INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Large-Scale Visual Speech Recognition

Brendan Shillingford[1*], *Yannis Assael*[1*], *Matthew W. Hoffman*[1], *Thomas Paine*[1], *Cían Hughes*[1], *Utsav Prabhu*[2], *Hank Liao*[2], *Hasim Sak*[2], *Kanishka Rao*[2], *Lorrayne Bennett*[1], *Marie Mulville*[1], *Misha Denil*[1], *Ben Coppin*[1], *Ben Laurie*[1], *Andrew Senior*[1], *Nando de Freitas*[1]

[1] DeepMind
[2] Google

Abstract

This work presents a scalable solution to continuous visual speech recognition. To achieve this, we constructed the largest existing visual speech recognition dataset, consisting of pairs of transcriptions and video clips of faces speaking (3,886 hours of video). In tandem, we designed and trained an integrated lipreading system, consisting of a video processing pipeline that maps raw video to stable videos of lips and sequences of phonemes, a scalable deep neural network that maps the lip videos to sequences of phoneme distributions, and a phoneme-to-word speech decoder that outputs sequences of words. The proposed system achieves a word error rate (WER) of 40.9% as measured on a held-out set. In comparison, professional lipreaders achieve either 86.4% or 92.9% WER on the same dataset when having access to additional types of contextual information. Our approach significantly improves on previous lipreading approaches, including variants of *LipNet* and of *Watch, Attend, and Spell* (WAS), which are only capable of 89.8% and 76.8% WER respectively.

Index Terms: visual speech recognition, lipreading.

1. Introduction and motivation

Visual speech recognition could positively impact the lives of hundreds of thousands of patients with speech impairments worldwide. For example, in the U.S. alone 103,925 tracheostomies were performed in 2014 [1], a procedure that can result in a difficulty to speak (disphonia) or an inability to produce voiced sound (aphonia). Assisting people with speech impairments is a key motivating factor behind this work. Deep learning techniques have allowed for significant advances in lipreading over the last few years [1–6]. However, these approaches have often been limited to narrow vocabularies, and relatively small datasets [1, 4, 6]. Often the approaches focus on single-word classification [7–20] and do not attack the continuous recognition setting. In this paper, we contribute a novel method for large-vocabulary continuous visual speech recognition and we report substantial reductions in word error rate (WER) over the state-of-the-art approaches.

We propose a novel lipreading system (Figure 1), which transforms raw video into a word sequence. The first component of this system is a data processing pipeline used to create the *Large-Scale Visual Speech Recognition* (LSVSR) dataset used in this work, distilled from YouTube videos and consisting of transcriptions, and their phoneme sequences, paired with video clips of faces speaking (3,886 hours). The creation of the dataset alone required a non-trivial combination of computer vision and machine learning techniques. At a high level this process takes as input raw video and annotated audio segments, filters and preprocesses them, and produces a collection of aligned phoneme and lip frame sequences. In contrast to previous work, our pipeline uses landmark smoothing, a blurriness filter, an improved speaking classifier network and outputs phonemes.

Next, this work introduces a new neural network architecture for lipreading, which we call *Vision to Phoneme* (V2P), trained to produce a sequence of phoneme distributions given a sequence of video frames. In light of the large scale of our dataset, the network design has been highly tuned to maximize predictive performance subject to the strong computational and memory limits of modern GPUs in a distributed setting, where we found that techniques such as group normalization [21] were key. Our approach is the first to combine a deep learning phoneme-based visual speech recognition model with phoneme-to-word decoding techniques.

Finally, this entire lipreading system results in an unprecedented WER of 40.9% as measured on our dataset's held-out set. In comparison, professional lipreaders achieve 86.4% or 92.9% WER on this dataset, depending on the amount of context given. Similarly, previous state-of-the-art approaches such as variants of *LipNet* [1] and of *Watch, Attend, and Spell* (WAS) [2] demonstrated WERs of only 89.8% and 76.8% respectively.

2. Related work

While there is a large body of literature on automated lipreading, much of the early work focused on single-word classification and relied on substantial prior knowledge [22]. For example, [23] predicted continuous sequences of tri-visemes using a traditional HMM model with visual features extracted from a codebook of clustered mouth region images. The predicted visemes were used to distinguish sentences from a set of 150 possible sentences. Furthermore, [24] predict words and sequences digits using HMMs, [25] introduce multi-stream HMMs, and [26] improve the performance by using visual features in addition to the lip contours. Later, [27] used coupled HMMs to jointly model audio and visual streams to predict sequences of digits. [28] used HMMs for sentence-level speech recognition in noisy environments of the IBM ViaVoice dataset by fusing handcrafted visual and audio features. For further details, we refer the reader to the survey material of [22, 29, 30].

