SIGNER-INDEPENDENT FINGERSPELLING RECOGNITION WITH DEEP NEURAL NETWORK ADAPTATION

*Taehwan Kim, Weiran Wang, Hao Tang, Karen Livescu**

Toyota Technological Institute at Chicago, USA

{taehwan,weiranwang,haotang,klivescu}@ttic.edu

ABSTRACT

We study the problem of recognition of fingerspelled letter sequences in American Sign Language in a signer-independent setting. Fingerspelled sequences are both challenging and important to recognize, as they are used for many content words such as proper nouns and technical terms. Previous work has shown that it is possible to achieve almost 90% accuracies on fingerspelling recognition in a signer-dependent setting. However, the more realistic signer-independent setting presents challenges due to significant variations among signers, coupled with the dearth of available training data. We investigate this problem with approaches inspired by automatic speech recognition. We start with the best-performing approaches from prior work, based on tandem models and segmental conditional random fields (SCRFs), with features based on deep neural network (DNN) classifiers of letters and phonological features. Using DNN adaptation, we find that it is possible to bridge a large part of the gap between signer-dependent and signer-independent performance. Using only about 115 transcribed words for adaptation from the target signer, we obtain letter accuracies of up to 82.7% with frame-level adaptation labels and 69.7% with only word labels.

Index Terms— American Sign Language, fingerspelling, deep neural network, adaptation, segmental CRF

1. INTRODUCTION

Automatic sign language recognition is a nascent technology that has the potential to improve the ability of Deaf and hearing individuals to communicate, as well as Deaf individuals' ability to take full advantage of modern information technology. For example, online sign language video blogs and news[1] are currently almost completely unindexed and unsearchable as they include little accompanying annotation.

Research on this problem has included both speech-inspired approaches and computer vision-based techniques, using either/both video and depth sensor input [1, 2, 3, 4, 5, 6, 7, 8, 9]. We focus on recognition from video, for applicability to existing recordings. Before the technology can be applied "in the wild", it must overcome challenges posed by visual nuisance parameters (e.g., lighting, occlusions) and signer variation. Annotated data sets for this problem are scarce, in part due to the need to recruit signers and skilled annotators.

We consider American Sign Language (ASL), in particular the fingerspelling component: the spelling out of a word as a sequence of handshapes or hand trajectories corresponding to individual letters. Fig. 1 gives example fingerspelling sequences. Fingerspelling accounts for roughly 12-35% of ASL [10] and is typically used for proper nouns or borrowings from English, which can often be the most important content words. Some aspects of fingerspelling can be characterized through the phonology of handshape [11, 12], which can be described in terms of phonological features. Most prior research on fingerspelling recognition has focused on constrained tasks such as single-letter or handshape classification or word recognition from a known vocabulary [13, 14, 15, 16, 17, 18, 19]. For the unconstrained letter sequence recognition problem, Kim et al. [7, 8] obtained $\sim 90\%$ average letter accuracies in a signer-dependent setting, using either tandem hidden Markov models (HMMs) or segmental conditional random fields (SCRFs), with features from neural network classifiers of letters and phonological features. That work used the largest video data set of which we are aware containing unconstrained, connected fingerspelling, consisting of four signers each signing 600 word tokens for a total of ~ 350k image frames.

In this paper we consider the problem of *signer-independence in unconstrained fingerspelling sequence recognition*, in the context of limited training data. Prior work has address signer adaptation for large-vocabulary German Sign Language recognition [9], but to our knowledge this paper is the first to address adaptation for fingerspelling. We investigate approaches to signer-independence including speed normalization and neural network adaptation. The adaptation techniques are largely borrowed from speech recognition research, but the application is quite different in that the overall amount of data is much smaller and the types of variation are different. We find that the simple signer normalization is ineffective, while DNN adaptation is very effective.

*This research was supported by NSF grant NSF-1433485. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

[1]E.g., http://ideafnews.com, http://aslized.org.

