Interspeech 2018
2-6 September 2018, Hyderabad

Building large-vocabulary speaker-independent lipreading systems

Kwanchiva Thangthai[1,2], *Richard Harvey*[1]

[1]School of Computing Sciences, University of East Anglia, Norwich, UK
[2]National Electronics and Computer Technology Center, Thailand

k.thangthai@uea.ac.uk, r.w.harvey@uea.ac.uk

Abstract

Constructing a viable lipreading system is a challenge because it is claimed that only 30% of information of speech production is visible on the lips. Nevertheless, in small vocabulary tasks, there have been several reports of high accuracies. However, investigation of larger vocabulary tasks is much rarer.

This work examines constructing a large vocabulary lipreading system using an approach based-on Deep Neural Network Hidden Markov Models (DNN-HMMs). We tackle the problem of lipreading an unseen speaker. We investigate the effect of employing several steps to pre-process visual features. Moreover, we examine the contribution of language modelling in a lipreading system where we use longer n-grams to recognise visual speech. Our lipreading system is constructed on the 6000-word vocabulary TCD-TIMIT audiovisual speech corpus. The results show that visual speech recognition can definitely reach 50% word accuracy on large vocabularies. We actually achieved a mean of 53.83% measured via three-fold cross-validation on the speaker independent setting of the TCD-TIMIT corpus using bigrams.

Index Terms: lipreading, deep neural network (DNN)

1. Introduction

Speechreading may be a natural way of silent speech communication between humans but, compared to the audio signal, the video signal is impoverished. Various works including [1, 2, 3, 4] estimate that only about 30% of speech production information is visible on the lips as the vocal cords, nasal cavity, oral cavity are mostly hidden. This leads to *homopheneous* words which look the same on the lips but sound different (words such as bat /bæt/ and mat /mæt/ are often perceived to be identical by lip readers).

Recent advances in computer vision, speech processing and machine learning ought to feed-in to better lipreading systems. Of these advances, deep-learning is the most prominent. In a small vocabulary task, for example, lipreading systems via convolutional neural network (CNN) features and attention-based encoder-decoders achieved word accuracies of 97% [5] on the GRID dataset [6]. An end-to-end lipreading system using Long-Short Memory (LSTM) networks on the OuluVS2 [7] dataset achieved 84.5% phrase accuracy [8]. However, in larger vocabulary tasks, the accomplishment of lipreading is much lower even if a complex deep learning approach has been employed. In the MV-LRS task [9], the word accuracy of lipreading is reported as 43.6% in frontal view, and 37.2% in profile view using sequence-to-sequence LSTMs. In the LRS task [5], a lipreading system achieved 49.8% word accuracy using a system called Watch Attend and Spell (WAS) which involves CNNs and multiple LSTMs.

In this work, we employ hybrid DNN-HMMs (as in [10, 11, 12]) but here, the visual feature is based-on a Deep AutoEncoder (DAE) [13, 14, 15]. The results are further processed using an n-gram language model.

2. Data

The most used of the large-vocabulary datasets is TCD-TIMIT [16] which a publicly-available audio-visual continuous speech corpus that has a 6019 word vocabulary recorded from 59 talkers (the volunteer set) and three professional lip speakers comprising over seven hours of speech data. This is a read-speech corpus captured in a studio environment. The video is recorded in two views: frontal and 30° view. We use the only frontal view from the volunteer set. Each talker reads 98 sentences selected from TIMIT. We also follow the provided lists of non-overlapping utterances for training and evaluation in two scenarios: speaker-dependent (SD) and speaker-independent (SI) scenarios. In SD where training and test speakers are overlapped, there are 3752 utterances for training and 1736 utterances for evaluation. In SI, the training set contains 3822 utterances of 39 speakers, and the evaluation set contains 1666 utterances of 17 speakers. Statistics of the volunteer set are available in Table 1.

