Improving computer lipreading via DNN sequence discriminative training techniques

Kwanchiva Thangthai, Richard Harvey

School of Computing Sciences, University of East Anglia, Norwich, UK

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

Abstract

Although there have been some promising results in computer lipreading, there has been a paucity of data on which to train automatic systems. However the recent emergence of the TCD-TIMIT corpus, with around 6000 words, 59 speakers and seven hours of recorded audio-visual speech, allows the deployment of more recent techniques in audio-speech such as Deep Neural Networks (DNNs) and sequence discriminative training.

In this paper we combine the DNN with a Hidden Markov Model (HMM) to the, so called, hybrid DNN-HMM configuration which we train using a variety of sequence discriminative training methods. This is then followed with a weighted finite state transducer. The conclusion is that the DNN offers very substantial improvement over a conventional classifier which uses a Gaussian Mixture Model (GMM) to model the densities even when optimised with Speaker Adaptive Training. Sequence adaptive training offers further improvements depending on the precise variety employed but those improvements are of the order of 10% improvement in word accuracy. Putting these two results together implies that lipreading is moving from something of rather esoteric interest to becoming a practical reality in the foreseeable future.

Index Terms: visual-only speech recognition, computer lipreading

1. Introduction

Notwithstanding long-term interest in human lipreading [1] and several sci-fi predictions that computer lipreading would be an easily realisable reality[1] it turns out to have been a tricky problem. And this is despite the observation that lipreading would be of practical use in silent speech interfaces or of utility to people who had lost the use of their vocal chords. So far then, many researchers have concentrated on audio-visual recognition, often in noise. Noise causes the audio recognizer to fail and lipreading then helps. In terms of advancing the art of lipreading, there is now a library of techniques based around DNNs that might be useful, but the lack of data has been discouraging to people who wish to map those methods across to the visual domain. Hence many lipreading system still report on isolated digits or letters, or small vocabulary tasks such as digit strings or 100-word vocabulary tasks.

Table 1 shows some more realistic and recent data. It also shows the best performance, measured as word accuracy, for isolated (I) and continuous speech recognition (C) task. Also shown are the number of talkers. The first system to report on a large vocabulary task, IBM ViaVoice [11], was devised in 2000 and reports a word accuracy of 48.92% on a 10,400 word vocabulary. Unfortunately, it was not a full lipreading system since it used the visual model to rescore a lattice produced from noisy audio. More recently, using data recorded from the BBC news, the LRW task, the best result was 84.5% accuracy but was achieved on a small vocabulary and isolated words. For larger vocabularies the word accuracy drops as does, often, the number of talkers. For example in RM-3000 and LiLiR, a continuous lipreading task with a DNN-HMM hybrid architecture, for the single-speaker 1000-vocabulary, 3000-word-utterance database, RM-3000 accuracies of 76.14% [4, 13] and 85.67% [5] are reported. For the AVICAR data [9], which consists of isolated-words, connected-digit and continuous speech tasks, the word accuracies range between 24.53% and 33% on combined 4-camera using multi-stream HMM. Recently [3], reports using an end-to-end deep learning system with 49.8% word accuracy on 4960 hours of BBC news audiovisual speech data (the LRS task). The data contain 118k utterances recorded from thousands of speakers with a vocabulary size 17,428 words. They also report on a similar task, a professional lipreader performed only 26.2% word accuracy [2].

Table 1: *Medium-sized lipreading databases*

Corpus	ASR task	Talker	Vocab size	Utt	Best word accuracy (%)
LRW [2]	I	>100	500	500k	84.50 [3]
RM-3000 [4]	C	1	1000	3k	84.67 [5]
LiLiR [6]	C	12	1000	2.4k	∼53.00 [7]
AVICAR[8]	I and C	100	1356	59k	∼33.00 [9]
TCD-TIMIT [10]	C	59	5958	5.4k	N/A
IBM ViaVoice [11]	C	290	10400	18k	48.92 [11, 12]
LRS [3]	C	>1000	17428	118k	49.80 [3]

