LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING

Yannis M. Assael[1,†], **Brendan Shillingford**[1,†], **Shimon Whiteson**[1] **& Nando de Freitas**[1,2,3]
Department of Computer Science, University of Oxford, Oxford, UK [1]
Google DeepMind, London, UK [2]
CIFAR, Canada [3]
{yannis.assael,brendan.shillingford,
shimon.whiteson,nando.de.freitas}@cs.ox.ac.uk

ABSTRACT

Lipreading is the task of decoding text from the movement of a speaker's mouth. Traditional approaches separated the problem into two stages: designing or learning visual features, and prediction. More recent deep lipreading approaches are end-to-end trainable (Wand et al., 2016; Chung & Zisserman, 2016a). However, existing work on models trained end-to-end perform only word classification, rather than sentence-level sequence prediction. Studies have shown that human lipreading performance increases for longer words (Easton & Basala, 1982), indicating the importance of features capturing temporal context in an ambiguous communication channel. Motivated by this observation, we present LipNet, a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. To the best of our knowledge, LipNet is the first end-to-end sentence-level lipreading model that simultaneously learns spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 95.2% accuracy in sentence-level, overlapped speaker split task, outperforming experienced human lipreaders and the previous 86.4% word-level state-of-the-art accuracy (Gergen et al., 2016).

1 INTRODUCTION

Lipreading plays a crucial role in human communication and speech understanding, as highlighted by the McGurk effect (McGurk & MacDonald, 1976), where one phoneme's audio dubbed on top of a video of someone speaking a different phoneme results in a third phoneme being perceived.

Lipreading is a notoriously difficult task for humans, specially in the absence of context[1]. Most lipreading actuations, besides the lips and sometimes tongue and teeth, are latent and difficult to disambiguate without context (Fisher, 1968; Woodward & Barber, 1960). For example, Fisher (1968) gives 5 categories of visual phonemes (called *visemes*), out of a list of 23 initial consonant phonemes, that are commonly confused by people when viewing a speaker's mouth. Many of these were asymmetrically confused, and observations were similar for final consonant phonemes.

Consequently, human lipreading performance is poor. Hearing-impaired people achieve an accuracy of only $17 \pm 12\%$ even for a limited subset of 30 monosyllabic words and $21 \pm 11\%$ for 30 compound words (Easton & Basala, 1982). An important goal, therefore, is to automate lipreading. Machine lipreaders have enormous practical potential, with applications in improved hearing aids, silent dictation in public spaces, security, speech recognition in noisy environments, biometric identification, and silent-movie processing.

Machine lipreading is difficult because it requires extracting spatiotemporal features from the video (since both position and motion are important). Recent deep learning approaches attempt to extract those features end-to-end. Most existing work, however, performs only word classification, not sentence-level sequence prediction.

[†]These authors contributed equally to this work.

[1]LipNet video: https://youtube.com/playlist?list=PLXkuFIFnXUAPIrXKgtIpctv2NuSo7xw3k

In this paper, we present LipNet, which is to the best of our knowledge, the first *end-to-end sentence-level* lipreading model. As with modern deep learning based automatic speech recognition (ASR), LipNet is trained end-to-end to make sentence-level predictions. Our model operates at the character-level, using spatiotemporal convolutional neural networks (STCNNs), recurrent neural networks (RNNs), and the connectionist temporal classification loss (CTC) Graves et al. (2006).

Our empirical results on the GRID corpus (Cooke et al., 2006), one of the few public sentence-level datasets, show that LipNet attains a 95.2% sentence-level word accuracy, in a overlapped speakers split that is popular for benchmarking lipreading methods. The previous best accuracy reported on an aligned word classification version of this task was 86.4% (Gergen et al., 2016). Furthermore, LipNet can generalise across unseen speakers in the GRID corpus with an accuracy of 88.6%.

We also compare the performance of LipNet with that of hearing-impaired people who can lipread on the GRID corpus task. On average, they achieve an accuracy of 52.3%, in contrast to LipNet's $1.69\times$ higher accuracy in the same sentences.

Finally, by applying saliency visualisation techniques (Zeiler & Fergus, 2014; Simonyan et al., 2013), we interpret LipNet's learned behaviour, showing that the model attends to phonologically important regions in the video. Furthermore, by computing intra-viseme and inter-viseme confusion matrices at the phoneme level, we show that almost all of LipNet's few erroneous predictions occur within visemes, since context is sometimes insufficient for disambiguation.

2 Related Work

In this section, we outline various existing approaches to automated lipreading.

Automated lipreading: Most existing work on lipreading does not employ deep learning. Such work requires either heavy preprocessing of frames to extract image features, temporal preprocessing of frames to extract video features (e.g., optical flow or movement detection), or other types of handcrafted vision pipelines (Matthews et al., 2002; Zhao et al., 2009; Gurban & Thiran, 2009; Papandreou et al., 2007; 2009; Pitsikalis et al., 2006; Lucey & Sridharan, 2006; Papandreou et al., 2009). The automated lipreading literature is too vast to adequately cover, so we refer the reader to Zhou et al. (2014) for an extensive review.