Until recently generalization across speakers and extraction of motion features have been considered open problems [22, 7, 1]. Advances in deep learning have made it possible to overcome these limitations, but most works still focus on single-word classification [7–20]. LipNet [1] was the first end-to-end model to tackle sentence-level lipreading by predicting character sequences. The model combined spatiotemporal convolutions with gated recurrent units (GRUs) and was trained using the connectionist temporal classification (CTC) [31] loss func-

[*] These authors contributed equally to this work.
[1] Health Care Utilization Project Network, "Hospital inpatient national statistics", `http://hcupnet.ahrq.gov`, 2014.

http://dx.doi.org/10.21437/Interspeech.2019-1669

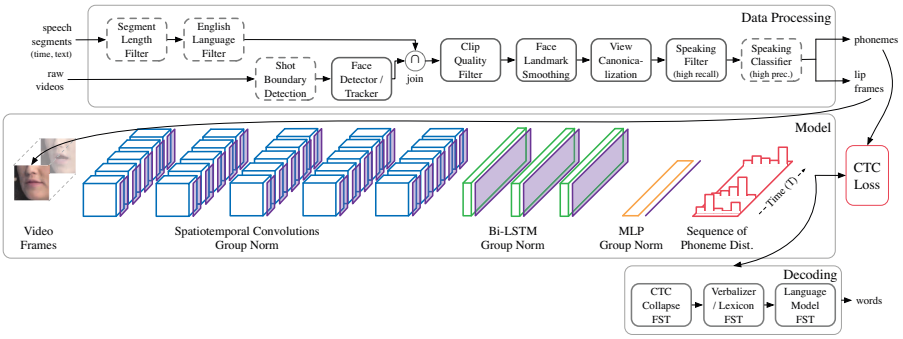

Figure 1: *The full visual speech recognition system introduced consists of a data processing pipeline that generates lip and phoneme clips from YouTube videos, and a phoneme recognition model combined with a phoneme-to-word decoding module used for inference.*

tion. LipNet was evaluated on the GRID corpus [32], a limited grammar and vocabulary dataset, where it achieved 4.8% and 11.4% WER in overlapping and unseen speaker evaluations respectively. By comparison, the performance of competent human lipreaders was 47.7%. LipNet is the closest model to our neural network. Similar architectures were subsequently introduced by [4–6]. [2] were the first to use sequence-to-sequence models with attention to tackle audio-visual speech recognition with a real-world dataset. The model "Watch, Listen, Attend and Spell" (WLAS), consists of a visual (WAS) and an audio (LAS) module. To evaluate WLAS, the authors created LRS, the largest dataset at that point with 246 hours of clips from BBC news broadcasts. The authors introduced an efficient video processing pipeline, and reported 50.2% WER, with the performance of professional lipreaders being 87.6% WER. [3] extended the work to multi-view sentence-level lipreading, but both [2, 3] pre-learn features with the audio-video synchronization classifier of [33], and fix these features to compensate for the large memory requirements of their attention networks. Contemporaneously with our work, [34] presented LRS3-TED, a dataset generated from English language talks available online. Using pre-learned features [30] presented sequence-to-sequence and CTC, character-level self-attention transformer models achieving a WER of 57.9% and 61.8% respectively.

Table 1: *Continuous visual speech recognition datasets.*

Dataset	Utter.	Hours	Vocab
GRID	33,000	28	51
IBM ViaVoice	17,111	35	10,400
MV-LRS	74,564	$\sim$155	14,960
LRS	118,116	$\sim$246	17,428
LRS3-TED	$\sim$165,000	$\sim$475	$\sim$57,000
LSVSR	**2,934,899**	**3,886**	**127,055**

In contrast to these, V2P predicts a sequence of phoneme distributions which are then fed into a decoder to produce a sequence of words. This flexible design enables us to easily accommodate very large vocabularies, that can be extended without retraining. Unlike previous work [2, 3, 30], V2P is memory and computationally efficient removing the need for pre-trained features. Finally, our dataset, LSVSR, is an order of magnitude larger than existing sentence-level visual speech recognition datasets ([2, 3, 28]; Table 1), and the content is more diverse and varied as it consists of general YouTube videos rather than solely news broadcasts, conference talks, or recording studios.

