Comparing phonemes and visemes with DNN-based lipreading

Kwanchiva Thangthai[1]
k.thangthai@uea.ac.uk
Helen L Bear[2]
helen@uel.ac.uk
Richard Harvey[1]
r.w.harvey@uea.ac.uk

[1] School of Computing Sciences
University of East Anglia
Norwich, UK

[2] Computer Science & Informatics
University of East London
London, UK.

Abstract

There is debate if phoneme or viseme units are the most effective for a lipreading system. Some studies use phoneme units even though phonemes describe unique short sounds; other studies tried to improve lipreading accuracy by focusing on visemes with varying results. We compare the performance of a lipreading system by modeling visual speech using either 13 viseme or 38 phoneme units. We report the accuracy of our system at both word and unit levels. The evaluation task is large vocabulary continuous speech using the TCD-TIMIT corpus. We complete our visual speech modeling via hybrid DNN-HMMs and our visual speech decoder is a Weighted Finite-State Transducer (WFST). We use DCT and Eigenlips as a representation of mouth ROI image. The phoneme lipreading system word accuracy outperforms the viseme based system word accuracy. However, the phoneme system achieved lower accuracy at the unit level which shows the importance of the dictionary for decoding classification outputs into words.

1 Introduction

As lipreading transitions from GMM/HMM-based technology to systems based on Deep Neural Networks (DNNs) there is merit in re-examining the old assumption that phoneme-based recognition outperforms recognition with viseme-based systems. Also, given the greater modeling power of DNNs, there is value in considering a range of rather primitive features such as Discrete Cosine Transform (DCT) [1] and Eigenlips [6] which had previously been disparaged due to their poor performance.

Visual speech units divide into two broad categories; phonemes and visemes. A phoneme is the smallest unit of speech that distinguishes one word sound from another [3]. Therefore it has a strong relationship with an acoustic speech signal. In contrast, a viseme is the basic visual unit of speech that represents a gesture of the mouth, face and visible parts of the teeth and tongue, the visible articulators. Generally speaking, mouth gestures have less variation than sounds and several phonemes may share the same gesture so a class of visemes may contain many different phonemes. There are many choices of visemes [5] and Table 1 shows one of those mappings [18].

Table 1: Neti [18] Phoneme-to-Viseme mapping.

Consonants			Vowels			Silence		
Viseme	TIMIT phonemes	Description	Viseme	TIMIT phoneme	Description	Viseme	TIMIT phoneme	Description
/A	/l/ /el/ /r/ /y/	Alveolar-semivowels	/V1	/ao/ /ah/ /aa/ /er/ /oy/ /aw/ /hh/	Lip-rounding based vowels	/S	/sil/ /sp/	Silence
/B	/s/ /z/	Alveolar-fricatives	/V2	/uw/ /uh/ /ow/	"			
/C	/t/ /d/ /n/ /en/	Alveolar	/V3	/ae/ /eh/ /ey/ /ay/	"			
/D	/sh/ /zh/ /ch/ /jh/	Palato-alveolar	/V4	/ih/ /iy/ /ax/	"			
/E	/p/ /b/ /m/	Bilabial						
/F	/th/ /dh/	Dental						
/G	/f/ /v/	Labio-dental						
/H	/ng/ /g/ /k/ /w/	Velar						

2 Developing DNN-HMM based lipreading system

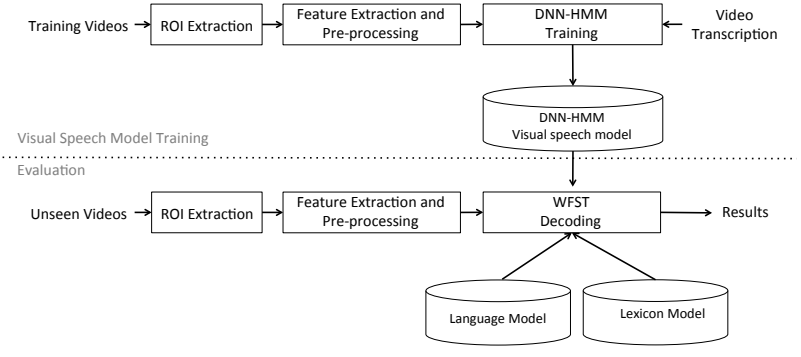

Figure 1: Lipreading system construction techniques.