Table 1: *Statistics from the volunteer-speakers dataset of TCD-TIMIT corpus [16].*

TCD-TIMIT statistics	**Volunteer set**
Total number of speakers	59
Total number of sentences	5,488
Total number of unique phonemes	38
Total number of phonemes tokens	213,115
Total number of unique words	5,958
Total number of word tokens	47,503
Average number of phonemes per sentence	38.83
Average number of words per sentence	8.65
Average number of phonemes per word	4.48

3. Lipreading system

Our lipreading systems are developed on the Kaldi toolkit [18]. We build lipreading systems via a weighted finite state transducer (WFST) decoder where we use hybrid DNN-HMMs [17] instead of GMM-HMMs. The WFST decoder is comprised of HMMs (H), phoneme context-dependency (C), lexicon model (L), and n-gram language model (G), called collectively the HCLG decoding-graph. To decode lipreading, the first-pass decoder generates a word lattice containing a possible set of words that matched the input lip signal; then the final result comes from the lattice re-scoring via a language model. We utilise the Deep autoencoder (DAE) method to extract a static feature from a cropped lip image. Deep autoencoders are feed-forward neural networks that learn non-linear mapping to reconstruct the input with minimum errors. We then transform the visual features

10.21437/Interspeech.2018-2112

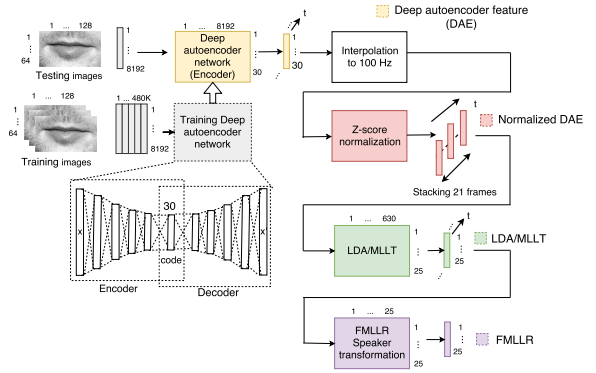


Figure 1: *Feature extraction and feature processing methods in the lipreading system.*

Figure 2: *Examples of the original lip ROIs (a) taken from speaker 02M in the TCD-TIMIT corpus and its reconstruction via (b)* 30*-dimensional DAE, (c)* 44*-dimensional DCT.*

to represent linguistic units using LDA/MLLT, and we also use FMLLR to transform feature space for each specific speaker.

3.1. Deep autoencoder and feature processing

Our visual feature is an appearance-based so extracted from the greyscale pixels of lip regions-of-interest (ROIs) which are available in the TCD-TIMIT corpus. The deep autoencoder (DAE), an unsupervised technique, reduces the 64 $\times$ 128 pixels grey-scale lip ROI to a 30-dimensional feature vector. The network structure is separated into two parts: decoder and encoder. The layer in the middle which usually contains the small number of units, i.e. 30 hidden-units, is a low-dimensional representation that is trained to yield the best reconstruction of the output. The DAE network, shown in Figure 1, composes of 11 hidden layers where the units in the encoder layer are (1024, 512, 256, 128, 64) and the units in the decoder layer are (64, 128, 256, 512, 1024) and 30 units in the code layer. We use ReLU activation function in each unit in the hidden layers and use linear unit in the code layer. The DAE model is trained on 480k images obtained from a training set and optimised with the mean square error (MSE) loss-function via Adam optimisation algorithm [19] using 50 epochs and mini-batch size 256. Reconstructions of lip ROIs from the 30-dimensional DAE feature, are shown in Figure 2 (b). They are higher image quality than reconstruction from the 44-dimensional discrete cosine transform (DCT) feature (c) as presented in TCD-TIMIT baseline [16].

We use three feature processing steps: (1) z-score normalization, (2) Linear discriminant analysis and maximum likelihood linear transform (LDA/MLLT), (3) and Feature space maximum likelihood linear regression (FMLLR) [20] transformation. Previous reports in lipreading use different sizes of LDA context window i.e. $\pm$3 [10, 21], $\pm$7 [22, 11], and 40 dimensions were retained. In LDA/MLLT, we use dynamic information covering 21 frames window (stacking $\pm$10 frames) and retain 25 dimensions since it obtains the best result in our preliminary test. The FMLLR feature also retains 25 dimensions.

