Dysarthric speech recognition is a challenging task as dysarthric data is limited and its acoustics deviate significantly from normal speech. Model-based speaker adaptation is a promising method by using the limited dysarthric speech to fine-tune a base model that has been pre-trained from large amounts of normal speech to obtain speaker-dependent models. However, statistic distribution mismatches between the normal and dysarthric speech data limit the adaptation performance of the base model. To address this problem, we propose to re-initialize the base model via meta-learning to obtain a better model initialization. Specifically, we focus on end-to-end models and extend the model-agnostic meta learning (MAML) and Reptile algorithms to meta update the base model by repeatedly simulating adaptation to different dysarthric speakers. As a result, the re-initialized model acquires dysarthric speech knowledge and learns how to perform fast adaptation to unseen dysarthric speakers with improved performance. Experimental results on UASpeech dataset show that the best model with proposed methods achieves 54.2% and 7.6% relative word error rate reduction compared with the base model, respectively, and it is comparable with the model without finetuning and the model directly fine-tuned.

The main contributions of this paper are: (1) We propose a new strategy for dysarthric speech recognition via meta-learning based model re-initialization which combines model re-initialization and meta-learning to address the problem of dysarthric speech recognition; (2) We explore the use of end-to-end ASR models pre-trained on large amounts of normal speech data and further adapted into speaker-dependent models which produced encouraging results.

1. Introduction

Dysarthria denotes a set of speech disorders related with neurological conditions and diseases such as traumatic brain injury, stroke, Parkinson’s disease or amyotrophic lateral sclerosis, which cause disturbances in muscular control over the speech production [1]. As a result, dysarthric speech exhibits very different acoustic characteristics from normal speech, and automatic speech recognition (ASR) systems trained on normal speech suffer from marked performance degradation when applied to dysarthric speech. The problem of modelling is exacerbated by possibly large inter- and intra-speaker variations inherent in dysarthric speech, and scarcity of dysarthric training data further increases the challenges in the development of robust ASR systems for dysarthric speakers.

To address these problems, recent research efforts [2] have explored the use of end-to-end ASR models pre-trained on large amounts of normal speech data and further adapted into speaker-dependent models which produced encouraging results.

Abstract

Dysarthric speech recognition is a challenging task as dysarthric data is limited and its acoustics deviate significantly from normal speech. Model-based speaker adaptation is a promising method by using the limited dysarthric speech to fine-tune a base model that has been pre-trained from large amounts of normal speech to obtain speaker-dependent models. However, statistic distribution mismatches between the normal and dysarthric speech data limit the adaptation performance of the base model. To address this problem, we propose to re-initialize the base model via meta-learning to obtain a better model initialization. Specifically, we focus on end-to-end models and extend the model-agnostic meta learning (MAML) and Reptile algorithms to meta update the base model by repeatedly simulating adaptation to different dysarthric speakers. As a result, the re-initialized model acquires dysarthric speech knowledge and learns how to perform fast adaptation to unseen dysarthric speakers with improved performance. Experimental results on UASpeech dataset show that the best model with proposed methods achieves 54.2% and 7.6% relative word error rate reduction compared with the base model, respectively, and it is comparable with the model without finetuning and the model directly fine-tuned.

1. Introduction

Dysarthria denotes a set of speech disorders related with neurological conditions and diseases such as traumatic brain injury, stroke, Parkinson’s disease or amyotrophic lateral sclerosis, which cause disturbances in muscular control over the speech production [1]. As a result, dysarthric speech exhibits very different acoustic characteristics from normal speech, and automatic speech recognition (ASR) systems trained on normal speech suffer from marked performance degradation when applied to dysarthric speech. The problem of modelling is exacerbated by possibly large inter- and intra-speaker variations inherent in dysarthric speech, and scarcity of dysarthric training data further increases the challenges in the development of robust ASR systems for dysarthric speakers.