Conventional techniques to model visual speech are based around Hidden Markov Models (HMMs) [22]. The aim of the model is to find the most likely word or unit sequence corresponding to the visual observation. HMMs comprise two probability distributions: the transition probability and the probability density function (PDF) associated with the continuous outputs. The transition probabilities represent a first-order Markov process. The PDF of speech feature vectors is modeled by a Gaussian Mixture Model (GMM) that is parameterised by the mean and the variance of each component.

There are some weaknesses of GMM, that have been found in acoustic modeling [12]. First, it is statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. Second, in order to reduce the computational cost by using a diagonal rather than a full covariance matrix, uncorrelated features are needed. These deficiencies motivate the consideration of alternative learning techniques.

The deep network structure can be considered as a feature extractor by using the number of neurons in multiple hidden layers to learn the essential patterns from the input features [16]. In addition, the backpropagation algorithm [11] with its appropriate learning criterion is essentially optimizing the model to fit to the training data discriminatively.

However, to decode a speech signal, temporal features and models that can capture the sequential information in speech such as an observable Markov sequence in the HMM is still necessary. Thus arises the DNN-HMM hybrid structure in which the DNN is used instead of the GMM in the HMM. The method essentially combines the advantages from these two algorithms.

2.1 Feature extraction

The literature provides a variety of feature extraction methods, often combined with tracking (which is essential if the head of the talker is moving). Here we focus on features that have been previously described as "bottom-up" [17] meaning that they are derived directly from the pixel data and require only a Region-Of-Interest, or ROI. Figure 2 illustrates a typical ROI taken from the TCD-TIMIT dataset described later in Section 4.1 plus two associated feature representations which we now describe.

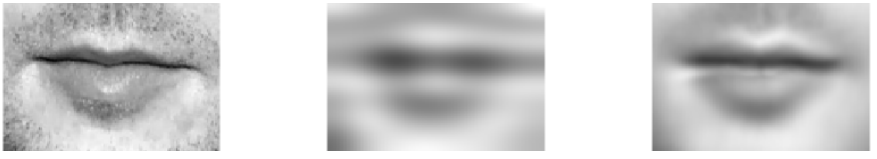

Figure 2: Comparing the original ROIs image (left) and its reconstruction via 44-coefficient DCT (middle) and 30-coefficient Eigenlip (right).

2.1.1 Discrete Cosine Transform (DCT)

The DCT is a popular transform in image coding and compression. DCT aims to represent the frequency domain of signal periodically and symmetrically using the cosine function. In particular, the DCT is a part of the Fourier Transform family but contains only the real part (Cosine). Because of its popularity most modern processors execute it very quickly (roughly $\mathcal{O}(N)$ for modern algorithms) so this also explains its ubiquity. For strongly correlated Markov processes the DCT approaches the Karhunen-Loeve transform in its compaction efficiency. Possibly this explains its popularity as a benchmark feature [18]. Here we use DCT II with zigzag property [26], which means that the first elements of the feature vector contain the low-frequency information. The resulting feature vector has 44 dimensions.

2.1.2 Eigenlip

The Eigenlip feature is another appearance-based approach[15]. The Eigenlips feature has been generated via Principal Component Analysis (PCA) [25]. Here, we use PCA to extract the Eigenlip feature from grey-scale lip images, where we retain only 30-dimensions of PCA. To construct the PCA, 25-ROIs images of each training utterance were randomly selected to be the set of training images. Almost about $100k$ images in total were used to compute the Eigenvector and Eigenvalue and use it for extracting training and testing utterances. Only 30 dimensions of principal components with high variation were retained.

2.2 Feature transformation

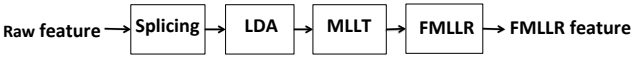

Figure 3: FMLLR feature pre-processing pipeline.