Notably, Goldschen et al. (1997) were the first to do visual-only sentence-level lipreading using hidden Markov models (HMMs) in a limited dataset, using hand-segmented phones. Later, Neti et al. (2000) were the first to do sentence-level audiovisual speech recognition using an HMM combined with hand-engineered features, on the IBM ViaVoice (Neti et al., 2000) dataset. The authors improve speech recognition performance in noisy environments by fusing visual features with audio ones. The dataset contains 17111 utterances of 261 speakers for training (about 34.9 hours) and is not publicly available. As stated, their visual-only results cannot be interpreted as visual-only recognition, as they are used as rescoring of the noisy audio-only lattices. Using a similar approach, Potamianos et al. (2003) report speaker independent and speaker adapted 91.62%, 82.31% WER in the same dataset respectively, and 38.53%, 16.77% WER in the connected DIGIT corpus, which contains sentences of digits.

Furthermore, Gergen et al. (2016) use speaker-dependent training on an LDA-transformed version of the Discrete Cosine Transforms of the mouth regions in an HMM/GMM system. This work holds the previous state-of-the-art on the GRID corpus with a speaker-dependent accuracy of 86.4%. Generalisation across speakers and extraction of motion features is considered an open problem, as noted in (Zhou et al., 2014). LipNet addresses both of these issues.

Classification with deep learning: In recent years, there have been several attempts to apply deep learning to lipreading. However, all of these approaches perform only word or phoneme classification, whereas LipNet performs full sentence sequence prediction. Approaches include learning multimodal audio-visual representations (Ngiam et al., 2011; Sui et al., 2015; Ninomiya et al., 2015; Petridis & Pantic, 2016), learning visual features as part of a traditional speech-style processing pipeline (e.g. HMMs, GMM-HMMs, etc.) for classifying words and/or phonemes (Almajai et al., 2016; Takashima et al., 2016; Noda et al., 2014; Koller et al., 2015), or combinations thereof (Takashima et al., 2016). Many of these approaches mirror early progress in applying neural networks for acoustic processing in speech recognition (Hinton et al., 2012).

Chung & Zisserman (2016a) propose spatial and spatiotemporal convolutional neural networks, based on VGG, for word classification. The architectures are evaluated on a word-level dataset BBC TV (333 and 500 classes), but, as reported, their spatiotemporal models fall short of the spatial architectures by an average of around 14%. Additionally, their models cannot handle variable sequence lengths and they do not attempt sentence-level sequence prediction.

Chung & Zisserman (2016b) train an audio-visual max-margin matching model for learning pre-trained mouth features, which they use as inputs to an LSTM for 10-phrase classification on the OuluVS2 dataset, as well as a non-lipreading task.

Wand et al. (2016) introduce LSTM recurrent neural networks for lipreading but address neither sentence-level sequence prediction nor speaker independence.

Garg et al. (2016) apply a VGG pre-trained on faces to classifying words and phrases from the MIRACL-VC1 dataset, which has only 10 words and 10 phrases. However, their best recurrent model is trained by freezing the VGGNet parameters and then training the RNN, rather than training them jointly. Their best model achieves only 56.0% word classification accuracy, and 44.5% phrase classification accuracy, despite both of these being 10-class classification tasks.

Sequence prediction in speech recognition: The field of automatic speech recognition (ASR) would not be in the state it is today without modern advances in deep learning, many of which have occurred in the context of ASR (Graves et al., 2006; Dahl et al., 2012; Hinton et al., 2012). The connectionist temporal classification loss (CTC) of Graves et al. (2006) drove the movement from deep learning as a component of ASR, to deep ASR systems trained end-to-end (Graves & Jaitly, 2014; Maas et al., 2015; Amodei et al., 2015). As mentioned earlier, much recent lipreading progress has mirrored early progress in ASR, but stopping short of sequence prediction.

LipNet is the first end-to-end model that performs sentence-level sequence prediction for visual speech recogntion. That is, we demonstrate the first work that takes as input as sequence of images and outputs a distribution over sequences of tokens; it is trained end-to-end using CTC and thus also does not require alignments.

Lipreading Datasets: Lipreading datasets (AVICar, AVLetters, AVLetters2, BBC TV, CUAVE, OuluVS1, OuluVS2) are plentiful (Zhou et al., 2014; Chung & Zisserman, 2016a), but most only contain single words or are too small. One exception is the GRID corpus (Cooke et al., 2006), which has audio and video recordings of 34 speakers who produced 1000 sentences each, for a total of 28 hours across 34000 sentences. Table 1 summarises state-of-the-art performance in each of the main lipreading datasets.

Table 1: Existing lipreading datasets and the state-of-the-art accuracy reported on these. The size column represents the number of utterances used by the authors for training. Although the GRID corpus contains entire sentences, Gergen et al. (2016) consider only the simpler case of predicting isolated words. LipNet predicts sequences and hence can exploit temporal context to attain much higher accuracy. Phrase-level approaches were treated as plain classification.