To address these problems, recent research efforts [2] have explored the use of end-to-end ASR models pre-trained on large amounts of normal speech data and further adapted into speaker-dependent models which produced encouraging results.
2. Related work

2.1. Dysarthric speech recognition

Model adaptation techniques have been successfully applied in dysarthric speech recognition [10]. The maximum a-posteriori (MAP) adaptation in traditional GMM-HMM enables intra-speaker variability modelling to improve ASR performance [11] and enhance the speaker adaptive training (SAT) [12]. Various training and adaptation techniques have also been proved effective in hybrid DNN-HMM systems [13]. Besides, end-to-end architectures have been introduced [2, 14] and reported to significantly improve the accuracy of ASR by adapting the pre-trained base model [2]. In this paper, we focus on end-to-end models to validate the effectiveness of our proposed methods.

2.2. Meta-learning

Meta-learning aims to teach a model how to learn various tasks, so that it can solve new tasks with few samples. Hence, meta-learning is suitable for dysarthric speech recognition, especially due to data scarcity in dysarthric speech for speaker adaptation. Recent work has proposed different meta-learning methods [4-8], among which we employ MAML and Reptile for model re-initialization due to their simplicity and effectiveness [7, 8].

Meta-learning has already been used in speech-related fields, such as speaker recognition [15], low-resource and code-switched ASR [16, 17], speaker adaptation for normal and accented speech recognition [18-20]. There is still much room for investigating speaker adaptation for dysarthric speech recognition. However, meta-training an end-to-end model from scratch requires large amounts of data [18-20], and is generally not achievable for dysarthric speech data which is difficult to obtain. Therefore, this paper proposes to meta-train a base model and further explore the use of Reptile as a novelty.

3. Proposed methods

In this section, the dysarthric speaker adaptation problem is first formulated. Then two meta-learning methods, i.e., MAML and Reptile, are proposed for re-initialization of a pre-trained base model to obtain a good initialization with improved adaptation performance to unseen dysarthric speakers. In addition, we propose an effective way by using per-task batch normalization running statistics to stabilize meta-learning.

3.1. Problem formulation

Assume the number of dysarthric speakers is $I$, and the paired data for $i$th speaker consists of the adaptation data $(X_{i,a}, Y_{i,a})$ and test data $(X_{i,a}, Y_{i,a})$, where $X$ is the speech feature, and $Y$ is the corresponding character label. Given an end-to-end base model $f(X; \theta)$ with the parameters $\theta$ pre-trained with large amounts of normal speech [2], the model adaptation for $i$th speaker aims to fine-tune $\theta$ over the adaptation data $(X_{i,a}, Y_{i,a})$

$$adapt\left(f, \theta, (X_{i,a}, Y_{i,a})\right) \rightarrow \theta_i,$$  

(1)

so that the adapted model $\theta_i$ can achieve better performance on the test data $(X_{i,a}, Y_{i,a})$. The performance can be measured by the loss function $L(Y_{i,a}, f(X_{i,a}; \theta))$, which is the character Connectionist Temporal Classification (CTC) loss or Cross-Entropy (CE) loss in the context of end-to-end ASR. Due to the different acoustic characteristics between the normal speech and dysarthric speech, the performance of adaptation to dysarthric speakers by using the base model is limited. To alleviate this issue, we propose to employ MAML and Reptile to re-initialize the base model $\theta$ to obtain a better initialization $\theta^*$ with powerful generalization capacity to unseen dysarthric speakers. The detailed MAML/Reptile based model re-initialization is outlined in Algorithm 1.

3.2. MAML based model re-initialization

To apply MAML, adaptation to each dysarthric speaker is treated as a task. First, let the meta-model $\theta^*$ be initialized with the base model $\theta$. For $i$th task, i.e., adaptation to $i$th speaker, training data $(X_{i,a}, Y_{i,a})$ and validation data $(X_{i,a}, Y_{i,a})$ are sampled from the adaptation data $(X_{i,a}, Y_{i,a})$. Then the inner-loop and outer-loop optimization are respectively performed for individual tasks learning and base model meta-updating [7].

In the inner-loop optimization, for $i$th task, $\theta^*$ is initialized with $\theta^*$ and adapted over $(X_{i,a}, Y_{i,a})$ with $J$ gradient descent updates, where one gradient descent update with learning rate $\alpha$ is performed by

$$\theta^* \leftarrow \theta - \alpha \nabla_{\theta} L(Y_{i,a}, f(X_{i,a}; \theta))$$

(2)

In the outer-loop optimization, we employ the first-order MAML [7] to update the meta-model $\theta^*$ with learning rate $\eta$ by using the loss of adapted models over the validation data

$$\theta^* \leftarrow \theta^* - \eta \frac{1}{J} \sum_{i=1}^{J} \nabla_{\theta} L(Y_{i,a}, f(X_{i,a}; \theta))$$

(3)

MAML integrates the adaptation process into its framework by first simulating speaker-specific adaptation, then meta-updates the model $\theta^*$. As a result, MAML optimizes for adaptation to dysarthric speakers, which produces a good initialization that can be rapidly adapted to unseen dysarthric speakers.

