Cross Lingual Transfer Learning for Zero-Resource Domain Adaptation

CROSS LINGUAL TRANSFER LEARNING FOR ZERO-RESOURCE DOMAIN ADAPTATION

Alberto Abad[1,2] *Peter Bell*[2] *Andrea Carmantini*[2] *Steve Renals*[2]

[1]INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal
[2]Centre for Speech Technology Research, University of Edinburgh, UK

ABSTRACT

We propose a method for zero-resource domain adaptation of DNN acoustic models, for use in low-resource situations where the only in-language training data available may be poorly matched to the intended target domain. Our method uses a multi-lingual model in which several DNN layers are shared between languages. This architecture enables domain adaptation transforms learned for one well-resourced language to be applied to an entirely different low-resource language. First, to develop the technique we use English as a well-resourced language and take Spanish to mimic a low-resource language. Experiments in domain adaptation between the conversational telephone speech (CTS) domain and broadcast news (BN) domain demonstrate a 29% relative WER improvement on Spanish BN test data by using only English adaptation data. Second, we demonstrate the effectiveness of the method for low-resource languages with a poor match to the well-resourced language. Even in this scenario, the proposed method achieves relative WER improvements of 18-27% by using solely English data for domain adaptation. Compared to other related approaches based on multi-task and multi-condition training, the proposed method is able to better exploit well-resource language data for improved acoustic modelling of the low-resource target domain.

***Index Terms*—** acoustic modelling, domain adaptation, multi-lingual speech recognition

1. INTRODUCTION

In automatic speech recognition (ASR), the problem of building acoustic models that behave robustly in different usage domains is still an open research challenge, despite the emergence of deep neural network (DNN) models. Several approaches have been proposed in recent years to adapt well-trained DNNs from a *source* domain to a new *target* domain, perhaps with limited training data. Examples include data augmentation strategies [1]; the use of auxiliary features such as i-vectors [2], posterior or bottleneck features [3, 4] trained on source-domain data; adapting selected parameters [5, 6, 7]; adversarial methods [8]; as well as simple yet effective approaches such as applying further rounds of training to DNNs initialised on source data.

The common ground in the vast majority of these works is that some transcribed data – even if usually a limited amount – from the target domain is available for adaptation of the acoustic models. This assumption, reasonable for well-resourced languages (WR), may not hold in the case of low-resourced languages (LR) for which even the amount of data available in the source domain may be very limited, and it is expensive or impractical to arrange for transcription of data from a new domain.

This is the scenario tackled in the IARPA MATERIAL programme[1]. The programme seeks to develop methods for searching speech and text in low-resource languages using English queries. In particular, ASR systems must operate on diverse multi-genre data, including telephone conversations, news and topical broadcasts. However, the only manually annotated training data available is from the telephone conversations domain.

One approach to this problem is to collect a corpus of untranscribed data from the target domain in the LR language (for example, by web-crawling) and use an ASR system built for the source domain to create an automatic transcription, which is then used to train domain-adapted models. This semi-supervised approach to DNN training has been successfully used eg. [9, 10, 11]. However, the technique requires careful confidence-based data selection, and is very sensitive to the performance of the source system on the target data. Another drawback, when rapid deployment to a new domain is required, is the need to run computationally expensive decoding on large quantities of data in order to harvest sufficient quantities of training material.

In this work, inspired by the challenges posed by the MATERIAL programme, we adopt a completely different approach: we explore whether it is possible to transfer a specific domain transform learned in a WR language to a LR language for which no target training data is available at all, in other words, is a method for adaptation between two given domains portable across languages? We thus aim to improve the performance of a LR ASR system in a new *target* domain by using only data of a WR language in both the *source* and *target* domains. To this end, we propose an adaptation scheme that uses multi-lingual AM training to enable cross-lingual sharing of domain adaptation techniques. Then, based on the hypothesis that initial layers of a DNN encode language-independent acoustic characteristics, we are able to transfer the adapted layers learned for one target domain from one language to another.

