SOURCE DOMAIN DATA SELECTION FOR IMPROVED TRANSFER LEARNING TARGETING DYSARTHRIC SPEECH RECOGNITION

Feifei Xiong[⋆] *Jon Barker*[⋆] *Zhengjun Yue*[⋆] *Heidi Christensen*[⋆†]

[⋆] Speech and Hearing Group (SPandH), Dept. of Computer Science, University of Sheffield, UK
[†] Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, UK

ABSTRACT

This paper presents an improved transfer learning framework applied to robust personalised speech recognition models for speakers with dysarthria. As the baseline of transfer learning, a state-of-the-art CNN-TDNN-F ASR acoustic model trained solely on source domain data is adapted onto the target domain via neural network weight adaptation with the limited available data from target dysarthric speakers. Results show that linear weights in neural layers play the most important role for an improved modelling of dysarthric speech evaluated using UASpeech corpus, achieving averaged 11.6% and 7.6% relative recognition improvement in comparison to the conventional speaker-dependent training and data combination, respectively. To further improve the transferability towards target domain, we propose an *utterance*-based data selection of the source domain data based on the entropy of posterior probability, which is analysed to statistically obey a Gaussian distribution. Compared to a *speaker*-based data selection via dysarthria similarity measure, this allows for a more accurate selection of the potentially beneficial source domain data for either increasing the target domain training pool or constructing an intermediate domain for incremental transfer learning, resulting in a further absolute recognition performance improvement of nearly 2% added to transfer learning baseline for speakers with moderate to severe dysarthria.

Index Terms— Transfer learning, data selection, entropy, posterior probability, dysarthric speech recognition

1. INTRODUCTION

The high inter- and intra-speaker variability inherent in dysarthric speech [1, 2] severely hinders the application of state-of-the-art automatic speech recognition (ASR) systems usually constructed using typical speech (cf. [3, 4]). Recent progresses in ASR performance have been achieved mainly via the use of deep neural networks (DNNs) [5], which require a large amount of data for a satisfactory recognition performance. However, the difficulty in collecting dysarthric speech data [6] means that obtaining sufficient data is a major challenge.

Research has been conducted for developing dysarthric speech recognition systems by better modelling of the dysarthric speech variability (cf. [7, 8]), or by focusing on dysarthric speech data collection [9, 10, 11, 12]. There has also been increasing interest in recent years in porting the advances from DNNs seen for mainstream ASR to that of dysarthric ASR, particularly through making best use of the often limited amount of dysarthric data. For instance, to reduce the influence caused by dysarthric speech variability in tandem ASR systems, bottleneck features have been proposed via DNNs using a large amount out-of-domain data [13], or convolutional neural networks have been employed [14]. Various advanced forms of DNN architecture were tested in [15] for both tandem and hybrid ASR systems for dysarthric speech. To fully exploit DNN-based acoustic modelling, data augmentation was successfully applied based on speed and tempo perturbation in the signal domain for dysarthric speech recognition [16].

In this paper, we first investigate the use of transfer learning (cf. [17]) to adapt DNN models towards specific target speakers in personalised dysarthric speech recognition. This is motivated by the observation that it is common to have access to a large amount of data from typical speakers (out-of-domain), but a much smaller amount of data from dysarthric speakers (in-domain), and in effect, that out-of-domain data is not entirely unrelated to the in-domain dysarthric data due to the shared lexical knowledge. Transfer learning is capable of transferring knowledge from one (source) domain to another related (target) domain to avoid re-building a new ASR system from scratch, which is desirable for dysarthric speech recognition, as i) the small amount of target data from a specific dysarthric speaker could easily lead to over-fitting during speaker-dependent training; ii) data combination scheme [18] like multi-condition or multi-style training may not always be feasible to use due to the probable large bias between the source domain and the target domain data. Actually, transfer learning has already been successfully applied to speaker adaption using DNNs in ASR system (cf. [19, 20]), however, to our best knowledge, this work is the first study to apply transfer learning to dysarthric speech recognition. DNN weight adaptation [21] is employed to form the transfer learning baseline, and its effectiveness in terms of the transferred layers will be compared to the data combination scheme with speaker-independent and dependent scenarios evaluated using the UASpeech corpus [11].

