SAT-LHUC: SPEAKER ADAPTIVE TRAINING FOR LEARNING HIDDEN UNIT CONTRIBUTIONS

Pawel Swietojanski and Steve Renals

The Centre for Speech Technology Research, University of Edinburgh
{p.swietojanski, s.renals}@ed.ac.uk

ABSTRACT

This paper extends learning hidden unit contributions (LHUC) unsupervised speaker adaptation with speaker adaptive training (SAT). Contrary to other SAT approaches, the proposed technique does not require speaker-dependent features, the generation of auxiliary generative models to estimate or extract speaker-dependent information, or any changes to the speaker-independent model structure. SAT-LHUC is directly integrated into the objective and jointly learns speaker-independent and speaker-dependent representations. We demonstrate that the SAT-LHUC technique can match feature-space regression transforms for matched narrow-band data and outperform it on wide-band data when the runtime distribution differs significantly from training one. We have obtained 6.5%, 10% and 18.5% relative word error rate reductions compared to speaker-independent models on Switchboard, AMI meetings and TED lectures, respectively. This corresponds to relative gains of 2%, 4% and 6% compared with non-SAT LHUC adaptation. SAT-LHUC was also found to be complementary to SAT with feature-space maximum likelihood linear regression transforms.

***Index Terms*—** SAT, Deep Neural Networks, LHUC

1. INTRODUCTION

Acoustic model (AM) adaptation aims to normalise the mismatch between training and runtime data distributions owing to the acoustic variability among speakers as well as other distortions introduced by the channel or acoustic environment. Speaker adaptive training (SAT) [1, 2], initially proposed for Gaussian mixture models (GMMs), aims to build a canonical acoustic model that is adjusted to the particular characteristics of speakers using linear transforms (operating in either model space [3] or feature space [4]) and found by maximising the likelihood of adaptation data under the model. Those techniques are often referred to as Maximum Likelihood Linear Regression (MLLR) transforms and the feature-space variant (fMLLR) has been successfully applied to speaker adaptive training of Deep Neural Network (DNN) acoustic models [5] often bringing significant improvements in accuracy [6, 7, 8, 9].

Here we are primarily concerned with direct speaker adaptive training of DNN parameters. Contrary to test-only adaptation approaches [10, 11, 12, 13, 14, 15, 16, 17], SAT may offer a more tunable canonical DNN model which is able to perform normalisation better than test-only adaptation. At the same time, we are interested in investigating the possibility of SAT training without using auxiliary features (such as i-vectors [18, 19, 9, 20]), bottleneck features [21, 22]) or additional speaker-dependent (SD) parameters that are added to the speaker-independent (SI) model and retuned in a separate SAT phase [23, 24, 25, 26, 27, 20, 28].

This paper builds on the recently introduced DNN model-based speaker adaptation technique of learning hidden unit contributions (LHUC) [15, 16]. In LHUC, an amplitude parameter is introduced for each hidden unit, tied on a per-speaker basis, and estimated in supervised [15] or unsupervised [16] fashion, the latter using first-pass alignments. This technique has resulted in significant reductions in WER, when tested using the TED talks datasets from the IWSLT evaluation, and was complementary to fMLLR [16]. Here, we extend this approach to speaker adaptive training (SAT-LHUC) in which SI and SD LHUC transforms are estimated during training.

2. LHUC AND SPEAKER ADAPTIVE TRAINING

A speaker independent DNN consists of multiple hidden layers, each implementing some non-linear transformations. Each individual hidden unit acts as an adaptive basis function that learns to recognise certain patterns in the previous layer. The learning process for the DNN is driven by a single objective, with the hidden units driven to specialize and become complementary to each other, in order improve the objective. To explain different patterns in the training data the hidden units learn some joint representation of the problem the model was tasked to solve. However, when the model is applied to unseen data, the relative importance of the hidden units may no longer be optimal. LHUC, given adaptation data, rescales the contributions (amplitudes) of the hidden units in the model without actually modifying their feature receptors (Fig. 2).