For the development of the proposed cross-lingual domain adaptation approach, it is more convenient to select a pair of languages for which data is available from both source and target domains in each language, enabling oracle experiments to be carried out. Hence, in this work we initially use English as the WR language and pretend

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL) contract #FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Alberto Abad was supported by FCT –Fundação para a Ciência e a Tecnologia– under project UIDB/50021/2020 and by PT2020 funds under the project "Unbabel Scribe: AI-Powered Video Transcription and Subtitle" with the contract number: 038510.

[1] https://www.iarpa.gov/index.php/research-programs/material

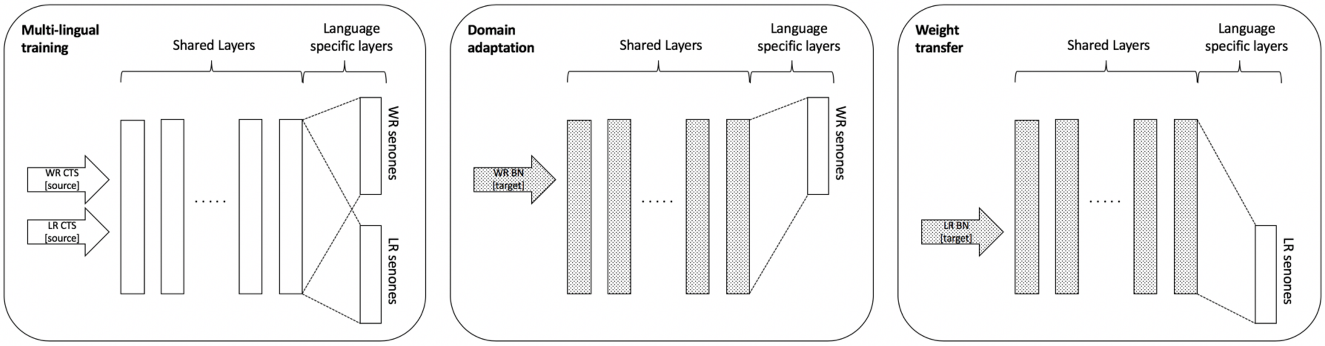

Fig. 1. Steps of the proposed cross-lingual domain adaptation scheme: 1) multi-lingual training; 2) adaptation of the shared parameters using WR data in the target domain; and 3) weight transfer of the domain adapted shared layers to the original LR final, language-dependent layers.

that Spanish is an LR language. As in the real MATERIAL task, we choose conversational telephone speech (CTS) as the source domain and use broadcast news (BN) as the target domain. We then explore the effectiveness of the proposed method on two of the actual MATERIAL target languages: Tagalog and Lithuanian.

The rest of this paper is organized as follows. Section 2 describes the proposed cross-lingual domain adaptation approach. Then, experimental setup, including corpora and details on the architecture of the developed ASR systems, is reported in section 3. Finally, experimental evaluation is presented in section 4 before the final concluding remarks.

2. CROSS-LINGUAL DOMAIN ADAPTATION

The main objective of this work is to propose an adaptation scheme for DNN-based acoustic models that allows for cross-lingual domain adaptation from a LR language system trained for one *source* domain into a *target* domain using solely adaptation data of a WR language. Considering that the role of the DNN is to learn a non-linear mapping between the acoustic features (e.g. MFCC) and phoneme-related classes (e.g. senones), it is a common interpretation that the initial layers of a phonetic network are expected to encode lower-level acoustic information, while deeper layers codify more complex cues closer to phonetic classes [12]. Thus, following this interpretation, we hypothesize that the initial layers of a phonetic network encode basic acoustic information that is independent of the language and the task at hand, while the later ones are specific to each language. Under this interpretation, we suggest that the modifications that could be applied to the initial layers of a phonetic network in any language to adapt to the specific characteristics of a new domain should be similar (and transferable) among different languages. To take advantage of this possibility, it is necessary to design a network architecture in which parameter transforms can be meaningfully shared among the LR and WR languages, followed by a set of final language specific layers. This solution can be attained through multi-task learning and it is the backbone of our proposed scheme. As can be seen in Figure 1, the process consists mostly on three steps: 1) a multi-lingual network is trained with data of both the LR and WR language in the source domain (left); 2) shared layers are adapted using WR data of the target domain (center); and 3) the adapted shared layers are transferred to the original LR language network resulting in a domain adapted version of the LR network (right).