Next, to further improve the efficiency of transfer learning for personalised dysarthric speech recognition, it is crucial to have sufficient *good* training data to force the senone distribution (DNN posterior probability) from the source domain data to be close to that from the target domain data. To this end, we propose a novel data selection strategy to actively pick up the potentially *good* data from available source domain data to add to the training data of a particular dysarthric target speaker. The approach is motivated by analysis of the entropy of the posterior probabilities of a specific speaker-dependent DNN model. It is observed that across speech portions, these entropies follow a Gaussian distribution and could then serve as a selector of *good* data for improved transfer learning through data combination in a re-training or via an incremental learning chain [22] to move towards the target domain. Further, compared to the speaker-based data selection in our earlier work [6]

This research has been supported by the DeepArt Project sponsored by Google, as well as been partly supported under the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Training Network for PAthological Speech processing; Grant Agreement No. 766287).

which showed that data from other speakers with similar severity of dysarthria are not necessarily guaranteed to improve ASR performance due to very individual speech characteristics of different speaker with dysarthria, the proposed data selection performing in utterance-based mode provides dedicated selection on source domain data to avoid negative transfer.

In the remainder of this paper, the transfer learning framework is briefly introduced in Section 2, together with the UASpeech database and its ASR setup. The proposed data selection method is introduced in Section 3, and experimental results will be presented in Section 4 before Section 5 concludes the paper.

2. SYSTEM OVERVIEW

Fig. 1 depicts the processing chain of transfer learning in terms of data and model interaction for dysarthric speech recognition. The limited target domain data is from a target speaker with specific severity of dysarthria, while the source domain data is usually comprised of typical speech in transfer learning baseline, which generates a personalised target domain DNN model for recognition. The conventional data combination scheme and speaker-dependent training under common ASR setup are used to test the effectiveness of transfer learning. When the proposed data selection is performed on source domain data pool (usually data from other dysarthric speakers), transfer learning will be processed again after the selected data has been added to target domain data (motivated by active learning [23]) for a re-training, or be done via incremental learning [22] (2-step transfer learning with an intermediate domain).

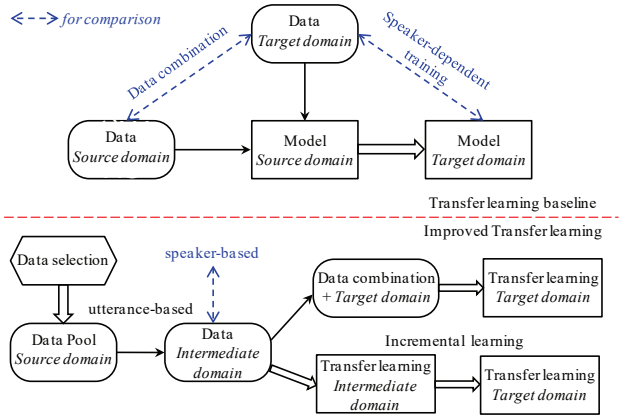

Fig. 1. Processing chain of transfer learning baseline and its improved version with the proposed data selection scheme.

2.1. Data Description

The UASpeech corpus [11] is employed to construct the source and the target data. It consists of data from 15 dysarthric speakers with cerebral palsy and 13 control (typical) speakers. There are 3 blocks of words for each speaker, and following previous published work (e.g., [4]), CTL (typical) and DYS (dysarthric) datasets are divided into training using blocks 1 and 3, and test data with block 2. DYS speakers were grouped in four severity levels based on a subjective estimate of perceptual speech intelligibility ratings (cf. [11]), namely *Severe* (speaker label as 'M04', 'F03', 'M12', 'M01'), *Moderate-Severe* ('M07', 'F02', 'M16'), *Moderate* ('M05', 'M11', 'F04') and *Mild* ('M09', 'M14', 'M10', 'M08', 'F05', cf. Fig. 4). Three baseline acoustic models are constructed as follows:

- CTL: the training data only contains the typical speech from 13 control speakers (total duration of 22.7 hours), which can be considered as a general trained model that might be widely available in public. CTL will serve as source domain model in transfer learning baseline in the following experiments;
- Speaker-independent (SI): the training data contains the other 14 DYS speakers (with duration of approximated 3 hours for each speaker) except for the particular target dysarthric speaker in the test stage. Hence, SI model is specific for each dysarthric speaker, i.e., personalised. Due to the dysarthria similarity to some extent, the SI training data will serve as source domain data pool for data selection;
- Speaker-dependent (SD): the training data is only comprised of the data from the target DYS speaker (target domain), split into training and test set. SD model is also personalised.