LHUC modifies h^l_j, the hidden unit output of unit j in layer l, using a speaker-dependent amplitude function:

$$h^l_j = \xi(r^{l,s}_j) \circ \psi(\mathbf{w}^l_j \mathbf{x} + b^l_j) \,. \qquad (1)$$

$r^s_j \in \mathbb{R}$ is an adaptable speaker-dependent parameter, re-parametrised by a function $\xi : \mathbb{R} \rightarrow \mathbb{R}^+$, where s is the speaker. $\mathbf{w}^l_j$ is the jth column of the corresponding weight matrix $\mathbf{W}^l \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{h}}}$, b^l_j denotes the bias, ψ is the hidden unit activation function, and $\circ$ denotes a Hadamard product.

In the original formulation of LHUC, for test-only adaptation, the speaker-dependent parameters $\theta^s_{LHUC} = \{\{r^{l,s}_j\}^{d^l_{\mathbf{h}}}_{j=1}\}^L_{l=1}$, and the speaker-independent parameters $\theta_{SI} = \{\{\mathbf{w}^l_j, b^l_j\}^{d^l_{\mathbf{h}}}_{j=1}\}^L_{l=1}$ were separately optimised. During training, the hidden units were estimated speaker-independently and were re-scaled by $\xi(r^s)$, with the speaker dependent-amplitude parameters θ^s_{LHUC} estimated using adaptation data. In this work, we use speaker-specific information to learn hidden unit amplitudes during training. The motivation for this approach, termed SAT-LHUC, is that it will lead to hidden units which can learn different behaviours for different speakers, for

This research was supported by EPSRC Programme Grant EP/I031022/1, *Natural Speech Technology* (NST). The NST research data collection may be accessed at http://datashare.is.ed.ac.uk/handle/10283/786.

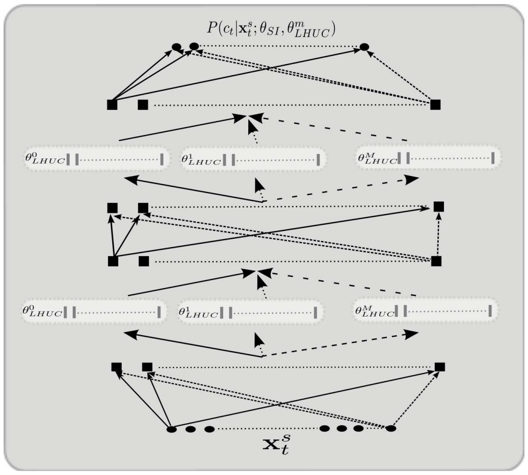


Fig. 1. Schematic of SAT-LHUC training.

example learning that a feature is harmful for some speakers but useful for others (look at Fig.3 for an illustration). Likewise, similar properties (once learned) can be exploited during adaptation to unseen speakers resulting in better speaker-adapted models.

To perform SAT training with LHUC, we use the following objective:

$$\mathcal{L}_{SAT}(\theta_{SI},\theta_{SD}) = -\sum_{t\in D}\log P(c_t|\mathbf{x}_t^s;\theta_{SI};\theta^{m_t}_{LHUC}) \qquad (2)$$

where s denotes the sth speaker, $m_t \in \{0, s\}$ selects the SI or SD LHUC transforms from $\theta_{SD} \in \{\theta^0_{LHUC}, ..., \theta^S_{LHUC}\}$ for each data-point separately (i.e. at the frame level, cf. Fig. 1) based on a Bernoulli distribution parametrised by γ hyper-parameter that determines the overall SI/SD data ratio, as follows:

$$k_t \sim \text{Bernoulli}(\gamma) \qquad (3)$$

$$m_t = \begin{cases} s & \text{if } k_t = 0 \\ 0 & \text{if } k_t = 1 \end{cases} \qquad (4)$$

3. EXPERIMENTAL SETUPS

We have evaluated SAT-LHUC using three different corpora: the TED talks corpus [29] following the IWSLT evaluation protocol (`www.iwslt.org`), the Switchboard corpus of conversational telephone speech [30] (`ldc.upenn.edu`) and the AMI meetings corpus [31, 32] (`corpus.amiproject.org`). Unless explicitly stated otherwise, the models share a similar structure across the tasks – DNNs with 6 hidden layers (2,000 units in each) and a sigmoid non-linearity. The output logistic regression layer models the distribution of context-dependent clustered tied states [33]. The features are presented in 11 (±5) frame long context windows.

