Improving Speaker-Independent Lipreading with Domain-Adversarial Training

Michael Wand, Jürgen Schmidhuber

The Swiss AI Lab IDSIA, USI & SUPSI, Manno-Lugano, Switzerland

`michael@idsia.ch`, `juergen@idsia.ch`

Abstract

We present a *Lipreading* system, i.e. a speech recognition system using only visual features, which uses *domain-adversarial training* for speaker independence. Domain-adversarial training is integrated into the optimization of a lipreader based on a stack of feedforward and LSTM (Long Short-Term Memory) recurrent neural networks, yielding an end-to-end trainable system which only requires a very small number of frames of *untranscribed* target data to substantially improve the recognition accuracy on the target speaker. On pairs of different source and target speakers, we achieve a relative accuracy improvement of around 40% with only 15 to 20 seconds of untranscribed target speech data. On multi-speaker training setups, the accuracy improvements are smaller but still substantial.

Index Terms: Lipreading, Deep Neural Networks, Long Short-Term Memory, Domain Adaptation

1. Introduction

Lipreading is the process of understanding speech by using solely visual features, i.e. images of the lips of a speaker. In communication between humans, lipreading has a twofold relevance [1]: First, visual cues play a role in spoken conversation [2]; second, hearing-impaired persons may use lipreading as a means to follow verbal speech.

With the success of computer-based speech recognition over the past decades, automatic lipreading has become an active field of research as well, with pioneering work by Petajan [3], who used lipreading to augment conventional acoustic speech recognition, and Chiou and Hwang [4], who were the first to perform lipreading without resorting to any acoustic signal at all. Since 2014, lipreading systems have systematically begun to use neural networks at part of the processing pipeline [5, 6] or for end-to-end-training [7, 8, 9]. In our previous work [7], we proposed a fully neural network based system, using a stack of fully connected and recurrent (LSTM, Long Short-Term Memory) [10, 11] neural network layers.

The scope of this paper is the introduction of state-of-the-art methods for *speaker-independent lipreading* with neural networks. We evaluate our established system [7] in a cross-speaker setting, observing a drastic performance drop on unknown speakers. In order to alleviate the discrepancy between training speakers and unknown test speaker, we use *domain-adversarial training* as proposed by Ganin and Lempitsky [12]: *Untranscribed* data from the target speaker is used as additional training input to the neural network, with the aim of pushing the network to learn an intermediate data representation which is domain-agnostic, i.e. which does not depend on whether the input data comes from a source speaker or a target speaker. We evaluate our system on a subset of the GRID corpus [13], which contains extensive data from 34 speakers and is therefore ideal for a systematic evaluation of the proposed method.

2. Related work

Lipreading can be used to complement or augment speech recognition, particularly in the presence of noise [3, 14], and for purely visual speech recognition [4, 15, 5]. In the latter case, ambiguities due to incomplete information (e.g. about voicing) can be mitigated by augmenting the video stream with *ultrasound* images of the vocal tract [16]. Visual speech processing is an instance of a *Silent Speech interface* [17]; further promising approaches include capturing the movement of the articulators by electric or permanent magnetic articulography [18, 19], and capturing of muscle activity using electromyography [20, 21, 22, 23].

Versatile lipreading features have been proposed, such as Active Appearance Models [24], Local Binary Patterns [25], and PCA-based *Eigenlips* [26] and *Eigentongues* [27]. For tackling speaker dependency, diverse scaling and normalization techniques have been employed [28, 29]. Classification is often done with Hidden Markov Models (HMMs), e.g. [30, 15, 31, 32]. Mouth tracking is done as a preprocessing step [32, 15, 5]. For a comprehensive review see [33].

Neural networks have early been applied to the Lipreading task [34], however, they have become widespread only in recent years, with the advent of state-of-the-art learning techniques (and the necessary hardware). The first deep neural network for lipreading was a seven-layer convolutional net as a preprocessing stage for an HMM-based word recognizer [5]. Since then, several end-to-end trainable systems were presented [7, 8, 9]. The current state-of-the-art accuracy on the GRID corpus is 3.3% error [9] using a very large set of *additional* training data; so their result is not directly comparable to ours.