2.2. ASR Setup

The hybrid DNN-HMM ASR training is used, and the alignment for DNN senones (context-dependent phonemic states) is provided by an auxiliary GMM-HMM training, where 13-dimensional MFCCs incorporating a spliced context window of length 9 frames are used, and these are subsequently transformed into a 40-dimensional vector via linear discriminant analysis and maximum likelihood linear transform. In addition, speaker adaptive training is employed based on feature-space maximum likelihood linear regression (cf. [24, 25]). A uniform language model is generated based on the transcriptions of speech files, as well as a word grammar network containing a silence model followed by a single word, denoted as '$<$ sil $>$ word'.

The factored form of time delay neural networks (TDNN-F) [26] incorporating convolutional neural networks (CNNs) is used as state-of-the-art DNN architecture. It contains 6 CNN layers at the bottom, fed with 40-dimensional log-mel-spectrogram features, and 9 following TDNN-F layers, as well as one linear layer before the output layer, trained with lattice-free maximum mutual information criterion [27]. Note that this linear layer is similar to linear hidden network (LHN) [28], which instead was added as a new layer for speaker adaption. Further, the linear bottleneck of each TDNN-F layer [26] allows for additional transfer beside the sole linear layer. The learning rates for training with 4 epochs are chosen initially from 0.001 ending at 0.0002 for CTL and SI, and from 0.0005 to 0.0001 for SD (with smaller amount of data), respectively. All setups use 3-fold speed perturbation with SoX resampling algorithm [29]. For more detailed experimental setup, the reader is directed to our released Kaldi scripts[1].

3. DATA SELECTION

Data selection aims to actively choose the samples from an available data pool that could assist the target domain data to further improve the transferability of transfer learning. It is hypothesised that the selected data share a similar distribution to the target domain data in terms of the DNN senone distribution. Entropy analysis of the DNN posterior probability has been shown to be capable of providing a strong correlation to the final recognition accuracy but without the complex decoding process (cf. [30]). The posterior probability $P(s,t)$ with s,t as (monophone) state and frame index, respectively, is calculated by a DNN forward-pass and an accumulation that maps

[1]We have released our Kaldi scripts for this paper's experiments in `https://github.com/ffxiong/uaspeech/s5_transfer`

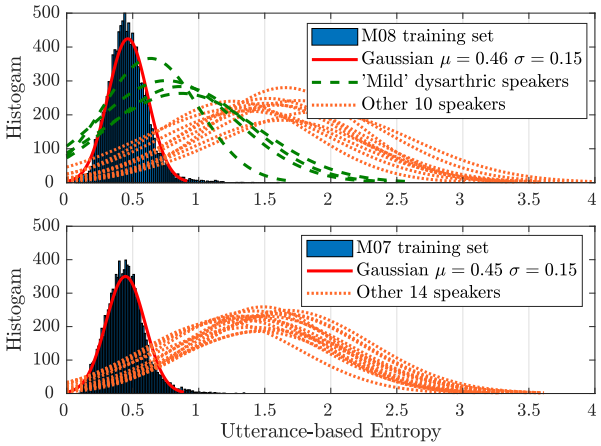

Fig. 2. Histogram (fitting to a Gaussian distribution) of utterance-based entropy of training sets from all 15 DYS speakers w.r.t. SD model of two specific dysarthric speakers, 'M08' and 'M07' (cf. Section 2.1) for example.

context-dependent senones to monophones, and the entropy $\mathcal{E}(t)$ of $P(s,t)$ is determined by

$$\mathcal{E}(t) = -\sum_{s}(P(s,t)\cdot \log_2 P(s,t))\,. \tag{1}$$

Utterance-based entropy $\overline{\mathcal{E}(t)}$ is calculated as the average value over frames t, and the silence portion is omitted by introducing a threshold of 0.05 due to the emphasis of the entropy w.r.t. speech phones.

It is observed that utterance-based entropy of the target data follows a Gaussian distribution when its SD model is used, as clearly illustrated in Fig. 2, where two dysarthric speakers 'M07' (*Moderate-Severe*) and 'M08' (*Mild*) are presented for illustration. Model-based selection can be straightforwardly derived when the mean μ and the standard deviation σ are determined based on the available target data. It is worthwhile noting that μ (0.44 to 0.46) and σ (0.14 to 0.15) behave almost consistent for all SD models with different speakers, indicating that one Gaussian distribution with $\mu = 0.45$ and $\sigma = 0.15$ is sufficient to model the entropy of DNN posterior probability from target domain data. Thereafter, the utterances with entropy values located inside the Gaussian distribution will be selected as potential *good* data for improved transfer learning process (cf. Fig. 1), represented as

$$|(\overline{\mathcal{E}(t)} - \mu)/\sigma| < x\,, \tag{2}$$

where x derives from x-sigma rule $(\mu \pm x\sigma)$ denoting different ranges of Gaussian distribution.