The raw features are 30-dimensional Eigenlip and 44-dimensional DCT features. These raw features are then normalized by subtracting the mean of each speaker and the 15 consecutive frames are spliced onto the feature to add dynamic information. Second, Linear Discriminant Analysis (LDA) [8] and Maximum Likelihood Linear Transform (MLLT) [9] are applied to reduce and map the features to a new space to minimize the within-class distance and maximize the between-class distance, where the class is the HMM-state, whilst simultaneously maximizing the observation likelihood in the original feature space.

Finally, feature-space Maximum Likelihood Linear Regression (fMLLR) [20],[9], also known as the feature-space speaker adaptation technique, is employed to normalize the variation within a speaker. These new 40-dimensional fMLLR features are used as inputs (labeled as feature pre-processing pipeline in Figure 3) to the subsequent machine learning. The use of LDA is quite commonplace in lipreading and is derived in the HiLDA framework [19]. MLLT and fMLLR are commonplace in acoustic speech recognition but have only recently been applied to visual speech recognition [7] albeit on smallish datasets.

2.3 Visual speech model training

Our DNN-HMM visual speech model training involves all five successive stages. Here, we detail the development of the visual speech model that we employ in this work including all steps and parameters.

2.3.1 Context-Independent Gaussian Mixture Model (CI-GMM)

The first step is to initialise a model by creating a simple CI-GMM model. This step creates a time alignment for the entire training corpus by simply constructing a mono phoneme/viseme model that contains 3-state GMM-HMMs for each speech unit.

The CI-GMMs are trained on the raw features, which are DCT and PCA, along with its first and second derivative coefficients ($\Delta + \Delta\Delta$). We use 3-state GMM-HMMs on each visual speech unit. Instead of setting the fixed number for increasing Gaussian mixture, we have set the maximum number of Gaussian to be 1000 so that each state will keep increasing independently until their variances reach the maximum. When the training process starts, the time alignment of training data is equally segmented and updated in every iteration for the first ten iterations, then updated every two iterations until a maximum of 40 iterations.

2.3.2 Context-Dependent Gaussian Mixture Model (CD-GMM)

The context-dependent viseme models (CD-GMMs) is specified on the same feature as in CI-GMM system. Here we use tied-state of 3-context visual speech model, where the tied-states are obtained from the data-driven approach tree-clustering [21]. We have specified the maximum number of leaf nodes to be 2000, which limits the number of states. The maximum number of Gaussians is set to $10K$. The training iterations continue until there is convergence which in practice is fewer than 35 iterations. We realign every 10 iterations.

2.3.3 CD-GMM with LDA-MLLT feature transformation

This training step also uses CD-GMMs, but trained on the LDA-MLLT features. The 40-dimensional LDA-MLLT features are formed by splicing 15 frames of the current frame (seven on the left and seven on the right) then reducing, via LDA, to 40 dimensions per

frame. This compact set of LDA-MLLT feature parameterizes to the 40-dimension that best associates with the visual speech unit and also comprises the dynamic of visual speech over 150ms. Again, the different set of tied-state CD-GMM has been constructed considered to the current feature. The maximum number of leaf nodes is set to $2,500$, and the total number of Gaussians is $15K$. This step utilises the equivalent number of training iterations and the realignment as those used in the previous step.

2.3.4 CD-GMM with Speaker Adaptive Training (SAT)

In a Speaker Adaptive Training (SAT) system, the CD-GMM are built on an fMLLR transformation on top of LDA-MLLT features by estimating a transform for each speaker. The same training process in the preceding step is then applied on the 40-dimensions of fMLLR feature, where the number of leaf nodes and Gaussian are identical.

2.3.5 Context-Dependent Deep Neural Networks (CD-DNN)

We construct the CD-DNNs model on the hybrid DNN-HMMs architecture. The CD-DNNs are trained and optimized by minimizing frame-based cross-entropy between the prediction and the PDF target. The PDF refers to the tied-state context-dependent label, which is generated from the SAT system, that aligned every frame. The feature we adopted for all DNN training is based on LDA+MLLT+fMLLR features with mean and variance normalization.