When it comes to deep learning there are three different approaches: 1. the deep feature with conventional HMM as in [14, 15]; 2. the end-to-end approach as in [2, 3]; 3. the hybrid DNN-HMM approach with FST decoder. However in [16] there seems to be little performance difference between the approaches. In this paper we choose the hybrid approach since it allows us to investigate the effect of sequence discriminative training. It is known that in audio, sequence discriminative training has gained a substantial 3-17% relative performance of a speech transcription system [17], 7-9% in [18] and 8.4% in [19] versus Cross Entropy (CE) training. Unlike CE, which minimizes the expected frame error individually, sequence discriminative training takes the inter-frame sequence into consideration. Here, we investigate three types of the sequence discriminative training criterion; Maximum Mutual Information (MMI), Minimum Phone Error (MPE) and state-level Minimum Bayes Risks (sMBR). We deploy the hybrid DNN with a fully automatic data driven approach which means no audio-to-phoneme alignment is needed.

[1] The 1968 movie "2001: a space odyssey" for example.

[2] This disparity between human and automatic systems has also been reported in [6].

2. Corpus and dataset

Looking at Table 1 and eliminating datasets that are either isolated words (I) or not available (IBM ViaVoice or LRS) we are left with TCD-TIMIT [10] as the largest available with 59 talkers and 3 professional lip speakers comprising over seven hours of speech data. The video is recorded in two views: frontal and 30° view captured in a studio environment with Sony PMW-EX3 cameras and wireless clip-on microphone. We use only the frontal view of 59 volunteer talkers. Each talker read 98 sentences selected from TIMIT. The majority of talkers (56) have an Irish accent. The remaining three talkers are removed as prescribed in [10]. Thus the total number of utterances used in this paper are 5488 captured from 56 speakers. 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 detailed in Table 2. [10]

Table 2: *Details of training set and evaluation set in TCD-TIMIT corpus (volunteer speakers) used in our experiments*

Dataset		Number of speaker			# Utter	# Word token	# Vocab
		Male	Female	Total			
SD	Train	29	27	56	3752	30739	4959
	Eval	29	27	56	1736	14360	3511
SI	Train	22	17	39	3822	31290	5180
	Eval	7	10	17	1666	13809	3388
Total		29	27	56	5488	47503	5958

provides a preliminary report of the accuracy using 12 viseme classes: the best results were 34.54% and 34.77% viseme accuracy in SD and SI respectively. However, authors vary considerably in their choice of viseme, so for comparative studies, viseme classlifcation and accuracy are highly problematic [20] so, here we use word accuracy.

3. Visual speech DNN-HMMs Training

This section describes the visual speech modeling method using the hybrid DNN-HMM structure. For acoustic speech recognition, the hybrid DNN-HMM structure is known to provide significant performance gains over the standard GMM-HMM [21, 22, 23]. There have also been some preliminary applications to lipreading systems [7, 5].

In the DNN-HMM hybrid approach, let $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_T$ be the T sequence of feature vectors extracted from each video and $\mathbf{w}$ be a word sequence that represented by a language model. The likelihood of input sequence can be computed by

$$p(\mathbf{X}|\mathbf{w}) = \prod_{t=1}^{T} p(\mathbf{x}_t|s_t)p(s_t|s_{t-1}), \tag{1}$$

where $p(\mathbf{x}_t|s_t)$ denotes the emission probability and $p(s_t|s_{t-1})$ is the transition probability obtained from the HMMs state transition. The emission probability can be approximated by $p(\mathbf{x}_t|s_t) = p(s_t|\mathbf{x}_t)p(\mathbf{x}_t)/p(s_t)$, via a Gaussian Mixture Model (GMM), in which case we have the conventional HMM speech recognition architecture or, via a DNN. To estimate the DNN's posterior $p(s|\mathbf{x}_t)$ on each state of an utterance u, the DNN uses a pseudo log-likelihood obtained via the softmax activation function

$$p(s|\mathbf{x}_{ut}) = \frac{\exp\{a_{ut}(s)\}}{\sum_{s'} \exp\{a_{ut}(s')\}}, \tag{2}$$

where $a_{ut}(s)$ refers to an activation of state s at the output layer. In which case, the pseudo log-likelihood of the visual speech model is

$$\log p(\mathbf{x}_{ut}|s) = \log p(s|\mathbf{x}_{ut}) - \log p(s), \tag{3}$$

DNNs used in visual speech modeling have conventionally been trained to optimize the cross-entropy (CE) between the prediction and the target HMM-state labels using mini-batch Stochastic Gradient Descent (mini-batch SGD) optimization and error back-propagation (BP) algorithms [24], to provide the posterior probability estimated of the HMM states. The HMM-state alignments are obtained from a GMM-HMM training process. Here, we use BP to minimize the cross-entropy between the predicted output and the HMM-state target. This is similar to DNN-HMM training in acoustic speech recognition as in [25].

