Which phoneme-to-viseme maps best improve visual-only computer lip-reading?

No Author Given

No Institute Given

Abstract. A critical assumption of all current visual speech recognition systems is that there are visual speech units called visemes which can be mapped to units of acoustic speech, the phonemes. Despite there being a number of published maps it is infrequent to see the effectiveness of these tested, particularly on visual-only lip-reading (many works use audio-visual speech). Here we examine 120 mappings and consider if any are stable across talkers. We show a method for devising maps based on phoneme confusions from an automated lip-reading system, and we present new mappings that show improvements for individual talkers.

1 Introduction

Phonemes are the discriminate sounds of a language [1] and the visual equivalent, although not precisely defined, are the visemes; [2–4]. A working definition of a viseme is a set of phonemes that have identical appearance on the lips. Therefore a phoneme falls into one viseme class but a viseme may map to many phonemes: a many-to-one mapping. In computer lip-reading there are several possibilities for Phoneme-to-Viseme (P2V) mappings and some are listed in, for example, [5] Tables 2.3 and 2.4. Such mappings are often consonant-only mappings [6, 3, 7, 8]; or devised from single-talker data (so are talker-dependent [9]) or devised from highly stylised vocabularies ([10] for example). These are useful starting points but a P2V mapping should cover all phonemes. So here we consider the possibility of using combinations of the various known mappings which cover the consonants (listed in Table 2) and which cover vowels (Table 1). In total we use 15 consonant maps and eight vowel maps, all of these are paired with each other to produce 120 P2V maps to test.

2 Dataset and Data Preparation

We use the AVLetters2 (AVL2) dataset [11], to train and test recognisers based upon the 120 P2V mappings. This dataset is British-English talkers reciting the alphabet seven times. We use four talkers for training which involves tracking their faces with Active Appearance Models (AAMs) [12] and extracting combined shape and appearance features. We select AAM features because they are known to out-perform other feature methods in machine visual-only lip-reading [13].

Table 1. Vowel Viseme:Phoneme maps

Classification	Viseme phoneme sets
Bozkurt [14]	{/ei/ /ʌ/} {/ei/ /e/ /æ/} {/ɜ/} {/i/ /ɪ/ /ə/ /y/} {/u/ /ʊ/ /w/} {/ɑʊ/} {/ɔ/ /ɑ/ /ɔɪ/ /əʊ/}
Disney [15]	{/ʊ/ /h/} {/ɛə/ /i/ /ai/ /e/ /a/} {/u/} {/ʊə/ /ɔ/ /ɔə/}
Hazen [4]	{/ɑʊ/ /ʊ/ /u/ /əʊ/ /ɔ/ /w/ /ɔɪ/} {/ʌ/ /ɑ/} {/æ/ /e/ /ai/ /ei/} {/ə/ /ɪ/ /i/}
Jeffers [16]	{/ɑ/ /æ/ /ʌ/ /ai/ /e/ /ei/ /ɪ/ /i/ /ɔ/ /ə/ /ɪ/} {/ɔɪ/ /ɔ/} {/ɑʊ/} {/ɜ/ /əʊ/ /ʊ/ /u/}
Lee [17]	{/i/ /ɪ/} {/e/ /ei/ /æ/} {/ɑ/ /ɑʊ/ /ai/ /ʌ/} {/ɔ/ /ɔɪ/ /əʊ/} {/ʊ/ /u/}
Montgomery [18]	{/i/ /ɪ/} {/e/ /æ/ /ei/ /ai/} {/ɑ/ /ɔ/ /ʌ/} {/ʊ/ /ɜ/ /ə/}{/ɔɪ/} {/i/ /hh/} {/ɑʊ/ /əʊ/} {/u/ /u/}
Neti [19]	{/u/ /ʊ/ /əʊ/} {/æ/ /e/ /ei/ /ai/} {/ɪ/ /i/ /ə/} {/ɔ/ /ʌ/ /ɑ/ /ɜ/ /ɔɪ/ /ɑʊ/ /ɦ/}
Nichie [20]	{/u/} {/ʊ/ /əʊ/} {/ɑʊ/} {/i/ /ʌ/ /ɪ/} {/ʌ/} {/i/ /æ/} {/e/ /ɪə/} {/u/} {/ə/ /ei/}

Figure 1 shows the count of the 29 phonemes in training component of AVL2 with the silence phoneme omitted. As is often the case, the rare phonemes in British English are not represented [13]. The division of these phoneme across viseme classes will vary with each different map. P2V mappings are contractive which is illustrated in Table 3 which lists the ratio of phonemes to visemes (excluding silence and phonemes not handled by that mapping). Thus, in Table 3, the Woodward map covers 24 consonant phonemes to four viesmes and has a confusion factor (CF) of $4/24 = 0.167$, whereas Jeffers vowels maps cover 23 phonemes which are mapped to eight visemes.

