LIPREADING WITH LONG SHORT-TERM MEMORY

Michael Wand, Jan Koutník, Jürgen Schmidhuber

The Swiss AI Lab IDSIA, USI & SUPSI

ABSTRACT

Lipreading, i.e. speech recognition from visual-only recordings of a speaker's face, can be achieved with a processing pipeline based solely on neural networks, yielding significantly better accuracy than conventional methods. Feedforward and recurrent neural network layers (namely Long Short-Term Memory; LSTM) are stacked to form a single structure which is trained by back-propagating error gradients through all the layers. The performance of such a stacked network was experimentally evaluated and compared to a standard Support Vector Machine classifier using conventional computer vision features (Eigenlips and Histograms of Oriented Gradients). The evaluation was performed on data from 19 speakers of the publicly available GRID corpus. With 51 different words to classify, we report a best word accuracy on held-out evaluation speakers of 79.6% using the end-to-end neural network-based solution (11.6% improvement over the best feature-based solution evaluated).

Index Terms— Lipreading, Long Short-Term Memory, Recurrent Neural Networks, Image Recognition

1. INTRODUCTION

It is well-known that humans understand speech not only by listening, but also by taking visual cues into account [1]. Hearing-impaired persons are in fact able to comprehend human speech by purely visual *lipreading*, i.e. by processing visual information from a speaker's lips and face. Consequently, research on making lipreading available to electronic speech recognition and processing systems has been of interest for some decades, with pioneering work done by Petajan [2]. His PhD thesis proposed to use lipreading to augment conventional automatic speech recognition (ASR), yet later researchers started to perform *purely visual* speech recognition [3], which is also the goal of this study. Lipreading systems typically consist of (at least) *feature extraction* and *classification*. The feature extraction can become quite complex: Many recent lipreading systems, e.g. [4, 5], use a lip tracking system as a first stage, followed by versatile image features such as Active Appearance Models [6] or

Local Binary Patterns [7]. Classification is frequently done with Support Vector Machines (SVMs), e.g. [7], or Hidden Markov Models (HMMs), e.g. [4, 5, 8, 9].

Our aim is to replace the complete visual speech recognition pipeline with a compact neural network architecture. Neural networks (NNs) have become increasingly popular in conventional speech recognition, first as feature extractors in an HMM-based architecture [10–12], more recently replacing the entire processing chain [13]. For the latter, the *Long Short Term Memory* (LSTM; [14]) architecture is typically used. Consequently, our approach to the lipreading problem uses a NN that chains feed-forward layers and LSTM layers, described in detail in subsection 4.2. Manual feature extraction is no longer required. The NN inputs are now the raw mouth images, as is common in modern computer vision tasks, but stands in stark contrast e.g. to [5, 7].

2. RELATED WORK

Lipreading has been used as a complementary modality for speech recognition from noisy audio data [2, 15], as well as for purely visual speech recognition [3, 16, 17]. The latter gives rise to a *Silent Speech interface*, which is defined as a system "enabling speech communication to take place when an audible acoustic signal is unavailable" [18]. Silent Speech technology has a large number of applications: It allows persons with certain speech impairments (e.g. laryngectomees, whose voice box (*larynx*) has been removed) to communicate, as well as enabling confidential and undisturbing communication in public places [18]. Further uses of lipreading have been proposed, e.g. automatic speech extraction from surveillance videos and its interpretation for forensic purposes [5]. Lipreading has been augmented with *ultrasound* images of the tongue and vocal tract [19–21]. Furthermore, there are Silent Speech interfaces based on very different principles, like speech recognition from electromyography [22–25] or (electro-)magnetic articulography [26].

NNs have been used in speech recognition as feature extractors in HMM-based speech recognizers [10, 11]. Neural networks, LSTMs in particular, started to replace larger parts of the speech processing chain previously dominated by HMMs. An end-to-end neural network system [27, 28] finally outperformed HMM-based systems and achieved the best performance (16% error) on the large Switchboard Hub5'00

Fig. 1. Two randomly chosen example frames from the GRID corpus with highlighted mouth area.

speech recognition benchmark [29].