4. RESULTS AND DISCUSSION

4.1. Effect of transfer learning

Motivated by the findings in [31] that the first layers of neural network usually learn general features, while last layers transit features to be specific to a particular task. It is therefore of interest to analyse the impact of individual neural network layer during transfer learning via weight adaptation. From preliminary experiments, compared to the source model which was trained with 4 epochs, 1 epoch during transfer learning is sufficient to avoid an over-fitting. With the small amount of target data, the learning rate in transfer learning is best reduced by half, and it is also suggested that the untransferred

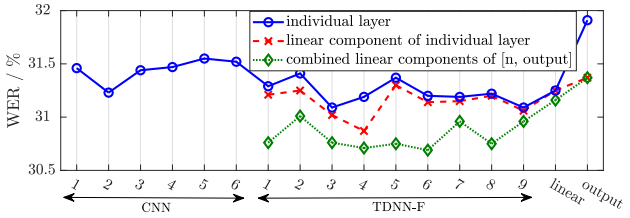

Fig. 3. Impact of the transferred layers in terms of individual layer, only linear component of individual layer (not in CNN layers), and the combined linear components from layer n gradually to the output layer (denoted as *[n, output]*).

layers still need to be re-trained with an even smaller learning rate (quarter) rather than being absolutely frozen (learning rate equals 0).

As shown in Fig. 3, it seems that hidden layers in TDNN-F layers are more transferable than CNN layers, particularly than the output layer. This is probably because CNN layers act more like feature extraction layer and the output layer is forced to be updated using the limited target data with a sparse distribution w.r.t. the large amount of senone-based states, easily leading to negative transfer [17]. Performance can be improved when only the linear component in each hidden TDNN-F layer is transferred, indicating that neural network weight adaptation is mainly accomplished via linear transfer, which performs similar to LHN speaker adaptation (cf. Section 2.2). Further, it is preferable to gradually add the transferred layers starting from the output layer towards the input layer, and it is found that it could be a good choice to transfer all the linear components in neural networks.

With the fine-tuned neural network parameters for transfer learning, the performance comparison between different models, in terms of the DYS test data from all 15 dysarthric speakers, is summarized in Table 1. In general, without target SD data for speakers with moderate and severer dysarthria, ASR systems would not achieve acceptable performance. SI outperforms CTL except for *Mild* group, indicating that data from other DYS speakers usually improve the recognition accuracy for the DYS speakers with moderate or severer dysarthria. On the other hand, a weak CTL model probably exhibits more transferability than SI model, resulting in the best overall performance. It also shows that transfer learning outperforms data combination, except for *Severe* group that might be too dissimilar to source model so that a brute-force transfer might be not the best option. Note that the best overall result in Table 1 is better than the result 37.5% in [32] or 34.8% in [13], and is comparable to the result 30.6% reported in [15].

Table 1. Averaged word error rates (WERs) for 4 DYS Groups in terms of different acoustic models. Data combination (denoted as '+') is done by combining source domain data (CTL/SI) and target domain data (SD) for a re-training from scratch. Transfer learning (denoted as '→') is achieved using target domain data (SD) to re-train CTL or SI model.

Systems	Severe	Mod.-Severe	Moderate	Mild	Overall
CTL	97.21	78.49	56.35	19.26	56.80
SI	90.75	71.27	51.86	32.40	57.66
SD	70.94	33.72	31.43	14.60	34.79
SD + CTL	67.14	34.43	25.68	13.31	32.42
SD + SI	63.02	30.90	28.15	18.90	33.29
SD→CTL	68.24	33.15	22.84	10.35	**30.76**
SD→SI	65.68	36.80	27.15	17.18	34.28