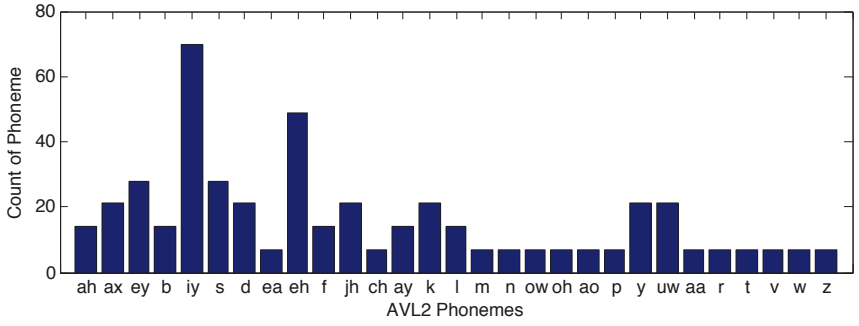

Fig. 1. Phoneme histogram of AVLetters-2 dataset

Table 2. Consonant Viseme:Phoneme maps

Classification	Viseme phoneme sets
Binnie [6]	{/p/ /b/ /m/} {/f/ /v/} {/θ/ /ð/} {/ʃ/ /ʒ/} {/k/ /g/} {/w/} {/r/} {/l/ /n/} {/t/ /d/ /s/ /z/}
Bozkurt [14]	{/g/ /ɦ/ /k/ /ŋ/} {/l/ /d/ /n/ /t/} {/s/ /z/} {/tʃ/ /ʃ/ /dʒ/ /ʒ/} {/r/} {/θ/ /ð/} {/f/ /v/} {/p/ /b/ /m/}
Disney [15]	{/p/ /b/ /m/} {/w/} {/f/ /v/} {/θ/} {/l/} {/d/ /t/ /z/ /s/ /r/ /n/} {/ʃ/ /tʃ/ /j/} {/y/ /g/ /k/ /ŋ/}
Finn [21]	{/p/ /b/ /m/} {/θ/ /ð/} {/w/ /s/} {/k/ /h/ /g/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/y/} {/z/} {/f/} {/v/} {/t/ /d/ /n/ /l/ /r/}
Fisher [3]	{/k/ /g/ /ŋ/ /m/} {/p/ /b/} {/f/ /v/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/t/ /d/ /n/ /θ/ /ð/ /z/ /s/ /r/ /l/}
Franks [7]	{/p/ /b/ /m/} {/f/} {/r/ /w/} {/ʃ/ /dʒ/ /tʃ/}
Hazen [4]	{/l/} {/r/} {/y/} {/b/ /p/} {m} {/s/ /z/ /h/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/ŋ/} {/f/ /v/} {/t/ /d/ /θ/ /ð/ /g/ /k/}
Heider [22]	{/p/ /b/ /m/} {/f/ /v/} {/k/ /g/} {/ʃ/ /tʃ/ /dʒ/} {/n/ /t/ /d/} {/l/} {/r/} {/θ/}
Jeffers [16]	{/f/ /v/} {/r/ /q/ /w/} {/p/ /b/ /m/} {/θ/ /ð/} {/tʃ/ /dʒ/ /ʃ/ /ʒ/} {/g/ /k/ /ŋ/} {/s/ /z/} {/d/ /l/ /n/ /t/}
Kricos [9]	{/p/ /b/ /m/} {/f/ /v/} {/w/ /r/} {/t/ /d/ /s/ /z/} {/l/} {/θ/ /ð/} {/ʃ/ /ʒ/ /tʃ/ /dʒ/} {/k/ /n/ /j/ /h/ /ŋ/ /g/}
Lee [17]	{/d/ /t/ /s/ /z/ /θ/ /ð/} {/g/ /k/ /n/ /ŋ/ /l/ /y/ /ɦ/} {/f/ /v/} {/r/ /w/} {/dʒ/ /tʃ/ /ʃ/ /ʒ/} {/p/ /b/ /m/}
Neti [19]	{/l/ /r/ /y/} {/s/ /z/} {/t/ /d/ /n/} {/ʃ/ /ʒ/ /dʒ/ /tʃ/} {/f/ /v/} {/ŋ/ /k/ /g/ /w/} {/p/ /b/ /m/} {/θ/ /ð/}
Nichie [20]	{/p/ /b/ /m/} {/f/ /v/} {/ɯ/ /w/} {/s/ /z/} {/ʃ/ /ʒ/ /tʃ/ /j/} {/t/ /d/ /n/} {/y/} {/θ/} {/l/} {/k/ /g/ /ŋ/} {/ɦ/} {/r/}
Walden [8]	{/p/ /b/ /m/} {/f/ /v/} {/θ /ð/} {/ʃ/ /ʒ/} {/w/} {/s/ /z/} {/r/} {/t/ /d/ /n/ /k/ /g/ /j/} {/l/}
Woodward [23]	{/t/ /d/ /n/ /l/ /θ/ /ð/ /s/ /z/ /tʃ/ /dʒ/ /ʃ/ /ʒ/ /j/ /k/ /g/ /h/} {/p/ /b/ /m/} {/f/ /v/} {/w /r/ /ɯ/}

