DECODING VISEMES: IMPROVING MACHINE LIP-READING

Helen L. Bear and Richard Harvey

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

ABSTRACT

To undertake machine lip-reading, we try to recognise speech from a visual signal. Current work often uses viseme classification supported by language models with varying degrees of success. A few recent works suggest phoneme classification, in the right circumstances, can outperform viseme classification. In this work we present a novel two-pass method of training phoneme classifiers which uses previously trained visemes in the first pass. With our new training algorithm, we show classification performance which significantly improves on previous lip-reading results.

***Index Terms*—** visemes, weak learning, visual speech, lip-reading, recognition, classification

1. INTRODUCTION

In machine lip-reading, the classification of an utterance from a visual-only signal, there are many obstacles to overcome. Some, such as pose [1, 2], motion [3, 4] and resolution [5] have been studied and measured, including the selection of a phoneme-to-viseme mapping [6, 7]. However, visemes are not precisely defined. Many working definitions have been offered such as; "A set of phonemes that have identical appearance on the lips" [7] or "A visual equivalent of a phoneme" [8]. However, there are challenges with using viseme labelled classifiers including: the homophone effect, not enough training data per class, and the consequential lack of differentiation between classes when there are too many visemes within a set. More recently, there is evidence that viseme labels may not be needed at all because with enough data, classifiers based on phoneme labels can outperform viseme classification [9, 10]. As phonemes are well studied, this idea is attractive. However, others have tested small numbers of visual units: visemes and found they also give acceptable results [11, 12]. It would be very helpful to be able to systematically vary the number visual units and hence devise optimal strategies for learning.

The rest of this paper is as follows; a summary of the analysis into the effect of varying the quantity of visemes in a set on lip-reading performance presented in [13] is followed by a short test on unit selection effects between classifier and its supporting network, the results of these are used to introduce the hypothesis for applying weak learning during classifier training. A full description of the experimental setup to test the hypothesis is included before analysis of results and conclusions.

2. BACKGROUND

A systematic study into varying the number of visemes was conducted in [13] which generated viseme sets of varying size. HTK [14] was used to build Hidden Markov Model (HMM) classifiers for every viseme in each set. We initialised a set of HMMs (`HCompV`), that were trained (and retrained) using `HERest` during which there were options to tie any required model states together (e.g. for short pause models) (`HHed`) or to force align the HMMs to a time-aligned ground truth (`HVite`) before producing a classification output. The output of classification was supported by a word bigram model created with `HBuild` and `HLStats`. Finally, this classification output was compared to the ground truth to measure its efficacy (`HResults`) which we measured using Correctness, C.

$$C = \frac{N - D - S}{N}, \tag{1}$$

where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors and N the total number of labels in the reference transcriptions [14].

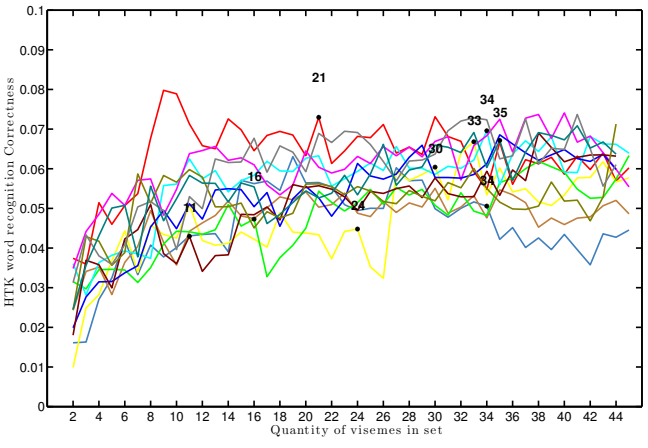


Fig. 1: Viseme correctness as the quantity of visemes changes in a set of classifiers for 12 LiLiR speakers. Results from [13].

Figure 1 shows our previous results [13], derived using the algorithm described in [7]. The algorithm works by merging visemes. For example, a label set with 44 visemes has been obtained from the label set of 45 visemes. At each merging stage we measure the difference in correctness compared to the previous set. Significant differences in Figure 1 are shown with black dots where the number represents the size of the significant set.

In Figure 1 the performance of classifiers with few visemes is poor due to the large number of homophones. An example of a homophone in the data are the words "port" and "bass". Using Speaker 1's 10-viseme P2V map these both become '$/v5/$ $/v9/$ $/v7/$' i.e. a single identifier for identifying two distinct words. Thus distinguishing between "port" and "bass" is impossible. Large numbers of visemes do not appear to further improve the correctness, probably because, as has been observed before, many phonemes look similar on the lips [15]. Looking at Figure 1 there appears to be a sweet spot where optimality might be found between visemes set sizes from 11 to 36.