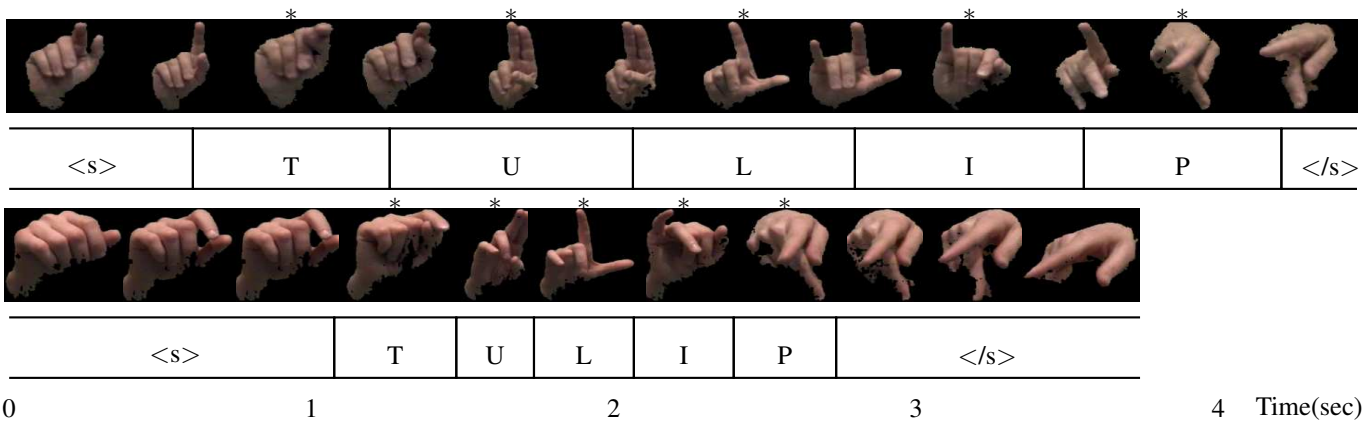

Fig. 1. Images and ground-truth segmentations of the fingerspelled word 'TULIP' produced by two signers. Image frames are sub-sampled at the same rate from both signers to show the true relative speeds. Asterisks indicate manually annotated peak frames for each letter. "<s>" and "</s>" denote non-signing intervals before/after signing.

2. METHODS

The task is to convert a video (a sequence of images), as in Fig. 1, to a sequence of letters. The segmentation into letters is unknown, so this is a sequence prediction task analogous to connected phone or word recognition. We start with the recognition approaches that have achieved the best prior results on this task [7, 8], with updates for improved performance with deeper neural networks. We next briefly describe the recognizers, the neural network classifiers and adaptation.

2.1. Recognizers

The first recognizer is a tandem model [20] based on [7]. Frame-level features are fed to neural network classifiers, one of which predicts the frame's letter label and six others which predict handshape phonological features.[2] Classifier outputs are concatenated with the image features, after a dimensionality reduction, and input to a hidden Markov model (HMM) recognizer with Gaussian mixture observation densities.

The second recognizer is a segmental CRF (SCRF) model based on [8]. SCRFs [21, 22] are conditional log-linear models with feature functions that can be based on variable-length segments of input frames, allowing for great flexibility in defining feature functions. As in [8], we use an SCRF to rescore lattices produced by a baseline frame-based recognizer (in this case, the tandem model). We use the same feature functions as in [8], which include language model features, a feature that measures agreement with the baseline recognizer, means of letter/phonological feature neural network classifier outputs over each segment, and "peak detection" features that measure the dynamics of each segment.

Finally, we also use a first-pass decoding SCRF from Tang *et al.* [23], which is independent of any frame-based recognizer. We use the same feature functions as in [23], namely average DNN outputs over each segment, samples of DNN outputs within the segment, duration and bias, all lexicalized.

[2]See [11, 12, 7] for details of the phonological features.

2.2. DNN adaptation

The DNNs are first trained in a signer-independent way on all but the test signer, using an L2-regularized cross-entropy loss. The inputs are the image features concatenated over a multi-frame window, which are fed through several fully connected layers followed by a softmax output layer. Inspection of data such as Fig. 1 reveals the main sources of signer variation: speed, hand appearance, and non-signing motion variation before/after signing. The speed variation is large, with a factor of 1.8 between the fastest and slowest signers. In the absence of adaptation data, we consider a simple speed normalization: We augment the training data with resampled image features, at 0.8x and 1.2x the original frame rate.