We deliberately omit the following phonemes from some mappings; /si/ (Disney), /axr/ /en/ /el/ /em/ (Bozkirt), /axr/ /em/ /epi/ /tcl/ /dcl/ /en/ /gcl/ kcl/ (Hazan), and /axr/ /em/ /el/ /nx/ /en/ /dx/ /eng/ /ux/ (Jeffers) because these are American diacritics which are not appropriate for a British English phonetic dataset. Note that all 29 phonemes in AVL2 appear across the existing P2V maps, but no mapping uses all of these phonemes. Missing phonemes from a viseme map are grouped into a garbage viseme (/gar/) to ensures we measure only the performance of the previously described viseme sets. That is, we are not creating a new map by defining new visemes within an existing map.

3 Recognition Method

Our ground truth for measuring correct recognition is a viseme transcription produced by converting a phonetic transcript of the training data to viseme

labels assuming the mapping being tested (Tables 1 & 2). Using HTK [24], we build viseme-level Hidden Markov Model (HMM) recognisers with five states and five mixture components per state. We implement a leave-one-out seven-fold cross validation. Seven folds are selected as we have seven utterances of the alphabet per talker in AVL2. The HMMs are initialised using 'flat start' training and re-estimated eight times and then force-aligned using HTK's `HVite`. Training is completed by re-estimating the HMMs three more times.

4 Comparison of current P2V maps results

We measure recognition performance of the HMMs by correctness, C, as there are no insertion errors to consider at the word level (AVL2 contains isolated words). Correctness is measured using:

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

where S is the number of substitution errors, D is the number of deletion errors and N the total number of labels in the reference transcriptions.

Word recognition is less accurate than viseme recognition. However, viseme recognition performance is not a fair test since each viseme set has a different number of visemes. Instead, words are a common comparator that can be cross-referenced from each viseme set, and ultimately it is the difference between sets that we are interested in rather than the absolute level of performance.

Figure 2 shows mean word correctness ± one standard error over all talkers for each consonant map along the x-axis paired with each vowel map. Figure 3 shows the same but for each vowel map along the x-axis paired with each consonant map. Both x-axes are ordered by the mean correctness. This means we can see clearly that the 'best' performing map for both consonants and vowels are from Lee (as this is left-most on the x-axis) for all talkers.

Comparing the consonant P2V maps in Figure 2 we see that the Disney vowels are significantly worse than all others when paired with all consonant maps. Over the other vowels there is overlap with the majority of error bars suggesting little significant difference over the whole group but Lee [17] and Bozkurt [14] vowels are consistently above the mean and above the upper error bar for Disney [15], Jeffers [16] and Hazen [4] vowels. In comparing the vowel P2V maps in Figure 3, Lee[17] and Hazen [4] are the best consonants by a margin above the mean whereas Woodward [23] and Franks [7] vie for bottom performance. The best performance in terms of correctness is of a combination of vowels from Lee and consonants from Jeffers but close second best is a combination of Lee's consonants and vowels and this has a much smaller error bar.