The cross-entropy objective is a frame-level training criterion for classification tasks and usually provides significant performance over standard GMM-HMM acoustic modeling in speech recognition. In visual speech model training, we use the frame level alignment generated from a context-dependent GMM-HMM system and the initial DNN-HMM parameters via stacking Restricted Boltzmann Machines (RBMs) pretraining [25]. We use CE to fine-tune the DNN parameters.The CE objective function is defined as

$$\mathcal{F}_{CE} = -\frac{1}{T}\sum_{u=1}^{U}\sum_{t=1}^{T_u}\sum_{s} l_{ut}(s)\log p(s|\mathbf{x}_{ut}), \tag{4}$$

where the T here is the total number of frames from all training utterances and $l_{ut}(s)$ is the Kronecker delta of the target state.

4. Sequence-discriminative training

CE is the most common objective function to construct a classification based DNN-HMM model but it is based on a frame-by-frame comparison. For lipreading where co-articulation and context are important, effective training of a DNN-HMM model implies consideration of a longer window. Sequence-discriminative training techniques fine-tune the existing DNN parameters, initially trained by CE, by using sequence-level criteria which take into consideration the HMM topology and language model. There are some reports in speech recognition system that apply the sequence-discriminative training in DNN acoustic model [18, 26] and also RNN-LSTM acoustic model [19, 27]. This work examines three criteria for sequence-discriminative training of the DNN visual speech model: maximum mutual information (MMI); state-level minimum Bayes risk (sMBR) and minimum phone error (MPE).

The MMI training criterion [28, 29] aims to maximize the mutual information between the distributions of observation and the reference word sequences. Let $\mathbf{X}_u$ represent the sequence of visual features and w_u is the word reference in an utterance u. MMI attempts to maximise

$$\mathcal{F}_{MMI} = \sum_{u} \log \frac{p(\mathbf{X}_u|S_u)^k P(\mathbf{w}_u)}{\sum_{\mathbf{w}} p(\mathbf{X}_u|S_{\mathbf{w}})^k P(\mathbf{w})}, \tag{5}$$

where S_u is the state sequence corresponding to the correct word $\mathbf{w}_u$ and k is the model scaling factor. The sum of denominator is practically computed from a decoding lattice instead of all the possible word to enhance the computational efficiency, where the decoding lattice generates via the weak language model. We also apply the frame rejection proposed by [18] to avoid infinite gradients, caused by missing words in the denominator lattice.

The sMBR/MPE training criteria aim to minimize the expected error, measured at state-level (sMBR [30]) or phone-level (MPE, [31]), between the sequence of visual features and the word sequence of each training utterance. Specifically, sMBR/MPE attempts to minimize

$$\mathcal{F}_{MBR/MPE} = \sum_u \log \frac{\sum_{\mathbf{w}} p(\mathbf{X}_u|S_{\mathbf{w}})^k P(\mathbf{w}) A(\mathbf{w}, \mathbf{w}_u)}{\sum_{\mathbf{w}'} p(\mathbf{X}_u|S_{\mathbf{w}'})^k P(\mathbf{w}')}, \tag{6}$$

where $A(\mathbf{w}, \mathbf{w}_u)$ is the raw accuracy between the word sequence $\mathbf{w}$ and the reference $\mathbf{w}_u$. The raw accuracy refers to the number of correct state labels in sMBR and the phone labels in MMI.

5. Decoding lipreading

A language model (LM) and lexicon model are essential to a speech recognizer and lipreading system. The lexicon model used in this work is the Irish accent phoneme pronunciation dictionary provided in TCD-TIMIT corpus that contains 156,516 word entries adapted from the CMU dictionary. The LM helps discriminate similar input patterns of words found in the lexicon and also reduce the search cost. For the LM, we use a the statistical based n-gram model where we train a word bi-gram from TCD-TIMIT provided text. To make it fair we use only text provided in the training set, thus we have two bi-gram LMs; one for SD and one for SI. We know already that longer n-grams mean better performance but extending the number of word n-grams can be too strict and will lead to difficulty in finding an appropriate parameter for visual speech modeling. Here we evaluate our LM by computing the perplexity of SD (35.16) and SI (33.10) evaluation sets against their LM with no out-of-vocabulary words found in both cases.