3. Dataset collection

Our dataset is extracted from public YouTube videos using large-scale parallel processing. Our pipeline builds on [35] which extracts audio clips paired with the relevant transcripts, yielding 140,000 hours of segments. We fetch the corresponding video, apply various filtering stages, after which 2% of the clips remain. After processing we obtain pairs of video and phoneme sequences, where videos are represented as identically-sized frames (128×128) stacked in the time dimension. Here, we describe the components of the pipeline.

Length filter, language filter. The duration of each clip is limited to 1-12 seconds, and non-English transcripts are filtered out using a language classifier. For evaluation, we further remove the utterances containing < 6 words. Sequences of X-SAMPA phonemes (40 plus silence) are obtained converting the transcripts to phonemes and then using forced alignment [35].

Raw videos, shot boundary detection, face detection. Constant spatial padding in each video segment is eliminated before a standard thresholding color histogram classifier identifies and removes segments containing shot boundaries. FaceNet [36] is used to detect and track face landmark locations.

Clip quality filter. We further remove blurry clips, clips including faces with an eye-to-eye width of less than 80 pixels, and clips with frame rates lower than 23fps. Frame rates above 30fps are downsampled. We allow a range of input frame rates as the effect is similar to different speaking paces.

Face landmark smoothing. The segments are processed by a face landmark tracker and the resulting landmark positions are smoothed using a temporal Gaussian kernel. Intuitively, this simplifies learning filters for the 3D convolution layers by reducing spatiotemporal noise. Empirically, our preliminary studies showed smoothing was crucial for achieving optimal performance. Next, following previous literature [2], we keep segments where the face yaw and pitch remain within $\pm 30°$.

View canonicalization. We obtain canonical faces using a reference face model, apply an affine transformation on the landmarks, and isolate the area around the lips.

Speaking filter. Using the smoothed landmarks, minor lip movements and non-speaking faces are discarded by thresholding the standard deviation of the upper-lower lip distance, a computationally cheap filter with high recall.

Speaking classifier. As a final step, we build *V2P-Sync*, a neural network architecture to verify the alignment of audio and video inspired by [33, 37]. However, V2P-Sync takes advantage of face landmark smoothing and processes longer time segments. Furthermore, spatiotemporal convolutions instead of spatial-only convolutions [33], and view canonicalization and

higher resolution inputs (128×128 vs 100×60) as compared to [37], prove useful. V2P-Sync takes a log mel-spectrogram and 9 grayscale video frames and separately produces an embedding for each, and is trained with a max-margin loss [33], resulting in a per-sample test accuracy of 81.2%. By applying V2P-Sync over 100 segments using a sliding window and aggregating scores, our per-video effective accuracy is increased.

4. V2P architecture

This work introduces the V2P model, which consists of a *3d convolutional module* for extracting spatiotemporal features from a given video clip, and a *temporal module* which aggregates them over time and outputs a sequence of phoneme distributions. Given input videos and target phoneme sequences the model is trained using the *CTC* loss. Finally, at test-time, a *decoder* based on finite state transducers (FSTs) is used to produce a word sequence given a sequence of phoneme distributions.

Neural network architecture. To explicitly address motion feature extraction, we designed a vision module based on VGG and made it volumetric, which proved crucial in our preliminary empirical evaluation and has been established in previous literature [1]. One of the main challenges in training a large vision module is finding an effective balance between performance and the imposed constraints of GPU memory. Our vision module consists of 5 convolutional layers with $[64, 128, 256, 512, 512]$ filters. A 2×2 max pooling was applied on the spatial dimensions of all layers but the fourth. To further reduce the memory footprint we limit the number of convolutional filters in these layers, and since the frame is centered around the lips, we omit spatial padding. As we can only fit 2 batch elements per GPU, we distribute training across 64 workers, for a total batch size of 128. To alleviate the communication cost of between-worker batch normalization statistics collection, and avoid the noisy alternative of within-worker statistics, we instead use group normalization [21].