In Table 3 we present data to suggest the best performing vowel P2Vs have a ratio of phonemes-to-visemes around 0.44 (top four CF mean = 0.44), and the better performing consonant maps have a CF of approximately 0.41 (top four CF mean = 0.41) so the better P2V is $<\sim 2$ phonemes per viseme.

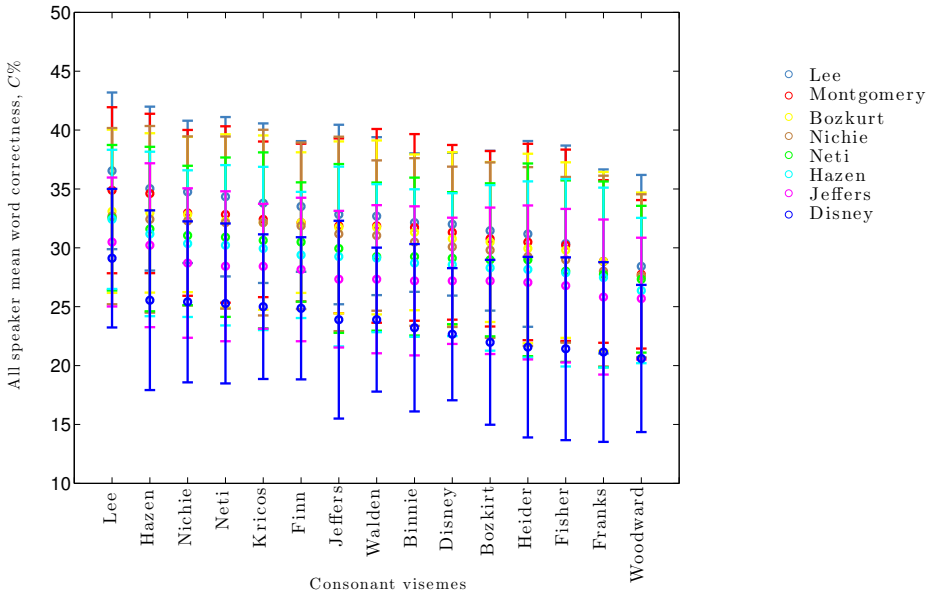

Fig. 2. Talker-dependent mean word recognition ± one standard error over all four talkers comparing consonant P2V maps paired with all vowel mappings

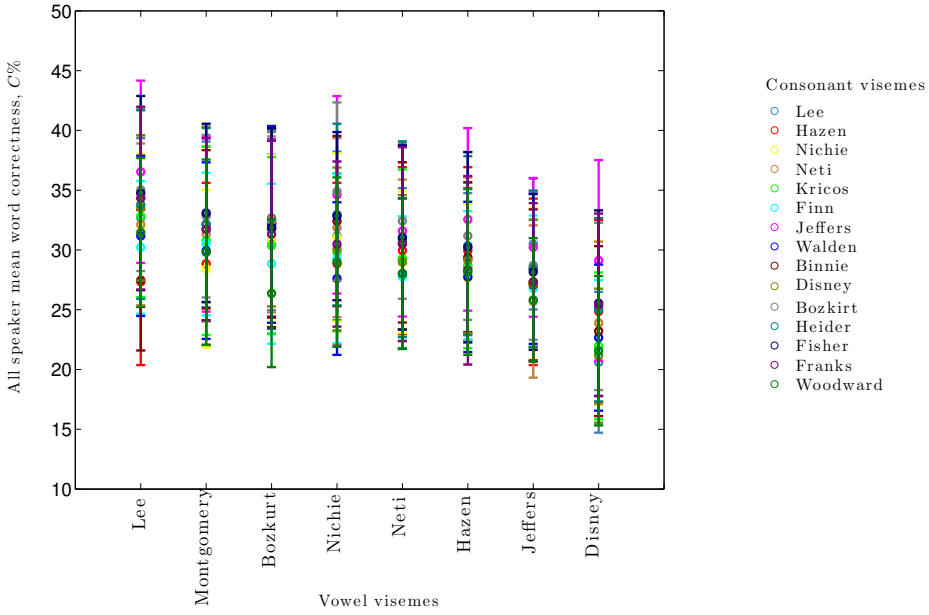

Fig. 3. Talker-dependent mean word recognition ± one standard error over all four talkers comparing vowel P2V maps paired with all consonant mappings