In *domain adaptation*, it is assumed that a learning task exhibits a domain shift between the training (or *source*) and test (or *target*) data. This can be mitigated in several ways [35]; we apply *domain-adversarial training* [12], where an intermediate layer in a multi-layer network is driven to learn a representation of the input data which is optimized to be domain-agnostic, i.e. to make it difficult to detect whether an input sample is from the source or the target domain. A great advantage of this approach is the end-to-end trainability of the entire system. For a summary of further approaches to domain adaptation with neural networks, we refer to the excellent overview in [12].

3. Data and preprocessing

We follow the data preprocessing protocol from [7]. We use the GRID corpus [13], which consists of video and audio recordings of 34 speakers (which we name *s1* to *s34*) saying 1000 sentences each. All sentences have 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 alternative words is given in parentheses. There are 51 distinct words; alternatives are randomly distributed so that

Figure 1: *Two randomly chosen example frames from the GRID corpus with highlighted mouth area.*
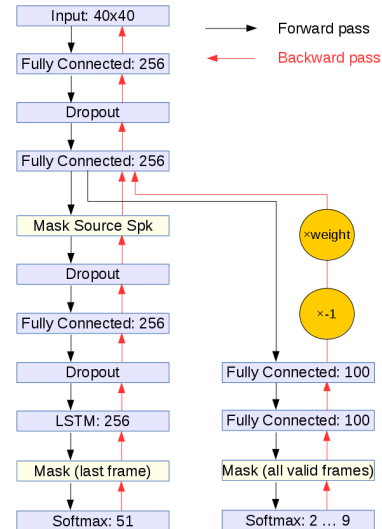

Figure 2: *Optimal network topology for adversarial training, with a common part (top), word classifier (bottom left, only on source speakers), and speaker classifier (bottom right). Note that the gradient of the speaker classifier is inverted, and that the contribution of the adversarial network is configurable.*

context cannot be used for classification. Each sentence has a length of 3 seconds at 25 frames per second, so the total data per speaker is 3000 seconds (50 minutes). Using the annotations contained in the corpus, we segmented all videos at word level, yielding 6000 word samples per speaker.

We experiment on speakers s1–s19: speakers 1-9 form the *development* speakers, used to determine optimal parameters; speakers 10–19 are the *evaluation* speakers, held back until the final evaluation of the systems. The data from each speaker was randomly subdivided into training, validation, and test sets, where the latter two contain five samples of each word, i.e. a total of $51 \cdot 5 = 255$ samples each. The training data is consequently highly unbalanced: For example, each letter from "a" to "z" appears 30 times, whereas each color appears 240 times.

We converted the "normal" quality videos (360×288 pixels) to greyscale and extracted 40×40 pixel windows containing the mouth area, as described in [7]. The frames were contrast-normalized and z-normalized over the training set, *independently* for each speaker. Unreadable videos were discarded.

All experiments have one dedicated *target* speaker on which this experiment is evaluated, and one, four, or eight *source* speakers on which supervised training is performed. Speakers are chosen consecutively, for example, the experiments on *four* training speakers on the development data are (s1 ... s4) → s5, (s2 ... s5) → s6, ⋯, (s9, s1, s2, s3) → s4, where → separates source and target speakers. We also compute baseline results on single speakers. The data sets of each speaker are used as follows: *Training* data is used for supervised training (on the source speakers) and unsupervised adaptation (on the target speaker). *Validation* data is used for early stopping, the network is evaluated on the *test* data.

4. Methods and System Setup

The system is based on the lipreading setup from [7], reimplemented in Tensorflow [36]. Raw 40×40 lip images are used as input data, *without* any further preprocessing except normalization. We stack several fully connected feedforward layers, optionally followed by Dropout [37], and one LSTM recurrent layer to form a network which is capable of recognizing sequential video data. The final layer is a softmax with 51 word targets. All inner layers use a tanh nonlinearity. During testing, classification is performed on the *last* frame of an input word, the softmax output on all previous frames is discarded. Similarly, during training, an error signal is backpropagated (through time and through the stack of layers) only from the last frame of each training word sample.