3. DATA

For comparable experiments, we select the same 12 speakers from the dataset [16] presented in [13]. For the seven male and five female speakers, each utters 200 sentences from [15]. Individual speakers were tracked using Active Appearance Models (AAMs) [17] and the extracted features consist of concatenated shape and appearance information representing only the mouth area of the face.

4. METHOD

In previous work, we essentially examined two different algorithms. In the first, the data were labelled with phonemes, we use `HCompV` to initialise the phoneme classifiers, and 11 repetitions of `HERest` to train the classifiers. This system had the advantage that the output was a sequence of phonemes, but the disadvantage that phoneme models are hard to train. The alternative was to use a smaller number of visemes. The data were labelled with the visemes, and we learned the viseme classifiers in the same way, `HCompV` followed by `HERest`. Our new method is a hybrid. We initially learn the visemes, these trained visemes then become the starting point phoneme classifiers (we know the mapping from the visemes to the phonemes for all sets of visemes). We now train the the phoneme models via repeated applications of `HERest`, thus we have obtained phoneme models but with a new initialisation based upon what was learned for the visemes. This process is illustrated in Figure 2. In this example $p1$, $p2$ and $p4$ are associated with $v1$, so are initialised as replicas of HMM $v1$. Likewise $p3$ and $p5$ are initialised as replicas of $v2$. We now retrain the phoneme models using the same training data.

In full; we initialise *viseme* HMMs with `HCompV`. Our prototype HMM is based upon a Gaussian mixture of five components and three states [18]. These are re-estimated 11 times over with `HERest`, including both short pause model state tying (between re-estimates 3 & 4 with `HHed`), and forced alignment between re-estimates 7 & 8 with `HVite`. This is steps 1 & 2 in Figure 2. But before classification, these viseme HMM definitions are used as initialised definitions for phoneme labelled HMMs (Figure 2 step 3). The respective viseme HMM definition is used for all the phonemes in its relative phoneme-to-viseme mapping. These phoneme HMMs are retrained and used for classification. This amendment to training is analogous with weak learning. We complete classification twice. First with a phoneme bigram network, second with a word bigram network. For both we apply a grammar scale factor of 1.0 and a transition penalty of 0.5 (based on [9]) with `HVite`. This is implemented using 10-fold cross-validation with replacement [19].

The advantage of our new training approach is that the phoneme classifiers have seen only positive cases therefore have good mode matching, the disadvantage is they are not exposed to negative cases to the same degree as the visemes.

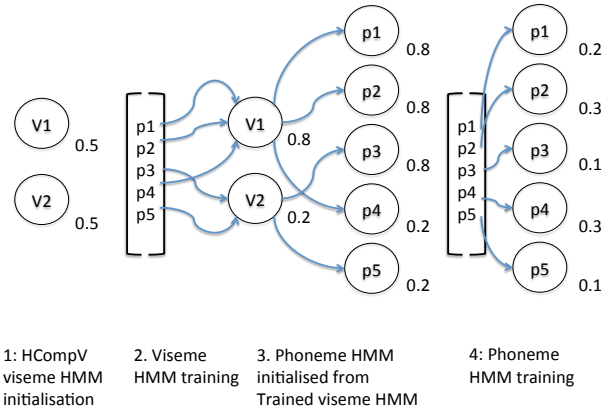


Fig. 2: Weak learning of visemes to initialise phoneme labelled classifiers.

4.1. Language network units

The systems under study in this paper have two components. The first component, the classifier takes the raw data and attempts to estimate a probable string of units. The second component, the language model, modifies that string on the basis of knowledge of how the units are co-located in the training data. In practice of course, these two components work together and there is no intermediate uncorrected string.

Here we are considering the problem of what the classification unit should be: a viseme? A phoneme? Or a word? But we also must consider how the language model should work. Should we use n-grams of phonemes? Visemes? Or words?

The further confusion is the unit on which we measure correctness. It is possible, for example, to build a word classifier followed by a bigram word network measured in terms of its viseme correctness. Such a system would be bizarre but is none-the-less possible. Table 1 shows some of the more sensible possibilities. The first row of Table 1 is a viseme classifier followed by a viseme bigram network with a viseme correctness of 0.0231. In Table 1 correctness is always measured by the units of the classifier. The dashed lines group different correctness units. The top group show viseme correctness which can be compared against each other, the second group show phoneme correctness and the bottom, word correctness.

In our data we have a large vocabulary (approximately 1000 words), so we eliminate word level classifiers as impractical. This leaves us with viseme classifiers for which the viseme word network is the worst performing so we do not consider this option either. For convenience the same data are plotted in Figure 3 with error bars of one standard error.

Table 1: Unit selection pairs for HMMs & language networks.