Table 3. Confusion Factors for each viseme map tested

Consonant Map	V:P	CF	Mean C	Vowel Map	V:P	CF	Mean C
Woodward	4:24	0.16	27.52	Jeffers	3:19	0.16	27.74
Disney	6:22	0.18	28.74	Neti	4:20	0.20	29.76
Fisher	5:21	0.23	28.86	Hazen	4:18	0.22	29.27
Lee	6:24	0.25	31.55	Disney	4:11	0.36	23.71
Franks	5:17	0.29	27.83	Lee	5:14	0.36	32.35
Kricos	8:24	0.33	29.46	Bozkurt	7:19	0.37	31.17
Jeffers	8:23	0.35	29.28	Montgomery	8:19	0.42	31.23
Neti	8:23	0.35	30.67	Nichie	9:15	0.60	31.13
Bozkurt	8:22	0.36	28.67	-	-	-	-
Finn	10:23	0.43	29.43	-	-	-	-
Walden	9:20	0.45	29.93	-	-	-	-
Binnie	9:19	0.47	29.43	-	-	-	-
Hazen	10:21	0.48	32.33	-	-	-	-
Heider	8:16	0.50	28.47	-	-	-	-
Nichie	18:33	0.54	30.94	-	-	-	-

5 New viseme mappings

Given that Lee [17] provides the best pairing of the existing phoneme to viseme maps, we now ask if there are alternatives that can perform better? Our first approach is to find talker-dependent P2V maps based upon phoneme confusion matrices generated by a visual-only automated recognition system using phoneme HMM classifiers. Where a phoneme is only ever correctly identified as itself (true positives on the confusion matrix diagonal), this is quickly allocated to be a viseme of that single phoneme.

Now we address the remaining phonemes which have been confused. The first candidate for viseme class 1 is a subset of Phonemes: $V_1 = \{\phi_1, \phi_2, \phi_{M_1}\}$ such that every pair, (ϕ_i, ϕ_j) in V_1 has $N_{ij} > 0$. V_1 is chosen as the largest such set. V_2, which is the second viseme set, is determined in the same way from the remaining phonemes until all phonemes are accounted for. Within this process phonemes are grouped into a viseme class only if *all* of the phonemes within the candidate group are mutually confused. Once a phoneme has been assigned to a viseme class, it is no longer considered for grouping and so any possible other viseme combinations that include this phoneme are discarded.

Our phoneme recognition produces confusions between consonant and vowel phonemes so we make two types of map, one that permits vowel and consonant phonemes to be mixed within the same viseme and a second which restricts visemes to be vowel or consonant phonemes only. These P2V maps for each talker are in Table 4. These are the "tightly confused" maps because all phonemes within each viseme have been confused with each other in the phoneme recognition.

These viseme sets will contain spurious phonemes that cannot be grouped into a viseme because they are not confused with *all* of the phonemes of the viseme. This leaves some single-phoneme visemes (e.g. /u/ in Talker 1 with mixed

Table 4. Tightly confused phoneme talker-dependent visemes. The score in brackets is the ratio of phonemes to visemes