3.3. Reptile based model re-initialization

Reptile differs from MAML with a simpler yet effective meta-model updating manner [8] by removing gradients calculation in the outer-loop optimization. Specifically, for $i$th task, i.e., adaptation to $i$th speaker, only training data $(X_{i,a}, Y_{i,a})$ is sampled from the adaptation data, then inner-loop optimization is performed like MAML to obtain $\theta_i$. After obtaining all adapted models $\theta_i$ $(i=1, 2, \ldots, I)$, the meta-model $\theta^*$ is updated via

| Algorithm 1 MAML/Reptile based model re-initialization |
| --- |
| **Input:** outer-loop and inner loop optimization steps $K$ and $J$, number of dysarthric speakers $I$, base model $\theta$, adaptation data $(X_{i,a}, Y_{i,a})$, learning rate $\alpha$ and $\eta$, meta-learning type $1$: $\theta^* \leftarrow \theta$
| 2: for $k \in \{1, 2, \ldots, K\}$ do
| 3: for $i \in \{1, 2, \ldots, I\}$ do
| 4: $\theta_i \leftarrow \theta$
| 5: Sample training data $(X_{i,a}, Y_{i,a})$ from $(X_{i,a}, Y_{i,a})$
| 6: for $j \in \{1, 2, \ldots, J\}$ do
| 7: Speaker adaptation via the gradient descent:
| $\theta_i \leftarrow \theta_i - \alpha \nabla_{\theta_i} L(Y_{i,a}, f(X_{i,a}; \theta_i))$
| 8: end for
| 9: end for
| 10: if meta-learning type is MAML then
| 11: Sample validation data $(X_{i,a}, Y_{i,a})$ $(i=1,2,\ldots,I)$ from $(X_{i,a}, Y_{i,a})$ and meta-update $\theta^*$:
| $\theta^* \leftarrow \theta^* - \eta \frac{1}{J} \sum_{i=1}^{J} \nabla_{\theta} L(Y_{i,a}, f(X_{i,a}; \theta_i))$
| 12: else if meta-learning type is Reptile then
| $\theta^* \leftarrow \theta^* - \eta \frac{1}{J} \sum_{i=1}^{J} (\theta_i - \theta^*)$
| 13: end if
| 14: end for
| 15: return re-initialized model $\theta^*$


16 dysarthric speakers and 13 normal speakers, where only the
Following the previous data splits [10], we utilize blocks 1 and
Mild. The dysarthric speech of each speaker has 3 blocks.

4 severity levels: Severe, Moderate Severe, Moderate and

3 as adaptation data for model re-initialization and speaker
adaptation, and block 2 as testing data.

4.1. ASR Models
Two different end-to-end architectures are used for ASR
models: (1) Listen, Attend and Spell (LAS) model [25], which
is an attention based sequence-to-sequence model that maps
acoustic features to characters. LAS contains an encoder (a 6-
layer VGG extractor followed by 5-layer BLSTM with 512
units per direction), 300-dimensional location-aware attention
[26] and a decoder that is a single-layer LSTM with 512 units.

(2) QuartzNet [27], which is an improved version of Jasper
model [28] and achieves comparable performance to Jasper but
with much less parameters. The QuartzNet contains 15 blocks
with 5 sub-blocks per block, each sub-block consists of a 1-D
separable convolution, batch normalization, ReLU and dropout.
The inputs of ASR models are 64-dimensional mel-spectrogram,
and the targets are 29-dimensional (26 English lower-case
letters, a punctuation symbol, a space and a blank token).