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 [32]. The decoding graph represents the components of a speech recognizer as an FST with weights (a WFST). It contains a set of context-dependent states with weighted arcs between the individual states. The weights are the incorporation of the visual speech model and the LM scores. The visual speech model score comes from the DNN and the LM score from the bi-gram LM. The graph is generated by HCLG $= \min(\det(H \cdot C \cdot L \cdot G))$ where $\cdot$ means WFST composition of HMM structure (H), phonetic context-dependency (C), lexicon (L) and language model or grammar (G). Since FSTs are finite-state machines, they operate on symbols where the input symbols correspond to context-dependent HMM states and the output symbols are words. To decode, each arc of HCLG is traversed for each input feature and state-level arcs are created for the visual speech cost and the graph costs so called a lattice. The detail of lattice generation can be found in [33]. Beam width pruning is applied every 25 frames where we use 13.0 for the Viterbi pruning beam 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 LM with the scaling factor over the range 5-15. Only the lowest word error rates after LM re-scoring are used.

6. Experiments and results

In our experiments, we evaluate the word accuracy of lipreading systems in two scenarios: the speaker-dependent scenario where all 56 speakers are seen in the training and speaker-independent where we evaluate on speakers unseen during training. All experiments use the phoneme unit since this is known to give the better word accuracy than visemes [7, 20].

6.1. Visual speech features

There is much discussion as to the optimal feature for lipreading (see [34] for example). Here we are not too concerned with the feature vector since the dataset already provides a region-of-interest (ROI) which is meant to contain the lips so we select a data-compressive image-based feature known as eigenlips. The supplied ROIs have slight variations in size by frame so we scale them to 128×256 pixels, the eigenlip feature is then extracted via Principal Component Analysis (PCA). Thirty dimensions are retained covering 85% of the principal component variances. To construct the PCA, 25-ROIs images of each training utterance are randomly selected to be the set of training images. There are around 4000 training utterances hence around 100k images in total were used to compute the eigen-analysis.

6.2. Beseline DNN model trained on CE

The CD-DNNs are trained and optimized by minimizing frame-based cross-entropy between the prediction and the target PDFs which are the "tied-state context-dependent label". These PDFs are generated from the Speaker Adaptive Training (SAT) system, then aligned into every frame. There are 1798 PDFs used in SD, and 1756 PDFs used in SI. The feature which we adopted for all DNN training process is based on a 40-dimensional feature-space MLLR (fMLLR, known as constrained MLLR [35]) feature with mean and variance normalization, where the fMLLR obtained via LDA+MLLT (LDA followed by a maximum-likihood linear transform) projection of 15 frames spliced of Eigenlip feature. 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 temporal splicing of $\pm n$ consecutive frames where $n = (0, ..., 6)$. The model is initialized by a stacking of RBMs 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 0.1. We use the minibatch SGD for fine-tuning with the 256 minibatch size. The results are reported as the mean of 10-fold cross-validation word accuracy where 10% of whole training set were used as a development set in each training fold. We evaluate the performance of models on the speaker-dependent (SD) and speaker-independent (SI) sets from TCD-TIMIT.

Table 3: *Lipreading word accuracy with varieties of machine learning. The feature pre-processing in each GMM training step is similar to [7].*

Model	Feature processing	Feature dim (±frame splicing)	Word accuracy (%)	
			SD	SI
CI-GMM	Δ + ΔΔ	90	1.04	3.84
CD-GMM	Δ + ΔΔ	90	5.79	4.76
CD-GMM	LDA-MLLT	40	19.53	19.56
CD-GMM SAT	FMLLR	40	28.79	24.57
DNN	FMLLR	40 (±0)	44.73	39.00
		120 (±1)	47.61	43.52
		200 (±2)	**48.89**	43.52
		280 (±3)	48.61	**43.61**
		400 (±4)	48.35	42.58
		440 (±5)	48.74	42.97
		520 (±6)	47.66	42.62