Classifier units	Network units	Classifier unit, C
Viseme	Viseme	0.0231
Viseme	Phoneme	0.1914
Viseme	Word	0.0851
Phoneme	Phoneme	0.1980
Phoneme	Word	0.1980
Word	Word	0.1874

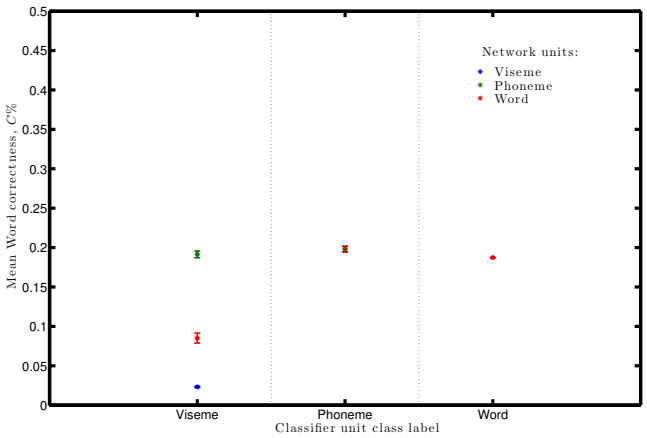


Fig. 3: Effects of support network unit choice with varying HMM classifier units measured in all speaker mean correctness, C.

5. RESULTS

Figure 4 shows the mean speaker-dependent correctness. We examine two configurations, one is phoneme classification where we measure phoneme correctness. These are the top two data series in Figure 4 (in green and pink), and the other is word classification where we measure word correctness. These are the lower two data series in Figure 4 in blue and red. Word correctness guessing is duplicated from [13] and is plotted in orange.

In the top two series, both have bigram phoneme networks, the lower of these two series uses a viseme classifier as in [13], and the upper our new phonemes denoted WLT. The lower pair of series use bigram word networks and again show the difference between visemes and our new method of training phoneme classifiers.

The situation in Figure 4 is summarised in Table 2. For hard to classify speakers, the new model training method gives a significant improvement.

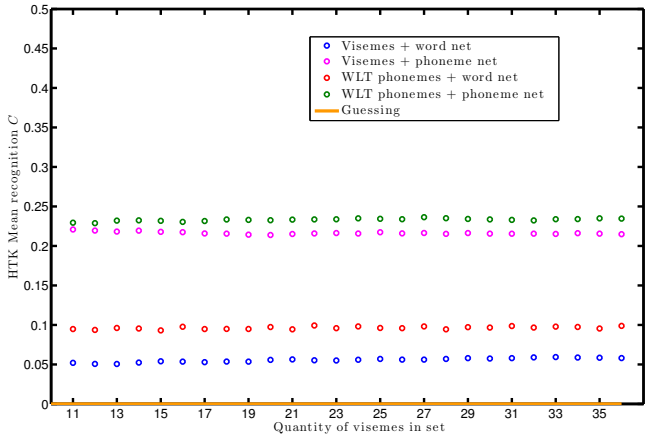
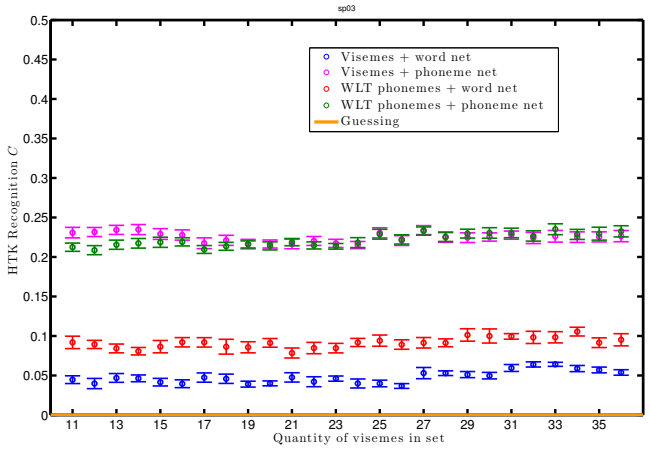


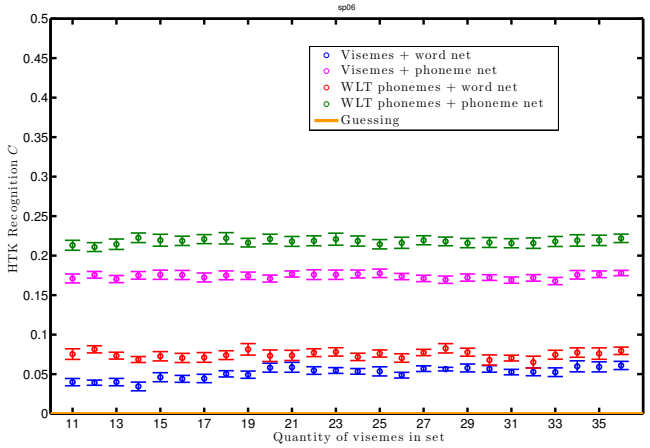
Fig. 4: HTK Correctness C for both types of classifier with either a phoneme or a word language model averaged over all 12 speakers.