4.1.1. Experimental settings
Two different end-to-end architectures are used for ASR
models: (1) Listen, Attend and Spell (LAS) model [25], which
is an attention based sequence-to-sequence model that maps
acoustic features to characters. LAS contains an encoder (a 6-
layer VGG extractor followed by 5-layer BLSTM with 512
units per direction), 300-dimensional location-aware attention
[26] and a decoder that is a single-layer LSTM with 512 units.

(2) QuartzNet [27], which is an improved version of Jasper
model [28] and achieves comparable performance to Jasper but
with much less parameters. The QuartzNet contains 15 blocks
with 5 sub-blocks per block, each sub-block consists of a 1-D
separable convolution, batch normalization, ReLU and dropout.
The inputs of ASR models are 64-dimensional mel-spectrogram,
and the targets are 29-dimensional (26 English lower-case
letters, a punctuation symbol, a space and a blank token).

4.1. Dataset
UASpeech [9] and Librispeech [24] datasets are used in our
experiments. Librispeech only contains normal speech, and we
employ 960h training data for training and dev-clean data as
validation data to select best base model. UASpeech contains
16 dysarthric speakers and 13 normal speakers, where only the
dysarthric speech is used for experiments. Based on subjective
speech intelligibility ratings, dysarthric speakers are grouped
into 4 severity levels: Severe, Moderate Severe, Moderate and
Mild. The dysarthric speech of each speaker has 3 blocks.

3 as adaptation data for model re-initialization and speaker
adaptation, and block 2 as testing data.

4.1. ASR Models
Two different end-to-end architectures are used for ASR
models: (1) Listen, Attend and Spell (LAS) model [25], which
is an attention based sequence-to-sequence model that maps
acoustic features to characters. LAS contains an encoder (a 6-
layer VGG extractor followed by 5-layer BLSTM with 512
units per direction), 300-dimensional location-aware attention
[26] and a decoder that is a single-layer LSTM with 512 units.

(2) QuartzNet [27], which is an improved version of Jasper
model [28] and achieves comparable performance to Jasper but
with much less parameters. The QuartzNet contains 15 blocks
with 5 sub-blocks per block, each sub-block consists of a 1-D
separable convolution, batch normalization, ReLU and dropout.
The inputs of ASR models are 64-dimensional mel-spectrogram,
and the targets are 29-dimensional (26 English lower-case
letters, a punctuation symbol, a space and a blank token).

4.1. Training and evaluation
For the base models training, LAS is trained with the combined
CTC and CE loss by using Adadelta optimizer [29] with
learning rate of 1 and batch size 16, QuartzNet is trained with
CTC loss by using Novagrad optimizer [28] with learning rate
of 0.05 and batch size 16. The final base LAS and QuartzNet
achieves 10.5% and 6.8% word error rate (WER) on dev-clean
split of Librispeech with greedy decoding.

For the model re-initialization, leave-one-subject-out cross
validation is adopted, i.e., all dysarthric speakers except for the
target speaker are used for MAML/Reptile based model re-
initialization, while the speech of the target speaker is used for
adaptation and testing. In the inner-loop optimization, updating
steps J for both LAS and QuartzNet are 5 with learning rate \( \eta \)
of 1e-4. In the outer-loop optimization, the updating steps K for
both LAS and QuartzNet are 150, while the learning rate \( \eta \) is
1 and 3e-4 for LAS and QuartzNet with the same optimizers as
the base models training, respectively. For comparison with
Reptile, joint training based model re-initialization with adap-
tation data of all dysarthric speakers is implemented, where the
same learning rate and optimizers as the outer-loop optimi-
ization of MAML are used, the model is re-initialized for only 1
epoch, as no improvements are observed with more epochs.

For the speaker adaptation, the total adaptation epochs for
each speaker is 10. For LAS, the learning rate of Adadelta
remains unchanged as 1. For QuartzNet, the initial learning rate
of Novagrad is 3e-4, then decayed by a half every epoch when
the adaptation epoch is larger than 5.

For the testing, we adopt weighted finite-state transducers
based decoding [30] by incorporating the lexicon and language
models into CTC/CE decoding with beam search size as 5,
where the lexicon and uniform language models are generated
based on the transcriptions of UASpeech following previously
published work [10].

To show the effectiveness of the proposed methods, we also
demonstrate the results achieved by a hybrid DNN-HMM
system [13], which is one of the state-of-the-art dysarthric
speech recognition systems on UASpeech dataset.