Classification	P2V mapping - permitting mixing of vowels and consonants
Talker1	{/ʌ/ /ai/ /i/ /n/ /əʊ/} {/b/ /e/ /ei/ /y/ } {/d/ /s/} {/tʃ/ /l/}
(CF:0.48)	{/t/} {/w/} {/f/} {/k/} {/ə/ /v/} {/dʒ/ /z/} {/ɑ/ /u/}
Talker2	{/ə/ /ai/ /ei/ /i/ /s/} {/e/ /v/ /w/ /y/} {/l/ /m/ /n/} {/ʌ/ /f/}
(CF: 0.44)	{/z/} tʃ/} {/t/} {/ɑ/} {/əʊ/ /u/} {/dʒ/ /k/} {/b/ /d/ /p/}
Talker3	{/ei/ /f/ /n/} {/d/ /t/ /p/} {/b/ /s/} {/l/ /m/} {/ə/ /e/} {/i/}
(CF: 0.68)	{/ɑ/} {/dʒ/} {/əʊ/} {/z/} {/y/} {/tʃ}/ {/ai/} {/ʌ/} {/ɑ/} {/dʒ/} {/k/ /w/} {/əʊ/} {/z/} {/v/} {/u/}
Talker4	{/ʌ/ /ai/ /i/ /ei/ } {/m/ /n/} {/ə/ /e/ /p/} {/k/ /w/} {/d/ /s/}
(CF: 0.64)	{/f/} {/v/} {/ɑ/} {/z/} {/tʃ/} {/b/} {/əʊ/} {/dʒ/ /t/} {/b/} {/əʊ/} {/l/} {/u/}
Classification	P2V mapping - restricting mixing of vowels and consonants
Talker1	{/ʌ/ /i/ /əʊ/ /u/} {/ɑ/ /ei/} {/ə/ /e/ /ei/} {/d/ /s/ /t/ } {/tʃ/ /l/}
(CF:0.50)	{/k/} {/z/} {/w/} {/f/} {/m/ /n/} {/dʒ/ /v/} {/b/ /y/}
Talker2	{/ai/ /ei/ /i/ /u/} {/əʊ/} {/ə/} {/e/} {/ʌ/} {/ɑ/} {/v/ /w/} {/k/}
(CF: 0.58)	{/d/ /b/} {/t/} {/tʃ/} {/l/ /m/ /n/} {/dʒ/ /p/ /y/} {/f/ /s/}
Talker3	{/ei/ /i/} {/ai/} {/ə/ /e/} {/ʌ/} {/d/ /p/ /t/} {/l/ /m/} {/k/ /w/}
(CF: 0.68)	{/tʃ/} {/əʊ/} {/y/} {/u/} {/ɑ/} {/z/} {/b/ /s/} {/v/} {/dʒ/} {/f/ /n/}
Talker4	{/ʌ/ /ai/ /i/ /ei/} {/ə/ /e/} {/m/ /n/} {/k/ /l/} {/dʒ/ /t/} {/b/}
(CF: 0.65)	{/əʊ/} {/y/} {/u/} {/ɑ/} {/w/} {/f/} {/v/} {/tʃ/} {/d/ /s/}

vowel and consonant phonemes), so our second approach relaxes the condition requiring confusion with all of the phonemes. We execute a second pass through the viseme sets. Any single-phoneme viseme classes are then permitted to merge with existing multi-phoneme classes if they share any confusions with that class. In the event that a phoneme has multiple class confusions it is merged with the class with the greatest confusion. We term these the "loosely confused" maps. Again we do two sets with vowel and consonant phonemes both mixed and separate. The final P2V maps are in Table 5 for four talkers.

Looking at Tables 4 and 5 there are no identical visemes with each map type between talkers, this confirms our variability of individual talker visual speech (excluding the true positive single phoneme visemes). We observe that none of the new visemes match the previously suggested visemes in the comparison study (Tables 1 and 2), e.g. the most common previous viseme was {/p/ /b/ /m/} and this is never created with our new method.

Figure 4 shows the word recognition performance using both the tightly confused map and the loosely confused map for each talker. Also shown is the performance using the Lee map as a benchmark. For Talker 1 no new viseme map significantly improves upon the benchmark performance, but we do see significant improvements for both Talker 2 and Talker 4 and a minor improvement within the error bars for Talker 3. For Talkers 2 and 3, both types of the split vowels and consonant maps demonstrate improvement on the benchmark, and for Talker 4 the tightly confused split vowels and consonants shows a significant

Table 5. Loosely confused phoneme talker-dependent visemes. The score in brackets is the ratio of phonemes to visemes