Optimization is performed by minimizing the multi-class cross-entropy using stochastic gradient descent applying Tensorflow's `MomentumOptimizer` with a momentum of 0.5, a

learning rate of 0.001, and a batch size of 8 sequences. The network weights are initialized following a truncated normal distribution with a standard deviation of 0.1. In order to compensate for the unbalanced training set, each training sample is weighted with a factor inversely proportional to its frequency. Early stopping (with a patience of 30 epochs) is performed on the validation data of the *source* speakers.

Adversarial training [12] is integrated as follows. At the second feedforward layer, we attach a further network which performs framewise *speaker* classification on source and target speakers. For this purpose, each training batch of 8 word sequences is augmented by eight additional word sequences from the target speaker, for which *no word label is used*, and *no gradient is backpropagated* from the word classifier. On the extended batch of 16 sequences, the "adversarial" network performs framewise speaker classification. This network follows a standard pattern (two feedforward layers with 100 neurons each plus a softmax layer with 2, 5, or 9 speaker outputs) and is trained jointly with the word classifier, with a configurable weight. If there are more word sequences from the source speaker(s) than from the target speaker, target sequences are repeated.

So far, this describes a joint classifier for two different tasks (speaker and word classification), resembling Caruana's Multi-task training [38]. The power of the adversarial network comes from a simple twist: The backpropagated gradient from the adversarial network is *inverted* where it is fed into the main branch of the network, causing the lower branch to perform gradient *ascent* instead of descent. Since the speaker classification part of the system learns to classify speakers, the inverted gradient fed into the "branching" layer causes the joint part of the network to learn to *confuse* speakers instead of separating them. The speaker classifier and the joint network work for opposite objectives (hence, "adversarial"); an idea first presented in the context of *factorial codes* [39]. Figure 2 shows a graphical overview of the system: The joint part is at the top, at the bottom are word classifier (left) and speaker classifier (right).

Table 1: *Baseline word accuracies on single speakers, averaged over the development set, with standard deviation. Layer types are **FC** (fully connected feedforward), **DP** (Dropout), and **LSTM**, followed by the number of neurons/cells. ∗ marks the (reimplemented and recomputed) best system from [7].*

Network	Training acc.	Test acc.
FC128-LSTM128-LSTM128 ∗	100.0% ±0.0%	78.5% ±5.6%
FC128-FC128-LSTM128	100.0% ±0.0%	79.5% ±5.8%
FC256-FC256-LSTM256	100.0% ±0.0%	79.4% ±5.7%
FC256-FC256-FC256-LSTM256	100.0% ±0.0%	79.0% ±5.6%
FC256-DP-FC256-DP-FC256-DP-LSTM256	96.4% ±1.9%	83.3% ±5.7%

5. Experiments and Results

5.1. Baseline Lipreader

The first experiment deals with establishing a baseline for our experiments, building on prior work [7]. We run the lipreader as a *single-speaker* system with different topologies, optionally using Dropout (always with 50% dropout ratio) to avoid overfitting the training set. Adversarial training is not used (i.e. the weight in figure 2 is set to zero). Table 1 shows the resulting test set accuracies averaged over the development speakers.

Without using Dropout, the accuracy on the test set is ∼79%. Note in particular that the baseline cannot substantially be improved by increasing the layer size or adding more layers. We remark that not only the average accuracy across speakers, but also the accuracies for every single speaker hardly vary.

The situation changes when Dropout is used: Now our best average accuracy is 83.3%, which is in line with results reported in literature (the most recent best result is 86.4% word accuracy [40], but with a different training/test data split). This best system, which is employed in the remainder of this paper, uses three feedforward layers each followed by Dropout, with 256 neurons each, followed by the LSTM layer with 256 LSTM cells, and the softmax layer. Thus the system is larger and has more layers than the baseline system, which is indeed made possible by the Dropout regularizer.