4. Visualizing visual speech features

We analyse visual representations using t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisation techniques, introduced by [23]. T-SNE is a tool for visualising high-dimensional data. It reduces feature dimension where it groups the similarity data points and distances between the dissimilarity, by computed the similarity matrix of data points and converting it into a joint probability which is then minimised via the Kullback-Leibler divergence between the joint probabilities of low-dimensional and original high-dimensional data. To observe the distribution of features, we use t-SNE to visualise the 30-dimensional DAE visual feature and the 39-dimensional MFCC acoustic feature. There are 2841 data points of utterance *Don't ask me to carry an oily rag like that* extracted from six speakers: three male and three female.

4.1. Understanding visualization output via Fisher-ratio (F-ratio) analysis

We measure the t-SNE output by computing the magnitude of the class discriminant ratio using Fisher's Ratio analysis by considering the speaker class label and a linguistic class label. The class discriminant ratio is the ratio of the between-classes variance and the within-classes variance. The class discriminant ratio can be computed by

$$Class\ F-ratio = \frac{S_B}{S_W}, \tag{1}$$

where S_B is the between-classes covariance matrix and S_W is the within-classes covariance matrix. The definition of covariance matrices are:

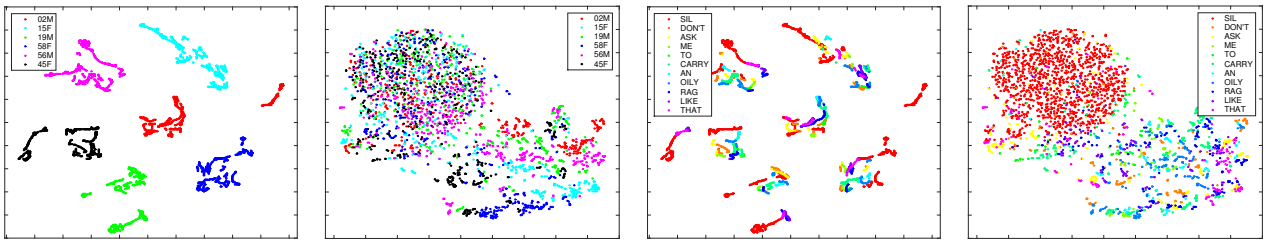
$$S_B = \sum_c (\mu_c - \bar{x})(\mu_c - \bar{x})^T, \tag{2}$$

$$S_W = \sum_c \sum_{i \epsilon c} (x_i - \mu_c)(x_i - \mu_c)^T, \tag{3}$$

where μ_c refers to the mean of each class and x_i refers to each data point. This F-ratio identifies the goodness of the data clustered regarding the provided class label. In our case we also use this ratio to identify whether the data are clustered by speaker similarity or linguistic similarity. For instant, if the F-ratio of speakers is higher than the F-ratio of linguistic units, we assume that the feature represents speaker similarity and vice versa. Note that we compute three F-ratios using speaker label (spk), and two linguistic labels which are word label (w) and phonetic label (ph).

4.2. Comparison of visual and auditory representations.

Figure 3 illustrates the difference of the original representation between the visual and acoustic speech features. In Figure 3 (a) and (b) the data points are labelled (with colours) by speakers, while in (c) and (d) the data points are labelled (coloured) by words. In these data, variations in the environment, and phonological patterns are controlled so variation is attributable to speaker identity. The acoustic and visual features look quite different – in the visual we have clusters from different speaker identities, but the acoustic features have largely removed identity as we can see from the mixed-colour clusters. This is confirmed by the speaker F-ratio (3.56). In other words, features



(a) DAE visual feature Speaker F-ratio=3.56

(b) MFCC feature Speaker F-ratio=0.08

(c) DAE visual feature Word F-ratio=0.03 Phonetic F-ratio=0.03

(d) MFCC feature Word F-ratio=1.55 Phonetic F-ratio=2.38

Figure 3: *T-SNE plots of visual speech and auditory speech features coloured by speakers (a,b) and words (c,d). The class F-ratios of speaker, word and phonetic are provided underneath the plot.*

