Multi-accent Speech Separation with One Shot Learning

Kuan Po Huang[1*], Yuan-Kuei Wu[2*], Hung-yi Lee[3]
[123]National Taiwan University
[1]Graduate Institute of Computer Science and Information Engineering
[23]Graduate Institute of Communication Engineering
{r09922005, f07942100, hungyilee}@ntu.edu.tw [*]

Abstract

Speech separation is a problem in the field of speech processing that has been studied in full swing recently. However, there has not been much work studying a multi-accent speech separation scenario. Unseen speakers with new accents and noise aroused the domain mismatch problem which cannot be easily solved by conventional joint training methods. Thus, we applied MAML and FOMAML to tackle this problem and obtained higher average Si-SNRi values than joint training on almost all the unseen accents. This proved that these two methods do have the ability to generate well-trained parameters for adapting to speech mixtures of new speakers and accents. Furthermore, we found out that FOMAML obtains similar performance compared to MAML while saving a lot of time.

1 Introduction

Speech separation has been a well-known task to solve in the speech processing field. Many model architectures mentioned in Section 2 have been proposed and achieved high performance. This suggests that deep learning based methods are suitable for the speech separation task.

Despite having promising results, the generalizability of these models is still questionable. The performance of switching to different datasets or environments is not guaranteed. A straightforward solution is to exhaustively collect data under all kinds of environment settings and train a model with these data jointly. Although this may sound reasonable, it is difficult to always consider every situation during training. To make sure that models can be quickly adapted to mixtures spoken by new speakers with not many samples, meta-learning comes to the rescue. Meta-learning has

been widely applied on different speech tasks, especially on speech recognition mentioned in Section 2. Nonetheless, there is not much work that applied meta-learning on the speech separation task. In our previous work, (Wu et al., 2020), we first proposed to solve the speech separation problem with meta-learning. Their setting is viewing utterance mixtures of two different speakers as a meta task. These speakers have the same accents. However, we hope that a speech separation model can have the ability to adapt to mixtures with accents never seen before. Thus, besides the setting of two different speakers forming a meta task, we also added a setting that meta tasks with speakers of same accents form an accent task set. Section 4 and 5.1 describe more about the dataset and task construction procedure.

Our contributions are listed below:

- To our best knowledge, we are the first to conduct speech separation experiments on a multi-accent dataset.

- We applied meta-learning to help improve the multi-accent speech recognition task.

The remaining sections of this paper are organized as follows. In Section 2, we give a brief overview of existing works related to speech separation and meta-learning. In Section 3, we elaborate the problem formulation of speech separation in detail. In Section 4, we list out the two phases of MAML, including the meta training phase and meta testing phase. Additionally, we show how FOMAML is modified from MAML. The experimental setup, dataset, and model we used are presented in Section 5. Finally, results and conclusions are given in Section 6 and 7.

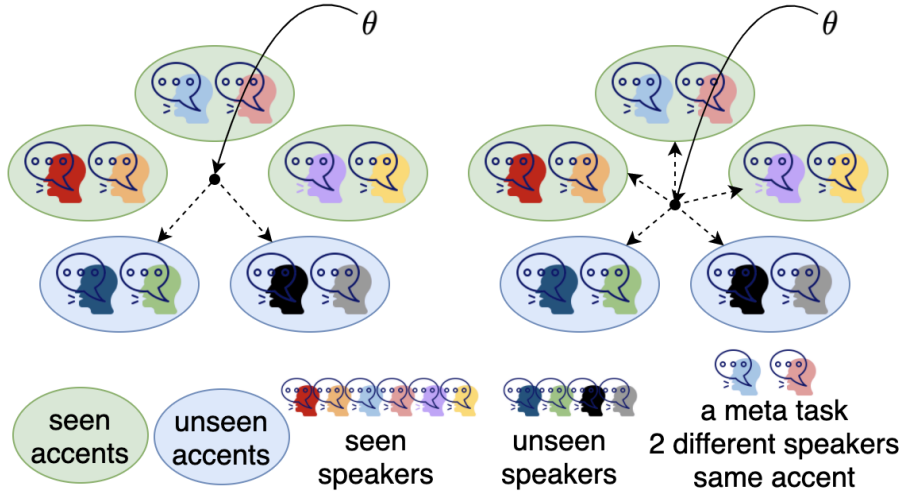[*]The two first authors made equal contributions.