Method	Dataset	Size	Output	Accuracy
Fu et al. (2008)	AVICAR	851	Digits	37.9%
Hu et al. (2016)	AVLetter	78	Alphabet	64.6%
Papandreou et al. (2009)	CUAVE	1800	Digits	83.0%
Chung & Zisserman (2016a)	OuluVS1	200	Phrases	91.4%
Chung & Zisserman (2016b)	OuluVS2	520	Phrases	94.1%
Chung & Zisserman (2016a)	BBC TV	> 400000	Words	65.4%
Gergen et al. (2016)	GRID	29700	Words*	86.4%
LipNet	GRID	28775	**Sentences**	**95.2%**

We use the GRID corpus to evaluate LipNet because it is sentence-level and has the most data. The sentences are drawn from the following simple grammar: $command^{(4)} + color^{(4)} + preposition^{(4)} + letter^{(25)} + digit^{(10)} + adverb^{(4)}$, where the number denotes how many word choices there are for each of the 6 word categories. The categories consist of, respectively, {bin, lay, place, set}, {blue, green, red, white}, {at, by, in, with}, $\{A, \ldots, Z\} \backslash \{W\}$, {zero, ..., nine}, and

{again, now, please, soon}, yielding 64000 possible sentences. For example, two sentences in the data are "set blue by A four please" and "place red at C zero again".

3 LIPNET

LipNet is a neural network architecture for lipreading that maps variable-length sequences of video frames to text sequences, and is trained end-to-end. In this section, we describe LipNet's building blocks and architecture.

3.1 SPATIOTEMPORAL CONVOLUTIONS

Convolutional neural networks (CNNs), containing stacked convolutions operating spatially over an image, have been instrumental in advancing performance in computer visions tasks such as object recognition that receive an image as input (Krizhevsky et al., 2012). A basic 2D convolution layer from C channels to C' channels (without a bias and with unit stride) computes

$$[\mathrm{conv}(\mathbf{x}, \mathbf{w})]_{c'ij} = \sum_{c=1}^{C} \sum_{i'=1}^{k_w} \sum_{j'=1}^{k_h} w_{c'ci'j'} x_{c,i+i',j+j'},$$

for input $\mathbf{x}$ and weights $\mathbf{w} \in \mathbb{R}^{C' \times C \times k_w \times k_h}$ where we define $x_{cij} = 0$ for i, j out of bounds. Spatiotemporal convolutional neural networks (STCNNs) can process video data by convolving across time, as well as the spatial dimensions (Karpathy et al., 2014; Ji et al., 2013). Hence similarly,

$$[\mathrm{stconv}(\mathbf{x}, \mathbf{w})]_{c'tij} = \sum_{c=1}^{C} \sum_{t'=1}^{k_t} \sum_{i'=1}^{k_w} \sum_{j'=1}^{k_h} w_{c'ct'i'j'} x_{c,t+t',i+i',j+j'}.$$

3.2 GATED RECURRENT UNIT

Gated Recurrent Unit (GRU) (Chung et al., 2014) is a type of recurrent neural network (RNN) that improves upon earlier RNNs by adding cells and gates for propagating information over more time-steps and learning to control this information flow. It is similar to the Long Short-Term Memory (LSTM) RNN (Hochreiter & Schmidhuber, 1997). We use the standard formulation:

$$\begin{aligned}
[\mathbf{u}_t, \mathbf{r}_t]^T &= \mathrm{sigm}(\mathbf{W}_z \mathbf{z}_t + \mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{b}_g) \\
\tilde{\mathbf{h}}_t &= \tanh(\mathbf{U}_z \mathbf{z}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \\
\mathbf{h}_t &= (\mathbf{1} - \mathbf{u}_t) \odot \mathbf{h}_{t-1} + \mathbf{u}_t \odot \tilde{\mathbf{h}}_t
\end{aligned}$$

where $\mathbf{z} := \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ is the input sequence to the RNN, $\odot$ denotes element-wise multiplication, and $\mathrm{sigm}(r) = 1/(1 + \exp(-r))$. We use a bidirectional GRU (Bi-GRU) as introduced by Graves & Schmidhuber (2005) in the context of LSTMs: one RNN maps $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\} \mapsto \{\overrightarrow{\mathbf{h}_1}, \ldots, \overrightarrow{\mathbf{h}_T}\}$, and another $\{\mathbf{z}_T, \ldots, \mathbf{z}_1\} \mapsto \{\overleftarrow{\mathbf{h}_1}, \ldots, \overleftarrow{\mathbf{h}_T}\}$, then $\mathbf{h}_t := [\overrightarrow{\mathbf{h}_t}, \overleftarrow{\mathbf{h}_t}]$. The Bi-GRU ensures that $\mathbf{h}_t$ depends on $\mathbf{z}_{t'}$ for all t'. To parameterise a distribution over sequences, at time-step t let $p(u_t|\mathbf{z}) = \mathrm{softmax}(\mathrm{mlp}(\mathbf{h}_t; \mathbf{W}_{mlp}))$, where mlp is a feed-forward network with weights $\mathbf{W}_{mlp}$. Then we can define the distribution over length-T sequences as $p(u_1, \ldots, u_T|\mathbf{z}) = \prod_{1 \le t \le T} p(u_t|\mathbf{z})$, where T is determined by $\mathbf{z}$, the input to the GRU. In LipNet, $\mathbf{z}$ is the output of the STCNN.