2.1. Multi-lingual training

Multi-task learning [13] refers to the process of simultaneously learning multiple tasks from a single data set that contains annotations for different tasks. Typically, the network architecture consists of some initial layers that are shared by the multiple tasks and some final task-specific layers, one for each considered task. Back-propagation is applied for each task alternatively using all the training data propagated through both the task-specific and shared layers. This type of learning provides an improved regularization effect of the resulting networks compared to conventional single task learning approaches and has been used successfully in single-language acoustic modelling [14, 15].

This type of approach has been also successfully applied in ASR using data from multiple languages to learn multi-lingual networks, in which each task objective corresponds to the phoneme (senone) classification of the different languages [16, 17]. In general, multi-lingual learning has shown to be particularly beneficial when languages with limited training resources are involved. In this work, we use multi-lingual learning to train an initial network using data from both the LR and the WR languages in the source domain. Hence, the source multi-language network has a set of shared layers followed by two language specific (LR and WR) set of final layers.

2.2. Domain adaptation

Given the multi-lingual architecture described previously, we adapt the shared components of the network using target data of the WR language whilst freezing the remaining language-specific layers. By doing so, our expectation is that the kind of transformations that the new adapted network will learn will be language-independent and will mostly be related with the particular acoustic characteristics of the new data. Keeping the upper layers frozen ensures that the newly adapted lower layers of the network will continue to be appropriate as inputs for the layers specific to the LR language, despite no LR data being used for the adaptation.

In general, one may explore any of the well-researched strategies for network weight adaptation, such as LHUC [6] or LIN [18]. In this work, given that substantial quantities of target domain data are available in the WR language, we simply adapt the weights of a selected subset of layers of the shared network (initialized with the weights learned in the multi-lingual learning stage) through simple backpropagation updates with a varying number of data training epochs and an appropriately chosen learning rate.

3. EXPERIMENTAL SETUP

3.1. Corpora

In all experiments we take conversational telephone speech (CTS) as the source domain, for which transcribed training data is available for all languages. For English and Spanish, we train on data from the Fisher corpus[19] (~200 hours and ~163 hours respectively). Note that for the former, we use only a subset of the full corpus. For Tagalog and Lithuanian, we use data from the IARPA Babel full language packs (80 hours and 40 hours respectively). CTS data is all sampled at 8khz. In the MATERIAL task, the target domain is a mixture of broadcast news (BN) and topical broadcast (TB) domain, both with wideband 16khz audio. We approximate this target in English and Spanish by using broadcast news (BN) data from HUB4 [20] with ~150 hours of English data, used for adaptation, and ~30 hours of Spanish, used for oracle experiments only. For each corpus we use the standard evaluation sets: the 1997 HUB4 English Evaluation set is used for English BN; and for the Tagalog and Lithuanian, we use the BN and TB "Analysis" test sets provided by the MATERIAL programme.

3.2. System description

The Kaldi toolkit [21] has been used for the development of all the ASR systems. To obtain the set of language-specific senones and frame-level phonetic alignments needed for training the DNNs, initial HMM-GMM systems have been built for each language and domain. HMM-GMM training follows the conventional recipes in Kaldi, consisting on several stages of refinement from monophone to context-dependent models trained on LDA+MLLT+fMLLR features [22]. HMM-DNN ASR systems share a common input feature representation of 43 dimensions corresponding to 40 high resolution MFCCs components plus 3 additional pitch and voicing related features [23]. Neither side speaker information (i-vector), nor speed perturbation data augmentation have been used in these experiments. Note that all data is downsampled to 8 kHz to match the sampling rate of the CTS source domain data. Hence, all the systems reported in this work have been trained and evaluated at 8kHz.

The acoustic models are TDNN networks trained with frame-level cross-entropy loss criterion [24]. Network architectures consist of a stack of 7 TDNN hidden layers, each containing 650 units, with RELU activation functions. These correspond to the shared language-independent layers of the network. For each language, a pre-final fully connected layer of 650 units with RELU activations and a final softmax layer is appended. The size of the output layers correspond to the size of the outputs of the single language networks of the source CTS domain. During training, the samples of each language are not scaled, thus no compensation for different training data sizes is performed.