For TED we follow the recipe described in [34]. In this work however, compared to [34, 16], our systems benefit from better language models developed for our IWSLT–2014 systems [35]: in particular, we rescore using a 4-gram language model estimated from 751 million words. The baseline TED AMs are trained on unadapted PLP features with first and second temporal derivatives. We report the results on `tst2010` and `tst2013` sets. The latter is more challenging due to larger speaker variability as well as the need for automatic segmentation.

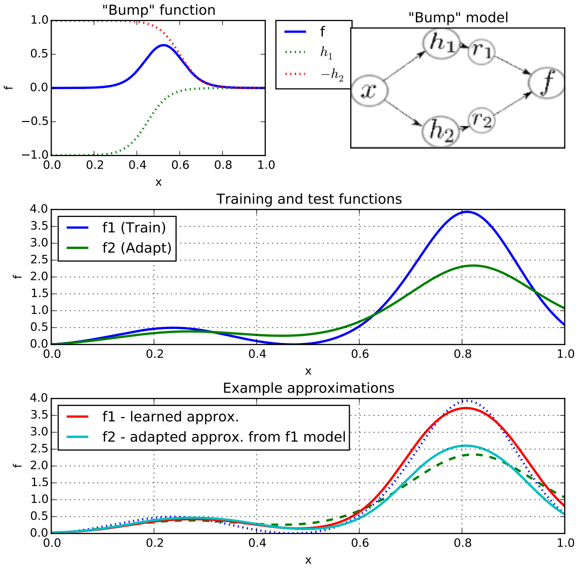


Fig. 2. Example illustration on how LHUC performs adaptation (best view in color). A “bump” model with two hidden units can approximate “bump” functions (top). To learn function f_2 given training data f_1 (middle), we splice two “bump” functions together (4 hidden units, one input/output) to learn an approximation of function f_1. Let us assume that we want to adapt to f_2 using LHUC scalers. We plot the model optimised to f_1 and adapted to f_2 by adjusting only LHUC parameters (bottom).

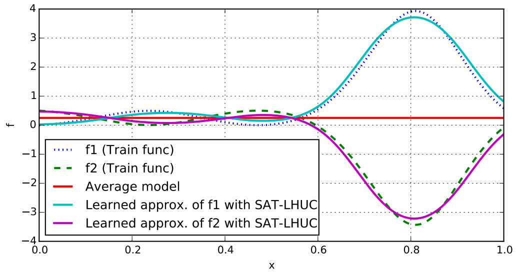
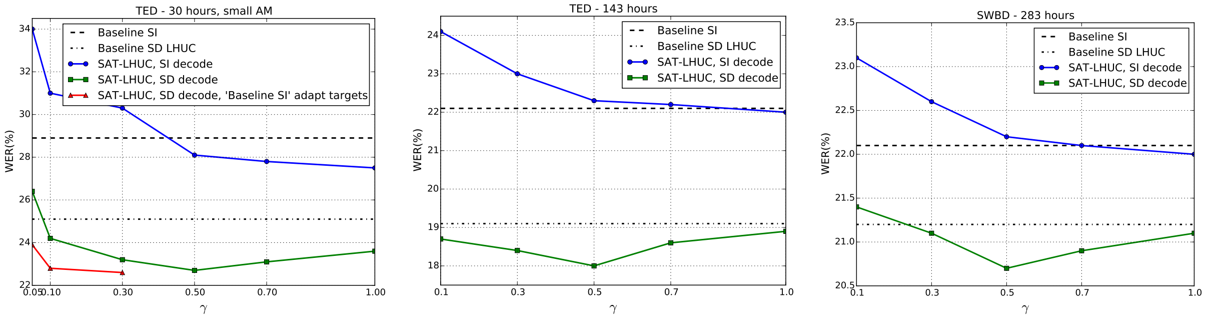


Fig. 3. Example illustration showing how SAT-LHUC can improve a learned representation. Assume we want to approximate both f_1 and f_2 with the similarly constrained (4 hidden units) model from Fig. 2. Again, it is possible with two sets of SAT-LHUC parameters for f_1 and f_2.