4.2. Experimental results

4.2.1. Performance of different models and strategies
Table 1 shows the performance comparison of different
combinations of models and strategies. First, we can observe
that compared with base models, the end-to-end ASR benefits
from the speaker adaptation, and dramatic improvements can be

\[
\mathbf{\theta^*} \leftarrow \mathbf{\theta} - \eta \frac{1}{L} \sum_{i=1}^{L} (\mathbf{\theta^*} - \mathbf{\theta})
\]
Figure 2. WER (%) versus ratios of training data for QuartzNet

Table 1. WER (%) of different combinations of models and strategies, strategies are followed by the base models: ‘+Adapt’ means perform speaker adaptation based on base models, ‘+M & Adapt’ means first perform M-based model re-initialization and then speaker adaptation based on the re-initialized model.

| Systems          | Severity levels | Overall |
|------------------|-----------------|---------|
|                  | Severe | Mod.Severe | Moderate | Mild |         |
| DNN-HMM [13]     | 67.4   | 31.4       | 23.6     | 12.2 | 30.6     |
| Base LAS         | 98.8   | 90.7       | 82.9     | 47.2 | 76.4     |
| +Adapt           | 73.7   | 43.2       | 34.9     | 13.0 | 37.9     |
| +Joint & Adapt   | 67.8   | 41.9       | 36.2     | 19.2 | 38.6     |
| +MAML & Adapt    | 70.0   | 38.2       | 31.1     | 13.2 | 35.2     |
| +Reptile & Adapt | 68.7   | 39.0       | 32.5     | 12.2 | 35.0     |
| Base QuartzNet   | 98.0   | 86.4       | 71.3     | 28.8 | 66.6     |
| +Adapt           | 76.4   | 36.1       | 24.3     | 7.1  | 33.0     |
| +Joint & Adapt   | 65.1   | 35.7       | 32.5     | 14.6 | 36.3     |
| +MAML & Adapt    | 70.0   | 33.3       | 24.5     | 7.2  | 30.6     |
| +Reptile & Adapt | 69.3   | 33.7       | 24.1     | 7.3  | 30.5     |

Figure 2. WER (%) versus ratios of training data for QuartzNet

achieved especially for speakers with Severe/Mod.Severe/Moderate dysarthria, which shows the significance of speaker adaptation for achieving good ASR performance. Second, QuartzNet outperforms LAS under the same strategy condition, where attribute it to that the base model with lower WER has more powerful capacity to generalize to dysarthric speakers and facilitates the adaptation process to obtain better fine-tuned models. Third, the proposed MAML/Reptile based model re-initialization improves the adaptation performance with lower overall WER, where the best system (QuartzNet+Reptile) achieves 54.2% and 7.6% relative WER reduction compared with the base model without finetuning and direct adaptation on the base model, respectively, which shows that meta-learning based model re-initialization can effectively acquire speech knowledge across dysarthric speakers, which enhances the generalization capacity of the re-initialized model to unseen dysarthric speakers. Besides, joint training based model re-initialization can only improve the adaptation performance for speakers with Severe or Mod.Severe dysarthria, as the joint training aims to find a global minimizer of CTC/CE loss of all dysarthric speakers, and hence focuses more on minimizing the larger losses of speakers with more severe dysarthria, leading to better initializations for those speakers. Finally, we can observe that most end-to-end ASR systems can achieve comparable performance with the hybrid DNN-HMM system, which can be even outperformed by the base QuartzNet with Reptile, this further demonstrates the effectiveness of the proposed MAML/Reptile based model re-initialization to achieve good dysarthric speech recognition performance.

4.2.2. Impact of training data size

To verify the advantages of the proposed methods for speaker adaptation with limited data, we vary the amount of training data, i.e., different ratios of total adaptation data are used for finetuning. WER results based on QuartzNet are reported in Figure 2, we can observe that the WER of all methods decreases with the increased training data as expected, the proposed MAML and Reptile based methods can achieve fast adaptation and lower WER compared with 'Base+Adapt' and 'Base+Joint&Adapt', significant WER reduction can be achieved by the proposed methods when the ratio is small with fewer training data, this shows the MAML/Reptile based re-initialization enables the model to effectively acquire cross-dysarthric-speaker knowledge, which facilitates the speaker adaptation to improve the ASR performance with limited data.