3.3 CONNECTIONIST TEMPORAL CLASSIFICATION

The connectionist temporal classification (CTC) loss (Graves et al., 2006) is widely used in modern speech recognition as it eliminates the need for training data that aligns inputs to target outputs (Amodei et al., 2015; Graves & Jaitly, 2014; Maas et al., 2015). Given a model that outputs a sequence of discrete distributions over the token classes (vocabulary) augmented with a special "blank" token, CTC computes the probability of a sequence by marginalising over all sequences that are defined as equivalent to this sequence. This simultaneously removes the need for alignments and addresses variable-length sequences. Let V denote the set of tokens that the model classifies at a single time-step of its output (vocabulary), and the blank-augmented vocabulary $\tilde{V} = V \cup \{\_\}$

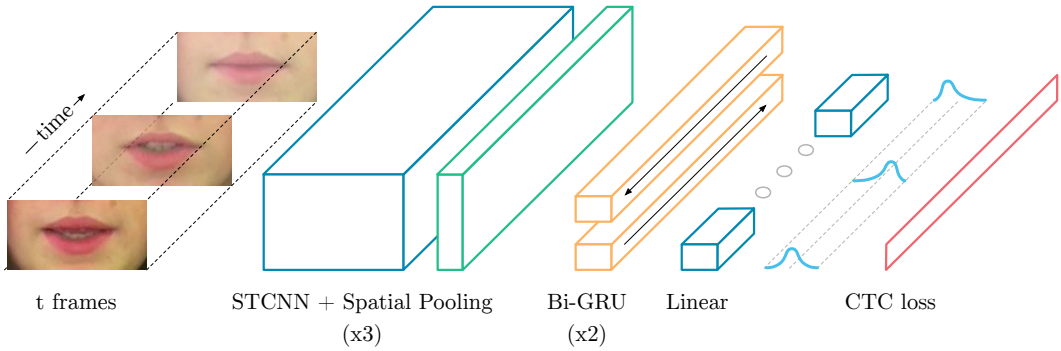


Figure 1: LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are processed by 2 Bi-GRUs; each time-step of the GRU output is processed by a linear layer and a softmax. This end-to-end model is trained with CTC.

where $\_$ denotes the CTC blank symbol. Define the function $\mathcal{B} : \tilde{V}^* \rightarrow V^*$ that, given a string over $\tilde{V}$, deletes adjacent duplicate characters and removes blank tokens. For a label sequence $y \in V^*$, CTC defines $p(y|\mathbf{x}) = \sum_{u \in \mathcal{B}^{-1}(y) \text{ s.t. } |u|=T} p(u_1, \ldots, u_T|\mathbf{x})$, where T is the number of time-steps in the sequence model. For example, if $T = 3$, CTC defines the probability of a string "am" as $p(aam) + p(amm) + p(\_am) + p(a\_m) + p(am\_)$. This sum is computed efficiently by dynamic programming, allowing us to perform maximum likelihood.

3.4 LipNet Architecture

Figure 1 illustrates the LipNet architecture, which starts with 3×(spatiotemporal convolutions, channel-wise dropout, spatial max-pooling). Subsequently, the features extracted are followed by two Bi-GRUs. The Bi-GRUs are crucial for efficient further aggregation of the STCNN output. Finally, a linear transformation is applied at each time-step, followed by a softmax over the vocabulary augmented with the CTC blank, and then the CTC loss. All layers use rectified linear unit (ReLU) activation functions. More details including hyperparameters can be found in Table 3 of Appendix A.

4 Lipreading Evaluation

In this section, we evaluate LipNet on the GRID corpus. The augmentation methods employed don't make use of external data and rely purely on the GRID corpus.

4.1 Data Augmentation

Preprocessing: The GRID corpus consists of 34 subjects, each narrating 1000 sentences. The videos for speaker 21 are missing, and a few others are empty or corrupt, leaving 32746 usable videos. We employ a split (unseen speakers; not previously used in the literature) holding out the data of two male speakers (1 and 2) and two female speakers (20 and 22) for evaluation (3971 videos). The remainder is used for training (28775 videos). We also use a sentence-level variant of the split (overlapped speakers) similar to Wand et al. (2016), where 255 random sentences from each speaker are used for evaluation. All remaining data from all speakers is pooled together for training. All videos are 3 seconds long with a frame rate of 25fps. The videos were processed with the DLib face detector, and the iBug face landmark predictor (Sagonas et al., 2013) with 68 landmarks coupled with an online Kalman Filter. Using these landmarks, we apply an affine transformation to extract a mouth-centred crop of size 100×50 pixels per frame. We standardise the RGB channels over the whole training set to have zero mean and unit variance.

Augmentation: We augment the dataset with simple transformations to reduce overfitting. First, we train on both the regular and the horizontally mirrored image sequence. Second, since the dataset provides word start and end timings for each sentence video, we augment the sentence-level training data with video clips of individual words as additional training instances. These instances have a

decay rate of 0.925. Third, to encourage resilience to varying motion speeds by deletion and duplication of frames, this is performed with a per-frame probability of 0.05. The same augmentation methods were followed in all proposed baselines and models.

4.2 BASELINES

To evaluate LipNet, we compare its performance to that of three hearing-impaired people who can lipread, as well as three ablation models inspired by recent state-of-the-art work (Chung & Zisserman, 2016a; Wand et al., 2016).

Hearing-Impaired People: This baseline was performed by three members of the Oxford Students' Disability Community. After being introduced to the grammar of the GRID corpus, they observed 10 minutes of annotated videos from the training dataset, then annotated 300 random videos from the evaluation dataset. When uncertain, they were asked to pick the most probable answer.

Baseline-LSTM: Using the sentence-level training setup of LipNet, we replicate the model architecture of the previous deep learning GRID corpus state-of-the-art (Wand et al., 2016). See Appendix A for more implementation details.