Table 3 presents the baseline results using the DNN model optimized on CE with various dimensions of the FMLLR input feature compared to the GMM-SAT model. We clearly see the learning ability of the DNN systems even with no spliced features ($n = 0$) by a 15.94% increase in accuracy on SD (from 28.79% to 44.73%) and 14.43% on SI (from 24.57% to 39.00%). The benefit of augmenting the neighboring context frames brings further improvement in the accuracy (at least 2.88% on SD and 3.58% on SI) compared to using the current frame alone. However, increased splicing does not monotonically increase performance. Here, the best performance of baseline SD is 48.89% with ± 2 context and SI is 43.61% with ± 3 context. We use ± 3 as a splicing context for the further experiments because it archives the best performance on SI which is the more realistic task.

6.3. Sequence discriminative training experiments

We conduct experiments on sequence discriminative training on top of the DNN model initiallised by CE via the three training criteria, sMBR, MPE, and MMI. First, decoding lattices and alignments of training data are needed. Here, the DNN trained on the CE criterion has been used as a seed model to decode training utterances by utilizing a unigram language model. The DNN model trained on CE are used for generating the posterior probability, then the raw state accuracy of each sentence in the lattice is computed. These steps are essential because it gives us the actual performance of the current visual speech model by decoding the training data itself with fewer constraints in the language model so we can identify the errors that need to be improved via sequence discriminative training criteria discussed in Section 4. We set the learning rate to 1×10^{-5}, while acoustic-scale and LM-scale are 0.1 and 1.0 respectively as in [18].

Table 4: *Comparisons of three sequence-discriminative training criteria sMBR, MPE, and MMI with and without DNN realignment against the DNN baseline (280 dimensional FMLLR). The DNN realignment means updating the target label derived from the alignments generated by the DNN baseline model.*

Model	Training objective	DNN realignment	Word accuracy (%)	
			SD	SI
Beseline	Cross-entropy	-	48.61	43.61
sMBR	CE + sMBR	Yes	53.96	48.32
		No	53.89	48.16
MPE	CE + MPE	Yes	54.18	48.48
		No	53.47	48.55
MMI	CE + MMI	Yes	51.75	45.83
		No	51.76	46.36

Table 4 shows the word accuracy before and after applying sequence discriminative training. The significant improvement can be seen in all cases compared to CD-DNN trained on CE with no significant different between with/without DNN alignment update.

We also examine the word accuracy when increasing the number of training iterations. Results in Figure 1 illustrate the performance variation with training iteration and alignment update. The 0th iteration means CE. For the SD configuration, sMBR and MPE have small changes after the sixth-iteration, while MMI still increases. However, the best result of SD is 56.59 % obtained from the 10-iteration of sMBR (7.98% higher than CE). For the SI configuration the highest word accuracy is 51.29% at the eighth-iteration of sMBR (7.68% higher than CE).

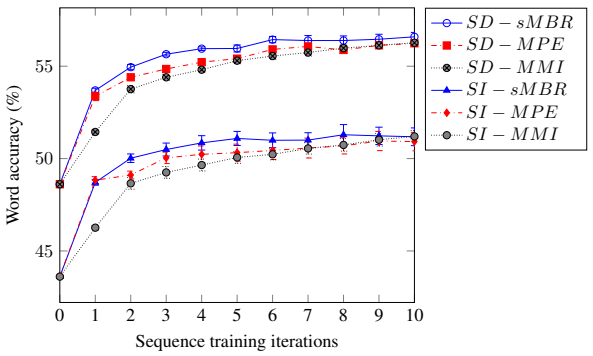

Figure 1: *Comparison of lipreading performance of SD and SI systems among three discriminative training criterion; sMBR, MPE, and MMI when we increase the training iterations. The best performance of SD is 56.59% on the 10th-iteration of sMBR and that of SI is 51.29% on 8th-interaton of sMBR. (Note: 0th-iteration means baseline DNN)*

7. Conclusions

We have built a successful lipreading system using DNNs and sequence discriminative training. Comparing our result with the baseline system, a conventional HMM, we see that performance has increased from around 4% word accuracy to around 51% in speaker independent mode. Looking in more detail, significant improvements are obtained using FMLLR, the DNN rather than a GMM, some temporal stacking ($n = \pm 1, \pm 2$) and the use of sequence discriminative training. The sequence discriminative training converges quickly (two or three iterations) but the method does not matter very much (sMBR, MPE, MMI). We think that if we had more data then the methods would differ and possibly more iterations would give grater benefit.