4.2.3. Learning curves

In addition, we demonstrate the learning curves of different methods in Figure 3, i.e., the WER versus training epochs. We can observe that the performance of ‘Base+Joint&Adapt’ nearly saturates in the first several training epochs, as the data for joint training based model re-initialization is also used for finetuning, fast convergence can be achieved for speaker adaptation. Contrarily, without using the data of target dysarthric speaker for model re-initialization, MAML and Reptile based methods can achieve fast adaptation and lower WER compared with ‘Base+Adapt’ and ‘Base+Joint&Adapt’, which validates the effectiveness of the proposed methods for obtaining an appropriate model initialization with powerful generalization capacity to unseen dysarthric speech.

5. Conclusions

This paper focuses on applying meta-learning to find a good end-to-end model initialization to dysarthric speech recognition. Given a base model pre-trained from large scale normal speech data, MAML and Reptile are proposed to meta-update the base model by incorporating across-dysarthric-speakers knowledge into the re-initialized model, which allows rapid speaker adaptation to unseen dysarthric speakers. Speaker adaptation results on UASpeech dataset show that compared with the base LAS and QuartzNet models, the best re-initialized models can achieve 7.7% and 7.6% relative WER reduction, respectively. Our future work focuses on extending the proposed methods for hybrid system, e.g., DNN-HMM, to achieve better performance.

6. Acknowledgements

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14208817).
7. References

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, pp. 246-269, 1969.

[2] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, *et al.*, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Interspeech*, 2019, pp. 784-788.

[3] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceeding of Fourth International Conference on Spoken Language Processing*. ICSLP’96, 1996, pp. 1137-1140.

[4] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842-1850.

[5] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, *et al.*, "Learning to learn by gradient descent by gradient descent," in *Advances in neural information processing systems*, 2016, pp. 3981-3989.

[6] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations*, 2017.

[7] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017, pp. 1126-1135.

[8] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, *et al.*, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[10] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[11] H. V. Sharma and M. Hasegawa-Johnson, "State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition," in *Proceedings of the NAACL HLT workshop on speech and language processing for assistive technologies*, 2010, pp. 72-79.

[12] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proceedings of SLPAT: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 65-71.

[13] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, *et al.*, "Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus," in *Interspeech*, 2018, pp. 2938-2942.

[14] Y. Takashima, T. Takiguchi, and Y. Arik, "End-to-end dysarthric speech recognition using multiple databases," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6395-6399.

[15] S. M. Kye, Y. Jung, H. B. Lee, S. J. Hwang, and H. Kim, "Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs," *arXiv preprint arXiv:2004.02863*, 2020.

[16] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta Learning for End-to-End Low-Resource Speech Recognition," *arXiv preprint arXiv:1910.12094*, 2019.

[17] G. I. Winata, S. Cahaaywijaya, Z. Lin, Z. Liu, P. Xu, and P. Fung, "Meta-Transfer Learning for Code-Switched Speech Recognition," *arXiv preprint arXiv:2004.14228*, 2020.

[18] O. Klejch, J. Fainberg, and P. Bell, "Learning to adapt: a meta-learning approach for speaker adaptation," in *Interspeech*, 2018, pp. 867-871.

[19] O. Klejch, J. Fainberg, P. Bell, and S. Renals, "Speaker Adaptive Training using Model Agnostic Meta-Learning," *arXiv preprint arXiv:1910.10605*, 2019.

[20] G. I. Winata, S. Cahaaywijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, *et al.*, "Learning Fast Adaptation on Cross-Accented Speech Recognition," *arXiv preprint arXiv:2003.01901*, 2020.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448-456.

[22] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Advances in neural information processing systems*, 2017, pp. 1945-1953.

[23] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?," in *Advances in Neural Information Processing Systems*, 2018, pp. 2483-2493.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206-5210.

[25] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Advances in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960-4964.

[26] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577-585.

[27] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, *et al.*, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6124-6128.

[28] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, *et al.*, "Jasper: An end-to-end convolutional neural acoustical model," in *Interspeech*, 2019, pp. 71-75.

[29] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[30] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-End Low-Resource Speech Recognition," *arXiv preprint arXiv:2012.5701*, 2012.