INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Model-Agnostic Fast Adaptive Multi-Objective Balancing Algorithm for Multilingual Automatic Speech Recognition Model Training

Jiabin Xue, Tieran Zheng, Jiqing Han

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
xuejiabin@hit.edu.cn, zhengtieran@hit.edu.cn, jqhan@hit.edu.cn

Abstract

This paper regards multilingual automatic speech recognition model training as a multi-objective problem because learning different languages may conflict, necessitating a trade-off. Most previous works on multilingual ASR model training mainly used data sampling to balance the performance of multiple languages but ignore the conflicts between different languages, resulting in an imbalance in multiple languages. The language-specific parameters of the multilingual ASR model are updated by the single language gradients while the update of the shared parameter is jointly determined by the gradient of every language on its shared parameter, namely shared gradient. Therefore, we propose a model-agnostic fast adaptive (MAFA) multi-objective balancing algorithm to balance multiple languages by avoiding the mutual interferences between their shared gradients. In the algorithm, based on the decrease in the training loss, we dynamically normalize the shared gradient magnitudes representing the speed of learning to balance the learning speed. To evenly learn multiple languages, the language with the worst performance is selected, and a balancing gradient nearest to the normalized gradient of the selected language and positively correlated with other normalized ones is obtained to eliminate the mutual interferences. The model trained by MAFA outperforms the baseline model on the Common Voice corpus.

Index Terms: multilingual automatic speech recognition, multi-objective balance optimization, model-agnostic

1. Introduction

In recent years, various neural network-based models have been intensively studied and achieved state-of-the-art performance on automatic speech recognition (ASR) tasks, e.g., hidden markov model-based hybrid ASR models [1–4] and deep sequence model-based end-to-end (E2E) ASR models [5–16]. There are approximately 7 thousand languages worldwide [17], while most models can only recognize a single language. Thus, building an ASR model capable of recognizing multiple languages has attracted great attention from many researchers.

The building process of the multilingual ASR model is equivalent to solving a multi-objective optimization problem if we regard the learning of each language as an independent objective. The simplest and most direct way to build the multilingual ASR model is collecting lots of mixture training samples of various languages and constructing a training sample set by the collected samples and then training the model based on the entire set [17]. Moreover, a proxy objective minimizing the mean of the language-specific objective functions is optimized based on the mini-batch, a randomly selected sample set of multiple languages. However, the number of speakers in different languages are different and the language with more speakers is easier to collect, resulting in sample imbalance [18]. And many same pronunciations correspond to the different spellings in different languages, resulting in mutual interferences between multiple objectives, which further leads to a problem that it is difficult for the aforementioned proxy objective to obtain an optimal model parameter for all languages [19]. Therefore the method usually causes an imbalance performance on multiple languages, i.e., some languages have good performance while others have poor performance [18, 20, 21].

Nowadays, there are a large number of works on balancing multiple languages, which can be divided into two categories, i.e., model-agnostic and model-dependent methods. Considering the problem of sample imbalance, most of the previous model-agnostic works attempted to employ data sampling. The training samples of different languages were randomly sampled with equal probability [18, 20, 21]. In the small number of remaining works, fixed language-specific weights were also employed to scale the contributions of the languages with different counts of training samples [22]. In addition, many model-dependent works attempted to balance multiple languages by extending the model architecture. A language representation was employed as an additional input to convert the universal multilingual ASR model into an expert on each language [23–26]. Another architecture extension used to handle language imbalance was the adapter module, in which a language-specific adapter module was added after every layer of the global multilingual ASR model [27]. Although the model-dependent methods can achieve better performance in most cases than the previous one, its application range is relatively limited. Therefore, in this paper, we focus on the universal model-agnostic methods.

The learning of different languages has different difficulty levels since each has its narrative rules of logic and grammar, e.g., the difference between Chinese and English in the active and passive voice [28]. One of the well-known human learning approaches is deliberate practice [29–31], which shows that people should expend more energy on learning difficult things rather than the easier or familiar ones. Considering there may be many conflicts between different languages, we attempt to evenly learn multiple languages by letting the model spend more time learning the language with the worst performance, i.e., complex language, without influencing other ones.