Baseline-2D: Based on the LipNet architecture, we replace the STCNN with spatial-only convolutions similar to those of Chung & Zisserman (2016a). Notably, contrary to the results we observe with LipNet, Chung & Zisserman (2016a) report 14% and 31% poorer performance of their STCNNs compared to the 2D architectures in their two datasets.

Baseline-NoLM: Identical to LipNet, but with the language model used in beam search disabled.

4.3 PERFORMANCE EVALUATION

To measure the performance of LipNet and the baselines, we compute the word error rate (WER) and the character error rate (CER), standard metrics for the performance of ASR models. We produce approximate maximum-probability predictions from LipNet by performing CTC beam search. WER (or CER) is defined as the minimum number of word (or character) insertions, substitutions, and deletions required to transform the prediction into the ground truth, divided by the number of words (or characters) in the ground truth. Note that WER is usually equal to classification error when the predicted sentence has the same number of words as the ground truth, particularly in our case since almost all errors are substitution errors.

Table 2 summarises the performance of LipNet compared to the baselines. According to the literature, the accuracy of human lipreaders is around 20% (Easton & Basala, 1982; Hilder et al., 2009). As expected, the fixed sentence structure and the limited subset of words for each position in the GRID corpus facilitate the use of context, increasing performance. On the unseen speakers split, the three hearing-impaired people achieve 57.3%, 50.4%, and 35.5% WER respectively, yielding an average of 47.7% WER.

Table 2: Performance of LipNet on the GRID dataset compared to the baselines, measured on two splits: (a) evaluating on only unseen speakers, and (b) evaluating on a 255 video subset of each speakers' sentences.

	Unseen Speakers		**Overlapped Speakers**	
Method	**CER**	**WER**	**CER**	**WER**
Hearing-Impaired Person (avg)	–	47.7%	–	–
Baseline-LSTM	38.4%	52.8%	15.2%	26.3%
Baseline-2D	16.2%	26.7%	4.3%	11.6%
Baseline-NoLM	6.7%	13.6%	2.0%	5.6%
LipNet	$\mathbf{6.4}\%$	$\mathbf{11.4}\%$	$\mathbf{1.9}\%$	$\mathbf{4.8}\%$

For both unseen and overlapped speakers evaluation, the highest performance is achieved by the architectures enhanced with convolutional stacks. LipNet exhibits a $2.3\times$ higher performance in the overlapped compared to the unseen speakers split. For unseen speakers, Baseline-2D and LipNet achieve $1.8\times$ and $4.2\times$ lower WER, respectively, than hearing-impaired people.

The WER for unseen speakers Baseline-2D is 26.7%, whereas for LipNet it is $2.3\times$ lower, at 11.4%. Similarly, the error rate for overlapped speakers was $2.4\times$ lower for LipNet compared to Baseline-2D. Both results demonstrate the importance of combining STCNNs with RNNs. This performance difference confirms the intuition that extracting spatiotemporal features using a STCNN is better than aggregating spatial-only features. This observation contrasts with the empirical observations of Chung & Zisserman (2016a). Furthermore, LipNet's use of STCNN, RNNs, and CTC cleanly allow processing both variable-length input and variable-length output sequences, whereas the architectures of Chung & Zisserman (2016a) and Chung & Zisserman (2016b) only handle the former.

Baseline-LSTM exhibits the lowest performance, in both unseen and overlapped speakers, with 52.8% and 26.3% WER, respectively. Interestingly, although Baseline-LSTM replicates the architecture of Wand et al. (2016), and despite the numerous data augmentation methods, the model performs $1.3\times$ lower than the reported 79.6% word-level accuracy illustrating the difficulty of a sentence-level task even in a restricted grammar.

Finally, by disabling the language model, the Baseline-NoLM exhibits approximately $1.2\times$ higher WER than our proposed model.

4.4 Learned Representations

In this section, we analyse the learned representations of LipNet from a phonological perspective. First, we create saliency visualisations (Simonyan et al., 2013; Zeiler & Fergus, 2014) to illustrate where LipNet has learned to attend. In particular, we feed an input into the model and greedily decode an output sequence, yielding a CTC alignment $\hat{u} \in \tilde{V}^*$ (following the notation of Sections 3.2 and 3.3). Then, we compute the gradient of $\sum_t p(\hat{u}_t|\mathbf{x})$ with respect to the input video frame sequence, but unlike Simonyan et al. (2013), we use guided backpropagation (Springenberg et al., 2014). Second, we train LipNet to predict ARPAbet phonemes, instead of characters, to analyse visual phoneme similarities using intra-viseme and inter-viseme confusion matrices.

4.4.1 Saliency Maps

We apply saliency visualisation techniques to interpret LipNet's learned behaviour, showing that the model attends to phonologically important regions in the video. In particular, in Figure 2 we analyse two saliency visualisations for the words *please* and *lay* for speaker 25, based on Ashby (2013).