Figure 1: Illustration of joint training and meta-learning for multi-accent speech separation. The oval area is the accent task sets. Each accent task set contains multiple meta tasks. The solid lines are the pretraining process, joint training on the left, and meta-learning on the right. The dashed lines represent the adaptation paths from parameters θ to the unseen accents of unseen speakers. This figure is modified from Gu et al. (2018) and our previous work Wu et al. (2020).

2 Related Work

Speech Separation End-to-end separation models have shown great success in separating speech mixtures of the WSJ0-2mix dataset designed by (Hershey et al., 2016) which is generated from the WSJ0 corpus(Paul and Baker, 1992). (Luo and Mesgarani, 2018) came up with a time-domain audio separation network (TasNet) that takes waveforms as input to alleviate the separation model from dealing with time-frequency representations. They further proposed convolutional TasNet (Luo and Mesgarani, 2019) which substitutes the LSTM layers in TasNet with convolutional layers. This overcame the problem of long temporal dependencies of LSTM and reduced the model size. Before long, they came up will the Dual-path RNN model, which used intra- and inter-blocks to capture local and global information dependencies within the speech mixtures. (Nachmani et al., 2020) utilized the idea of Dual-path RNN and added a speaker identity loss to improve performance on separating mixtures with an unknown number of speakers. (Tzinis et al., 2020) proposed to use a separator constructed with U-ConvBlocks which can not only reduce the number of layers while still having high performance but also require less computational resources and time. This helped the model to more likely be used in real-time speech separation. (Zeghidour and Grangier, 2020) integrated speaker identity information into the separating process,

and obtained state-of-the-art performance.

Meta-learning Meta-learning has recently become a trend when it comes to solving multi-task problems. This training method has been widely applied in the computer vision field, for instance, (Vinyals et al., 2016; Rusu et al., 2018; Sun et al., 2019). Meta-learning is also used in the natural language processing field. (Gu et al., 2018) used MAML (Finn et al., 2017) for low-resource neural machine translation (NMT). Moreover, in the speech processing domain, some speech-related problems are solved with meta-learning, too. (Winata et al., 2020) applied meta-transfer learning on code-switched speech recognition. (Xiao et al., 2020; Hsu et al., 2020) applied meta-learning to solve the multilingual low-resource speech recognition problem. (Winata et al., 2019) also used MAML to adapt models to unseen accents on speech recognition. (Indurthi et al., 2019) adopted meta-learning algorithms to perform speech translation on speech-transcript paired low-resource data. (Chen et al., 2021) came up with some improvements of meta-learning to help the speaker verification task.

3 Speech Separation

In this work, we perform single channel speech separation. Given a mixture

$$\mathbf{x} = \sum_{c=1}^{C} \mathbf{s}_c \qquad (1)$$

where C is the number of speakers in mixture $\mathbf{x} \in \mathbf{R}^T$ and $\mathbf{s}_c \in \mathbf{R}^T$ are the ground truth sources. For speech separation, the goal is to estimate C sources $\{\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_C\} \in \mathbf{R}^T$ such that the estimates sources are as similar as the ground truth sources. The model we used in this work is Conv-TasNet (Luo and Mesgarani, 2019). In their work, the similarity of the estimated sources and ground truth sources are measured by scale-invariant signal-to-noise ratio (Si-SNR) shown in Eq.(4):

$$\mathbf{s}_{\text{proj}} = \frac{\mathbf{s} \cdot \hat{\mathbf{s}}}{\|\mathbf{s}\|^2} \mathbf{s} \qquad (2)$$

$$\text{error} = \hat{\mathbf{s}} - \mathbf{s}_{\text{proj}} \qquad (3)$$

$$\text{Si-SNR} = 10 \log_{10} \frac{\|\mathbf{s}_{\text{proj}}\|^2}{\|\text{error}\|^2} \qquad (4)$$