Massively parallel graphics processing units (GPUs) became available in the last few years. Since then, *Convolutional* NNs (CNNs) trained by gradient descent [30] dominate, e.g. [31], the area of image recognition, as well as related tasks like object detection and segmentation. The first CNN application in lipreading [17] uses the CNN as a preprocessor for an HMM-based sequence classifier.

3. THE GRID DATA CORPUS

Our experiments were performed using the GRID audiovisual corpus [32][1], consisting of video and audio recordings of 34 speakers saying 1000 sentences each. The total length of the recordings is 28 hours; two example video frames are shown in Figure 1. Each of the sentences has a fixed structure: *command(4) + color(4) + preposition(4) + letter(25) + digit(10) + adverb(4)*, for example "Place red at J 2, please", where the number of alternatives words is given in parentheses. A total of 51 different words are contained in the GRID corpus; the alternative words for each of the six sentence parts are distributed uniformly. The letter *W* is excluded because its pronunciation is vastly longer than for any other letter.

Sentences have a fixed length of 3 seconds at a frame rate of 25 frames per second. Each sentences thus spans across 75 frames. Video data bis available in "normal" and "high" quality; the normal quality video with 360×288 pixel resolution, converted to greyscale, was used. Unreadable videos, in particular for speaker number 8, were discarded. The frame-level alignments distributed with the corpus were used to obtain word level segmentations of the video, causing the training dataset to consist of $6 \cdot 1000 = 6000$ single words per speaker. The acoustic part of the GRID corpus was not used.

A 40×40 pixel window containing the mouth area from each video frame was extracted using the following procedure. The face area localized using the Mathematica `FindFaces[ ]` function was converted into the LAB color-space. The *A* component pixels were multiplied element-wise by a Gaussian matrix (with mean located at 30% of of the

image height along the middle column) and $\sigma = 500$ pixels, and rescaled into $[0, 1]$ interval. The center of mass of pixels that have value above 0.9 was considered as the mouth center. The face area size was inflated by factor of 1.5, scaled to 128 pixel width, and a window of 40×40 pixels with the mouth coordinates in the center was extracted. This patch was converted to greyscale, the contrast was maximized (i.e. all pixel values were remapped to $[0, 1]$ interval), and all the values in the complete dataset were standardized.

Speakers 1–19 from the entire GRID were used: speakers 1-9 form the development set that was used to determine the parameter settings; speakers 10–19 form the evaluation set, held back until the final evaluation of the systems. All experiments are *speaker-dependent*, i.e. training and test data for the classifiers were always taken from the same speaker. The results reported in this paper are averaged over the speakers. The data for each speaker was randomly divided into training, validation, and test sets, where the latter two contain five sample videos of each word, i.e. a total of $51 \cdot 5 = 255$ samples each. The training data is however highly unbalanced: For example, each letter from "a" to "z" appears 30 times, whereas each color appears 240 times.

4. METHODS

4.1. Baseline Feature Extraction and Classification

The NN-based lipreader was compared to a baseline SVM classifier using conventional features, namely *Eigenlips* [15], which were used as a baseline feature in [17], and *Histograms of Oriented Gradients* (HOG) [33] as a more complex feature which yielded good performance in preliminary experiments.

Eigenlip features are created from raw frames by computing the PCA decomposition on the training data and then transforming all the images by multiplication with the PCA matrix, retaining only a certain number of dimensions ordered by maximal variance. HOG is originally a feature extractor for object recognition [33]; it divides the image window into small spatial regions (cells) and accumulates a local 1-D histogram of gradient directions or edge orientations over the pixels in each cell. Histogram entries are normalized over larger spatial areas. The HOG features were obtained using the VLFeat library [34].

Since SVMs are not sequence classifiers, a single feature vector has to be computed from the sequence of frames representing each word, called *sequence feature vector*. Sequences vary in length. For example, a typical letter "a" is pronounced in 3–4 frames, whereas a longer word like "please" can occupy more than 10 frames. In the sequence feature vector, all frames are stacked while enforcing a specified vector length: frames are repeated if a sequence is shorter than this length, neighboring frames of longer sequences are averaged.

