INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Stimulated Deep Neural Network for Speech Recognition

Chunyang Wu[1], *Penny Karanasou*[1], *Mark J.F. Gales*[1], *Khe Chai Sim*[2]

[1]University of Cambridge
[2]National University of Singapore
{cw564, pk407, mjfg}@eng.cam.ac.uk, simkc@comp.nus.edu.sg

Abstract

Deep neural networks (DNNs) and deep learning approaches yield state-of-the-art performance in a range of tasks, including speech recognition. However, the parameters of the network are hard to analyze, making network regularization and robust adaptation challenging. Stimulated training has recently been proposed to address this problem by encouraging the node activation outputs in regions of the network to be related. This kind of information aids visualization of the network, but also has the potential to improve regularization and adaptation. This paper investigates stimulated training of DNNs for both of these options. These schemes take advantage of the smoothness constraints that stimulated training offers. The approaches are evaluated on two large vocabulary speech recognition tasks: a U.S. English broadcast news (BN) task and a Javanese conversational telephone speech task from the IARPA Babel program. Stimulated DNN training acquires consistent performance gains on both tasks over unstimulated baselines. On the BN task, the proposed smoothing approach is also applied to rapid adaptation, again outperforming the standard adaptation scheme.

Index Terms: Deep Neural Networks, Stimulated Learning, Speaker Adaptation

1. Introduction

In recent years, deep neural networks [1, 2, 3] (DNNs) have successfully been applied to acoustic models of state-of-the-art speech recognition systems. DNNs are a set of hidden layers with linear transformations and non-linear activations for making predictions. This approach allows complex data to be well modeled. However, the whole DNN remains a "black-box" and the behaviors of the activations can be hard to understand. In computer vision, several works [4, 5, 6] have been proposed to analyze neural network weights. However, this topic has rarely been investigated in speech recognition.

In order to enhance interpretation, stimulated learning [7] has been proposed to augment the DNN optimization with regularization at the hidden-layer activation level. A stimulation term is introduced to the training criterion to encourage the DNN activations in a region to be similar. For instance, by using a phonemic prior, the neurons in different regions are regularized to correspond to different phonemes. In this way, DNN activations can then be interpreted and visualized directly according to the underlying stimulated prior [7].

Another advantage of stimulated DNNs is that the information from its activation patterns can be utilized for DNN regularization and adaptation schemes. In speaker adaptation, limited amount of data are used to adapt the speaker-independent (SI) acoustic model to a speaker. Several methods have been proposed to impose interpretable structures on the DNN topology to effectively adapt the DNN. Transformation-based schemes [8, 9, 10] add additional linear layers as speaker-dependent (SD) transforms. In the DNN-CAT [11, 12] and the multi-basis adaptive neural network [13] models, speaker-dependent interpolation weights are introduced to combine different DNN modules to handle the acoustic distortions. The learning hidden unit contributions [14] (LHUC) and the parametric activation [15] adaptation methods introduce a scaling factor on each hidden-layer activation. In [16], the differentiable pooling technique is used to obtain the speaker-dependent compensation from a hidden-activation candidate pool. Although they can be used immediately on a stimulated DNN, the expressive activation properties are rarely taken advantage of on these approaches.

This paper investigates stimulated training of DNNs for both network regularization and robust adaptation. There are two major contributions of this paper. First, stimulated DNN regularization is applied to large vocabulary recognition tasks and achieves consistent gains. Second, based on the activation patterns obtained by stimulated learning, a smoothing approach is proposed to regularize the DNN adaptation schemes. In the standard LHUC adaptation method, the neurons are treated as independent items without considering their inter-relations: a large number of parameters are required to adapt, which is incapable of handling rapid adaptation scenarios with limited data. In the regularized LHUC approach, the information from a node's spatial neighbors in the stimulated DNN are utilized to robustly smooth LHUC scaling factors during the adaptation phase. This kind of auxiliary indicators is expected to help regularize the LHUC behaviors even when there is insufficient adaptation data. The experiments are conducted on two large vocabulary recognition tasks: a U.S. English broadcast news (BN) transcription task and a Javanese conversational telephone speech (CTS) task from the IARPA Babel program.

The rest of this paper is organized as follows. The stimulated DNN regularization is presented in Section 2. In Section 3, we propose the regularized LHUC adaptation approach on stimulated DNNs. Experimental results are reported in Section 4. This paper is concluded in Section 5.