If we have access to some labeled data from the test signer, but not a sufficient amount for training full signer-specific DNNs, we can apply adaptation. A number of DNN adaptation approaches have been developed (e.g., [24, 25, 26, 27]). We first consider two simple approaches based on linear input networks (LIN) and linear output networks (LON) [28, 29, 30], shown in Fig. 2. Most of the network parameters are fixed; only a limited set of weights at the input and output layers are learned. In the first approach (LIN+UP in Fig. 2), we apply a single affine transformation W_{LIN} to the static features at each frame (before concatenation) and feed the result to the trained signer-independent DNNs. We jointly learn W_{LIN} and adapt the last (softmax) layer weights by minimizing the same cross-entropy loss on the adaptation data, and "warm-start" the softmax layer with the learned signer-independent weights. The second approach (LIN+LON in Fig. 2) uses the same input adaptation layer, but rather than adapting the softmax weights, it removes the softmax output activation and adds a new softmax output layer W_{LON} for the test signer, trained jointly with the same cross-entropy loss. Finally, we also consider adaptation by fine-tuning; that is, updating all of the DNN weights on adaptation data starting from the signer-independent weights. The adaptation can

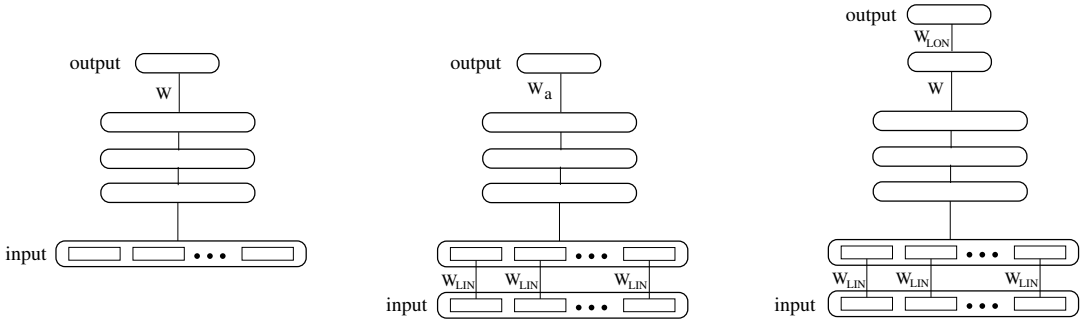

Fig. 2. Left: Unadapted DNN classifier; middle: adaptation via linear input network and output layer updating (LIN+UP); right: adaptation via linear input network and linear output network (LIN+LON).

use either ground-truth frame-level letter labels (using human annotation as described in [7]) or labels obtained by forced alignment if only word labels are available.

3. EXPERIMENTS

We use the ASL video data set of [8], comprising four signers each fingerspelling 600 word tokens consisting of two repetitions of a 300-word list, including common English words, names, and foreign words. Annotators marked the peak of articulation of each letter, and the annotations were converted to a "ground-truth" frame labeling by assuming that the letter boundaries occur mid-way between peaks. Following [8], the hand portion of each image is extracted via hand detection and segmentation using a signer-specific Gaussian color model, followed by suppression of irrelevant pixels. The extracted hand images are resized to 128×128 and Histogram of Gradient (HoG) [31] features are extracted using multiple spatial grids (4×4, 8×8, and 16×16), followed by dimensionality reduction with principal components analysis (PCA).

3.1. Frame classification

The initial unadapted signer-independent DNNs are trained on all but the test signer for each of the seven tasks (letters and the six phonological features). The input is the 128-dimensional HoG features concatenated over a 21-frame window, and the networks have three hidden layers of 3000 ReLUs [32]. Cross-entropy training is done with a weight decay penalty of 10^{-5} via stochastic gradient descent (SGD) over 100-sample minibatches for up to 30 epochs, with dropout [33] at a rate of 0.5 at each hidden layer, fixed momentum of 0.95, and initial learning rate of 0.01, which is halved when held-out accuracy stops improving. These hyperparameters were tuned on held-out (signer-independent) data in initial experiments, not reported here in the interest of space. We pick the best-performing epoch on held-out data.