The optimization method used is natural gradient stabilized stochastic gradient descent [25] with an exponentially decaying learning rate. The starting learning rate is set to 0.0015 and decays by a factor of 10 over the entire training. The baseline and multilingual AMs were trained for 3 epochs with a minibatch size of 256. The parameter change is limited to a maximum of 2 for each minibatch to avoid parameter explosion. In the adaptation stage of the proposed approach, the network is initialized with the multi-lingual network weights and the learning rate of the frozen layers is set to 0. The number of adaptation epochs is a varying parameter investigated in the results section. All the remaining configurations are identical to that of the multi-lingual training stage.

	Test condition			
	WR		LR	
	CTS source	BN target	CTS source	BN target
mono-ling BN AM	—	11.8	—	19.2
mono-ling CTS AM	22.6	19.6	32.3	40.0
multi-ling CTS AM	23.6	19.2	32.6	32.9

Table 1. WER (%) of the mono-lingual matched BN target domain trained systems (first row), the mono-lingual mismatched CTS source domain trained systems (second row) and the multi-lingual ASR system trained with WR and LR CTS source data (third row) for the different test domain conditions.

Since the focus of this work is on adaptation of the AM, domain matched language models are always used in decoding. That is, CTS source and BN target domain test data is decoded with corresponding language-specific CTS and BN LMs. While this seems to go against the assumption of working with low resource languages, text data is usually much easier to obtain compared to transcribed speech. CTS LMs are trained on the training transcriptions of the relevant corpora. BN LM for Spanish is also trained only on the training transcriptions, while the BN LM for English is trained using the transcriptions and additional text from the 1996 CSR HUB-4 Language Model and the North American News Text Corpus. BN LMs for Lithuanian and Tagalog are trained on around 30M words of web-crawled text from CommonCrawl and other online sources.

4. RESULTS

4.1. English and Spanish baseline systems

In this and the following two sections, we take Spanish to be the LR language; as always, the WR language is English. Table 1 shows, in the first row, word error rate (WER) performance of single language baseline systems using BN target in-domain acoustic models. In the case of the LR language, this is an oracle experiment, since we assume that in reality, no target domain data is available for this language. The second row shows results of CTS source domain acoustic models evaluated with both CTS and BN test data. As expected, one observes a large degradation when decoding LR BN target domain data with CTS acoustic models: about 20% absolute WER compared with the matched experiments. The performance drop can also be observed in the WR experiments, but it is not so acute, probably due to the increased amount of training data used in the WR system. Note that the objective of this work is to make the 40.0% of the LR baseline system when decoding target domain data as close as possible to the oracle 19.2% figure, without using any in-domain LR language training data.

The last row of Table 1 shows the WER performance of the multi-lingual system trained with LR and WR data of the CTS source domain. For both languages, the performance of the multi-lingual models on the CTS source domain is close to that obtained using the respective mono-lingual model. However, for the cross-domain test case, there is a remarkable improvement in the LR language performance from 40.0% to 32.9%. This is a 17.8% relative improvement in the BN target domain achieved by using only out-of-domain CTS WR data, thanks to the multi-lingual training scheme. While the benefits of multi-lingual training for LR were expected, it is very interesting to observe that these are much more significant in the cross-domain case. This may be partially explained due to the

	WR	LR
	BN target	BN target
mono-ling CTS AM (1)	19.6	40.0
multi-ling CTS AM (2)	19.2	32.9
proposed CL adapt AM (3)	14.5	**28.4**
multi-task CL AM (4)	12.4	29.1
multi-task CL + adapt AM (5)	12.3	29.1
multi-cond CL AM (6)	12.5	29.2
multi-cond CL + adapt AM (7)	12.2	29.1

Table 2. WER (%) of the mono-lingual AM (1), the multi-lingual AM (2), the proposed adapted AM (3) and alternative cross-lingual AM (4-7) obtained in the WR and LR BN target domain test sets.

considerably increased amount of data and variation to which this network is exposed compared to the LR baseline.