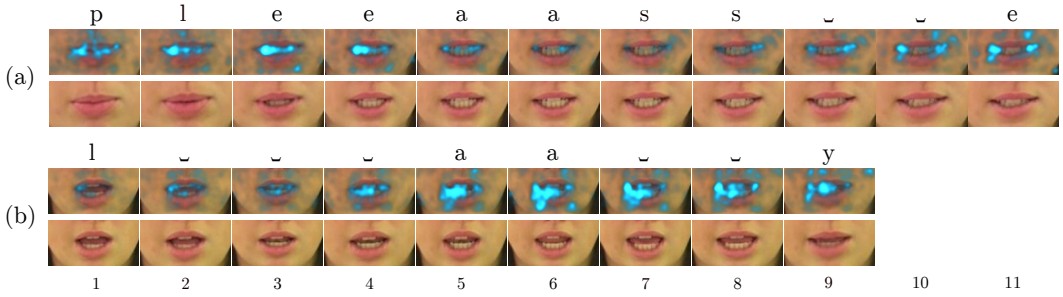


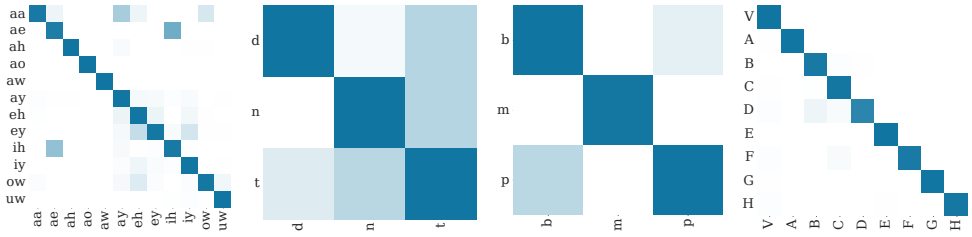
Figure 2: Saliency maps for the words (a) *please* and (b) *lay*, produced by backpropagation to the input, showing the places where LipNet has learned to attend. The pictured transcription is given by greedy CTC decoding. CTC blanks are denoted by '␣'.

The production of the word *please* requires a great deal of articulatory movement at the beginning: the lips are pressed firmly together for the bilabial plosive /p/ (frame 1). At the same time, the blade of the tongue comes in contact with the alveolar ridge in anticipation of the following lateral /l/. The lips then part, allowing the compressed air to escape between the lips (frame 2). The jaw and lips then open further, seen in the distance between the midpoints of the upper and lower lips, and the lips spread (increasing the distance between the corners of the mouth), for the close vowel /iy/ (frame 3–4). Since this is a relatively steady-state vowel, lip position remains unchanged for the rest of its duration (frames 4–8), where the attention level drops considerably. The jaw and the lips then close slightly, as the blade of the tongue needs to be brought close to the alveolar ridge, for /z/ (frames 9–10), where attention resumes.

Lay is interesting since the bulk of frontally visible articulatory movement involves the blade of the tongue coming into contact with the alveolar ridge for /l/ (frames 2–6), and then going down for the vowel /ey/ (frames 7–9). That is exactly where most of LipNet's attention is focused, as there is little change in lip position.

4.4.2 VISEMES

According to DeLand (1931) and Fisher (1968), Alexander Graham Bell first hypothesised that multiple phonemes may be visually identical on a given speaker. This was later verified, giving rise to the concept of a *viseme*, a visual equivalent of a phoneme (Woodward & Barber, 1960; Fisher, 1968). For our analysis, we use the phoneme-to-viseme mapping of Neti et al. (2000), clustering the phonemes into the following categories: Lip-rounding based vowels (V), Alveolar-semivowels (A), Alveolar-fricatives (B), Alveolar (C), Palato-alveolar (D), Bilabial (E), Dental (F), Labio-dental (G), and Velar (H). The full mapping can be found in Table 4 in Appendix A. The GRID corpus contain 31 out of the 39 phonemes in ARPAbet. We compute confusion matrices between phonemes and then



(a) Lip-rounding vowels (b) Alveolar (c) Bilabial (d) Viseme Categories

Figure 3: Intra-viseme and inter-viseme confusion matrices, depicting the three categories with the most confusions, as well as the confusions between viseme clusters. Colours are row-normalised to emphasise the errors.

group phonemes into viseme clusters, following Neti et al. (2000). Figure 3 shows the confusion matrices of the 3 most confused viseme categories, as well as the confusions between the viseme categories. The full phoneme confusion matrix is in Figure 4 in Appendix B.

Given that the speakers are British, the confusion between /aa/ and /ay/ (Figure 3a) is most probably due to the fact that the first element, and the greater part, of the diphthong /ay/ is articulatorily identical with /aa/: an open back unrounded vowel (Ferragne & Pellegrino, 2010). The confusion of /ih/ (a rather close vowel) and /ae/ (a very open vowel) is at first glance surprising, but in fact in the sample /ae/ occurs only in the word *at*, which is a function word normally pronounced with a reduced, weak vowel /ah/. /ah/ and /ih/ are the most frequent unstressed vowels and there is a good deal of variation within and between them, e.g. *priv<u>a</u>te* and *watch<u>e</u>s* (Cruttenden, 2014).

The confusion within the categories of bilabial stops /p b m/ and alveolar stops /t d n/ (Figures 3b-c) is unsurprising: complete closure at the same place of articulation makes them look practically identical. The differences of velum action and vocal fold vibration are unobservable from the front.

Finally, the quality of the viseme categorisation of Neti et al. (2000) is confirmed by the fact that the matrix in Figure 3d is diagonal, with only minor confusion between alveolar (C) and palato-alveolar (D) visemes. Articulatorily, alveolar /s z/ and palato-alveolar /sh zh/ fricatives are distinguished by only a small difference in tongue position: against the palate just behind the alveolar ridge, which is not easily observed from the front. The same can be said about dental /th/ and alveolar /t/.