On the evaluation speakers, the baseline system achieves an average accuracy of 78.3%, and the Dropout system is at 83.9% accuracy. This improvement is significant (one-tailed t-test with paired samples, $p = 2.38 \times 10^{-6}$).

The accuracies in a cross-speaker setting, again on the development speakers, are given in table 2. The accuracy decreases drastically, in particular when only one source speaker is used for training: On an unknown target speaker, the system achieves only an average 13.5% accuracy. The situation is clearly better when training data from multiple speakers is used, but even for eight training speakers, the average accuracy on an unknown speaker is only 37.8%. We also note that the test accuracy on the *source* speakers does not rise when data from multiple speakers is used, even though there is more training data. It appears that the additional data does not "help" the system to improve its performance. On an unknown speaker, however, training data from multiple speakers does improve performance, very probably the system learns to be more speaker-agnostic. A similar observation with a very different input signal was reported in [41].

Clearly, lipreading across different speakers is a challenging problem. In the remainder of this paper, we show how domain-adversarial training helps to tackle this challenge.

Table 2: *Baseline word accuracies on training across speakers without adaptation by domain-adversarial training, averaged over the development set, with standard deviation. The best network from table 1 (FC256-DP-FC256-DP-FC256-DP-LSTM256) was used.*

Number of training spk	Source spk		Target spk Test acc.
	Train acc.	Test acc.	
1	96.6% ±1.5%	81.9% ±6.4%	13.5% ±6.9%
4	89.3% ±1.9%	78.4% ±3.2%	31.2% ±8.0%
8	82.1% ±0.8%	74.5% ±1.0%	37.8% ±9.8%

5.2. Tuning of the Adversarial System

We now augment the baseline word classification network with adversarial training as described in section 4, thus making full use of the system shown in figure 2. For now, we use *all* sequences from the training set of the target speaker. As suggested in [12], we found it beneficial to gradually activate adversarial training: the weight of the adversarial part is set to zero at the beginning, every 10 epochs, it is raised by 0.2 until the maximum value of 1.0 is reached at epoch 50. The results of this experiment are shown in the upper two blocks of table 3, where it can be seen that adversarial training causes substantial accuracy improvement, particularly with only one source speaker: In this case, the accuracy rises by more than 40% relative, from 13.5% to 19.2%. In the case of four or eight source speakers, the accuracy improves by 13.1% resp. 12.2% relative. We tuned this system using various topologies for the adversarial part, as well as different weight schedules for adversarial training, finding rather consistent behavior. The only setting which is emphatically discouraged is starting with an adversarial weight greater than zero. See section 6 for further analysis.

5.3. Training with Very Little Target Data

While the presented system does not require supervised training data from the target speaker, we still use the entire training set of the target speaker. In practical applications, even unsupervised training data may only be sparsely available, so this setup is somewhat undesired.

Since the content of the target training sequences is irrelevant for the adversarial training, we may hypothesize that we could also do with a much smaller set of target training data. So as a final experiment, we reduce the number of training sequences for the target speaker. The training protocol remains as before; in particular, training is always performed on the full set of source sequences, target sequences are repeated as necessary.

Table 3: *Word accuracies and standard deviations for systems with adversarial training on* all *target sequences or a* subset of 50 *target sequences, on the development speakers.*

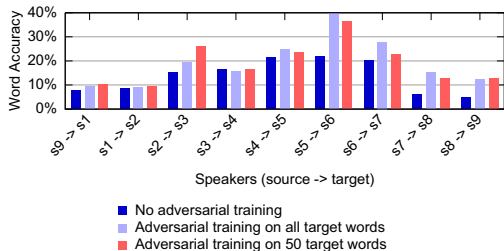| Adversarial Training on	Number of training spk	Target Test acc.	Relative Improvement
None	1	13.5% ±6.9%	-
	4	31.2% ±8.0%	-
	8	37.8% ±9.8%	-
All Target Sequences	1	19.2% ±10.0%	42.0%
	4	35.3% ±10.5%	13.1%
	8	42.4% ±13.0%	12.2%
50 Target Sequences	1	18.9% ±8.9%	40.0%
	4	32.9% ±10.5%	5.4%
	8	41.8% ±9.7%	10.6%