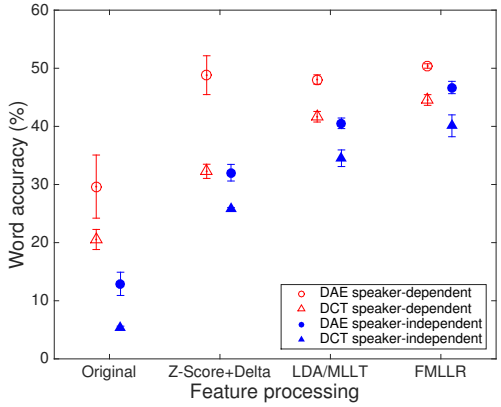


Figure 4: *Word accuracy (%) of speaker-dependent (red) and speaker-independent (blue) lipreading systems via DNN-HMMs models. Results are obtained from DAE feature (circle) and DCT feature (triangle) as a function of feature transformation methods with* ± 1 *standard error. Note that we show DCT results as for a reference.*

extracted from lip ROIs are highly speaker dependent. The red points in (c) and (d) show silence which is well-clustered in the acoustic and, again, separated by identity in the visual. In the acoustic data, a slight effect of speaker dependency can be found in the non-silence classes.

5. Experiments

The DNN-HMM visual speech models are trained on six hidden layers and with sigmoid non-linearity 2048 units per layer. We initialise the DNN parameters using the Restricted Boltzmann Machines (RBMs) pre-training and optimise via the standard cross-entropy (CE). For each input feature type, we splice ± 5 consecutive frames as dynamic features covering 11 frames context. We use the constrained time alignments that generated from visual-only GMM-HMMs with speaker adaptive training system (SAT) [24]. The evaluation task is the large vocabulary continuous speech using the TCD-TIMIT corpus. There are two scenarios: speaker dependent (seen speakers), and speaker independent (unseen speakers). We report the word accuracy on the mean of three-fold cross-validation (with ± 1 standard error) where we use the recommended set as the first-fold and we prepare another two-fold by retaining the similar proportion. The DAE feature is created from the training images for each particular cross-validation set and scenario.

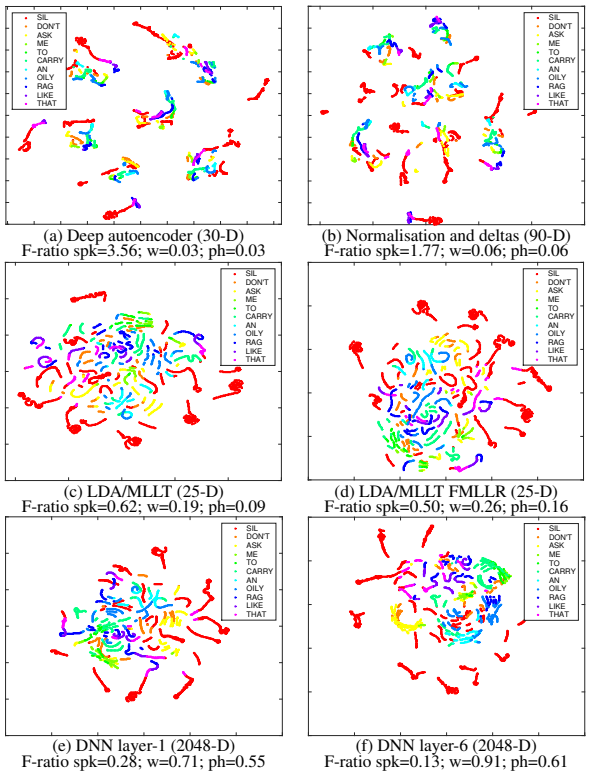
5.1. Effect of feature preprocessing

This experiment investigates the effect of feature transformation in DNN-HMM training. We utilize three feature transformation methods: (1) z-score normalization and deltas ($\Delta + \Delta\Delta$); (2) phonetic class discriminant feature via LDA/MLLT and (3) speaker-transformed features via FMLLR. In Figure 4, there is a clear trend of increasing word accuracy from each step of feature transformation. This trend applies to both DAE and DCT features. The original untransformed features have the lowest performance in both scenarios. The dependency of speaker identity (noted in Figure 3) is evident in the difference in performance between the SI and SD systems which is around 15%. Applying feature normalisation and deltas is highly beneficial to both SD and SI word accuracy, especially with DAE features where the word accuracy improves by almost 20% compared to the original results. Next, we found that LDA/MLLT and FMLLR help minimise the speaker identity effect and enhance word accuracy which is shown by the small gap between the performance on seen and unseen speakers. The explanation for the LDA/MLLT is that it transforms the feature space to satisfy the class discrimination which indirectly reduces speaker identity. The FMLLR is directly proposed to solve the speaker identity effect. Overall the highest word accuracy of speaker-independent lipreading is 46.69% using the DAE feature with the FMLLR transformation method.