The CD-DNNs model is trained on six hidden layers with 2048 neurons per layer, where we use the sigmoid non-linearity function in each neuron. The input layer is the fMLLR feature with temporally spliced 11 consecutive frames. The model is initialized by a stacking of Recurrent Boltzman Machines (RBM) with three iterations on a single-GPU machine. The learning rate for RBM training is 0.4 and applying L2 penalty (weight decay) at 0.0002. The learning rate for fine-tuning has been set to 0.008 with dropout of 0.1. We use the minibatch-Stochastic Gradient Descent (SGD) for fine-tuning with minibatch size of 256. We produce a development set for tuning the network by randomly selecting 10% of training data. Every DNN training iteration is required to have a cross-validation loss is lower than the previous training iteration. If a iteration is rejected then one retries with a new stochastic gradient descent parameter. The terminating condition is that the new loss is little different from the old loss (specifically we use a difference smaller than 0.001 of the loss as a suitable terminating condition).

2.4 Decode lipreading with WFST Decoder

Weighted Finite-state Transducer decoders have been increasingly used to decode speech signal in Large Vocabulary Continuous Speech Recognition (LVCSR) tasks and have also become a state-of-the-art decoder [13]. To decode a visual speech signal, we need a visual speech model, a language model, and a lexicon or as so called, a pronunciation dictionary.

Our lipreading decoder comprises the visual speech DNN-HMM model,the TCD-TIMIT pronunciation dictionary and the word bi-gram language model. We generate the decoding graph as a finite-state transducer (FST) via the Kaldi toolkit [21] .Beam width pruning is applied every 25 frames where we use 13.0 for the Viterbi pruning beam [24] and 8.0 for the lattice beam and the visual speech model scale is 0.1. The lattice that contains the entire surviving path is re-scored by applying the bigram language model with the scaling factor over the range $5-15$. Only the lowest word error rates after LM re-scoring are used.

3 Analysis of the pronunciation dictionary

Reducing the set of speech units, such as reducing a set of phonemes to a set of visemes, reduces the discriminative power of the classification model whilst increasing the complexity of pronunciation dictionary by increasing the volume of homophonic words. This suggests that word accuracy of a viseme based system will be lower than a phoneme based system. The counter argument is that visemes might be simpler to classify (because there are fewer of them and they are meant to be better matched to the visual signal) so there is clearly a trade-off between homopheny and unit accuracy [7].

Table 2: Example of phoneme and viseme dictionary with its corresponding IPA symbols.

Word Entry	IPA Symbol	Phoneme Dictionary	Viseme Dictionary
TALK	t ɔk	t ao k	C V1 H
TONGUE	t ʌŋ	t ah ng	C V1 H
DOG	d ɔg	d ao g	C V1 H
DUG	d ʌg	d ah g	C V1 H
CARE	k e r	k eh r	H V3 A
WELL	w e l	w eh l	H V3 A
WHERE	w e r	w eh r	H V3 A
WEAR	w e r	w eh r	H V3 A
WHILE	w ai l	w ay l	H V3 A

Table 2 shows examples of the homophoneme and homoviseme words that occur in the TCD-TIMIT dictionary. Figure 4 describes the homophone problem in two ways. On the left words are binned according to how many homophones they have. Thus the column labelled "1 occur" is the count of all unique words, the column labelled "2 occur" is the count of words that have one other homophone and so on. It is evident the switch to visemes causes more homophones particularly large numbers of high-multiplicity homophones. This effect can also be seen in the dictionary size (right of Figure 4). Homophones cause dictionary entries to merge so the visual dictionary is smaller than the acoustic one.

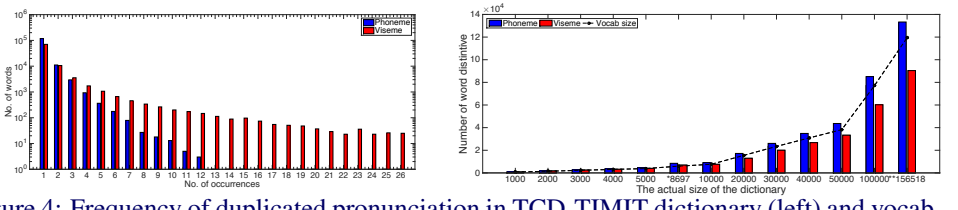

Figure 4: Frequency of duplicated pronunciation in TCD-TIMIT dictionary (left) and vocabulary size (right) for both phoneme and viseme units.