Figure 3: *Accuracies with and without adversarial training on pairs of one source and one target speaker, on the development set.*
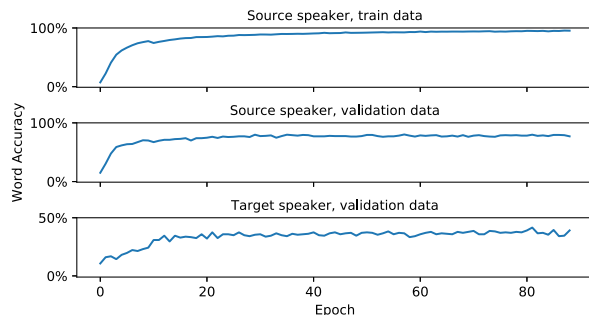

Figure 4: *Accuracy vs. epoch on different data sets with adversarial training, for speaker pair s5 → s6. Note that the target accuracy shows a substantial rise at epoch 10, where adversarial training sets in.*

The original number of 5490 target training sequences can be reduced to 50 sequences without a substantial loss of accuracy—this amounts to only 15-20 seconds of untranscribed target data. Results are shown in the lower block of table 3: For example, in the case of a single source speaker, the target accuracy drops to 18.9% instead of 19.2%. The improvement is lower when more source speakers are used. We hypothesize that this stems from the growing ratio between the number of source sequences and the number of target sequences.

Finally, figure 3 shows an accuracy breakdown for speaker pairs, i.e. for single-speaker supervised training. In eight out of nine cases, domain-adversarial training clearly outperforms the baseline system, often by a substantial margin. We also observe that the accuracy gain depends very much on the speaker pair.

5.4. Evaluation

We evaluate our result on the evaluation speakers, i.e. speakers 10–19 from the GRID corpus. The hypothesis to be tested states that adversarial training improves the accuracy of the cross-speaker lipreader trained on one, four, or eight source speakers, using either all target sequences or 50 target sequences. We use the one-tailed t-test with paired samples for evaluation.

Table 4 shows the resulting accuracies, relative improvements, and p-values. Improvements are significant in all cases in which the entire target speaker data is used. For 50 target sequences, significance can be ascertained only in the case of a single source speaker, but we always get some improvement.

We finally note that when applying such a system in practice, untranscribed data is accrued *continuously*: so the quality of the system on the target speaker could be improved continuously as well, without requiring any extra data collection.

Table 4: *Word accuracies, relative improvements, and p-values for systems with adversarial training, on the evaluation speakers. Significant results are marked with ∗.*

Adversarial Training on	Number of training spk	Target Test acc.	Relative Improvement	p-value
None	1	18.7%	-	-
	4	39.4%	-	-
	8	46.5%	-	-
All Target Sequences	1	25.4%	35.8%	0.0030∗
	4	43.6%	10.7%	0.0261∗
	8	49.3%	6.0%	0.0266∗
50 Target Sequences	1	24.1%	28.9%	0.0045∗
	4	41.5%	5.3%	0.1367
	8	47.0%	1.1%	0.3555

6. Analysis

In this section we attempt to shed light on the effect of domain-adversarial training. Figure 4 shows the progress of training for speakers s5 → s6 versus the training epoch, with adversarial training activated. The *source speaker* accuracies on validation and test set are ∼78%, almost unaffected by adversarial training. The *target speaker* accuracies are 39.1% on the validation set and 39.5% on the test set, our greatest single increase with adversarial training: without adversarial training, the target accuracy is less than 22%.

From the steady rise of the first curve, we see that the training progresses smoothly. This is the expected behavior for a well-tuned system. On the *validation* sets, the accuracy varies much less smoothly, with jumps of several percent points between epochs. We observed that this behavior is quite consistent for all systems, with or without adversarial training, and also for varying numbers of training speakers. Clearly the "error landscape" between training and validation data is very different, both within the same speaker and between different speakers.