The current multilingual ASR model includes a shared parameter utilized across all the languages and language-specific parameters used for the single languages. And these parameters are trained by the gradient-based backpropagation algorithm [32]. Moreover, each optimization objective biases the model toward the corresponding language by generating a gradient to the shared parameter, namely shared gradient, during the training process. Thus, we propose a novel multilingual ASR model training method, i.e., Model-Agnostic Fast Adap-

http://dx.doi.org/10.21437/Interspeech.2021-355

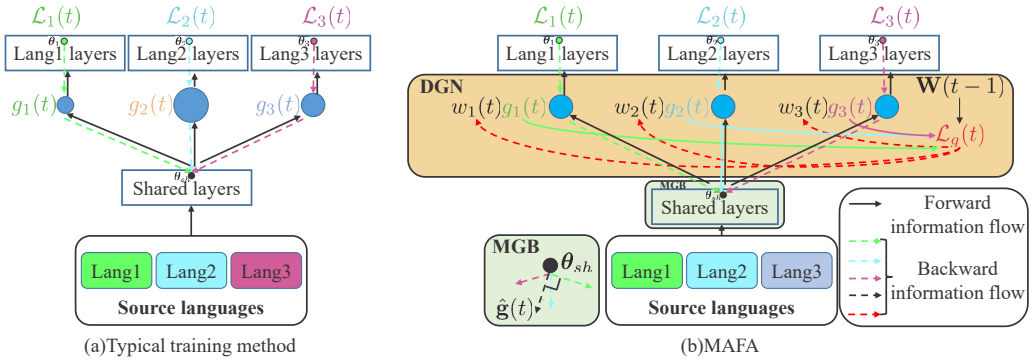


Figure 1: *An illustration of the typical and proposed training method. We assume that Lang3 is the complex language.*

tive (MAFA) multi-objective balancing algorithm, to make the model evenly learn multiple languages. Since the magnitude of the shared gradient represents the learning speed of its corresponding language. In the proposed method, the magnitudes of the shared gradients are first normalized by using a historical learning speed evaluated on the decrease in their training loss, which forces the model to learn multiple languages with the same speed. Then, to make the model increased focus on the complex language without mutual interferences, the complex one is selected according to the performance of the model on the languages included in the mini-batch. And the gradient nearest to the selected language gradient and in the form of acute angle with the rest gradients, which is called a balancing gradient, is employed to update the shared parameter, since the angles between the gradients indicate their interrelationships, e.g., the acute and obtuse angles indicate positive and negative correlations respectively. In this way, we can search for a pareto improvement [33] by improving the performance of the model on the complex language without negatively affecting the others at each training step. And a common optimal solution to multiple objectives, i.e., pareto solution [33], can be also obtained. There are two significant benefits by using the proposed MAFA to train the multilingual ASR model: 1) the model can learn multiple languages in a balanced manner; 2) a common optimal performance on multiple languages can be obtained.

2. Related work

There are recently many multi-objective balancing methods that seek to balance multiple task-specific optimization objectives. A gradient normalization was used to normalize the gradients of multiple objectives [34]. An attention mechanism was utilized in dynamic weight averaging to adjust the weights of multiple objectives dynamically [35]. The key performance metrics were employed to evaluate the difficulty of various tasks, and a higher weight was assigned to the more difficult task [36]. In contrast, [37] showed that the higher weight was assigned, the easier task was. Since the authors suggested that the easier the task was to learn, the more information it contained, and this method seemed more suitable when the task had noisy labeled data. But, these methods consider only balancing multiple objectives by adaptive weights, not avoiding their mutual interferences. Therefore, it is difficult for these methods to obtain the pareto solution. A frank-wolfe algorithm was used to obtain the pareto solution to multiple objectives [38]. However, it was not available for the neural network since solving its underlying optimization problem for the high-dimensional gradient requires a considerable expenditure of time and energy.

3. Methodology

3.1. Dynamic gradient normalization

We propose a dynamic gradient normalization (DGN) to balance the learning speed of multiple languages by normalizing the shared gradient magnitudes at each training step. The magnitudes are dynamically normalized according to the historical learning speed of the corresponding language.