4 Experiment methodology

4.1 Data and Benchmarks

We use the TCD-TIMIT [10] corpus containing 59 volunteer speakers. We chose this dataset because it is the largest vocabulary audio-visual speech corpus available in the public domain. The WFST operates on a vocabulary of almost 6,000 words from a dictionary of 160,000 entries. This dataset provides lists of non-overlapping utterances for training and

evaluation in two scenarios: speaker-dependent (SD) and speaker-independent (SI). In the SD scenario, visual models are trained on 3,752 utterances and evaluated on 1,736 utterances. Whereas in the SI experiment, 3,822 utterances from 39 talkers are in the training set and we evaluate on the remaining 17 talkers containing a total 1,666 utterances.

The TCD-TIMIT release includes a baseline viseme accuracy for both speaker dependent and speaker independent settings using Jeffers and Barley [14] visemes. The best viseme accuracy of recognizing 12 viseme units reported on TCD-TIMIT is 34.77% in speaker independent tests and 34.54% on speaker dependent tests. The context independent viseme models (referred to as mono-viseme in the paper) were trained on 44-coefficient DCT feature with 4-state HMMs and 20 Gaussian mixtures per state.

5 Results

5.1 Viseme-based lipreading experiment

One fundamental measure of the performance of an automatic lipreading system is viseme accuracy. Since the viseme recognizer requires no dictionary or language model, it is quicker to build and optimise.

Table 3 lists the accuracies achieved with our viseme based lipreading system. In comparison to the viseme accuracies benchmarked with the TCD-TIMIT corpus, our best SD viseme accuracy is 46.61% with Eigenlips, compared to 34.54%, an improvement of 12.07%. Our best SI viseme accuracy is 44.61% which improves on the benchmark 34.77% by 10.16%, again with the Eigenlips features.

Table 3: Viseme-based lipreading accuracy (%).

Model	Feature	Viseme accuracy (%)		Word accuracy (%)	
		SD	SI	SD	SI
CD-GMM + SAT	DCT	44.66	42.48	14.37	10.47
CD-DNN		43.67	38.00	23.89	9.17
CD-GMM + SAT	Eigenlips	45.59	44.61	16.71	12.15
CD-DNN		46.61	44.60	33.06	19.15

Word accuracy achieved with visemes, albeit lower than the viseme accuracy, also shows that Eigenlip features outperform the DCT: we achieved 33.06% in speaker dependent tests, and 19.15% in speaker independent tests.

5.2 Phoneme-based lipreading experiment

Table 4 shows the word and phoneme accuracies achieved with our phoneme-based lipreading system. This system achieved the most accurate lipreading with a word accuracy of 48.74%. It is interesting that with the phoneme recogniser, word accuracy is greater than phoneme accuracy, because in the viseme recogniser, this is vice versa.

Again, highest accuracy is achieved with Eigenlip features rather than DCT. One interesting observation apparent in Tables 3 and 4 is that the introduction of the DNN makes little difference to the unit accuracy but a bigger difference to a word accuracy for both DCT and eignlips features.

Table 4: Phoneme-based lipreading accuracy(%).

Model	Feature	Phoneme accuracy (%)		Word accuracy (%)	
		SD	SI	SD	SI
CD-GMM + SAT	DCT	28.22	27.37	21.88	17.72
CD-DNN		29.18	28.08	37.40	33.87
CD-GMM + SAT	Eigenlips	31.14	29.59	28.79	24.57
CD-DNN		33.44	31.10	**48.74**	**42.97**

5.3 Discussion

Figure 5 plots all of our experimental results comparing unit accuracies (along the x-axis) against the word accuracies (on the y-axis) along with errorbars showing ± 1 standard error.