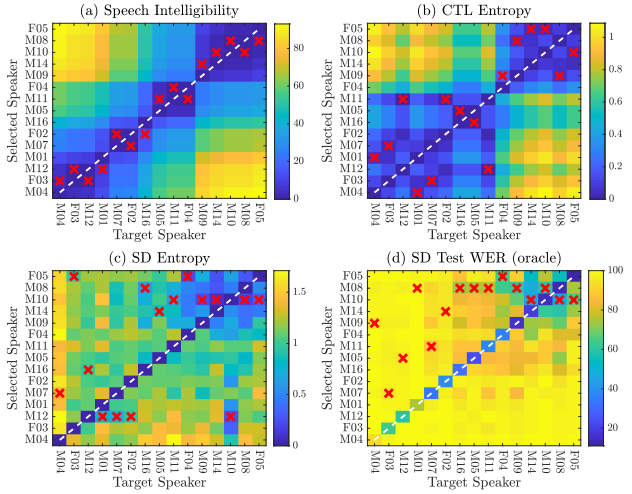

Fig. 4. Confusion matrix for 15 DYS speakers in terms of (a) mutual absolute differences of perceptual speech intelligibility; (b) mutual differences of speaker-based entropy using CTL model; (c) mutual difference of speaker-based entropy using individual SD model; (d) WERs using individual SD model with test data (oracle scenario).

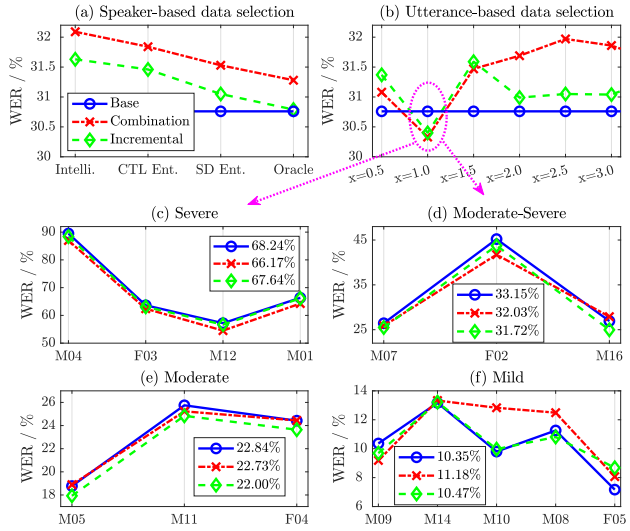

Fig. 5. Data selection for improved transfer learning in speaker-base and utterance-based modes. Base system is SD→CTL with best overall result in Table 1.

4.2. Effect of data selection

To investigate how to make good use of the available transcribed data from other dysarthric speakers for one specific target speaker, data selection is applied, both in terms of selecting data from a whole speaker (speaker-based) and from individual utterances (utterance-based). Intuitively, data from speakers with similar dysarthria severity could mutually benefit from each other in transfer learning, however it has been shown that this is not always the case [6]. We investigate a number of similarity measures: perceptual speech intelligibility scores, the speaker-based entropy value averaged over all the training utterances using CTL model and individual SD model, as well as an oracle case where WERs from the test data are used.

As seen on the plots of between-speaker similarity in Fig. 4, a higher similarity is seen between speakers with similar severity levels. In the plots, the speakers are ordered according to intelligibility [11], and hence high similarity scores concentrate along the diagonal line (the actual diagonal being the self-score of the target speaker). This is seen with respect to speech intelligibility and speaker-based entropy with the CTL model, because the underlying principle of these two estimation methods is based on the typical speech data. On the other hand, when the SD model is used for an individual dysarthric speaker, the confusion matrix becomes dispersive, particularly for *Moderate* to *Severe* speakers, indicating a large inter-speaker variability. This phenomenon was also observed in Fig. 2 where some similarity with other speakers is seen for the *Mild* speaker 'M08', but not for *Moderate-Severe* speaker 'M07'. The data from the speaker with the nearest similarity is selected for improved transfer learning, as shown in the lower panel of Fig. 1.

Fig. 5 (a) shows the results of adding data based on selecting from a whole speaker. It is seen that selection based on speech intelligibility and CTL model cannot provide further improvement compared to the base transferred model, indicating that such similarity measures solely using typical speech data is not sufficient for speakers with varying degrees of dysarthria. No further improvement obtained by the oracle case indicates that speaker-based data selection is not effective nor sufficient, since most of the utterance-based entropy is still out of the target distribution (cf. Fig. 2). Further, incremental learning consistently outperforms data combination.