The Conv-TasNet model is a mask-based model which consists of an encoder, separator, and decoder. The encoder encodes the mixture $\mathbf{x}$ to a latent space as shown in Eq.(5).

$$\mathbf{x}_{\text{enc}} = \text{enc}(\mathbf{x}) \qquad (5)$$

$\mathbf{x}_{\text{enc}} \in \mathbf{R}^{H \times T'}$ is the encoder output, where H is the dimension of the latent space and T' is the length of $\mathbf{x}_{\text{enc}}$. The separator then calculates C masks $\mathbf{m}_i \in \mathbf{R}^{H \times T'}$, $i \in \{1, \cdots, C\}$ based on $\mathbf{x}_{\text{enc}}$ shown in Eq.(6).

$$\mathbf{m}_i = \text{sep}(\mathbf{x}_{\text{enc}}) \qquad (6)$$

The masks are then multiplied with the encoder output, forming separated features $\mathbf{d}_i$ shown in Eq.(7),

$$\mathbf{d}_i = \mathbf{x}_{\text{enc}} \odot \mathbf{m}_i \qquad (7)$$

where $\odot$ is the element-wise multiplication. The separated features $\mathbf{d}_i$ can be viewed as source representations, and are further input to a decoder to estimate separated sources shown in Eq.(8).

$$\hat{\mathbf{s}}_i = \text{dec}(\mathbf{d}_i) \qquad (8)$$

At this point, before measuring the estimated sources with Si-SNR, there is a label permutation problem. An align between $\{\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_C\}$ and $\{\mathbf{s}_1, \cdots, \mathbf{s}_C\}$ needs to be decided. We used the utterance-level permutation invariant training(uPIT) method described in (Kolbæk et al., 2017) to solve this problem.

4 MAML

The procedure of MAML (Finn et al., 2017) is stated as follows. Given a set of multi-accent tasks $\mathcal{T} = \{\{\mathcal{T}_1^i\}_{i=1}^{tq_1}, \cdots, \{\mathcal{T}_K^i\}_{i=1}^{tq_K}\}$, where K is the number of accents. $\mathcal{T}_k = \{\mathcal{T}_k^i\}_{i=1}^{tq_k}$ is the accent task set containing tasks only with the kth accent and tq_k denotes the task quantity of the kth accent task set. The set of tasks $\mathcal{T}$ is split into the source task set $\mathcal{T}_{source}$ and the target task set $\mathcal{T}_{target}$. The model denoted as f, will be trained on the source task set $\mathcal{T}_{source}$ in the hope of having the ability to quickly adapt to the target task set $\mathcal{T}_{target}$.

4.1 Meta Training Phase

During the meta training phase, the MAML algorithm aims to find initialized parameters θ that can further be quickly adapted to new tasks. Moreover, these initialized parameters should be sensitive to the difference between two different tasks, such that adaptation of the initialized parameters can significantly improve the performance on new tasks sampled from the source task set $\mathcal{T}_{source}$. This is achieved by the inner loop and outer loop optimization. A batch of tasks $\tau_{source} = \{\tau_1, \cdots, \tau_b\}$ is sampled from $\mathcal{T}$ proportional to the task quantity of every accent task set, e.g., for an accent task set $\mathcal{T}_k$, the larger tq_k is, the more likely a task is to be sampled from it. Each task in τ_{source} is further split into a support set τ^{sup} and a query set τ^{qry}. The support set is used to adapt the model parameters by performing a one-step gradient decent, which is known as the inner loop shown in Eq.(9).