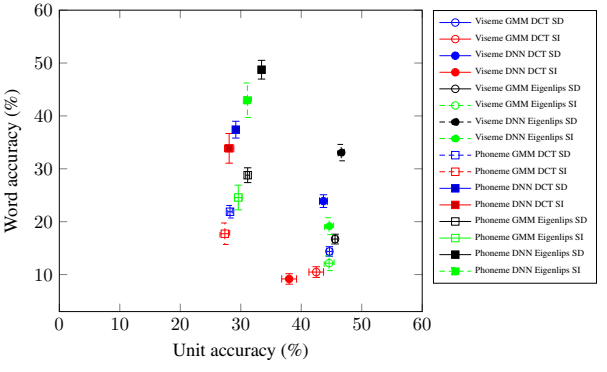

Figure 5: Lipreading system performance in GMM system.

Figure 5 has two clusters: one, in the bottom right, represents the viseme experiments and the other, on the upper left the phonemes. Here we are representing viseme classifiers with circles (filled represents the DNN, open the GMM) and the phonemes with squares (either filled or open depending on the classifier). The colours represent the various SI/SD or DCT/Eigenlips combinations.

The phoneme recogniser naturally obtained lower unit accuracy scores because it has three times more phoneme classes than viseme classes (13 to 38 respectively). But this does not mean that phoneme classes have less power to model a visual gesture. This is visualised in the confusion matrices in Figure 6 where the colour patterns are consistent between phoneme classes (on the left of Fig 6 and between viseme classes on the right of Fig 6).

We note that reducing the set of visual speech units also reduces the discriminant power of the classification model whilst increasing the complexity of pronunciation dictionary by increasing the volume of homophone (homoviseme) words. This suggests that word accuracy of a viseme based system will be less likely to outperform the phoneme based system.

One of the disadvantages of the DNN is that it is not easy to examine to internals of the network to discover from where it is getting its performance. However there is a clue in the previous observation which is that the DNN appears to make the most difference to word accuracy rather than unit accuracy. Visual speech is notorious for extensive co-articulation so the implication is that either there are significant differences in the window length between the GMM and the DNN or the DNN is better able to model co-articulation

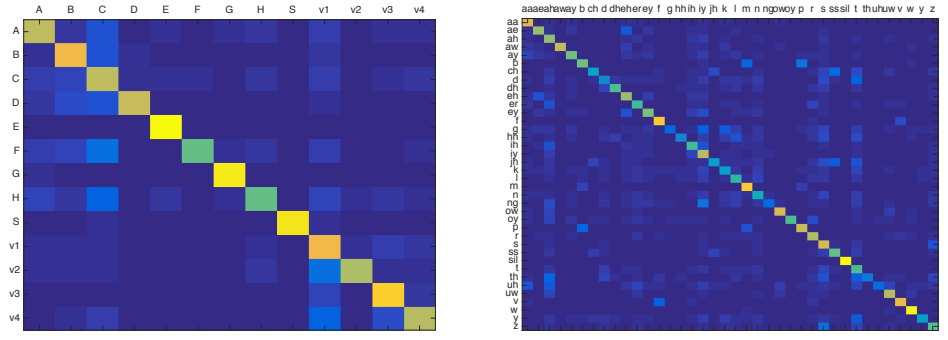

Figure 6: Comparison of visemes confusion matrix (left) vs phomemes confusion matrix (right).

than the GMM. Although there are variations in the window length, here the GMM has a slightly longer span of 150ms compared to 110ms for the DNN, it is latter explanation is the most likely. In other work [23] we were able to use identical features and we also found the DNN superior furthermore we know the DNN to be better able to learn data structured on non-linear manifolds so we believe this is the most likely explanation for the success of the DNN.

One caveat is that we have not optimised the scaling factor used in our language models so there is probably more performance to come when measured as word accuracy.

6 Conclusion

One observation we found is that DNN-HMM viseme recognizers can easily overfit to the training observations, this is shown in the performance disparity between SD and SI configurations. It could potentially be interesting to use visemes as an initialisation for phoneme recognition in a hierarchical training method similar to that in [4] in the future.