5 CONCLUSIONS

We proposed LipNet, the first model to apply deep learning to end-to-end learning of a model that maps sequences of image frames of a speaker's mouth to entire sentences. The end-to-end model eliminates the need to segment videos into words before predicting a sentence. LipNet requires neither hand-engineered spatiotemporal visual features nor a separately-trained sequence model.

Our empirical evaluation illustrates the importance of spatiotemporal feature extraction and efficient temporal aggregation, confirming the intuition of Easton & Basala (1982). Furthermore, LipNet greatly outperforms a human lipreading baseline, exhibiting $4.1\times$ better performance, and 4.8% WER which is $2.8\times$ lower than the word-level state-of-the-art (Gergen et al., 2016) in the GRID corpus.

While LipNet is already an empirical success, the deep speech recognition literature (Amodei et al., 2015) suggests that performance will only improve with more data. In future work, we hope to demonstrate this by applying LipNet to larger datasets, such as a sentence-level variant of that collected by Chung & Zisserman (2016a).

Some applications, such as silent dictation, demand the use of video only. However, to extend the range of potential applications of LipNet, we aim to apply this approach to a jointly trained audio-visual speech recognition model, where visual input assists with robustness in noisy environments.

ACKNOWLEDGMENTS

This work was supported by an Oxford-Google DeepMind Graduate Scholarship, the EPSRC, and CIFAR. We would also like to thank: NVIDIA for their generous donation of DGX-1 and GTX Titan X GPUs, used in our experiments; Áine Jackson, Brittany Klug and Samantha Pugh for helping us measure the experienced lipreader baseline; Mitko Sabev for his phonetics guidance; Odysseas Votsis for his video production help; and Alex Graves and Oiwi Parker Jones for helpful comments.

REFERENCES

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

D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. *arXiv preprint arXiv:1512.02595*, 2015.

P. Ashby. *Understanding phonetics*. Routledge, 2013.

J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016a.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016b.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

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*, 120(5):2421–2424, 2006.

A. Cruttenden. *Gimson's pronunciation of English*. Routledge, 2014.

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*, 20(1): 30–42, 2012.

F. DeLand. The story of lip-reading, its genesis and development. 1931.

R. D. Easton and M. Basala. Perceptual dominance during lipreading. *Perception & Psychophysics*, 32(6): 562–570, 1982.

E. Ferragne and F. Pellegrino. Formant frequencies of vowels in 13 accents of the british isles. *Journal of the International Phonetic Association*, 40(01):1–34, 2010.

C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research*, 11(4):796–804, 1968.

Y. Fu, S. Yan, and T. S. Huang. Classification and feature extraction by simplexization. *IEEE Transactions on Information Forensics and Security*, 3(1):91–100, 2008.

A. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

S. Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa. Dynamic stream weighting for turbo-decoding-based audiovisual ASR. In *Interspeech*, pp. 2135–2139, 2016.

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

A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pp. 1764–1772, 2014.

A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pp. 369–376, 2006.

M. Gurban and J.-P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. *IEEE Transactions on Signal Processing*, 57(12):4765–4776, 2009.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

S. Hilder, R. Harvey, and B.-J. Theobald. Comparison of human and machine-based lip-reading. In *AVSP*, pp. 86–89, 2009.

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. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

D. Hu, X. Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3574–3582, 2016.

S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

D. E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 10(Jul):1755–1758, 2009.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

P. Lucey and S. Sridharan. Patch-based representation of visual speech. In *HCSNet workshop on use of vision in human-computer interaction*, pp. 79–85, 2006.

A. L. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng. Lexicon-free conversational speech recognition with neural networks. In *NAACL*, 2015.

I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000.

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

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.

K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Lipreading using convolutional neural network. In *INTERSPEECH*, pp. 1149–1153, 2014.

G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition. In *Workshop on Multimedia Signal Processing*, pp. 264–267, 2007.

G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

S. Petridis and M. Pantic. Deep complementary bottleneck features for visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2304–2308. IEEE, 2016.

V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation. In *Interspeech*, 2006.

G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR Workshop*, 2014.

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

Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, and K. Nakazono. Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss. *Interspeech*, pp. 277–281, 2016.

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

M. F. Woodward and C. G. Barber. Phoneme perception in lipreading. *Journal of Speech, Language, and Hearing Research*, 3(3):212–222, 1960.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265, 2009.

Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen. A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605, 2014.

A ARCHITECTURE DETAILS

In this appendix, we provide additional details about the implementation and architecture.

A.1 IMPLEMENTATION

LipNet is implemented using Torch, the warp-ctc CTC library (Amodei et al., 2015), and Stanford-CTC's decoder implementation. The network parameters were initialised using He initialisation (He et al., 2015), apart from the square GRU matrices that were orthogonally initialised, as described in (Chung et al., 2014). The models were trained with channel-wise dropout (dropout rate $p = 0.5$) after each pooling layer and mini-batches of size 50. We used the optimiser Adam (Kingma & Ba, 2014) with a learning rate of 10^{-4}, and the default hyperparameters: a first-moment momentum coefficient of 0.9, a second-moment momentum coefficient of 0.999, and the numerical stability parameter $\epsilon = 10^{-8}$.

The CER and WER scores were computed using CTC beam search with the following parameters for Stanford-CTC's decoder: beam width 200, $\alpha = 1$, and $\beta = 1.5$. On top of that, we use a character 5-gram binarised language model, as suggested in (Graves & Jaitly, 2014).