$$\theta'_j \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\tau_j^{sup}}(f_\theta) \qquad (9)$$

where α is the learning rate. The goal of the inner loop is to minimize the loss of τ_j^{sup} with respect to f_θ. More concisely,

$$\theta'_j = \arg\min_\theta \mathcal{L}_{\tau_j^{sup}}(f_\theta) \qquad (10)$$

At this point, the sum of the query loss of each query set in τ_{source} is calculated by

$$\mathcal{L}_{qry} = \sum_{j=1}^b \mathcal{L}_{\tau_j^{qry}}(f_{\theta'_j}) \qquad (11)$$

The goal of the meta training phase is to minimize the total loss of the query sets. This is also performed by a one-step gradient decent, known as the outer loop shown in Eq.(12).

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{qry} \qquad (12)$$

4.2 Meta Testing Phase

During the meta testing phase, we perform a procedure (see Eq.(13)) similar to the inner loop in the meta training phase. This procedure adapts the parameters θ obtained in the meta training phase to the target tasks $\tau_{target} = \{\tau'_1, \cdots, \tau'_b\}$.

$$\theta_j \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\tau'^{sup}_j}(f_\theta) \qquad (13)$$

4.3 First-order MAML (FOMAML)

Eq.(14) is the calculation of the gradient in the outer loop, where $\mathcal{L}_{\tau^{qry}_j}$ is denoted as $\mathcal{L}^j$ for simplicity.

$$\nabla_\theta \mathcal{L}_{qry} = \nabla_\theta \sum_{j=1}^{b} \mathcal{L}^j(f_{\theta'_j}) = \sum_{j=1}^{b} \nabla_\theta \mathcal{L}^j(f_{\theta'_j})$$
$$(14)$$

When performing the outer loop during the meta training phase, high computational cost is needed to calculate the second-order derivatives with back-propagation. Eq.(15) is the first-order approximation of the second-order derivative,

$$\frac{\partial \mathcal{L}^j(f_{\theta'_j})}{\partial \theta^d} = \sum_{i=1}^{D} \frac{\partial \mathcal{L}^j(f_{\theta'_j})}{\partial \theta'^i_j} \frac{\partial \theta'^i_j}{\partial \theta^d} \approx \frac{\partial \mathcal{L}^j(f_{\theta'_j})}{\partial \theta'^d_j}$$
$$(15)$$

where θ is a D dimensional parameter, θ^d is the d-th dimension of θ and θ'^i_j is the i-th dimension of θ'_j. The difference between FOMAML and MAML is that this approximation is used instead of the second-order derivatives. Thus, compared to MAML, FOMAML can save a lot of computational time, resulting in a faster gradient calculation.

5 Experiments

5.1 Dataset

The multi-accent speech utterances are collected from the speech accent archive (Weinberger, 2014). This archive currently has more than 200 kinds of accents and 2939 samples. Each native or non-native speaker speaks the same English paragraph. We selected 123 accents that contain more than one speaker since we need utterances of two different speakers to generate mixtures. We split these accents into three sets, 85 accents for generating the training tasks and 19 accents each for generating the developing and testing tasks. The utterance of each speaker is split into segments with a duration of 4 seconds. For each accent, we construct meta tasks by following the task construction method
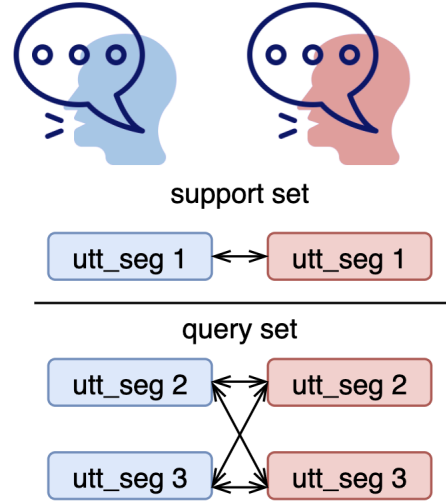