In case of Switchboard (SWBD) we use the Kaldi GMM recipe [36, 37], using Switchboard-1 Release 2 (LDC97S62). Our baseline unadapted acoustic models were trained on either MFCC or LDA/MLLT features. The results are reported on the full Hub5 00 set (LDC2002S09) to which we will refer as `eval2000`.

For AMI, we follow the Kaldi GMM recipe described in [38], which is using the so called AMI Full-ASR split on train, dev and eval sets. On this corpus we also train a separate set of models using mel-filter-bank (FBANK) features for which fMLLR transforms cannot be easily obtained, and as such, LHUC makes an interesting adaptation alternative.

The SAT related statistics for each of the above corpora are given in Table 1. Note, in this work we adapt to the headset or the side of a conversation, rather than the actual speaker: hence the number of clusters (or estimated transforms) during training can differ from the number of speakers.



(a) `tst2013` using 30hour-TED AM

(b) `tst2013` using Full-TED AM

(c) `eval2000` using SWBD MFCC AM

Fig. 4. WER(%) as a function of γ in equation (3).

Table 1. Corpus statistics related to SAT and adaptation. In parentheses we give the physical number of speakers.

	Training		Test	
Corpora	#Clusters	Time (h)	#Clusters	Time (h)
AMI	547 (155)	80	135 (36)	17.5
TED	788 (788)	143	39 (39)	9.0
SWBD	4804 (4000)	283	80 (80)	3.6

4. RESULTS

We first investigated the impact of the SI/SD ratio when training the DNN weights and the SI and SD LHUC transforms. The SI/SD ratio depends on γ, the hyper-parameter in eq (3). To speed-up the experimental turnaround we initially limited our experiments to the TED corpus with 30 hours training data, using smaller models (1,000 hidden units per layer). The segments for this limited condition were sampled in such a way that the number of speakers remained the same between the limited and full variants. Results of those experiments on `tst2013`, for different settings of $\gamma \in \{0.05, 0.1, 0.3, 0.5, 0.7, 1.0\}$, can be found in Fig. 4 (a). Note, when $\gamma = 0$ the SI transform would not be estimated; conversely for $\gamma = 1.0$ there would be only a single global SI transform. The latter case is a variant of parametrised sigmoid activations with a learnable amplitude during training [39].

The first observation one can draw from Fig. 4 is that the accuracy of the SAT-LHUC model and the SI decodes depends on the amount of data used to estimate the SI LHUC transforms during training – the less SI data that flows through SI LHUC transforms, the worse SI results are, with a dramatic decrease in first-pass accuracy when less than 30% of data is treated as speaker-independent ($\gamma < 0.3$). Conversely, increasing the SI/SD ratio to about 50% results in comparable accuracy to the standalone SI-trained model. This trend holds for other scenarios with more data, including Full-TED (i.e. 143 hours training data) (Fig. 4 (b)) and SWBD (Fig. 4 (c)).

The parametrised sigmoid function (for $\gamma = 1.0$) is particularly effective for data-constrained experiments (compare Fig. 4 (a) with (b) and (c)); for instance, on 30hour-TED the parametrised sigmoid model results in a WER of 27.5% while the conventional sigmoid model has a WER of 28.9%. This advantage diminishes for bigger models and more data.

Table 2. WER(%) on Ted Lectures. Feature-transform (FT) denotes fMLLR transforms.

System		IWSLT Test set	
Training	Decoding	`tst2010`	`tst2013`
Baseline speaker-independent systems			
SI	SI	15.0	22.1
SAT-LHUC	SI	15.2	22.3
Baseline speaker-adapted systems			
SAT-FT	FT	12.9	20.8
SI	LHUC	12.7	19.1
SAT-FT	FT-LHUC	11.8	18.5
Proposed speaker-adapted systems			
SAT-LHUC+FT	FT	12.7	21.0
SAT-LHUC	LHUC	12.4	18.0
SAT-LHUC-FT	FT-LHUC	11.6	17.6

In the second experiment we investigated how SAT-LHUC affects the accuracy of LHUC adapted systems. To do so we adapted SAT-LHUC models using the first pass adaptation targets obtained from the corresponding SAT-LHUC systems operating in SI mode. Here we can see that a speaker-dependent representation provides a more tunable canonical model. For example, on 30hour-TED an adapted SAT-LHUC $\gamma = 0.3$ system produced 8% relative lower WERs when compared to an adapted SI system (23.2% vs. 25.1%), regardless of the fact that the SAT-LHUC adaptation alignments were 1.4% absolute worse than its SI counterpart (30.3% vs. 28.9%).