We next consider DNN normalization and adaptation with different types and amounts of supervision. For LIN+UP and LIN+LON, we adapt by running SGD over minibatches of 100 samples with a fixed momentum of 0.9 for up to 20 epochs, with initial learning rate of 0.02 (which is halved when accuracy stops improving on the adaptation data). For

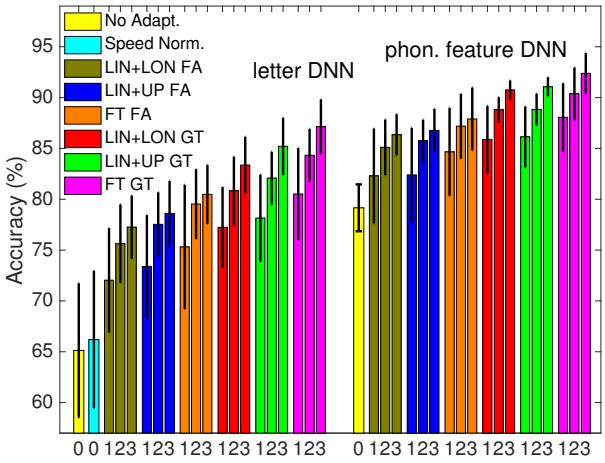

Fig. 3. DNN frame accuracies with and without normalization/adaptation. The horizontal axis labels indicate the amount of adaptation data (0, 1, 2, 3 = none, 5%, 10%, 20% of the test signer's data, corresponding to no adaptation, $\sim$ 29, $\sim$ 58, and $\sim$ 115 words). GT = ground truth labels; FA = forced alignment labels; FT = fine-tuning.

fine-tuning, we use the same SGD procedure as for the signer-independent DNNs. We pick the epoch with the highest accuracy on the adaptation data. The resulting frame accuracies are given in Fig. 3. In addition, Fig. 3 includes the result of speed normalization for the case of letter classification. Speed normalization provides consistent but very small improvements, while adaptation gives large improvements in all settings. LIN+UP slightly outperforms LIN+LON, and fine-tuning outperforms both LIN+UP and LIN+LON. For letter sequence recognition in the next section, we adapt via fine-tuning using 20% of the test signer's data.

Fig. 4 further analyzes the DNNs via confusion matrices. One of the main effects is the large number of incorrect predictions of the non-signing classes (<s>, </s>). We observe the same effect with the phonological feature classifiers. This may be due to the previously mentioned fact that non-linguistic gestures are variable and easy to confuse with signing when given a new signer's image frames. The confusion matrices show that, as the DNNs are adapted, this is the main type of error that is corrected.

	Tandem HMM					Rescoring SCRF					1st-pass SCRF				
Signer	1	2	3	4	**Mean**	1	2	3	4	**Mean**	1	2	3	4	**Mean**
No adapt.	45.9	45.3	37.4	42.5	**42.8**	47.4	48.8	38.9	43.7	**44.7**	44.7	46.7	27.5	38.6	**39.4**
Forced align.	69.8	71.5	60.4	63.9	**66.4**	70.5	74.0	61.8	65.5	**68.0**	75.6	75.1	63.5	64.5	**69.7**
Ground truth	78.0	87.0	68.4	78.6	**78.0**	77.6	86.5	70.5	78.6	**78.3**	84.8	89.4	75.1	81.6	**82.7**

Table 1. Letter accuracies (%) on four test signers.