Figure 2: Illustration of a meta task. For two different speakers with the same accent, we sample 3 utterance segments to form a meta task. Thus, there will be 9 mixtures. However, during training, we only sample one mixture to form the support set since our setting is one shot learning. The other 4 mixtures that do not contain the utterance segments in the support set are selected to form the query set.

described in (Wu et al., 2020). We select at most 12 speakers for each accent and generate speech mixtures for each pair of speakers with the same accents. Thus, there will be at most $\binom{12}{2} = 66$ meta tasks and at least $\binom{2}{2} = 1$ meta task for each accent. In each meta task, 3 utterance segments are selected from each speaker and mixed with an SNR level randomly selected between 0 to 5 dB and resampled at an 8kHz sample rate. This results in $3 \times 3 = 9$ speech mixtures in one meta task. Fig.(2) is an illustration describing the support set and query set of a meta task. Finally, for the training, developing, and testing set, 22.4, 3.8, and 3.9 hours of speech mixtures are generated.

5.2 Model

The model we used is Conv-TasNet (Luo and Mesgarani, 2019). It consists of an encoder, separator, and a decoder. The encoder is a 1-dim convolution, which transforms the input mixture into a representation. The separator then calculates two masks based on the encoder output. More specifically, it consists of R stacks of temporal convolutional networks (TCN). Each TCN layer consists of M 1-dim exponentially increasing dilated convolutional blocks. These M blocks each have a residual connection and a skip connection. The residual connection is the input of the next block and the

skip connection of all blocks are summed together, passing a parametric relu, linear projection, and a sigmoid function to produce two masks. The two masks are multiplied with the representation output from the encoder respectively and further input into the decoder to generate two separate waveforms of the two speakers. The decoder is also a 1-dim convolution. The configuration that we used is the one that obtained the best performance reported in (Luo and Mesgarani, 2019).
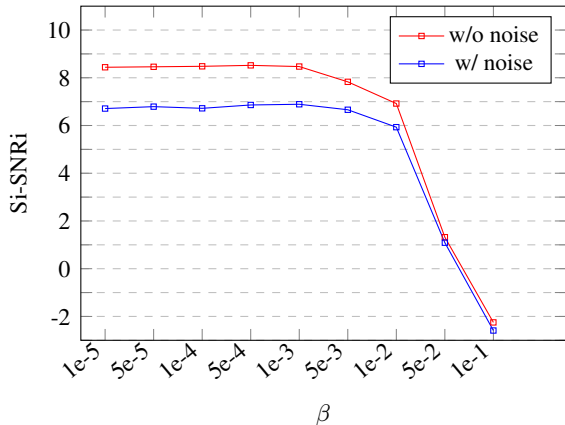

Figure 3: For fine-tuning after joint training, we evaluated the performance by adjusting the learning rate β in the range of 10^{-5} to 10^{-1}.

5.3   Joint Training and Transfer Learning

There are many other works such as (Chen et al., 2020; Tong et al., 2017), that try to solve the domain mismatch problem, where the source domain and target domain datasets do not have a similar distribution. Joint training refers to pretraining a model with different source domain data together. Transfer learning refers to adapting the pretrained model to some partial target domain data and testing the fine-tuned model on the target domain data. The most common adaptation method is fine-tuning. Moreover, the domain mismatch scenario has a low-resource problem if the target domain has only fewer data compared to the scale of the source domain data. There are also several works that tried to solve this problem, such as (Chen and Mak, 2015; Zoph et al., 2016; Hsu et al., 2020). Our jointly trained model is also based on this low-resource scenario.

5.4   MAML and FOMAML

To deal with the domain mismatch and low-resource problem, we applied MAML as our training method in the hope of performing better than

joint training. We set the number of the support set in each task as 1, meaning that the model needs to have the ability to adapt to a new task by only seeing one speech mixture of two new different speakers with a new accent never seen before. We also trained our model with FOMAML in order to know whether calculating gradients with first-order approximation still obtains relatively good performance compared to training with MAML.