The research leading to these results was supported by EPSRC grant EP/I031022/1 (Natural Speech Technology), the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012 and Singapore Ministry of Education Academic Research Fund Tier 2 (Official Project No: MOE2014-T2-1-068). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

http://dx.doi.org/10.21437/Interspeech.2016-580

2. Stimulated DNN Regularization

Deep neural networks are commonly integrated to the acoustic model of speech recognition systems: given a frame observation $\mathbf{x}$, a DNN is usually used to predict the posterior of the context-dependent target y

$$P(y=i|\mathbf{x}) = \frac{\exp\left(z_i^{(L)}\right)}{\sum_j \exp\left(z_j^{(L)}\right)}. \quad (1)$$

The DNN activation input $\mathbf{z}^{(l)}$ and output $\mathbf{h}^{(l)}$ are recursively defined as

$$\begin{aligned}\mathbf{z}^{(l)} &= \mathbf{W}^{(l)^T}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}, 1 \le l \le L \\ \mathbf{h}^{(l)} &= \boldsymbol{\sigma}\left(\mathbf{z}^{(l)}\right), 1 \le l < L \\ \mathbf{h}^{(0)} &= \mathbf{x}\end{aligned}$$

where $\boldsymbol{\sigma}(\cdot)$ represents the sigmoid function; L is the total number of layers in the neural network; $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the transformation parameters belonging to the l-th hidden layer; $\mathbf{x}$ is the input feature.

Define $\boldsymbol{\theta} = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots, \mathbf{W}^{(L)}, \mathbf{b}^{(L)}\}$ as the DNN parameters, standard training schemes try to minimize some criterion $\mathcal{L}(\boldsymbol{\theta})$ over a training set. For instance, the cross-entropy (CE) criterion is

$$\mathcal{L}_{ce}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t|\mathbf{x}_t;\boldsymbol{\theta}) \quad (2)$$

where T is the total of training samples; $\mathbf{x}_t$ and y_t are a training sample and its true context-dependent target. One of the issues with the standard training is that the hidden activations are not interpretable. This lack of interpretability can cause issues for network regularization and speaker adaptation as it is difficult to relate weights from the network to each other.

To address this problem, the stimulated leaning [7] approach has been proposed. The aim of stimulated training is to train DNNs where nodes with similar activation functions are grouped together in the spatial ordering, instead of forming an arbitrarily-ordered set of activations. In detail, a phone (or grapheme) dependent prior distribution is defined over the normalized activation function outputs for each of the layers. The nodes in each layer are reorganized into a grid, *e.g.*, a layer with 1024 nodes can form a 32×32 two-dimensional grid. In this way, each node i of a layer is represented as a point in a two dimensional *network-grid* space, denoted as $\mathbf{s}_i$. A point in this network-grid space is also associated with each phone p, denoted as $\mathbf{s}_p$. The phone positions can be determined via methods like t-SNE [17] over the acoustic feature means of the phonemes. It is then possible to define a normalized distance from every node to the correct phone position. These normalized distances are used as a prior over the distribution of activation function values for a layer. This prior encourages activation functions in the same locality to have the same normalized output.

To implement stimulated training, a regularization term $\mathcal{R}_{st}(\boldsymbol{\theta})$ is added to the training criterion $\mathcal{F}(\boldsymbol{\theta})$

$$\mathcal{F}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \eta_{st}\mathcal{R}_{st}(\boldsymbol{\theta}) \quad (3)$$

where $\mathcal{L}(\boldsymbol{\theta})$ is the standard training criterion; η_{st} determines the contribution of the stimulated prior $\mathcal{R}_{st}(\boldsymbol{\theta})$. Here $\mathcal{R}_{st}(\boldsymbol{\theta})$ is based on the KL-divergence of the prior distribution over the current distribution $g(\mathbf{s}_i, \hat{\mathbf{s}}_{p_t})$ and the normalized activation $\bar{h}_i^{(l)}(\mathbf{x}_t)$

$$\mathcal{R}_{st}(\boldsymbol{\theta}) = \frac{1}{T}\sum_t \sum_l \sum_i g(\mathbf{s}_i, \hat{\mathbf{s}}_{p_t}) \log\left(\frac{g(\mathbf{s}_i, \hat{\mathbf{s}}_{p_t})}{\bar{h}_i^{(l)}(\mathbf{x}_t)}\right) \quad (4)$$

where the two distributions are defined as:

1. *phone-specific activation distribution prior:* the normalized distance of a node and the current active-phone position:

$$g(\mathbf{s}_i, \hat{\mathbf{s}}_{p_t}) = \frac{\exp\left(-\frac{1}{2\sigma_{st}^2}||\mathbf{s}_i - \hat{\mathbf{s}}_{p_t}||_2^2\right)}{\sum_j \exp\left(-\frac{1}{2\sigma_{st}^2}||\mathbf{s}_j - \hat{\mathbf{s}}_{p_t}||_2^2\right)} \quad (5)$$

   where $\mathbf{s}_i$ is the position of the i-th node in the network-grid space; $\hat{\mathbf{s}}_{p_t}$ is the position in the network-grid space of the "correct" phoneme at time t; σ_{st} controls the sharpness of the prior surface.

2. *network activation distribution:* $\bar{h}_i^{(l)}(\mathbf{x}_t)$ is the normalized activation output for the i-th node on the l-th layer

$$\bar{h}_i^{(l)} = \frac{h_i^{(l)}\beta_i^{(l)}}{\sum_j h_j^{(l)}\beta_j^{(l)}}, \quad \beta_i^{(l)} = \sqrt{\sum_k {w_{ik}^{(l+1)}}^2} \quad (6)$$

   where $\beta_i^{(l)}$ is used to reflect the impact that the activation function has on the following layer $l+1$ and has been found to be important for stimulated training.

This form of prior can be applied to any form of network. In this work it is applied to a DNN trained using either the cross-entropy, or Minimum-Phone-Error (MPE) sequential training.

3. Stimulated DNN Adaptation

Recent progress in neural network adaptation introduce additional structures to the network as model parameters. In LHUC [14] and the parametric activation [15] adaptation approaches, a speaker-dependent scaling factor $\alpha_i^{(ls)}$ is introduced independently to every activation of all the hidden layers. Following [15],

$$\tilde{h}_i^{(ls)}(\mathbf{x}_t) = \alpha_i^{(ls)} h_i^{(l)}(\mathbf{x}_t) \quad (7)$$

where $\tilde{h}_i^{(ls)}$ denotes the adapted output of the i-th node on the l-th layer and s stands for the speaker index. Scaling factors are introduced per activation thus a large number of parameters are required to adapt. The lack of interpretable meanings among the activations causes that they are modeled as independent components, instead of groups based on functional similarities.

As illustrated in Figure 1(b), the network grid behaves as a smooth surface on each layer of a stimulated DNN. The nearby nodes in the spatial ordering are likely to perform analogously. Based on this phenomenon, a regularized LHUC model associated with the stimulated DNN is proposed, which aims at smoothing the adapted activation outputs by spatial neighbors. Define $\boldsymbol{\alpha}^{(s)}$ as a super-vector concatenating the LHUC scaling factors of activations in all hidden layers. The adaptation regularization term is defined as

$$\mathcal{R}_L(\boldsymbol{\alpha}^{(s)}) = \frac{1}{2T^{(s)}}\sum_l \sum_i \sum_j \left(q_{ij} \sum_{t\in\mathbb{I}^{(s)}} f_{ij}\left(\mathbf{x}_t; \boldsymbol{\alpha}^{(s)}\right)\right) \quad (8)$$

where $\mathbb{I}^{(s)}$ and $T^{(s)}$ are respectively the index set and the total of frames belonging to the s-th speaker; $f_{ij}(\cdot)$ is the squared difference between two activation outputs

$$f_{ij}\left(\mathbf{x}_t;\boldsymbol{\alpha}^{(s)}\right)=\left(\beta_i^{(l)}\tilde{h}_i^{(ls)}(\mathbf{x}_t)-\beta_j^{(l)}\tilde{h}_j^{(ls)}(\mathbf{x}_t)\right)^2,$$

the term $\beta_i^{(l)}$, as defined in Eq. 6, is also introduced to keep the consistency in stimulated learning. q_{ij} is the impact of a neighbor to the current node, measured by the normalized Euclidean distance between their positions on the stimulated grid

$$q_{ij}=\frac{1}{Q_i}\exp\left(-\frac{1}{2\sigma_L^2}||\mathbf{s}_i-\mathbf{s}_j||_2^2\right),$$

σ_L is the decay factor and $Q_i=\sum_j\exp\left(-\frac{1}{2\sigma_L^2}||\mathbf{s}_i-\mathbf{s}_j||_2^2\right)$.