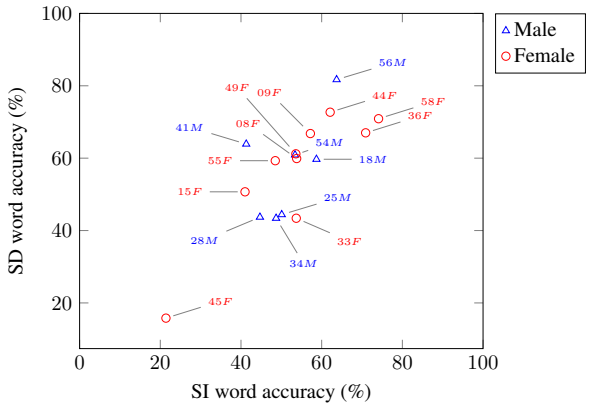

Figure 2: *Accuracy in SD vs SI for a variety of talkers.*

As usual with lipreading systems the identity of the talkers can be very significant, the range of talker accuracy in Figure 2 is between 20% and 80% accuracy: a wide range. We see this as an important clue that DNNs are better model of the complex density of lipreading features but are not yet capable of modeling that variation by identity. This therefore remains an important topic for future work.

8. References

[1] J. Bulwer, *Philocopus, or the Deaf and Dumbe Mans Friend*. Humphrey and Moseley, 1648.

[2] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[3] 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.

[4] D. Howell, "Confusion modelling for lip-reading," Ph.D. dissertation, University of East Anglia, 2015.

[5] K. Thangthai, R. W. Harvey, S. J. Cox, and B. Theobald, "Improving lip-reading performance for robust audiovisual speech recognition using DNNs," in *Auditory-Visual Speech Processing, AVSP 2015, Vienna, Austria, September 11-13, 2015*, 2015, pp. 127–131.

[6] Y. Lan, R. Harvey, and B. J. Theobald, "Insights into machine lip reading," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4825–4828.

[7] 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.

[8] B. Lee, M. Hasegawa-johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *in Proc. Conf. Spoken Language, Jeju, Korea*, 2004, pp. 2489–2492.

[9] A. Biswas, P. Sahu, and M. Chandra, "Multiple camera in car audio-visual speech recognition using phonetic and visemic information," *Comput. Electr. Eng.*, vol. 47, no. C, pp. 35–50, Oct. 2015.

[10] 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.

[11] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, "Audio visual speech recognition," IDIAP, Tech. Rep., 2000.

[12] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 165–168.

[13] D. Howell, S. Cox, and B. Theobald, "Visual units and confusion modelling for automatic lip-reading," *Image Vision Comput.*, vol. 51, no. C, pp. 1–12, Jul. 2016.

[14] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[15] S. Tamura, H. Ninomiya, N. Kitaoka, and S. Osuga, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," *APSIPA ASC2015*, 2015.

[16] Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, "An empirical exploration of CTC acoustic models," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2623–2627.

[17] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," in *in Proc. ICASSP*, 2013.

[18] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *in Proc. INTERSPEECH*, 2013.

[19] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," *entropy*, vol. 15, no. 16, pp. 17–18, 2014.

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

[21] A. R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan 2012.

[22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan 2012.

[23] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7398–7402.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 10 1986.

[25] 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.

[26] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," in *ICASSP 2013*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), January 2013.

[27] P. Voigtlaender, P. Doetsch, S. Wiesler, R. Schlter, and H. Ney, "Sequence-discriminative training of recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2100–2104.

[28] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, vol. 11, Apr 1986, pp. 49–52.

[29] S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, April 1993, pp. 491–494 vol.2.

[30] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3761–3764.

[31] D. Povey and P. C. Woodland, "Pminimum phone error," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[32] 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.

[33] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafit, S. Kombrink, P. Motlek, Y. Qian, K. Riedhammer, K. Vesel, and N. T. Vu, "Generating exact lattices in the WFST framework," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4213–4216.

[34] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading," in *International Conference on Auditory-Visual Speech Processing 2009*, 2009, pp. 102–106.

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