For fair comparison, the amount of selected data using the utterance-based data selection is capped so that the added data is of a similar size to the amount available in speaker-based data selection (training data amount from 1x DYS speaker, rather 14x used in SI system in Table 1) for target domain data. We also choose different ranges of Gaussian distribution according to $x = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ in Equation (2). Fig. 5 (b) shows that data with utterance-based entropy located in $\mu \pm \sigma$ ($x = 1.0$) contributes the most to the improved transfer learning both with respect to data combination and incremental learning. Furthermore, Fig. 5 (c)-(f) depict the WER of each dysarthric speaker grouped in four groups to pinpoint the individual advantage. For speakers with severe dysarthria, data combination outperforms incremental learning, resulting in an averaged 2.1% absolute WER reduction compared to the base transferred model. This also indicates that it might be difficult to generate an intermediate domain being close to target domain for speech variability from very severe dysarthria. On the other hand, for *Moderate-Severe* and *Moderate* groups, incremental learning is superior in general. However, no further transfer gain is observed for *Mild* group as the original CTL model is close enough to the target domain.

5. CONCLUSIONS

This paper investigated the use of transfer learning applied for improving personalised dysarthric speech recognition. Consistent transfer gain can be observed when source domain data is from typical speech for speakers with various severity of dysarthria, and transfer learning performs more effectively than conventional data combination and speaker-dependent training. To further optimise the use of available data from other dysarthric speakers, a data selection scheme has been proposed based on entropy distribution of DNN posterior probabilities. We found that speaker-based data selection via similarity measure easily raises negative transfer, and it is preferable to conduct utterance-based data selection within 1-sigma range of a determined Gaussian distribution to exploit further transfer gain, particularly for test scenarios with moderate to severe dysarthria. Further, incremental learning outperforms data combination in general, except for the case with very severe dysarthria.

6. REFERENCES

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, no. 3, pp. 462–496, 1969.

[2] B. Blaney and J. Wilson, "Acoustic variability in dysarthria and computer speech recognition," *Clinical Linguistics & Phonetics*, vol. 14, no. 4, pp. 307–327, 2000.

[3] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[4] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proceedings of Interspeech*, Portland, Oregon, USA, Sept. 2012, pp. 1776–1779.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling: Recognition of disordered speech with sparse data," in *IEEE Spoken Language Technology Workshop*, South Lake Tahoe, USA, Dec. 2014, pp. 254–259.

[7] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[8] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019.

[9] J. R. Deller, M. S. Liu, L. J. Ferrier, and P. Robichaud, "The Whitaker database of dysarthric (cerebral palsy) speech," *The Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[10] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, Oct. 1996, pp. 1962–1965.

[11] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of Interspeech*, Brisbane, Australia, Sept. 2008, pp. 1741–1744.

[12] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2011.

[13] H. Christensen, M. B. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Proceedings of Interspeech*, Lyon, France, Aug. 2013, pp. 3642–3645.

[14] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *IEEE International Conference on Signal Processing (ICSP)*, Hangzhou, China, Oct. 2014, pp. 505–509.

[15] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Y. LAM, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the CUHK dysarthric speech recognition system for the UASpeech corpus," in *Proceedings of Interspeech*, Hyderabad, India, Sept. 2018, pp. 2938–2942.

[16] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proceedings of Interspeech*, Hyderabad, India, Sept. 2018, pp. 471–475.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[18] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[19] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 7304–7308.

[20] Z. Huang, S. M. Siniscalchi, and C.-H. Lee, "A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition," *Neurocomputing*, vol. 218, pp. 448–459, 2016.

[21] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, Dec. 2017, pp. 279–286.

[22] S. Thrun and L. Pratt, Eds., *Learning to Learn*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[23] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[24] 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 *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HI, USA, July 2011.

[25] D. Povey and K. Yao, "A basis method for robust estimation of constrained MLLR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4460–4463.

[26] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of Interspeech*, Hyderabad, India, Sept. 2018, pp. 3743–3747.

[27] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proceedings of Interspeech*, San Francisco, USA, Sept. 2016, pp. 2751–2755.

[28] 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. 10-11, pp. 827–835, 2007.

[29] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, Sept. 2015, pp. 3586–3589.

[30] F. Xiong, J. Zhang, B. T. Meyer, H. Christensen, and J. Barker, "Channel selection using neural network posterior probability for speech recognition with distributed microphone arrays in everyday environments," in *CHiME Workshop on Speech Processing in Everyday Environments*, Hyderabad, India, Sept. 2018, pp. 19–24.

[31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 1–9.

[32] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proceedings of workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Dresden, Germany, Sept. 2015.