Classification P2V mapping - permitting mixing of vowels and consonants	
Talker1	{/b/ /e/ /ei/ /p/ /w/ /y/ /k/} {/ʌ/ /ai/ /f/ /i/ /m/ /n/ /əʊ/}
(CF:0.28)	{/dʒ/ /z/} {/ɑ/ /u/} {/d/ /s/ /t/} {/tʃ/ /l/} {/ə/ /v/}
Talker2	{/ɑ/ /ə/ /ai/ /ei/ /i/ /s/ /tʃ/} {/e/ /t/ /v/ /w/ /y/} {/l/ /m/ /n/}
(CF: 0.32)	{/ʌ/ /f/} {/z/} {/b/ /d/ /p/} {/əʊ/ /u/} {/dʒ/ /k/}
Talker3	{/ʌ/ /ai/ /ei/ /f/ /i/ /n/} {/ə/ /e/ /y/ /tʃ/} {/b/ /s/ /v/}
(CF: 0.40)	{/dʒ/} {/əʊ/} {/z/} {/l/ /m/ /u/} {/d/ /p/ /t/} {/k/ /w/} {/ɑ/}
Talker4	{/ʌ/ /ai/ /tʃ/ /i/ /ei/ } {/ɑ/ /m/ /u/ /n/} {/ə/ /e/ /p/ /v/ /y/}
(CF: 0.32)	{/dʒ/ /t/} {/k/ /l/ /w/} {/əʊ/} {/d/ /f/ /s/} {/b/}
Classification P2V mapping - restricting mixing of vowels and consonants	
Talker1	{/ʌ/ /i/ /əʊ/ /u/} {/ɑ/ /ai/} {/ə/ /e/ /ei/} {/b/ /w/ /y/}
(CF:0.47)	{/k/} {/z/} {/m/} {/l/} {/d/ /f/ /s/ /t/} {/tʃ/} {/dʒ/ /k/ /v/ /z/}
Talker2	{/ɑ/ /ʌ/ /ə/ /ai/ /ei/ /i/ /əʊ/ /u/} {/k/ /t/ /v/ /w/}
(CF: 0.29)	{/f/ /s/} {/tʃ/ /l/ /m/ /n/} {/dʒ/ /p/ /y/} {/b/ /d/} {/z/}
Talker3	{/ʌ/ /ai/ /i/ /ei/} {/ə/ /e/} {/b/ /s/ /v/} {/d/ /p/ /t/}
(CF: 0.56)	{/y/} {/dʒ/} {/əʊ/} {/z/} {/u/} {/ə/ /e/} {/l/ /m/} {/k/ /w/} {/f/ /n/} {/ɑ/} {/tʃ/}
Talker4	{/ʌ/ /ai/ /i/ /ei/} {/tʃ/ /k/ /l/ /w/} {/d/ /f/ /s/ /v/} {/m/ /n/}
(CF: 0.50)	{/f/} {/ɑ/} {/dʒ/ /t/} {/əʊ/} {/u/} {/y/} {/b/}

improvement. Comparing mixed consonant and vowel maps against split consonant and vowel maps, the split maps are always better than mixed maps for all talkers in this data. In comparing the loosely confused maps versus the tightly confused maps, the tight confusions are better for two out of our four talkers (Talkers 1 and 2) and equal for a third (Talker 1). These are talkers with highest confusion factor P2V maps (Tables 4 & 5). This is despite the tightly confused viseme set including single phoneme-viseme classes which can be confused with parts of the tightly confused classes.

6 Conclusions and Future work

We have completed a comprehensive experimental study of previously suggested P2V maps and shown that Lee [17] is the best of the previously published P2V maps. Puzzlingly the Lee mapping is not that popular among engineers of lip-reading systems so our finding should be of immediate use.

We have also outlined how it is possible to build phoneme-to-viseme maps in a systematic way using confusion matrices from real recognisers. We believe that this is a more principled approach than previous methods (including Lee's [17] whose method is bound by the Fisher [3] visemes) and also allows comparison between talkers using phonetic terminology. Further we have shown that the automatic method need do no worse than the Lee visemes and can exceed performance. We acknowledge that our dataset is still rather small and the sparsely represented phonemes are unlikely to be accurately modelled. In future we would