Table 1. Word Accuracy averaged over the Development Speakers 1–9 and Evaluation Speakers 10–19.

| Spk.	Method	Word Acc.	Std. Dev.
1–9	Eigenlips + SVM	68.4%	7.4%
	HOG + SVM	71.1%	6.7%
	LSTM	79.4%	4.3%
10–19	Eigenlips + SVM	70.6%	4.2%
	HOG + SVM	71.3%	4.3%
	LSTM	79.6%	4.3%

4.2. Neural Network Lipreader

Neural networks (NNs) consist of processing units (neurons) connected by trainable weighted connections. The neurons are typically organized in layers, which can be broadly distinguished as follows based on their connectivity: (1) feedforward NNs pass the input signal to the output neurons without allowing cyclic computations; (2) recurrent NNs wire the connections in a cycle which forms a temporal memory. NNs are, in the supervised case, typically trained by gradient descent, which is realized by error back-propagation through the layers followed by adjustment of the weights. In the case of recurrent NNs the error propagates along the time axis as well, referred to as back-propagation through time, implemented by unfolding the recurrent connections into a feedforward structure as deep as the length of the sequence. Such deep networks cause the gradient to explode or vanish [35], which can be fixed by replacing a single recurrent NN unit by a LSTM cell that avoids the problem by linear recurrent connection of a *cell* inside the LSTM unit [14]. The information flow through the LSTM cell is regulated by *input*, *output* and *forget* [36] gates using multiplicative connections. See [37] for detailed LSTM description, analysis and setup.

5. EXPERIMENTS

5.1. Experimental Setup

Two feature extractions (Eigenlips and HOG), both combined with the SVM classifier, were compared to the LSTM lipreader. All parameters were optimized on the development speakers 1–9. The error on all systems is reported on the speaker-dependent test set, no training error is reported.

In a series of experiments on the speakers 1–9, the optimal configuration of the feature extraction for the SVM experiments was determined, as well as the best neural network structure. The best PCA cutoff for the Eigenlip features was at 100 components, the best HOG cell size was 8. Best SVM recognition results were obtained with a linear kernel at a sequence feature vector length of 6 frames. Increasing the fea-
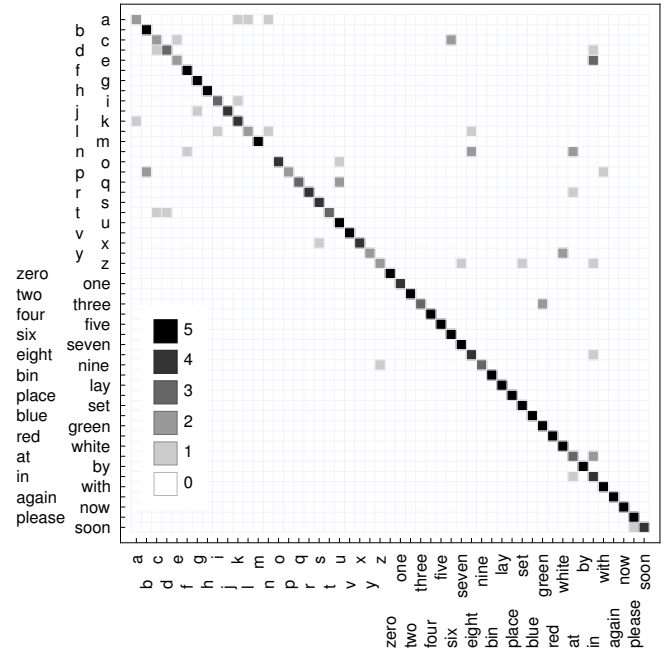
Fig. 2. Confusion matrix for speaker 7, using the neural network lipreader. Rows represent reference labels, columns represent the classification hypothesis. Greyscale level indicates the density of matching reference and hypothesis. The accuracy is highest on longer words and lowest on letters.

ture vector length did not improve the accuracy, almost certainly because among the short video sequences containing letters, where most errors occur (see subsection 5.2), 6 frames already cover the entire information available. Taking this observation into account, we hypothesize that a dedicated sequence classifier like an HMM would *not* substantially improve the classification accuracy on this corpus.

