /r/	/ng/	24	/p/ /r/ /ng/
SP05	17	/ae/	/eh/	16	/ae/ /eh/
SP06	35	/ae/ /ah/	/iy/	34	/ae/ /ah/ /iy/
SP09	12	/b/ /w/ /v/	/jh/ /hh/	11	/b/ /w/ /v/ /jh/ /hh/
SP12	36	/ah/	/ao/	34	/ah/ /ao/

Table 2: Viseme class merges which improve word recognition

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

Table 3: Phoneme correctness values for each speaker, these are on the right hand side of each respective subfigure in Figure 2

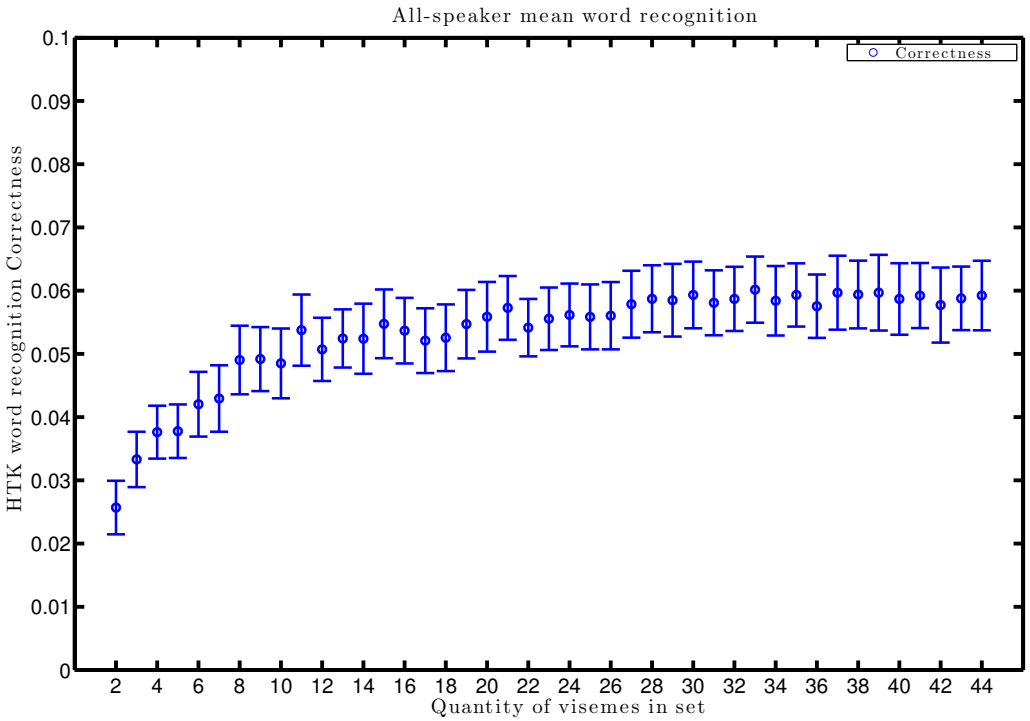


Figure 3: *Word recognition measured by correctness of the classifiers. Error bars show ± one standard error.*

to assert that better lip-reading can be achieved with phonemes than visemes. It is true that, generally speaking, larger numbers of visemes out-perform smaller numbers, but the curves in Figure 2 are far from monotonic. Even Figure 3, which is the mean performance over all speakers, is not monotonic.

There are a number of proposed phoneme-to-viseme maps in the literature, typically they generate between 10 and 20 visemes (see [1] for a summary) - the well known Lee set has six consonant visemes and five vowels [13]; Jeffers eight & three [14] and so on. Looking at Figures 2 & 3 there is certainly a rapid drop-off in performance for fewer than ten visemes but the region between ten and 20 contains the optimum viseme set for three out of the 12 speakers which is no more than chance. In other words, for each speaker there is an optimal number of visual units (shown by the best performing result in Figure 2) but that optimal number is not related to any of the conventional viseme definitions, nor is the number of phonemes. The correctness of the phoneme recognition for each speaker is shown in Table 3.

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

5. Conclusions

We have described a method that allows us to construct any number of visual units. We remind the reader that we are not proposing that our visemes are the best, our priority in this case is a method for enabling comparison of viseme sets in a controlled manner.

The presence of an optimum is a result of two competing effects. In the first, as the number of visemes shrinks the number of homophones rises and it becomes more difficult to recognise words (correctness drops). In the second, as the number of visemes rises we run out of training data to learn the subtle differences in lip-shapes (if they exist), so again, correctness drops.

Thus, the optimum number of visual units lies beween one and 45. In practice we see this optimum is between the number of phonemes and eight (which is the size of one of the smaller viseme sets).

For future work we are interested to extend these methods to work across speakers with a view to identify combinations of phonemes which can improve more than an single speaker.

6. References

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

[2] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

[3] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language and Hearing Research*, vol. 11, no. 4, p. 796, 1968.

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

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

[6] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[7] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004. [Online]. Available: http://www.springerlink.com/openurl.asp?

[8] Cambridge University, UK. (1997) BEEP pronounciation dictionary. [Online]. Available: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

[9] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

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

[11] D. L. Howell, *Confusion Modelling for Lip-Reading. PhD thesis*. University of East Anglia, 2014.

[12] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1082–1089, 2006.

[13] S. Lee and D. Yook, "Audio-to-visual conversion using hidden markov models," in *PRICAI 2002: Trends in Artificial Intelligence*. Springer, 2002, pp. 563–570.

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