4.2. Cross-lingual network adaptation results

The third row of Table 2 reports the performance of the proposed method in contrast to the mono-lingual (first row) and multi-lingual (second row) baselines. The proposed cross-lingual domain adaptation method has been tested for a varying number of training epochs (from 0.5 to 3) and adapted shared layers (from 1 to 6); the reported result corresponds to the best adaptation configuration, which is obtained when the 3 first hidden shared layers are adapted using all WR target domain data for 1 epoch of training. We observed in the complete set of experiments that results on LR data are not particularly sensitive to number of epochs or layers amongst those tested, ranging from 28.4% to 29.1% in all cases. However, as expected, this is not the case for the WR language results, in which performance tends to keep improving with increased number of adapted layers and epochs. Hence, an improvement in the WR case does not necessarily imply an improvement in the LR case. We conclude that there seems to be a limit on the amount of information that is transferable from the WR system to the LR system. Overall, we observe that by using WR CTS source domain data for multi-lingual learning we are able to improve from 40.0% to 32.9% and by then using WR BN target domain data and the proposed adaptation method, we further increase performance to 28.4% WER. This is an absolute 11.6% WER decrease, which is a recovery of around 50% of the performance loss due to the lack of LR training data in the CTS target domain when compared to a system fully trained with BN data (see Table 1). This is attained by using only WR data and no any additional LR data.

4.3. Comparison with similar cross-lingual approaches

In this section, the proposed approach is compared with two related cross-lingual information transfer methods: first, with an AM trained in a multi-task fashion, considering LR source, WR source and WR target as three separate tasks, referred to as *multi-task*; and second, with an AM trained in a multi-task and multi-condition fashion, considering the LR source as one task, and the mix of WR source and WR target as a second task, referred to here as *multi-cond*. Rows 4 and 6 in Table 2 report performance of these two cross-lingual alternative approaches. For these experiments, the exact same network architecture, training and decoding recipes as previously have been followed. Notice that, like in the proposed method, it is possible to use these networks as an initialization for further fine-tuning using only WR target domain data for a varying number of epochs and shared adaptation layers. Thus, rows 5 and 7 of Table 2 report results after fine-tuning adaptation for the best configuration of epochs and number of adapted layers in each case. For both approaches, the additional fine-tuning does not provide significant improvements. Performance converges already after the initial training. In fact, we observe that the best adaptation configuration is attained with the minimal number of epochs and adaptation layers. For any other configuration, performance oscillates in absolute differences of ± 0.1. Overall, the cross-lingual proposed scheme outperforms any of the other methods, being able to better leverage information from the WR data for improved LR acoustic modelling in the target domain.

	Tagalog			Lithuanian		
	BN	TB	Avg.	BN	TB	Avg.
mono-ling CTS (1)	53.2	58.7	57.3	45.6	43.0	44.0
multi-ling CTS (2)	46.5	52.2	50.7	38.2	36.5	37.1
proposed CL adapt (3)	41.9	48.5	**46.8**	31.6	32.1	**31.9**

Table 3. WER (%) of the single language AM (1), the multi-lingual AM (2) and the proposed adapted AM in the Tagalog and Lithuanian MATERIAL evaluation sets.

4.4. Experiments with MATERIAL languages

In this section we investigate the proposed cross-lingual adaptation approach considering two of the IARPA MATERIAL languages: Tagalog and Lithuanian. For these experiments, we keep the same network architecture and training and decoding recipes as previously, including the set of best parameters found for the proposed cross-lingual method: adaptation of the 3 first hidden shared layers for 1 epoch. The new languages are less related to English than Spanish, and the data available present a poorer match between source and target conditions. Despite the significant differences among languages and target domain conditions, results reported in Table 3 show that the proposed method is able to effectively exploit English data to improve ASR performance of LR languages in any of the wideband data sub-conditions. As expected, the proposed method is more effective in closer target conditions to those of the WR data: for the BN wideband sub-condition the relative WER improvements are 21.2% for Tagalog and 30.7% for Lithuanian; while the improvements for the TB wideband sub-condition are 17.4% for Tagalog and 25.3% for Lithuanian. Overall, the average relative WER improvements for the wideband conditions are 18.3% and 27.5% for the Tagalog and Lithuanian languages, respectively.