Finally, we investigated whether the inferior adaptation results for $\gamma < 0.3$ were caused by differences in learned representations or by lower quality adaptation targets. We used the adaptation targets of the 'Baseline SI' model (28.9%WER) and adapted SAT-LHUC models trained with $\gamma \in \{0.05, 0.1, 0.3\}$ on 30hour-TED. The results (Fig. 4 (a)) indicate that the reason for lower adaptation accuracies (compared to $\gamma = 0.5$ system) was mostly due to less accurate adaptation targets. Adapting the $\gamma = 0.3$ model with the 'Baseline SI' targets reduces the WERs of $\gamma = 0.3$ system to 22.6% (from 23.2%) – 2.5% absolute lower when compared the baseline SI LHUC system (25.1%) (both systems used the same adaptation targets) and 0.1% absolute lower than the best $\gamma = 0.5$ system. This further strengthens our claim that the SAT-LHUC models indeed learn a better and more tunable speaker-dependent representation, but its use is somehow limited by a necessary trade-off of managing a good SI first-pass model.

Fig. 4 (c) shows similar plot but for Switchboard data (more detailed discussion below) and one can observe a similar pattern, with $\gamma = 0.5$ being an optimal choice. This, in conjunction with another validation on AMI data, is a strong indicator that SAT-LHUC training with roughly half of the data-points being treated as speaker-

Table 3. WER(%) and relative WER change (WERR)(%) on Switchboard Hub00. Feature-transform (FT) denotes fMLLR transforms.

	System			Hub5'00				
	Training	Decoding	Features	SWB	CHE	TOTAL	WERR (%)	Baseline Sys. ID
Baseline speaker-independent models								
A	SI	SI	MFCC	15.8	28.4	22.1		
B	SI	SI	LDA/MLLT	15.2	28.2	21.7		
Baseline speaker-adapted systems								
C	SI	LHUC	MFCC	15.4	27.0	21.2	-4.5	A
D	SI	LHUC	LDA/MLLT	14.7	26.6	20.7	-4.6	B
E	SAT-FT	FT	LDA/MLLT	14.2	26.2	20.2	-7.0	B
F	SAT-FT	FT+LHUC	LDA/MLLT	14.2	25.6	19.9	-1.5	E
SAT Trained								
G	SAT-LHUC	LHUC	MFCC	14.8	26.5	20.7	-6.3 / -2.4	A / C
H	SAT-LHUC	LHUC	LDA/MLLT	14.6	25.9	20.3	-6.5 / -1.9	B / D
I	SAT-FT-LHUC	FT+LHUC	LDA/MLLT	14.1	25.6	19.9	-0.0	F

independent makes a good task-independent setting.

SAT-LHUC on TED: Table 2 presents more detailed comparisons of SAT-LHUC adaptation trained on 143 hours of TED talks. Most of our observations are based on `tst2013` which is larger and more challenging than `tst2010`. However, for the sake of comparability with our previous work on LHUC [16], we also report the results on `tst2010` which is better matched to the training data. Speaker-independent baselines are listed in the first block of Table 2 – we can see the SAT-LHUC model in SI decoding mode falls around 0.2% absolute behind the standalone SI model on both test sets. Then, the second block presents the adapted baselines including speaker adaptive training with fMLLR transforms applied at both training and decoding stages. Here we can observe 6–14% relative improvement from fMLLR transforms; surprisingly the improvement is smaller for the more mismatched data of `tst2013` in which scenario test-only LHUC performs significantly better – improving accuracy by 15.3% and 13.5% relative for `tst2010` and `tst2013`, respectively. The two approaches can be further combined resulting in an additional improvement of 5% relative compared to their standalone usage.

The third block of Table 2 presents the WERs of the proposed SAT-LHUC training scheme (section 2). We observe further gains for both sets: for instance, on `tst2013` SAT-LHUC gives 6% relative gain when compared to test-only LHUC and 13% gain when compared to an fMLLR transform. Not surprisingly, joint combination of SAT LHUC with SAT fMLLR bring further gains of about 2% relative on average.