The outputs of the convolutional stack are then fed into a temporal module which performs longer-scale aggregation of the extracted features over time. In constructing this component we evaluated a number of recurrent neural network and dilated convolutional architectures, as shown in in Section 5. The best architecture presented performs temporal aggregation using a stack of 3 bidirectional LSTMs with a hidden state of 768, interleaved with group normalization. The output of these LSTM layers is then fed through a final two-layer MLP ($1568 \rightarrow 768 \rightarrow 42$) to produce a sequence of conditionally independent phoneme distributions in addition to the CTC blank symbol, upon which we apply the CTC loss.

This model architecture is similar to LipNet [1], but differs in some crucial ways. In comparison to our work, LipNet used GRU and dropout, both of which we found to perform poorly in preliminary experiments. We instead use LSTM, group normalization, and a larger network with distributed training. Finally, while both models use CTC for training, crucially, V2P predicts phonemes rather than characters.

Connectionist temporal classification (CTC). CTC is a loss function for the parameterization of distributions over sequences of label tokens, without requiring alignments of the input sequence to the label tokens [31]. CTC models the probability of a label sequence by marginalizing over alignments of the label sequence to the output of the RNN, inserting a special blank symbol and collapsing consecutive identical labels.

Uncertainty in speech recognition arises mainly from two sources: uncertainty in perceived audio or visual input, and uncertainty in the words representing the sounds. Much previous work uses CTC to model characters given audio or visual input directly [1, 38], but this is problematic and exacerbated by the higher uncertainty for what is being spoken in lipreading. Thus, we take the conventional approach of using phonemes instead. Uncertainty in perception of the input is further increased for visual speech recognition since the information required to disambiguate some phonemes is not visible (e.g. the position of the tongue). The resulting visually similar phonemes are called visemes. Viseme groupings may differ per speaker [39] however, making it difficult to incorporate this knowledge in the design. Using phonemes allows the model to directly encode this uncertainty in its output, rather than implicitly inside the model. Word uncertainty is handled separately by the decoder.

Alternatively to using phonemes with CTC, some previous work solves this problem using RNN transducers [40] or sequence-to-sequence with attention [2], which jointly model all sources of uncertainty. However, the results of [41] suggest these models achieve similar performance to CTC; thus their advantage mainly lies in their size.

Decoding. As described earlier, our model produces a sequence of phoneme distribution followed by standard decoding using finite state transducers (FSTs) to arrive at word sequences [42, 43]. In our work we make use of a combination of three individual weighted FSTs: a *CTC postprocessing FST* removes duplicate symbols and CTC blanks, a *lexicon* FST maps input phonemes to output words, then a 5-gram language model (Katz backoff, 50 million n-grams, vocabulary size of 1 million) reweighs the resulting word sequences. Beam search is then performed on the CTC output distribution using this FST.

5. Evaluation

We examine the performance of V2P trained on LSVSR with hyperparameters tuned on a validation set. We evaluate it on a held-out test set roughly 37 minutes long, containing approximately 63,000 video frames and 7100 words. We also describe and compare against a number of alternative methods from previous work, and show that V2P gives significant performance improvements. Except for V2P-NoLM, all models used the same 5-gram word-level language model during decoding. To construct the validation and test sets we removed blurry videos by thresholding the variance of the Laplacian of each frame; we kept them in the training set as a form of data augmentation.

Professional lipreaders. We consulted a professional lipreading company to measure the difficulty of LSVSR and hence the impact V2P could have. We regenerated clips from a subset of our test set and cropped the whole head instead of the mouth region. Each video could be replayed $10\times$ at half or normal speed. Since the inherent ambiguity in lipreading necessitates relying on context, we conducted experiments both with and without context. For the former, we used clips with transcripts that had at least 6 words. For the latter, we used clips with at least 12 words, and presented to the lipreader the first 6 words, the title, and the category of the video, then asked them to transcribe the rest of the clip.

Audio-Ph. For an approximate bound on performance, we train a speech recognition model on the audio of the utterances. The architecture is based on Deep Speech 2 [38], but trained to predict phonemes rather than characters.

LipNet-Ch. We replicate the character-level CTC architecture of LipNet [1] and we use the FST decoding pipeline with a character-level language model as described in [42].