The input data for the SVM can become very high-dimensional, which we assume to be the reason why higher-degree polynomial SVM kernels yielded lower accuracy than linear kernels. Zhao et al. [7] report that second-degree polynomial SVM kernels perform the best, however their SVM classifier treats the sequential information differently.

The best LSTM lipreader consists of one feed-forward layer followed by two recurrent LSTM layers, 128 units (neurons / LSTM cells) each, and a softmax layer with 51 units that perform the word classification. The learning rate was set to 0.02, momentum was not used, and early stopping with a delay of 10 epochs was used. Weights were initialized from uniform distribution over the range $[-0.05, 0.05]$.

5.2. Results

The word level classification accuracies are summarized in Table 1. The LSTM lipreader yields statistically significant improvement (one-tailed t-test with $p = 0.05$) over the conventional features combined with the SVM classifier. The Eigenlip and HOG features perform similarly, both worse than the LSTM lipreader. On the Evaluation Speakers 10–19, the LSTM lipreader improves the accuracy by 11.6%, compared to the best conventional solution (HOG + SVM).

Figure 2 shows a typical confusion matrix on speaker 7 data, where classification was performed with the LSTM lipreader. The rows show reference word labels, the columns show hypotheses. The confusion on letters is far higher (upper part of the matrix) than on longer words (lower part of the matrix). For this speaker and configuration the accuracy on the letters is 69.8% (at 4% chance level), the accuracy on the non-letter words is 93.4% (at 3.8% chance level). The total accuracy is 82.0%.

The discrepancy between letters and other words is caused by three major factors: First, the letters from 'a' to 'z' are highly confusing even under optimal circumstances. In particular, this applies to voiced and voiceless versions of the same letter, like 'p' and 'b'. Consequently, *visemes* (visual units for recognition) frequently do not distinguish between such similarly-looking sounds at all (see e.g. [38] and the references therein). Second, single letters video sequences are often are very short, sometimes consisting of only 3-4 frames. This means that very little data is available and also that the letter pronunciation is highly influenced by adjacent sounds. This context does *not* help distinguishing different letters, since the parts of the GRID sentences are statistically independent from each other. We note that [17] report phone accuracy on a corpus which consists of *whole words*: here the context plays a great role in *improving* recognition.

The confusion matrices are qualitatively similar across all speakers and all experimental setups – the longer words are recognized with close to 100% accuracy, whereas confusion is highest on the letters.

6. CONCLUSION

This study shows that the neural network based lipreading system applied to raw images of the mouth regions achieves significantly better word accuracy than a system based on a conventional processing pipeline utilizing feature extraction and classification. The LSTM lipreader with a single feed-forward network, which learns the features automatically together with training the LSTM sequence classifier, consistently achieved almost 80% word accuracy in speaker-dependent lipreading.

The experiments, not described in this paper, also included the highly popular CNNs instead of the fully connected feed-forward layer, but the results did not improve.

One of the possible reasons is that the small, 40×40 pixel area already contains just enough information for the classification. Experiments with CNNs and large image sizes as well as evaluation of a speaker-independent LSTM lipreader are the subject of future experiments.

7. REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, no. 5588, pp. 746 – 748, 1976.

[2] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading) ," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1984.

[3] G. I. Chiou and J.-N. Hwang, "Lipreading from Color Video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192 – 1195, 1997.

[4] Y. Lan, R. Harvey, B.-J. Theobald, E.-J. Ong, and R. Bowden, "Comparing Visual Features for Lipreading," in *Proc. of the Int. Conference on Auditory-Visual Speech Processing*, 2009, pp. 102 – 106.