The effect of adversarial training is clearly observable: At epoch 10, where adversarial training becomes active (with 0.2 weight), the target accuracy jumps visibly, even though *the criterion for which the adversarial network is optimized is very different from the word accuracy which is plotted in the graph.* This is a remarkable success, even though it should be noted (compare figure 3) that on other speaker pairs, we obtain a much lower improvement by adversarial training.

7. Conclusion

In this study we have described how to apply *domain-adversarial training* to a state-of-the-art lipreading system for improved speaker independency. When training and test are performed on pairs of different speakers, the average improvement is around 40%, which is highly significant; this improvement even persists when the amount of untranscribed target data is drastically reduced to about 15-20 seconds. When supervised training data from several speakers is available, there is still some improvement, from a much higher baseline

8. Acknowledgements

9. References

[1] L. Woodhouse, L. Hickson, and B. Dodd, "Review of Visual Speech Perception by Hearing and Hearing-impaired People: Clinical Implications," *International Journal of Language and Communication Disorders*, vol. 44, no. 3, pp. 253 – 270, 2009.

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

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

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

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

[6] S. Petridis and M. Pantic, "Deep Complementary Bottleneck Features for Visual Speech Recognition," in *Proc. ICASSP*, 2016, pp. 2304 – 2308.

[7] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," in *Proc. ICASSP*, 2016, pp. 6115 – 6119.

[8] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv:1611.01599, 2016.

[9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," arXiv:1611.05358, 2016.

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

[11] 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.

[12] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. ICML*, 2015, pp. 1180 – 1189.

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

[14] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning Dynamic Stream Weights For Coupled-HMM-Based Audio-Visual Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 863 – 876, 2015.

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

[16] 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.

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

[18] J. Wang and S. Hahn, "Speaker-Independent Silent Speech Recognition with Across-Speaker Articulatory Normalization and Speaker Adaptive Training," in *Proc. Interspeech*, 2015, pp. 2415 – 2419.

[19] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A Silent Speech System based on Permanent Magnet Articulography and Direct Synthesis," *Computer Speech and Language*, vol. 39, pp. 67 – 87, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230815300255

[20] 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.

[21] M. Wand and T. Schultz, "Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation," in *Proc. Interspeech*, 2014, pp. 1189 – 1193.

[22] Y. Deng, J. T. Heaton, and G. S. Meltzner, "Towards a Practical Silent Speech Recognition System," in *Proc. Interspeech*, 2014, pp. 1164 – 1168.

[23] M. Wand and J. Schmidhuber, "Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition," in *Proc. Interspeech*, 2016, pp. 3032 – 3036.

[24] I. Matthews, T.Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," *IEEE Trans. on Pattern Analysis and Machine Vision*, vol. 24, no. 2, pp. 198 – 213, 2002.

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

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

[27] 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.

[28] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The Challenge of Multispeaker Lip-reading," in *Proc. AVSP*, 2008, pp. 179 – 184.

[29] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving Visual Features for Lip-reading," in *Proc. AVSP*, 2010.

[30] 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.

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

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

[33] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A Review of Recent Advances in Visual Speech Decoding," *Image and Vision Computing*, vol. 32, pp. 590 – 605, 2014.

[34] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. E. Hennecke, "Lipreading by Neural Networks: Visual Preprocessing, Learning and Sensory Integration," in *Proc. NIPS*, 1993, pp. 1027 – 1034.

[35] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge And Data Engineering*, vol. 22, no. 10, pp. 1345 – 1359, 2010.

[36] M. A. et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, software available from tensorflow.org.

[37] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving Neural Networks by Preventing Co-adaptation of Feature Detectors," Arxiv: 1207.0580v1, 2012.

[38] R. Caruana, "Multitask Learning," Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, 1997.

[39] J. Schmidhuber, "Learning Factorial Codes by Predictability Minimization," *Neural Computation*, vol. 4, no. 6, pp. 863 – 879, 1992.

[40] S. Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa, "Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR," in *Proc. Interspeech*, 2016, pp. 2135 – 2139.

[41] M. Wand and T. Schultz, "Session-independent EMG-based Speech Recognition," in *Proc. Biosignals*, 2011, pp. 295 – 300.