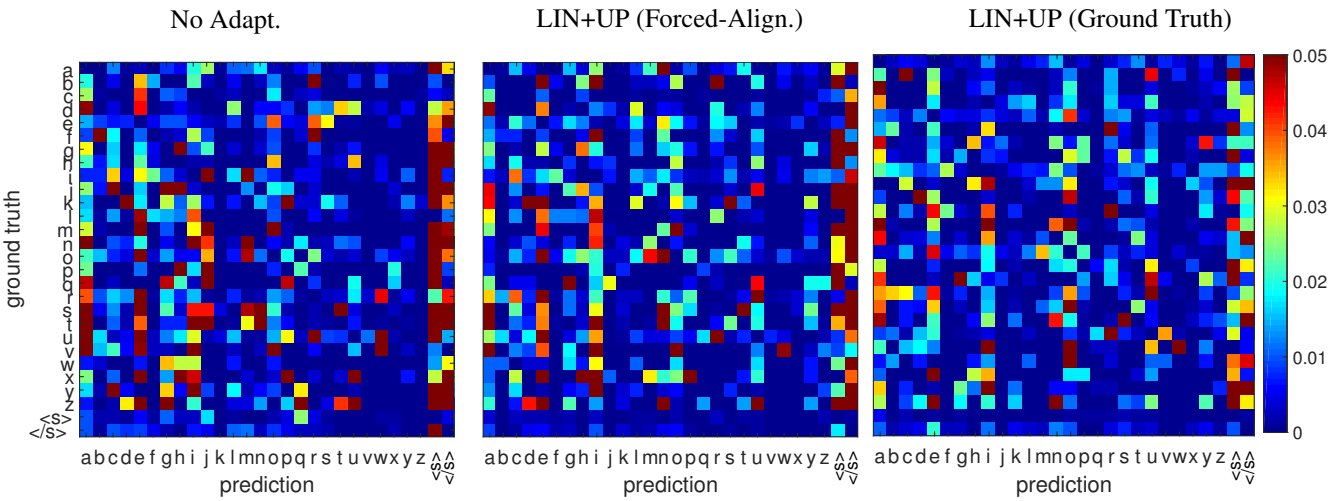

Fig. 4. Confusion matrices of DNN classifiers for one test signer (Signer 1). 20% of the test signer's data (115 words) was used for adaptation, and a disjoint 70% was used to compute confusion matrices. Each matrix cell is the empirical probability of the predicted class (column) given the ground-truth class (row). The diagonal has been zeroed out for clarity.

3.2. Connected letter recognition

In connected letter recognition, we measure performance via the letter accuracy, analogously to the word or phone accuracy in speech recognition. Table 1 shows the letter accuracies obtained with the tandem, rescoring SCRF, and first-pass SCRF models with DNN adaptation via fine-tuning, using different types of adaptation data. For all models, we do not retrain the models with the adapted DNNs, but tune several hyperparameters[3] on 10% of the test signer's data. The tuned models are evaluated on an unseen 10% of the test signer's remaining data; finally, we repeat this for eight choices of tuning and test sets, covering the 80% of the test signer's data that we do not use for adaptation, and report the mean letter accuracy over the test sets.

As shown in Table 1, without adaptation both tandem and SCRF models do poorly, achieving only roughly 40% letter accuracies, with the rescoring SCRF slightly outperforming the others (recall that signer-dependent recognition achieves about 90% letter accuracies [8]). With adaptation, however, performance jumps to up to 69.7% letter accuracy with forced-alignment adaptation labels and up to 82.7% accuracy with ground-truth adaptation labels. All of the adapted models perform similarly, but interestingly, the first-pass SCRF is slightly worse than the others before adaptation and better (by 4.4% absolute) after ground-truth adaptation. One hypothesis is that the first-pass SCRF is more dependent on the DNN performance, while the tandem model uses the original image features and the rescoring SCRF uses the tandem model hypotheses and scores. Once the DNNs are adapted, however, the first-pass SCRF outperforms the other models.

[3]See [7, 8, 23] for details of the tuning parameters.

4. CONCLUSION

In this study of signer-independent and adapted ASL fingerspelling recognition, we have seen that fingerspelling has great variability in speed, hand appearance, and appearance of non-signing gestures. We have improved performance on new signers via adaptation of DNNs in tandem and SCRF recognizers. Several DNN adaptation approaches are successful, with the largest improvements coming from simple fine-tuning on adaptation data. This approach improves letter accuracies from around 40% (unadapted) to up to 69.7% with weak word-level supervision and up to 82.7% with ground-truth frame labels for the adaptation data. While the models perform similarly, the best adapted model is a first-pass SCRF. The main DNN improvements come from resolving confusions between actual letters and the non-signing ("silence") class. Future work will continue to improve the models and adaptation approaches, as well as address other types of variability that are needed to port the models to video data "in the wild".