On both the frame-level and the sequential DNN systems in this paper, LHUC optimization using the frame-level cross-entropy criterion with the proposed regularization is investigated. The adaptation criterion $\mathcal{F}(\boldsymbol{\alpha}^{(s)})$ is given by

$$\mathcal{F}(\boldsymbol{\alpha}^{(s)})=\mathcal{L}_{ce}(\boldsymbol{\alpha}^{(s)})+\eta_L R_L(\boldsymbol{\alpha}^{(s)}) \tag{9}$$

where η_L penalizes the importance of the regularization term. The optimization of the regularized LHUC scaling factors can be performed in a stochastic gradient descend fashion. The essential gradients $\frac{\partial\mathcal{L}}{\partial\alpha_i^{(ls)}}$ and $\frac{\partial\mathcal{L}}{\partial h_i^{(l)}}$ are calculated by

$$\frac{\partial\mathcal{F}}{\partial\alpha_i^{(ls)}}=h_i^{(l)}\frac{\partial\mathcal{F}}{\partial\tilde{h}_i^{(ls)}}+\eta_L\beta_i^{(l)}h_i^{(l)}\sum_j f_{ij}^*$$

$$\frac{\partial\mathcal{F}}{\partial h_i^{(l)}}=\alpha_i^{(ls)}\frac{\partial\mathcal{F}}{\partial\tilde{h}_i^{(ls)}}+\eta_L\beta_i^{(l)}\alpha_i^{(ls)}\sum_j f_{ij}^*$$

where $\frac{\partial\mathcal{F}}{\partial\tilde{h}_i^{(ls)}}$ is recursively calculated by back-propagation and

$$f_{ij}^*=(q_{ij}+q_{ji})\left(\beta_i^{(l)}\tilde{h}_i^{(ls)}-\beta_j^{(l)}\tilde{h}_j^{(ls)}\right).$$

4. Experiments

Experiments were conducted on a U.S. English broadcast news (BN) task and a Javanese conversational telephone speech task from the IARPA Babel program. The relevant GMMs, DNNs and the proposed models were trained on an extended version of HTK Toolkit 3.5 [18].

4.1. Broadcast News

The training set for this task included the 144-hour 1996 & 1997 Hub-4 English Broadcast News Speech dataset (`LDC97S44`, `LDC98S71`), containing 288 shows with approximately 8k speakers. For performing evaluation, both the BN testsets 2.7-hour `Dev03` and 2.6-hour `Eval03` were used. The utterances of both testsets were processed by automatic segmentation and their averaged utterance durations were 10.7 and 10.9 seconds, respectively. Decoding was performed with the RT04 tri-gram language model [19]. The adaptation schemes were evaluated in a rapid utterance-level unsupervised fashion: the SI decoding hypotheses are used to estimate the LHUC parameters for each utterance.

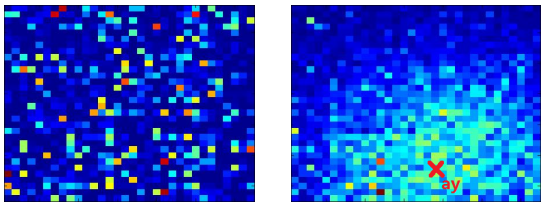
The baseline DNN cross-entropy system used the 468-dimensional PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$, processed by both global cepstral mean normalization (CMN) and cepstral variance normalization (CVN), in a temporal context window of 9 frames as the input feature. The neural network consisted of 5 hidden layers with 1024 nodes in each layer and the context-dependent targets were approximately 6k. Its parameters were initialized by the layer-wise discriminative pre-training and then optimized by back-propagation. 28 shows with about 600 speakers were randomly selected as the cross validation set. The well-trained CE DNN was subsequently used to generate the lattices of the training set, initialize the sequential MPE DNN and further tune for four iterations under the MPE criterion to obtain the baseline MPE DNN system. On both the CE & MPE DNN systems, we used their respective hypothesis alignment to adapt the DNN via LHUC for each utterance in the testsets.