SAT-LHUC on Switchboard: In contrast to other corpora, we have observed that test-only LHUC does not match the WERs obtained from SAT fMLLR models. Comparing the WERs of SI system B (21.7%) with test-only LHUC system D (20.7%) and the SAT trained baseline that utilised fMLLR feature-transforms, system E (20.2%) (Table 3), it is apparent that the improvement from test-only LHUC is comparable with other test-only adaptation techniques, e.g. feature-space discriminative linear regression [7], but neither matches the SAT fMLLR models. This could be due to the fact Switchboard is narrow-band and thus contains less information for discrimination between speakers [40], especially when estimating relevant statistics from small amounts of unsupervised adaptation data. Additionally, the Switchboard part of `eval2000` has a large overlap between training and test speakers – 36 out of 40 test speakers are observed in training [41], which limits the need for adaptation, but also enables models to learn much more accurate speaker characteristics during supervised speaker adaptive training.

Table 4. WER(%) on AMI.

Training	Decoding	Features	`dev`	`eval`
Baseline speaker-independent systems				
SI	SI	FBANK	26.5	29.1
SAT-LHUC	SI	FBANK	26.3	28.9
Speaker-adapted systems				
SI	LHUC	FBANK	25.6	27.1
SAT-LHUC	LHUC	FBANK	24.9	26.1
SI	FT	FMLLR	26.2	27.3
SAT-FT	FT+LHUC	FMLLR	25.6	26.2

The adaptation results of the SAT-LHUC model are given in Table 3 in row H (20.3%) where we almost match the SAT fMLLR baseline (20.2). We also observe that LHUC performs relatively better under more mismatched conditions – here Callhome (CHE) subset of `eval2000`– similar to what was found on TED. Note, we train two sets of models, one on MFCC features to stay compatible with test-only adaptation techniques reported in [7] as well as linear discriminant analysis (LDA) features based on which Kaldi SWBD recipe [36] estimates FMLLR transforms - which form our baseline for the SAT training.

SAT-LHUC on AMI: Table 4 gives the WERs when we applied SAT-LHUC to the AMI dataset. The SAT-LHUC system was trained with $\gamma = 0.5$ and was found to bring an average 3.2% relative WER reduction on top of LHUC applied to the SI trained model, or 8% relative reduction when compared to unadapted FBANK-trained models. The final numbers match a more complicated adaptation pipeline that adapts with FMLLR transforms followed by test-only LHUC adaptation.

5. CONCLUSIONS

We have proposed SAT-LHUC, an effective speaker adaptive training extension to the LHUC adaptation technique. SAT-LHUC does not require any auxiliary models or additional SAT training stages on top of the SI model to be effective, though it can be easily combined with other adaptation methods to bring further gains. The standalone variant is probably the simplest SAT approach proposed to date. This work is further extended in [42]; in the future we plan to evaluate whether the proposed form of SAT remains effective with other types of non-linearities (as is the case for LHUC adaptation [16]), and an extension to sequence discriminative training [43,44,36].

6. REFERENCES

[1] T Anastasakos, J McDonough, R Schwartz, and J Makhoul, "A compact model for speaker-adaptive training," in *Proc ICSLP*, 1996, pp. 1137–1140.

[2] MJF Gales, "Cluster adaptive training of hidden markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 417–428, 2000.

[3] CJ Leggetter and PC Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, pp. 171–185, 1995.

[4] MJF Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, April 1998.

[5] G Hinton, L Deng, D Yu, GE Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, TN Sainath, and B Kingsbury, "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, Nov 2012.

[6] A Mohamed, TN Sainath, G Dahl, B Ramabhadran, GE Hinton, and MA Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. ICASSP*, May 2011, pp. 5060–5063.

[7] F Seide, X Chen, and D Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc IEEE ASRU*, 2011.

[8] P Swietojanski, A Ghoshal, and S Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proc IEEE ICASSP*, 2013.

[9] G Saon, H Soltau, D Nahamoo, and M Picheny, "Speaker adaptation of neural network acoustic models using i-vectors.," in *Proc IEEE ASRU*, 2013, pp. 55–59.