5. REFERENCES

[1] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," in *Proc. Interspeech*, 2007.

[2] M. M. Zaki and S. I. Shaheen, "Sign language recognition using a combination of new vision based features," *Pattern Recognition Letters*, pp. 3397–3415, 2010.

[3] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," in *Proc. ECCV*, 2004.

[4] S. Liwicki and M. Everingham, "Automatic recognition of fingerspelled words in British Sign Language," in *Proc. 2nd IEEE Workshop on CVPR for Human Communicative Behavior Analysis*, 2009.

[5] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Model-level data-driven sub-units for signs in videos of continuous sign language," in *Proc. ICASSP*, 2010.

[6] C. Vogler and D. Metaxas, "Toward scalability in ASL recognition: Breaking down signs into phonemes," in *Proc. Gesture Workshop*, 1999.

[7] T. Kim, K. Livescu, and G. Shakhnarovich, "American Sign Language fingerspelling recognition with phonological feature-based tandem models," in *Proc. SLT*, 2012.

[8] T. Kim, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition with semi-Markov conditional random fields," in *Proc. ICCV*, 2013.

[9] J. Forster, O. Koller, C. Oberdörfer, Y. Gweth, and H. Ney, "Improving continuous sign language recognition: Speech recognition techniques and system design," in *Proc. SLPAT*, 2013.

[10] C. Padden and D. C. Gunsauls, "How the alphabet came to be used in a sign language," *Sign Language Studies*, 2004.

[11] D. Brentari, *A Prosodic Model of Sign Language Phonology*, MIT Press, 1998.

[12] R. E. Johnson and S. K. Liddell, "Toward a phonetic representation of signs: sequentiality and contrast," *Sign Language Studies*, vol. 11, no. 2, pp. 241–274, 2010.

[13] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios, "BoostMap: A method for efficient approximate similarity rankings," in *Proc. CVPR*, 2004.

[14] S. Ricco and C. Tomasi, "Fingerspelling recognition through classification of letter-to-letter transitions," in *Proc. ACCV*, 2009.

[15] G. Tsechpenakis, D. Metaxas, and C. Neidle, "Learning-based dynamic coupling of discrete and continuous trackers," *Computer Vision and Image Understanding*, vol. 104, no. 2–3, 2006.

[16] R. Bowden and M. Sarhadi, "A non-linear model of shape and motion for tracking finger spelt American sign language," *Image and Vision Computing*, vol. 20, no. 9–10, 2002.

[17] P. Goh and E. Holden, "Dynamic fingerspelling recognition using geometric and motion features," in *Proc. ICIP*, 2006.

[18] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos," *Journal of Machine Learning Research*, vol. 14, 2013.

[19] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *Proc. ICCV*, 2011.

[20] D. P. W. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. ICASSP*, 2001.

[21] S. Sarawagi and W. W. Cohen, "Semi-Markov conditional random fields for information extraction," in *NIPS*, 2004.

[22] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU*, 2009.

[23] H. Tang, W. Wang, K. Gimpel, and K. Livescu, "Discriminative segmental cascades for feature-rich phone recognition," in *Proc. ASRU*, 2015.

[24] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. ICASSP*, 2013.

[25] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proc. ICASSP*, 2013.

[26] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. SLT*, 2014.

[27] R. Doddipatla, M. Hasan, and T. Hain, "Speaker dependent bottleneck layer training for speaker adaptation in automatic speech recognition," in *Proc. Interspeech*, 2014.

[28] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. Eurospeech*, 1995.

[29] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. SLT*, 2012.

[30] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proc. Interspeech*, 2010.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005.

[32] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton, "On rectified linear units for speech processing," in *Proc. ICASSP*, 2013.

[33] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.