A.2 LIPNET ARCHITECTURE

The videos were processed with DLib face detector (King, 2009) and the iBug face shape predictor with 68 landmarks (Sagonas et al., 2013). The RGB input frames were normalised using the following per-channel means and standard deviations: $[\mu_R = 0.7136, \sigma_R = 0.1138, \mu_G = 0.4906, \sigma_G = 0.1078, \mu_B = 0.3283, \sigma_B = 0.0917]$.

Table 3 summarises the LipNet architecture hyperparameters, where T denotes time, C denotes channels, F denotes feature dimension, H and W denote height and width and V denotes the number of words in the vocabulary including the CTC blank symbol.

Table 3: LipNet architecture hyperparameters.

Layer	Size / Stride / Pad	Input size	Dimension order
STCNN	$3 \times 5 \times 5$ / $1, 2, 2$ / $1, 2, 2$	$75 \times 3 \times 50 \times 100$	$T \times C \times H \times W$
Pool	$1 \times 2 \times 2$ / $1, 2, 2$	$75 \times 32 \times 25 \times 50$	$T \times C \times H \times W$
STCNN	$3 \times 5 \times 5$ / $1, 2, 2$ / $1, 2, 2$	$75 \times 32 \times 12 \times 25$	$T \times C \times H \times W$
Pool	$1 \times 2 \times 2$ / $1, 2, 2$	$75 \times 64 \times 12 \times 25$	$T \times C \times H \times W$
STCNN	$3 \times 3 \times 3$ / $1, 2, 2$ / $1, 1, 1$	$75 \times 64 \times 6 \times 12$	$T \times C \times H \times W$
Pool	$1 \times 2 \times 2$ / $1, 2, 2$	$75 \times 96 \times 6 \times 12$	$T \times C \times H \times W$
Bi-GRU	256	$75 \times (96 \times 3 \times 6)$	$T \times (C \times H \times W)$
Bi-GRU	256	75×512	$T \times F$
Linear	$27 +$ blank	75×512	$T \times F$
Softmax		75×28	$T \times V$

Note that spatiotemporal convolution sizes depend on the number of channels, and the kernel's three dimensions. Spatiotemporal kernel sizes are specified in the same order as the input size dimensions. The input dimension orderings are given in parentheses in the input size column.

Layers after the Bi-GRU are applied per-timestep.

A.3 BASELINE-LSTM ARCHITECTURE

Baseline-LSTM replicates the setup of Wand et al. (2016), and is trained the same way as LipNet. The model uses two LSTM layers with 128 neurons. The input frames were converted to grayscale and were down-sampled to 50×25px, dropout $p = 0$, and the parameters were initialised uniformly with values between $[-0.05, 0.05]$.

B PHONEMES AND VISEMES

Table 4 shows the phoneme to viseme clustering of Neti et al. (2000) and Figure 4 shows LipNet's full phoneme confusion matrix.

Table 4: Phoneme to viseme clustering of Neti et al. (2000).

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

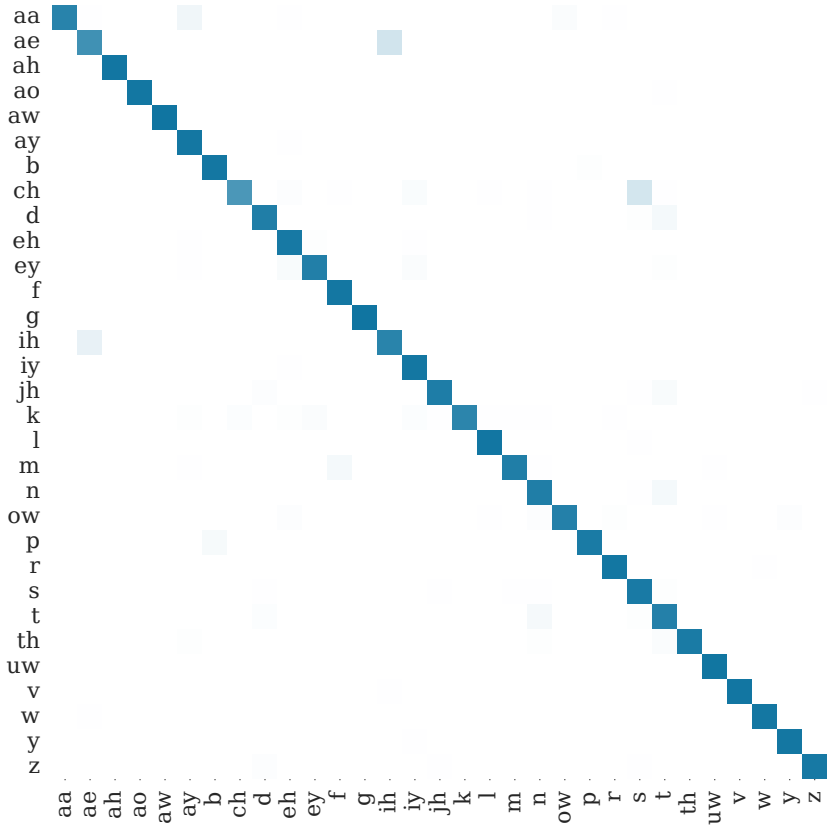


Figure 4: LipNet's full phoneme confusion matrix.