In order to setup the stimulated DNNs, the mono-phone 2D positions were firstly obtained via t-SNE [17] over the training-set averaged CMLLR [20] frames of the phonemes. They were then scaled to fit in the unit square $[0,1]\times[0,1]$. The network configuration of stimulated DNNs kept the same as that of the baselines. The prior sharpness factor σ_{st}^2 was empirically set as 0.1. The sequential MPE stimulated system were initialized as the CE stimulated one and then tuned for three iterations. The standard and regularized LHUC models were then evaluated on both the stimulated CE & MPE DNNs.

Table 1 reports the impact of the stimulating penalty η_{st} on CE systems. All the stimulated DNNs (*CE-Stimu*) penalized from 0.05 to 0.2 outperformed the unstimulated baseline (*CE*). The best system was achieved by that with 0.05, decreasing the word error rate (WER) by 0.8% in absolute value. In addition to the performance, a meaningful pattern was also obtained. As shown in Figure 1, we compared the first-hidden-layer activation grids of the unstimulated and stimulated ($\eta_{st}=0.05$) DNNs on a frame example belonging to the phone "ay". Because of the arbitrary ordering issue, there is no pattern in the grid of the unstimulated DNN. However on the stimulated one, the activation grid nicely corresponded to the stimulating pattern: the nodes around the "ay" location echoed higher activation values.

System	η_{st}	%WER
CE	0	12.7
CE-Stimu	0.05	**11.9**
	0.1	12.1
	0.15	12.5
	0.2	12.6

Table 1: Comparison of the Impact of η_{st} on Stimulated CE Systems on Dev03.



(a) Unstimulated (b) Stimulated $\eta_{st}=0.05$

Figure 1: Comparison of unstimulated and stimulated DNN activations on an "ay" frame.

The best stimulated system ($\eta_{st}=0.05$) was subsequently used to investigate the proposed regularized LHUC adaptation approach (*+regLHUC*). The distance decay σ_L^2 was fixed as

System	η_L	%WER
LHUC	–	12.4
regLHUC	0	11.6
	0.05	11.5
	0.1	**11.4**
	0.15	11.5
	0.2	11.5

Table 2: Comparison of the Impact of η_L in Utterance-Level Adaptation on Dev03.

0.01, according to empirical results. The impact of the LHUC regularization penalty η_L in utterance-level unsupervised adaptation is compared in Table 2. The best adaptation performance was obtained when η_L was set as 0.1, which outperformed the original LHUC method by 0.2% absolutely on WER. Then, η_L was fixed as 0.1 and a summary of rapid adaptation on the CE systems is given in Table 3, including the performance on the unseen testset `Eval03`. The regularized LHUC method on top

System	Dev03	Eval03
CE	12.7	10.7
+LHUC	12.4	10.6
+regLHUC	12.8	10.8
CE-Stimu	11.9	10.3
+LHUC	11.6	10.0
+regLHUC	**11.4**	**9.9**

Table 3: CE Utterance-Level Adaptation on Broadcast News.

of the stimulated DNN acquired improvement compared with the default LHUC (+*LHUC*) on both testsets. The regularized LHUC was also tested on the default CE baseline. However, since the arbitrary neighbors were unable to provide useful information, no enhancement was achieved on top of the unstimulated system.

This CE stimulated DNN was then used to train the MPE stimulated system. Table 4 reports the comparison of the performance of different MPE systems. Similar to the CE ones,

System	Dev03	Eval03
MPE	11.6	10.1
+LHUC	11.2	9.8
MPE-Stimu	11.2	9.8
+LHUC	10.9	9.5
+regLHUC	**10.6**	**9.4**

Table 4: MPE Utterance-Level Adaptation on Broadcast News.

the MPE stimulated DNN outperformed the unstimulated MPE baseline. The regularized LHUC on the stimulated system achieved the best performance as well, reducing the WER up to 5% relatively in contrast with the SI MPE stimulated system.

4.2. Javanese

The next experiments were conducted on the Javanese conversational telephone speech task from the IARPA Babel program (`IARPA-babel402b-v1.0b`). The Full Language Pack (FLP) of the Javanese (402) language was used. The training set consisted of approximately 60 hours from 720 speakers. The evaluation set was the 10.2-hour `Dev` set from 120 speakers and its segmentation was automatically processed with an average duration 2.1 seconds. Decoding was performed with a decoding tri-gram language model followed by confusion network.