We have added more evidence to the argument that phoneme classifiers can outperform those of visemes. Whilst there is still debate about visemes, we can not forget them, but given the evidence showing a significant improvement in word accuracy from the reduction in homophonic words in a pronunciation dictionary, we suggest that phonemes are the current optimal class labels for lipreading.

We have also illustrated the noticeable performance gain by changing visual representation from DCT to Eigenlips. The best word accuracy in this work is 48.74% on SD and 42.97% on SI achieved with the DNN-HMM phoneme unit recognizer trained on Eigenlip features. However, the disadvantage of Eigenlip feature is a learned linear mapping that needs to be trained.

Conventional systems have shown speaker independence to be a challenge, here with a novel DNN-HMM architecture, we have reduced the effect between these arrangements. We speculate that the success of the DNN is likely to do its ability to better model the effects of co-articulation which is a well known bugbear of human and machine lip-readers.

References

[1] Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[2] I. Almajai, S. Cox, R. Harvey, and Y. Lan. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2722–2726, March 2016. doi: 10.1109/ICASSP.2016.7472172.

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

[4] Helen L Bear and Richard Harvey. Decoding visemes: Improving machine lip-reading. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2009–2013. IEEE, 2016.

[5] Helen L Bear, Richard W Harvey, Barry-John Theobald, and Yuxuan Lan. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In *Advances in Visual Computing*, pages 230–239. Springer, 2014. doi: 10.1007/978-3-319-14364-4_22.

[6] Christoph Bregler and Yochai Konig. "Eigenlips" for robust speech recognition. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 2, pages II–669. IEEE, 1994.

[7] Stephen J Cox, Richard W Harvey, Yuxuan Lan, Jacob L Newman, and Barry-John Theobald. The challenge of multispeaker lip-reading. In *AVSP*, pages 179–184, 2008.

[8] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936.

[9] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75 – 98, 1998. ISSN 0885-2308.

[10] N. Harte and E. Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, May 2015. ISSN 1520-9210. doi: 10.1109/TMM.2015.2407694.

[11] Robert Hecht-Nielsen et al. Theory of the backpropagation neural network. *Neural Networks*, 1(Supplement-1):445–448, 1988.

[12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[13] Dominic Howell, Stephen Cox, and Barry Theobald. Visual units and confusion modelling for automatic lip-reading. *Image and Vision Computing*, 51:1–12, 2016.

[14] J Jeffers and M Barley. Lipreading (speechreading). *Charles C. Thomas, Springfield, IL*, page b10, 1971.

[15] M Kirby, F Weisser, and G Dangelmayr. A model problem in the representation of digital image sequences. *Pattern Recognition*, 26(1):63 – 73, 1993. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/0031-3203(93)90088-E. URL http://www.sciencedirect.com/science/article/pii/003132039390088E.

[16] Jianchang Mao and Anil K Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE transactions on neural networks*, 6(2):296–317, 1995.

[17] I Matthews, TF Cootes, JA Bangham, SJ Cox, and RW Harvey. Extraction of visual features for lipreading. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002. ISSN 0162-8828. doi: 10.1109/34.982900.

[18] Neti, Potamianos, Luettin, Matthews, Glotin, Vergyri, Sison, Mashari, and Zhou. Audio-visual speech recognition, 2000. Technical report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore.

[19] Gerasimos Potamianos, Juergen Luettin, and Chalapathy Neti. Hierarchical discriminant features for audio-visual LVCSR. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 165–168. IEEE, 2001.

[20] Daniel Povey and George Saon. Feature and model space speaker adaptation with full covariance Gaussians. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. The Kaldi speech recognition toolkit. In *In IEEE 2011 workshop*, 2011.

[22] Lawrence Rabiner and B Juang. An introduction to hidden Markov models. *IEEE ASSP magazine*, 3(1):4–16, 1986.

[23] Kwanchiva Thangthai and Richard Harvey. Improving computer lipreading via dnn sequence discriminative training techniques. In *INTERSPEECH 2017 (to be published)*, 2017.

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

[25] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[26] Inc. Xilinx. 2d discrete cosine transform (dct) v2.0 logicore product specification, 2002.