5.5   Experiment Settings

For both the joint training and MAML methods, we trained the model from randomly initialized parameters for 100 epochs with the Adam optimizer of 0.001 learning rate and 0.00001 weight decay. For the MAML methods, during the meta training phase, we set $\alpha = 0.01$. For joint training, we also fine-tuned the model parameters with the method in Eq.(13). We tested the fine-tuning learning rate β on the testing set, reported it in section 6, and used the learning rates that obtained the best performance for joint training as our baseline. However, for the models trained with MAML methods, the fine-tuning learning rate β is fixed at 0.01 since other values lead to significant performance degradation.

6   Results

6.1   Joint Training

For joint training, we tested the fine-tuning learning rate β on the testing set as shown in Fig.(3), and found out that $\beta = 5e-4$ obtained the best performance on the clean testing set, while $\beta = 1e-3$ obtained the best performance on the testing set with noise. We use these two experiment settings as our baseline.

6.2   MAML and FOMAML

Comparing models (d), (f) with model (b), we can see that MAML and FOMAML perform better than the joint training baseline. This suggests that the initial model parameters obtained by MAML and FOMAML have the better potential to be adapted to new unseen tasks. Besides, the standard deviation of the testing accent task sets of models (d) and (f) are both less than model (b). This implies that the performance of the models trained with MAML and FOMAML have small dispersion with respect to the mean Si-SNRi value of all the accents compared to the model jointly trained. From Fig.(4), we can see that model (b) performs better

	method	fine-tune	test w/o noise	test w/ noise
(a)	Joint Training	before	8.40 ± 2.25	6.67 ± 2.10
(b)		after	8.52 ± 2.20	6.89 ± 1.84
(c)	FOMAML	before	8.45 ± 3.19	6.66 ± 2.59
(d)		after	$\mathbf{10.13} \pm 2.12$	8.19 ± 1.62
(e)	MAML	before	-6.19 ± 1.38	-6.85 ± 1.31
(f)		after	10.11 ± 1.86	$\mathbf{8.26} \pm 1.52$

Table 1: Evaluation results of joint training and MAML methods on the testing accent task sets with and without noise. The two numbers in a cell denote the average Si-SNRi of all the testing tasks and the standard deviation of all the testing accent task sets.
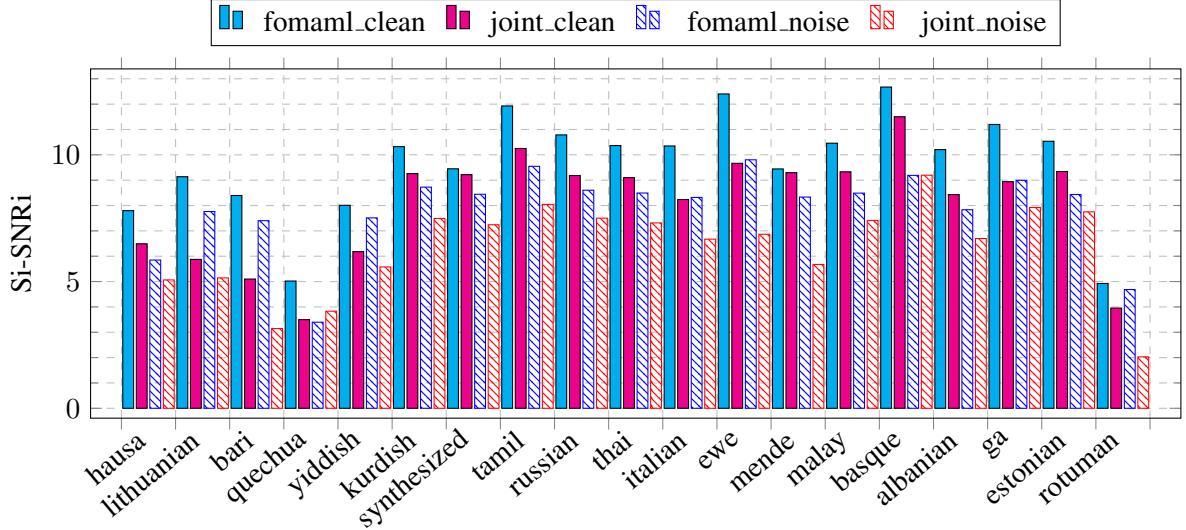

Figure 4: Evaluation results of each testing accent task set for model (b) and (d) in table 1.

on all accents when there is no noise involved and performs better on most of the accents when there is noise in the mixtures.