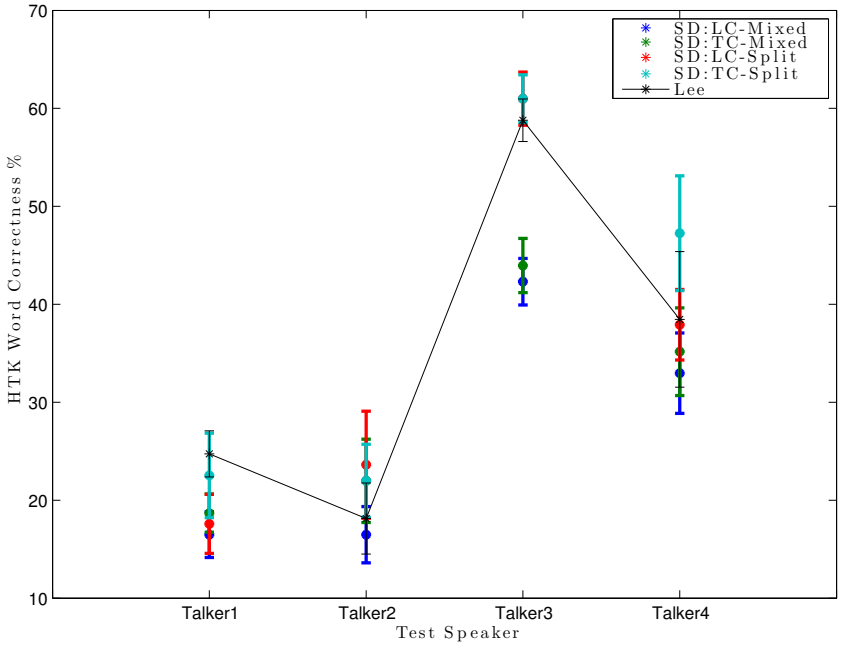

Fig. 4. HTK word Correctness using tightly confused and loosely confused viseme sets based on phoneme recognition confusions. SD = Speaker Dependent, LC = Loosely coupled, TC = Tightly Coupled, Mixed = Mixed vowels and consonant phonemes within viseme classes and Split = separated vowel and consonant visemes

like to extend this to full set of American and British phonemes but that will require a more extensive set of data.

References

1. Association, I.P.: Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press (1999)
2. Chen, T., Rao, R.R.: Audio-visual integration in multimodal communication. Proceedings of the IEEE **86** (1998) 837–852
3. Fisher, C.G.: Confusions among visually perceived consonants. Journal of Speech, Language and Hearing Research **11** (1968) 796
4. Hazen, T.J., Saenko, K., La, C.H., Glass, J.R.: A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In: Proceedings of the 6th International Conference on Multimodal Interfaces. ICMI '04, New York, NY, USA, ACM (2004) 235–242
5. Theobald, B.J.: Visual speech synthesis using shape and appearance models. PhD thesis, University of East Anglia (2003)

6. Binnie, C.A., Jackson, P.L., Montgomery, A.A.: Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. Journal of Speech and Hearing Disorders **41** (1976) 530
7. Franks, J.R., Kimble, J.: The confusion of english consonant clusters in lipreading. Journal of Speech, Language and Hearing Research **15** (1972) 474
8. Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C.J.: Effects of training on the visual recognition of consonants. Journal of Speech, Language and Hearing Research **20** (1977) 130
9. Kricos, P.B., Lesner, S.A.: Differences in visual intelligibility across talkers. The Volta Review (1982)
10. Owens, E., Blazek, B.: Visemes observed by hearing-impaired and normal-hearing adult viewers. Journal of Speech and Hearing Research **28** (1985) 381
11. Cox, S., Harvey, R., Lan, Y., Newman, J., Theobald, B.J.: The challenge of multi-speaker lip-reading. In: International Conference on Auditory-Visual Speech Processing, Citeseer (2008) 179–184
12. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60** (2004) 135–164
13. Cappelletta, L., Harte, N.: Phoneme-to-viseme mapping for visual speech recognition. In: ICPRAM (2). (2012) 322–329
14. Bozkurt, E., Erdem, C., Erzin, E., Erdem, T., Ozkan, M.: Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. Proc. of Signal Proc. and Communications Applications (2007) 1–4
15. Lander, J.: Read my lips: Facial animation techniques. http://www.gamasutra.com/view/feature/131587/read_my_lips_facial_animation_.php (2014) Accessed: 2014-01-28.
16. Jeffers, J., Barley, M.: Speechreading (lipreading). Thomas Springfield, IL: (1971)
17. Lee, S., Yook, D.: Audio-to-visual conversion using hidden markov models. In: PRICAI 2002: Trends in Artificial Intelligence. Springer (2002) 563–570
18. Montgomery, A.A., Jackson, P.L.: Physical characteristics of the lips underlying vowel lipreading performance. The Journal of the Acoustical Society of America **73** (1983) 2134
19. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J.: Audio-visual speech recognition. In: Final Workshop 2000 Report. Volume 764. (2000)
20. Nitchie, E.B.: Lip-Reading, principles and practise: A handbook for teaching and self-practise. Frederick A Stokes Co, New York (1912)
21. Finn, K.E., Montgomery, A.A.: Automatic optically-based recognition of speech. Pattern Recognition Letters **8** (1988) 159–164
22. Heider, F., Heider, G.M.: An experimental investigation of lipreading. Psychological Monographs **52** (1940) 124–153
23. Woodward, M.F., Barber, C.G.: Phoneme perception in lipreading. Journal of Speech, Language and Hearing Research **3** (1960) 212
24. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchec, V., Woodland, P.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department (2006)