LipNet-Ph. We train LipNet to predict phonemes and use

Table 2: *Phoneme and word error rates on LSVSR test set.*

Method	Param.	PER	WER
Prof. w/o context	–	–	92.9 ± 0.9
Prof. w/ context	–	–	86.4 ± 1.4
Audio-Ph	58M	12.5 ± 0.5	18.3 ± 0.9
LipNet-Ch	7M	–	93.0 ± 0.6
LipNet-Ph	7M	65.8 ± 0.4	89.8 ± 0.5
Seq2seq-Ch	15M	–	76.8 ± 0.8
LipNet-Large-Ph	40M	53.0 ± 0.5	72.7 ± 1.0
V2P-FullyConv	29M	41.3 ± 0.6	51.6 ± 1.2
V2P-NoLM	49M	33.6 ± 0.6	53.6 ± 1.0
V2P	**49M**	$\mathbf{33.6 \pm 0.6}$	$\mathbf{40.9 \pm 1.2}$

the same FST-based decoding pipeline and language model.

LipNet-Large-Ph. Recall from the earlier discussion that LipNet uses dropout, whereas V2P makes heavy use of group normalization, crucial for our small batches per worker. For a fair size-wise comparison, we introduce a replica of V2P, that uses GRUs, dropout, and no normalization.

Seq2seq-Ch. We reimplemented the previous state-of-the-art sequence-to-sequence character-level architecture, WAS [2]. Although their implementation was followed as closely as possible, training end-to-end quickly exceeded the memory limitations of modern GPUs. To work around this, Chung et al. [2] kept the convolutional weights fixed using a pretrained network from [33], which we were unable to use as their network inputs were processed differently. Instead, we replace the 2D convolutional network with the *improved* lightweight 3D visual processing network of V2P. As shown by [1, 20, 37], spatiotemporal aggregation of features benefits performance.

V2P-FullyConv. Identical to V2P, except the LSTMs are replaced with 6 dilated temporal convolution layers with a kernel size of 3 and dilation rates of [1,1,2,4,8,16].

V2P-NoLM. Identical to V2P, except during decoding the language model is replaced by a dictionary of 100k words which are weighted by their smoothed frequency in the training data.

5.1. Results

Table 2 shows the phoneme and word error rates for all of the models, and the number of parameters of each. The error rates are computed as the sum of the edit distances of the predicted and ground-truth sequence pairs divided by total ground-truth length. Bootstrap sampling is used to compute the standard error associated with each rate. These results show that the variant of LipNet tested in this work is approximately able to perform on-par with professional lipreaders, when additional context was provided, with WER of 86.4% and 89.8% respectively. Similarly, we see that the WAS variant provides a substantial reduction to this error, resulting in a WER of 76.8%. However, the full V2P method presented in this work is able to further halve the WER, obtaining 40.9% at testing time. Interestingly, we see that although the bi-directional LSTM provides the best performance, using a fully-convolutional network still results in performance that is significantly better than all previous methods. Finally, although we see that the full V2P model performs best, replacing the 5-gram language model with a unigram one results only in a drop of approximately 13 WER to 53.6%.

By predicting phonemes directly, we also sidestep the need to design phoneme-to-viseme mappings [39]. The inherent uncertainty is instead modelled directly in the predictive distribution. Finally, by differentiating the likelihood of the phoneme sequence with respect to the inputs using guided backpropagation we compute the saliency maps shown in Figure 2.

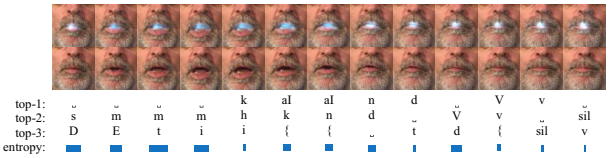

Figure 2: *Saliency map for "kind of" and the top-3 predictions of each frame and the entropy of the phoneme predictive distribution. The CTC blank character is represented by '␣'.*

To demonstrate the generalization power of V2P, we evaluate on LRS3-TED [34] and compare it to the TM-seq2seq model of [30]. Unlike LSVSR, LRS3-TED includes faces at angles between $\pm 90°$ instead of $\pm 30°$, and clips may be shorter than one second. Thus, we conducted two experiments. First, we evaluate performance on a subset of the LRS3-TED test set filtered according to the same protocol used to construct LSVSR, by removing clips with larger face angles and shorter lenght, and second, on the full unfiltered test set. Despite the fact that we do not train or fine-tune V2P on LRS3-TED, V2P achieves WERs of 47.0 ± 1.6 and 55.1 ± 0.9 respectively, outperforming TM-seq2seq's 57.9. V2P is able to generalize well, achieving state-of-the-art performance on datasets with different conditions on which it was not trained. Due to the difficulty of obtaining a continually front-on view of a face at a sufficiently high resolution without an individual's consent, the model is not suited for lipreading in scenarios such as surveillance.