This situation is further illustrated in Figure 5 which shows the t-SNE plot at various stages in the system. In (a) is show the visualisation before any normalisation, so we see the speaker identity clusters. In (b) the effect of normalisation and deltas is to slightly reduce the effects of speaker identity. In (c) and (d) which show the effect of LDA/MLLT and FMLLR the identity clusters have disappeared and word clusters are emerging. The effect of the DNN is illustrated in Figure 5 (e) and (f) which shows, as we move up the network the words become more strongly defined which can also be seen in the F-ratios.

5.2. Results of sequence discriminative training via sMBR

We employ state-level minimum Bayes risk (sMBR) sequence discriminative training method [25] to improve visual speech modeling, as it yielded 51.29% of SI word accuracy in lipreading in TCD-TIMIT corpus [11]. In Table 2, we achieve 53.83% word accuracy in the unseen scenario which is a 7.14% improvement from the DNN-HMMs. This is 2.54% higher than the earlier reported SI result based on the Eigenlips feature.



(a) Deep autoencoder (30-D)
F-ratio spk=3.56; w=0.03; ph=0.03

(b) Normalisation and deltas (90-D)
F-ratio spk=1.77; w=0.06; ph=0.06

(c) LDA/MLLT (25-D)
F-ratio spk=0.62; w=0.19; ph=0.09

(d) LDA/MLLT FMLLR (25-D)
F-ratio spk=0.50; w=0.26; ph=0.16

(e) DNN layer-1 (2048-D)
F-ratio spk=0.28; w=0.71; ph=0.55

(f) DNN layer-6 (2048-D)
F-ratio spk=0.13; w=0.91; ph=0.61

Figure 5: *T-SNE plots of DAE with various transformation methods and a representation inside a DNN layer presented with three levels of F-ratio - speaker (spk), word (w), and phonetic (ph).*

Table 2: *DAE lipreading results using DNN-HMMs and DNN-HMMs with* 10*-iterations sMBR training.*

sMBR training	Word accuracy (%) SD testset	SI testset
Without sMBR	50.39	46.69
With sMBR	57.36	53.83

5.3. Effect of language modelling

A language model (LM) can be used to constrain word combinations to form legitimate sentences or sentence fragments. It is usually learnt from the training text. We use five word n-gram language models: uniform prior; unigram; bigram (mainly used); trigram and 4gram. The term uniform prior means that we use no language model (a unigram model with uniform probabilities). The results in Table 3 illustrate that the n-gram order of language modeling contributes to noticeable changes in lipreading performance. Lipreading performance gets below 10% without an LM, but the word accuracy increases significantly to about 68% when we use trigrams and 4grams. These observations are consistent with what is known about human lipreaders who make considerable use of their linguistic and domain knowledge. We also evaluate if the language model dominates the lipreading performance by decoding a random noise vector. Results in the guessing column indicate that language modeling has successfully increase lipreading accuracy only in combination with a suitable associated visual input signal.

Table 3: *Word accuracy (%) of lipreading system decoded with difference language models.*

Word-based n-gram language model (LM)	Word accuracy (%) SI testset	Guessing
uniform prior LM	6.24	1.63
unigram LM	10.69	2.04
(currently use) bigram LM	53.83	2.02
trigram LM	67.69	2.03
4gram LM	68.45	2.02

6. Discussion and conclusions

Among other things, this study demonstrates that lipreading systems can be built via the conventional techniques of acoustic speech recognition system based-on DNN-HMMs and sMBR training. In a 6000-word vocabulary task, we achieved 53.83% word accuracy in the speaker independent scenario using DAE features and the FMLLR transformation method. This is the best known accuracy on the TCD-TIMIT data. The results and the visualization of each feature indicate that feature processing steps are relevant to gain speaker-independent lipreading accuracy because they reduce the influence of speaker identity found in the original space of the DAE feature.