5. CONCLUSIONS

This paper has demonstrated that it is possible to transfer domain adaptation of DNNs from one language to another, enabling adaption of a low-resourced language to be performed with absolutely no data from the target domain. This has been achieved thanks to a multi-lingual network architecture that allows for meaningful share of the parameter transforms among languages. In our experiments, the proposed cross-lingual domain adaptation approach outperforms other similar methods achieving up to a 29% relative WER improvement in the target domain when similar languages and source and target domain conditions are considered. Moreover, the proposed adaptation scheme also allows for remarkable WER improvements in the case of less favorable language and domain conditions. Future work will extend the method to sequence-trained models and also investigate the combination with other cross-lingual information transfer methods, such as bottle-neck features trained on multilingual multi-domain data, and SAT vector-based approaches (e.g. i-vectors).

6. REFERENCES

[1] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchinani, “Generation of large-scale simulated utterances in virtual rooms to train deep neural networks for far-field speech recognition in Google Home,” in *Proc. Interspeech*, 2017.

[2] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, “Reverberation robust acoustic modelling using i-vectors with time delay neural networks,” in *Proc. Interspeech*, 2015.

[3] S. Thomas, S. Ganapathy, and H. Hermansky, “Cross-lingual and multi-stream posterior features for low resource LVCSR systems,” in *Proc. Interspeech*, 2010.

[4] P.J. Bell, M. J. F. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P.C. Woodland, “Transcription of multi-genre media archives using out-of-domain data,” in *Proc. IEEE Workshop on Spoken Language Technology*, Dec. 2012.

[5] 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.

[6] P. Swietojanski, J. Li, and S. Renals, “Learning hidden unit contributions for unsupervised acoustic model adaptation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, August 2016.

[7] L. Samarakoon and K.C. Sim, “Factorized hidden layer adaptation for deep neural network based acoustic modelling,” *IEEE/ACM Transactions on audio, speech and language processing*, vol. 24, no. 12, pp. 2241–2250, 2016.

[8] Y. Shinohara, “Adversarial multi-task learning of deep neural networks for speech recognition,” in *Proc. Interspeech*, 2016, pp. 2369–2372.

[9] K. Vesely, M Hannemann, and L. Burget, “Semi-supervised training of deep neural networks,” in *Proc. ASRU*, 2013.

[10] T. Drugman, J. Pylkko, and R. Kneser, “Active and semi-supervised learning in asr: benefits on the acoustic and language models,” in *Proc. Interspeech*, 2016.

[11] Andrea Carmantini, Peter Bell, and Steve Renals, “Untranscribed web audio for low resource speech recognition,” in *Proc. Interspeech*, 2019.

[12] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, pp. 41–75, 1997.

[14] Michael L Seltzer and Jasha Droppo, “Multi-task learning in deep neural networks for improved phoneme recognition,” in *Proc. ICASSP*. IEEE, 2013, pp. 6965–6969.

[15] Peter Bell, Pawel Swietojanski, and Steve Renals, “Multitask learning of context-dependent targets in deep neural network acoustic models,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 238–247, 2017.

[16] Z. Tüske, R. Schlüter, and H. Ney, “Multi-lingual hierarchical MRASTA features for ASR,” in *Proc. Interspeech*, 2013.

[17] J.-T. Huang, J. Li, D. Yu, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *Proc. ICASSP*, 2013.

[18] 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.

[19] Christopher Cieri, David Miller, and Kevin Walker, “The fisher corpus: a resource for the next generations of speech-to-text,” in *LREC*, 2004, vol. 4, pp. 69–71.

[20] David Graff, Zhibiao Wu, Robert MacIntyre, and Mark Liberman, “The 1996 broadcast news speech and language-model corpus,” in *Proc. DARPA Workshop on Spoken Language technology*, 1997, pp. 11–14.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, Dec. 2011.

[22] Shakti P. Rath, D. Povey, K. Veselý, and J. Cernocký, “Improved feature processing for deep neural networks,” in *Proc. INTERSPEECH*, 2013.

[23] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *Proc. ICASSP*, 2014, pp. 2494–2498.

[24] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. Interspeech*, 2015, pp. 3214–3218.

[25] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur, “Parallel training of DNNs with natural gradient and parameter averaging,” *Proc. ICLR workshop*, 2015.