In the DNN training, the 77-dimensional multilingual bottleneck feature [21] on 11 Babel languages was used. It was processed by both side-level CMN & CVN and the DNN used it in a temporal context window of 9 frames as the input feature. The context-dependent targets were approximately 4k and the DNN configuration was $693 \times 1024^5 \times 4k$. DNN parameters were initialized in a layer-wise discriminative pre-training fashion and then optimized by back-propagation. 36 speakers were randomly selected as the cross validation set. The well-trained CE DNN was subsequently used to initialize the sequential MPE DNN and further tuned for three iterations under the MPE criterion to obtain the baseline MPE DNN system. For the stimulated DNNs, the graphemic [22] information, instead of phones, was used and the positions of the correct graphemes were given by the t-SNE method. Both CE & MPE stimulated DNNs were optimized in a similar fashion as the unstimulated ones.

The word error rate of stimulated DNNs on different η_{st} is summarized in Table 5. It showed a small but consistent gain over the standard training, on both the Viterbi decoding and that processed by confusion network (+*CN*). This indicated

System	η_{st}	%WER	
		Viterbi	+CN
CE	0	60.5	58.3
CE-Stimu	0.05	60.4	58.2
	0.1	60.1	58.0
	0.15	**60.0**	**57.9**
	0.2	60.6	58.2

Table 5: CE Performance of Stimulated DNNs on Javanese.

the effectiveness of the stimulated training using the graphemic information. The best stimulated system was achieved with $\eta_{st} = 0.15$. It was then used to train the stimulated MPE DNN

System	%WER	
	Viterbi	+CN
MPE	58.5	56.3
MPE-Stimu	**57.6**	**55.7**

Table 6: MPE Performance of Stimulated DNNs on Javanese.

and the comparison of MPE systems is given in Table 6. A consistent gain was also acquired by the stimulated DNN in contrast with the unstimulated MPE baseline. In this high-error-rate scenario, LHUC adaptation was not examined.

5. Conclusion

In this paper, stimulated training of DNNs for network regularization and robust adaptation has been investigated. Both schemes rely on generating a smooth "surface" over its activation functions. The approaches were evaluated on two large vocabulary speech recognition tasks: a U.S. English broadcast news task and a Javanese CTS task from the IARPA Babel program. Stimulated DNN training yields consistent performance gains on both tasks over unstimulated baselines. On the BN task, the smoothing approach is applied to rapid adaptation, outperforming the original LHUC scheme. Future work will look at stimulated learning in recurrent neural networks.

6. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[2] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] X. Chen, A. Eversole, G. Li, D. Yu, and F. Seide, "Pipelined back-propagation for context-dependent deep neural networks." in *INTERSPEECH*, 2012.

[4] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 427–436.

[5] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 5188–5196.

[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer vision–ECCV 2014*. Springer, 2014, pp. 818–833.

[7] S. Tan, K. C. Sim, and M. Gales, "Improving the interpretability of deep neural networks with stimulated learning," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 617–623.

[8] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[9] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[10] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems." in *INTERSPEECH*, 2010, pp. 526–529.

[11] T. Tan, Y. Qian, M. Yin, Y. Zhuang, and K. Yu, "Cluster adaptive training for deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4325–4329.

[12] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4535–4539.

[13] C. Wu and M. J. F. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4315–4319.

[14] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[15] C. Zhang and P. Woodland, "DNN speaker adaptation using parameterised sigmoid and relu hidden activation functions," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5300–5304.

[16] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised speaker adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4305–4309.

[17] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. Woodland, and C. Zhang, "The HTK book (for HTK version 3.5)," 2015.

[19] S. Tranter, M. Gales, R. Sinha, S. Umesh, and P. Woodland, "The development of the Cambridge University RT-04 diarisation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[20] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[21] Z. Tuske, D. Nolden, R. Schluter, and H. Ney, "Multilingual mrasta features for low-resource keyword search and speech recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7854–7858.

[22] M. J. F. Gales, K. M. Knill, and A. Ragni, "Unicode-based graphemic systems for limited resource languages," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 5186–5190.