6. Conclusions

We presented a novel, large-scale visual speech recognition system. Our system consists of a data processing pipeline used to construct a vast dataset—an order of magnitude greater than all previous approaches both in terms of vocabulary and the sheer number of example sequences. We described a scalable model for producing phoneme and word sequences from processed video clips that is capable of nearly halving the error rate of the previous state-of-the-art methods on this dataset, and achieving a new state-of-the-art in a dataset presented contemporaneously with this work. The combination of methods in this work represents a significant improvement in lipreading performance, a technology which can enhance automatic speech recognition systems, and which has enormous potential to improve the lives of speech-impaired patients worldwide.

7. Acknowledgements

We thank Andrew Zisserman for advising, Hagen Soltau for valuable resources, and Shane Agnew, Sean Legassick, Iason Gabriel, Dominic King, and Alan Karthikesalingam for helpful comments and contributions.

8. References

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," in *GPU Technology Conference*, 2017.

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

[3] J. S. Chung and A. Zisserman, "Lip reading in profile," in *British Machine Vision Conference*, 2017.

[4] A. Thanda and S. M. Venkatesan, "Audio visual speech recognition using deep recurrent neural networks," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Springer, 2017, pp. 98–109.

[5] A. Koumparoulis, G. Potamianos, Y. Mroueh, and S. J. Rennie, "Exploring ROI size in deep learning based lipreading," in *International Conference on Auditory-Visual Speech Processing*, 2017.

[6] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-ctc," in *International Conference on Automatic Face Gesture Recognition*. IEEE, 2018, pp. 548–555.

[7] M. Wand, J. Koutnik, and J. Schmidhuber, "Lipreading with long short-term memory," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2016, pp. 6115–6119.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696.

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *Interspeech*, 2014, pp. 1149–1153.

[11] C. Sui, M. Bennamoun, and R. Togneri, "Listening with your eyes: Towards a practical visual speech recognition system using deep Boltzmann machines," in *International Conference on Computer Vision*. IEEE, 2015, pp. 154–162.

[12] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," in *International Speech Communication Association*, 2015.

[13] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2016, pp. 2304–2308.

[14] O. Koller, H. Ney, and R. Bowden, "Deep learning of mouth shapes for sign language," in *ICCV Workshop on Assistive Computer Vision and Robotics*, 2015, pp. 85–91.

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

[16] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2016, pp. 2722–2726.

[17] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani *et al.*, "Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss," *Interspeech*, pp. 277–281, 2016.

[18] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *British Machine Vision Conference*, 2017.

[19] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *Interspeech*, 2017.

[20] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and lstms," *Computer Vision and Image Understanding*, vol. 176-177, pp. 22 – 32, 2018.

[21] Y. Wu and K. He, "Group normalization," in *European Conference on Computer Vision*, 2018, pp. 3–19.

[22] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.

[23] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous automatic speech recognition by lipreading," in *Motion-Based recognition*. Springer, 1997, pp. 321–343.

[24] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, "Speaker independent audio-visual database for bimodal ASR," in *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.

[25] G. Potamianos and H. P. Graf, "Discriminative training of hmm stream exponents for audio-visual speech recognition," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 6. IEEE, 1998, pp. 3733–3736.

[26] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *Conference on Image Processing*. IEEE, 1998, pp. 173–177.

[27] S. M. Chu and T. S. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *Interspeech*, 2000.

[28] 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.

[29] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in Visual and Audio-Visual Speech Processing*, vol. 22, p. 23, 2004.

[30] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[31] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning*, 2006, pp. 369–376.

[32] 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, 2006.

[33] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *ACCV Workshop on Multi-view Lip-reading*, 2016.

[34] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[35] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 368–373.

[36] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[37] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3d convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[38] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[39] H. L. Bear and R. Harvey, "Phoneme-to-viseme mappings: the good, the bad, and the ugly," *Speech Communication*, vol. 95, pp. 40 – 67, 2017.

[40] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer," in *Automatic Speech Recognition and Understanding Workshop*. IEEE, 2017, pp. 193–199.

[41] R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in *Interspeech*, 2017.

[42] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2015, pp. 167–174.

[43] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.