Table 4: *Examples of challenging to predict words and easy to predict words*

Examples of the difficult to predict words (less than 10% correct) High frequency words	Low frequency words
A (ah)	YET (y eh t)
IT (ih t)	TELL (t eh l)
I (ay)	DESSERT (d ih z er t)
IN (ih n)	ORDER (ao r d er)
IS (ih z)	DOCTOR (d aa k t er)
YOU (y uw)	SURE (sh uh r)
HE (hh iy)	MUSTARD (m ah s t er d)
DOES (d ah z)	CHURCH (ch er ch)
THEM (dh eh m)	MARINE (m er iy n)
AT (ae t)	HOUSE (hh aw s)
Examples of the easy to predict words (more than 90% correct) High frequency words	Low frequency words
SHE (sh iy)	SHORTAGE (sh ao r t ah jh)
CAN (k ae n)	BECOME (b ih k ah m)
YEAR (y ih r)	STEEP (s t iy p)
WOULD (w uh d)	BOB (b aa b)
OILY (oy l iy)	REDWOODS (r eh d w uh d z)
SMALL (s m ao l)	OUTDOORS (aw t d ao r z)
BROTHER (b r ah dh er)	BRIGHT (b r ay t)
MAKES (m ey k s)	WIRE (w ay er)
SELDOM (s eh l d ah m)	LET (l eh t)
OVER (ow v er)	INCREASES (ih n k r iy s ah z)

Table 4 gives example words which are relatively easy or difficult to lipread. There are two effects at play: firstly there is homopheniosity or the confusion of words because they have identical shapes on the lips and secondly there is the observation that certain sounds are more visible on the lips than others. Words such as “brother” and “makes” have bilabials at the start of the word which makes them easier to spot than “at” or“he”. Longer words are easier to lip-read than shorter ones, and the homophene effect means the classifier has to guess from a considerable number of alternatives (which might explain why some of the low-frequency difficult words still contain bilabials – albeit weakly enunciated bilabials such as found in “marine” or “mustard”).

Although accomplishment of computer lipreading is dependent on the degree of n-gram language model, we observe that language modelling does not dominate the entire lipreading decoder a fact verified by poor the results of decoding the random signal.

7. References

[1] G. H. Nicholls and D. L. Mcgill, "Cued speech and the reception of spoken language," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 2, pp. 262–269, 1982.

[2] I. d. l. R. R. Ortiz, "Lipreading in the prelingually deaf: what makes a skilled speechreader?" *The Spanish journal of psychology*, vol. 11, no. 2, pp. 488–502, 2008.

[3] N. Bauman, "Speechreading (lip-reading)," http://hearinglosshelp.com/blog/speechreading-lip-reading, 2011, accessed on 23rd March 2018.

[4] L. Southwick and M. Vacala, "Chapter 31 - patients with disabilities," in *Physician Assistant*, 4th ed., R. Ballweg, E. M. Sullivan, D. Brown, and D. Vetrosky, Eds. Philadelphia: W.B. Saunders, 2008, pp. 593 – 606.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[7] I. Anina, Z. Zhou, G. Zhao, and M. Pietikinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, May 2015, pp. 1–5.

[8] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMS," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2592–2596.

[9] J. S. Chung and A. Zisserman, "Lip reading in profile." BMVC, 2017.

[10] 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)*, March 2016, pp. 2722–2726.

[11] K. Thangthai and R. Harvey, "Improving computer lipreading via DNN sequence discriminative training techniques," in *Proc. Interspeech 2017*, 2017, pp. 3657–3661.

[12] M. H. Rahmani and F. Almasganj, "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features," in *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, April 2017, pp. 195–199.

[13] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A. R. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Interspeech 2010*, 2010, pp. 1692–1695.

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[15] K. Paleček, "Extraction of features for lip-reading using autoencoders," in *International Conference on Speech and Computer*. Springer, 2014, pp. 209–216.

[16] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[18] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[20] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[21] A. H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," in *Proc. Interspeech 2017*, 2017, pp. 3752–3756.

[22] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept 2003.

[23] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[24] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 2 - Volume 2*, ser. ICASSP '97. Washington, DC, USA: IEEE Computer Society, 1997.

[25] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, 2013, pp. 2345–2349.