[5] R. Bowden, S. Cox, R. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B.-J. Theobald, "Recent Developments in Automated Lip-reading," in *Proc. SPIE*, 2013.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Trans. on Pattern Anal. and Machine Intel.*, vol. 23, no. 6, pp. 681 – 685, 2001.

[7] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading With Local Spatiotemporal Descriptors," *IEEE Trans.s on Multimedia*, vol. 11, no. 7, pp. 1254 – 1265, 2009.

[8] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," in *Proc. Interspeech*, 2007, pp. 658–661.

[9] F. Tao and C. Busso, "Lipreading Approach for Isolated Digits Recognition Under Whisper and Neutral Speech," in *Proc. Interspeech*, 2014, pp. 1154 – 1158.

[10] H. Bourlard and N. Morgan, *Connectionist Speech Recognition. A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[11] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Proc. Magazine*, vol. 29, no. 6, pp. 82 – 97, 2012.

[12] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid Speech Recognition with Deep Bidirectional LSTM," in *Proc. ASRU*, 2013, pp. 273 – 278.

[13] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proc. ICASSP*, 2013, pp. 6645 – 6649.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comp.*, vol. 9, pp. 1735 – 1780, 1997.

[15] C. Bregler and Y. Konig, "'Eigenlips' for Robust Speech Recognition," in *Proc. ICASSP*, 1994, pp. 669 – 672.

[16] R. Bowden, S. Cox, R. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B.-J. Theobald, "Is Automated Conversion of Video to Text a Reality?" in *Proc. SPIE*, 2012.

[17] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using Convolutional Neural Network," in *Proc. Interspeech*, 2014, pp. 1149 – 1153.

[18] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. Gilbert, "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270 – 287, 2010.

[19] B. Denby and M. Stone, "Speech Synthesis from Real Time Ultrasound Images of the Tongue," in *Proc. ICASSP*, 2004, pp. I–685 – I–688.

[20] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue Feature Extraction for an Ultrasound-based Silent Speech Interface," in *Proc. ICASSP*, 2007, pp. I–1245 – I–1248.

[21] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips," *Speech Communication*, vol. 52, pp. 288 – 300, 2010.

[22] N. Sugie and K. Tsunoda, "A Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 7, pp. 485 – 490, 1985.

[23] M. S. Morse, S. H. Day, B. Trull, and H. Morse, "Use of Myoelectric Signals to Recognize Speech," in *Proc. 11th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, 1989, pp. 1793 – 1794.

[24] T. Schultz and M. Wand, "Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition," *Speech Comm.*, vol. 52, no. 4, pp. 341 – 353, 2010.

[25] M. Wand, M. Janke, and T. Schultz, "Tackling Speaking Mode Varieties in EMG-based Speech Recognition," *IEEE Transaction on Biomedical Engineering*, vol. 61, no. 10, pp. 2515 – 2526, 2014.

[26] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (Silent) Speech Recognition System for Patients Following Laryngectomy," *Medical Engineering and Physics*, vol. 30, pp. 419 – 425, 2008.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, pp. 369 – 376.

[28] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," in *Proc. ICML*, 2014, pp. 1764 – 1772.

[29] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "DeepSpeech: Scaling up end-to-end speech recognition," arXiv: 1412: 5567v1, 2014.

[30] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackeland, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, pp. 541 – 551, 1989.

[31] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A Committee of Neural Networks for Traffic Sign Classification," in *Proc. IJCNN*, 2011, pp. 1918 – 1921.

[32] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Soc. of America*, vol. 120, no. 5, pp. 2421 – 2424, 2006.

[33] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. CVPR*, vol. 1, 2005, pp. 886 – 893.

[34] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," http://www.vlfeat.org/, 2008.

[35] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001, ch. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

[36] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[37] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *CoRR*, vol. abs/1503.04069, 2015. [Online]. Available: http://arxiv.org/abs/1503.04069

[38] L. Cappelletta and N. Harte, "Viseme Definitions Comparison for Visual-only Speech Recognition," in *Proc. EUSIPCO*, 2011, pp. 2109 – 2113.