Table 2: Minimum, maximum, and range of mean correctness measured over all speakers for the various methods. Top of table shows word correctness, bottom of table phoneme correctness.

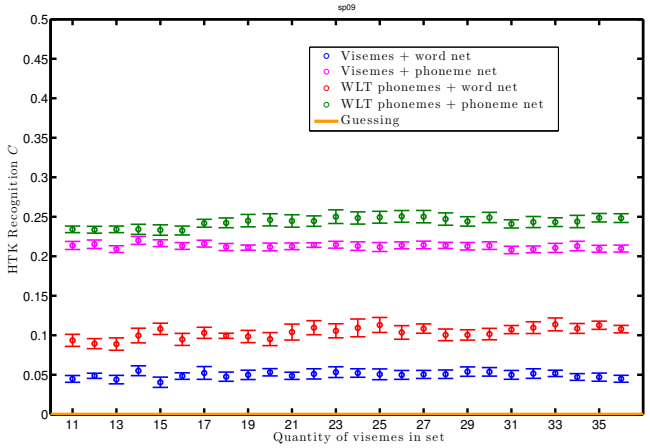
	Min	Max	Range
WLT phonemes + phoneme net	0.2253	0.2367	0.0114
Visemes + phoneme net	0.2036	0.2214	0.0179
Effect of WLT	0.0217	0.0153	–
WLT phonemes + word net	0.0905	0.0995	0.0090
Visemes + word net	0.0274	0.0601	0.0327
Effect of WLT	0.0631	0.0394	–



(a) Speaker 3



(b) Speaker 6



(c) Speaker 9

Fig. 5: HTK Correctness C for a variety of classifiers with either phoneme or word language models for three speakers.

Figures 5a, b & c show example performances for three speakers. Whilst not monotonic, these graphs are much smoother than the speaker-dependent graphs shown in [13]. Which is encouraging because it implies that our new algorithm is optimising its learning for each speaker-dependent phoneme-to-viseme mapping.

Figure 5 shows that, for certain numbers of visemes, and for certain speakers, the weak learning method gives improvement. However, with the right number of visemes for a particular speaker, the new method will always give a significant improvement.

Looking at Figure 5 there appeard to be a few regions where the new training method gives only marginal improvement. Not all speakers have these regions. We think the presence of these regions is associated with speakers that have more co-articulation than others. If this is true, then the phonemes are blurred together, the learning is more difficult and performance declines. We do not have enough speakers to make this anything other than speculation at this stage. Our own observation is that young people have more co-articulation than old people, but this is something for further investigation.

6. CONCLUSIONS

The choice of visual units in lip-reading has caused some debate. Some workers use visemes as adduced by for example Fisher [20] (in which visemes are a theoretical construct representing phonemes should look identical on the lips [10]). Others have noted that lip-reading using phonemes gives superior performance to visemes such as in [9].

Here, we supply further evidence to the more nuanced hypothesis first presented in [13], which is that there are intermediary units, which for convenience we call visemes, that can provide superior performances provided they are derived by an analysis of the data. A good number of visemes in a set is higher than previously thought.

In this paper we have presented a novel learning algorithm which shows improved performance for these new data-driven visemes by using them as an intermediate step in training phoneme classifiers. The essence of our method is to retrain the viseme models in a fashion similar to weak learning. This two-pass approach on the same training data has improved the training of phoneme labelled classifiers and increased the classification performance.

7. REFERENCES

[1] K. Kumar, Tsuhan Chen, and R.M. Stern, "Profile view lip reading," in *IEEE International Conference on Acoustics, Speech and Signal Processing. (ICASSP)*, 2007, vol. 4, pp. IV–429–IV–432.

[2] Yuxuan Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 432–437.

[3] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681 –685, June 2001.

[4] E.J. Ong and R. Bowden, "Robust facial feature tracking using shape-constrained multi-resolution selected linear predictors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1844–1859, 2011.

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

[6] Luca Cappelletta and Naomi Harte, "Phoneme-to-viseme mapping for visual speech recognition.," in *ICPRAM (2)*, 2012, pp. 322–329.

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

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

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

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

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

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

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

[14] Steve J Young, Gunnar Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev, and PC Woodland, "The HTK book version 3.4," 2006.

[15] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[16] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden, "Improving visual features for lip-reading," *International Conference on Audio-Visual Speech Processing (AVSP)*, vol. 7, no. 3, 2010.

[17] Iain Matthews and Simon Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[18] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 198 –213, Feb 2002.

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

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