By comparing models (d) and (f), we found out that these two training methods have similar performance. Model (d) has a slightly higher performance than model (f) under the circumstances that the mixtures are clean in the testing tasks, while model (d) has a slightly lower performance than model (f) under the circumstances that there is noise in the testing tasks. However, MAML requires more than 10 times the training time compared to FOMAML, indicating that the first-order approximation takes advantage over calculating the second-order derivatives by saving a lot of time while still obtaining similar performance. Moreover, FOMAML without fine-tuning (model (c)) has similar performance compared to the baseline model, and yet somehow, initialized parameters obtained by MAML (model (e)) do not have the ability to perform speech separation.

7 Conclusion

Our results show that MAML and FOMAML training methods are effective on multi-accent speech separation. More specifically, it is confirmed that these two methods are better than joint training when adapting to new speakers with new accents and even noisy environments. Besides, FOMAML is shown to be sufficient for dealing with the multi-accent speech separation task and can reduce a large amount of training time. Despite the fact that FOMAML outperforms joint training on the testing set, we can still see that the performance of each accent task set varies a lot from Fig.(4). This is probably due to the task-difficulty imbalance issue described in (Xiao et al., 2020), perhaps some speakers with special accents may be hard to separate. Thus, in the future, we will try to solve this problem with meta sampling methods mentioned in (Xiao et al., 2020).

References

Dongpeng Chen and Brian Kan-Wing Mak. 2015. Multitask learning of deep neural networks for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1172–1183.

Yafeng Chen, Wu Guo, and Bin Gu. 2021. Improved meta-learning training for speaker verification. *arXiv preprint arXiv:2103.15421*.

Yi-Chen Chen, Jui-Yang Hsu, Cheng-Kuang Lee, and Hung-yi Lee. 2020. Darts-asr: Differentiable architecture search for multilingual speech recognition and adaptation. *arXiv preprint arXiv:2005.07029*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2018. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. Meta learning for end-to-end low-resource speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7844–7848. IEEE.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2019. Data efficient direct speech-to-text translation with modality agnostic meta-learning. *arXiv preprint arXiv:1911.04283*.

Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913.

Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE.

Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266.

Eliya Nachmani, Yossi Adi, and Lior Wolf. 2020. Voice separation with an unknown number of multiple speakers. In *International Conference on Machine Learning*, pages 7164–7175. PMLR.

Douglas B Paul and Janet Baker. 1992. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2018. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412.

Sibo Tong, Philip N Garner, and Hervé Bourlard. 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In *Proc. of INTERSPEECH*, CONF.

Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis. 2020. Sudo rm-rf: Efficient networks for universal audio source separation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*.

Steven H. Weinberger. 2014. Speech accent archive.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. 2020. Meta-transfer learning for code-switched speech recognition. *arXiv preprint arXiv:2004.14228*.

Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2019. Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186, Florence, Italy. Association for Computational Linguistics.

Yuan-Kuei Wu, Kuan-Po Huang, Yu Tsao, and Hung-yi Lee. 2020. One shot learning for speech separation. *arXiv preprint arXiv:2011.10233*.

Yubei Xiao, Ke Gong, Pan Zhou, Guolin Zheng, Xiaodan Liang, and Liang Lin. 2020. Adversarial meta sampling for multilingual low-resource speech recognition. *arXiv preprint arXiv:2012.11896*.

Neil Zeghidour and David Grangier. 2020. Wavesplit: End-to-end speech separation by speaker clustering. *arXiv preprint arXiv:2002.08933*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.