[10] 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, pp. 2171–2174.

[11] V Abrash, H Franco, A Sankar, and M Cohen, "Connectionist speaker normalization and adaptation," in *Proc Eurospeech*, 1995, pp. 2183–2186.

[12] 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 IEEE SLT*, 2012.

[13] D Yu, K Yao, H Su, G Li, and F Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition.," in *Proc IEEE ICASSP*, 2013, pp. 7893–7897.

[14] H Liao, "Speaker adaptation of context dependent deep neural networks.," in *In Proc. ICASSP*. 2013, pp. 7947–7951, IEEE.

[15] O Abdel-Hamid and H Jiang, "Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition.," in *Proc. Interspeech*. pp. 1248–1252, ISCA.

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

[17] P Swietojanski and S Renals, "Differentiable pooling for unsupervised speaker adaptation," in *Proc. IEEE ICASSP*, 2015.

[18] N Dehak, PJ Kenny, R Dehak, P Dumouchel, and P Ouellet, "Front end factor analysis for speaker verification," *IEEE Trans Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2010.

[19] M Karafiat, L Burget, P Matejka, O Glembek, and J Cernozky, "iVector-based discriminative adaptation for automatic speech recognition," in *Proc IEEE ASRU*, 2011.

[20] Y Miao, H Zhang, and F Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proc. Interspeech*, 2014.

[21] P Bell, P Swietojanski, and S Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc IEEE ICASSP*, 2013.

[22] Y Liu, P Karanasou, and T Hain, "An investigation into speaker informed DNN front-end for LVCSR," in *Proc IEEE ICASSP*, 2015.

[23] JS Bridle and S Cox, "Recnorm: Simultaneous normalisation and classification applied to speech recognition," in *Advances in Neural Information Processing Systems 3*, 1990, pp. 234–240.

[24] J Trmal, J Zelinka, and L Müller, "On speaker adaptive training of artificial neural networks," in *Proc. Interspeech*, 2010.

[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 IEEE ICASSP*, 2013, pp. 4277–4280.

[26] C Wu and M Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *Proc. ICASSP*. 2015, IEEE.

[27] T Tan, Y Qian, M Yin, Y Zhuang, and K Yu, "Cluster adaptive training for deep neural network," in *Proc. ICASSP*. 2015, IEEE.

[28] M Delcroix, K Kinoshita, T Hori, and T Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Proc. ICASSP*. 2015, IEEE.

[29] M Cettolo, C Girardi, and M Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proc EAMT*, 2012, pp. 261–268.

[30] John J Godfrey, Edward C Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*. IEEE, 1992, pp. 517–520.

[31] J Carletta, "Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus.," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[32] S Renals, T Hain, and H Bourlard, "Recognition and understanding of meetings: The AMI and AMIDA projects," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'07*, Kyoto, 12 2007, IDIAP-RR 07-46.

[33] GE 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.

[34] P Swietojanski, A Ghoshal, and S Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proc. ICASSP*, 2013.

[35] P Bell, P Swietojanski, J Driesen, M Sinclair, F McInnes, and S Renals, "The UEDIN ASR Systems for the IWSLT 2014 Evaluation," in *Proc. IWSLT*, 2014.

[36] K Vesely, A Ghoshal, L Burget, and D Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, Lyon, France, August 2013.

[37] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlíček, Y Qian, P Schwarz, J Silovský, G Stemmer, and K Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, December 2011.

[38] P Swietojanski, A Ghoshal, and S Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. IEEE ASRU*, December 2013.

[39] C Zhang and PC Woodland, "Parameterised Sigmoid and ReLU Hidden Activation Functions for DNN Acoustic Modelling," in *Proc. Interspeech*, 2015.

[40] M Wester, Z Wu, and J Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *Proc. of Interspeech*, September 2015.

[41] J Fiscus, W M Fisher, A F Martin, M A Przybocki, and D S Pallett, "2000 NIST evaluation of conversational speech recognition over the telephone: English and Mandarin performance results," in *Proc. Speech Transcription Workshop*. Citeseer, 2000.

[42] P Swietojanski, J Li, and S Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *arXiv:1601.02828*, 2016.

[43] D Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, University of Cambridge, 2003.

[44] B Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. IEEE ICASSP*, 2009, pp. 3761–3764.