We define the dynamically normalized shared gradients:

$$\hat{\mathbf{G}}(t) = [\hat{\mathbf{g}}_1(t), \hat{\mathbf{g}}_2(t), \dots, \hat{\mathbf{g}}_N(t)], \tag{1}$$

where $\hat{\mathbf{g}}_n(t) = w_n(t)\mathbf{g}_n(t)$ is the normalized gradient, $\mathbf{G}(t) = [\mathbf{g}_1(t), \mathbf{g}_2(t), \dots, \mathbf{g}_N(t)]$ is the shared gradient set, $\mathbf{g}_n(t) = \frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}_{sh}}(t)$ and $\mathcal{L}_n(t)$ are the shared gradient and the objective function of the n-th language, $\boldsymbol{\theta}_{sh}$ is the shared parameter, N is the number of languages, and t is the training step. $\mathbf{W}(t) = [w_1(t), w_2(t), \dots, w_N(t)]$ is the trainable weight set of the shared gradients, and $w_n(t)$ is the trainable weight.

To obtain the normalized $\hat{\mathbf{G}}(t)$, the backpropagation algorithm [32] is employed to adjust $\mathbf{W}(t-1)$ according to the historical learning speed of the corresponding language. So the gradient label and the gradient objective function need to be defined. To generate the label in a natural way, we identify two evaluation indicators that measure the historical speed according to the decrease in training loss of the corresponding language n, i.e., training loss decreasing ratio l_n and historical training speed factor $s_n(t)$:

$$l_n(t) = \frac{\mathcal{L}_n(t)}{\mathcal{L}_n(0)}, \tag{2}$$

$$s_n(t) = \frac{l_n(t)}{\sum_{i=1}^{N} l_i(t)}. \tag{3}$$

The mean of the L^1-norm of $\mathbf{G}(t)$ is used as the common scale, and the gradient label is obtained by combing the common scale and its speed factor. The definitions of the gradient label and objective function are:

$$\mathcal{L}_g(t) = \sum_{n=1}^{N} (w_n(t-1)g_n(t) - \tilde{g}_n(t))^2, \tag{4}$$

$$where \quad \tilde{g}_n(t) = \bar{g}(t)(s_n(t))^{\alpha},$$

$$\bar{g}(t) = \frac{1}{N}\sum_{n=1}^{N} w_n(t-1)g_n(t).$$

$\mathcal{L}_g(t)$ is the gradient objective function, $g_n(t)$ and $\tilde{g}_n(t)$ are the L^1-norm and gradient label of the shared gradient of the language n, $\bar{g}(t)$ is the mean of all $g_n(t)$, and α is the hyperparameter. By (2) − (4), we can find that the defined indicators are inversely associated with the learning speed of the corresponding language, and the higher the value of $s_n(t)$, the higher the gradient magnitude should be for the n-th language to encourage the model to learn the language more quickly. Moreover, the size of α controls the strength of the gradient balancing.

Finally, we can obtain $\mathbf{W}(t)$ as follows:

$$\mathbf{W}(t) = \mathbf{W}(t-1) - \frac{\partial \mathcal{L}_g}{\partial \mathbf{W}}(t). \tag{5}$$

3.2. Multiple gradients balancing

The multiple gradients balancing (MGB) is proposed, in which the balancing gradient is obtained by finding a gradient nearest to the normalized gradient of the complex language k and positively correlated with the rest of the normalized gradients. This process can be expressed as:

$$\begin{aligned} \mathbf{g}^*(t) = \min_{\mathbf{g}(t)} \frac{1}{2} ||\hat{\mathbf{g}}_k(t) - \mathbf{g}(t)||_2^2 \\ s.t. \quad \tilde{\mathbf{G}}(t)^{\mathrm{T}} \mathbf{g}(t) \geq 0, \end{aligned} \tag{6}$$

where $\mathbf{g}^*(t)$ is the balancing gradient, and $\tilde{\mathbf{G}}(t)$ is the set of the rest normalized gradients.

To solve (6), it is converted into the lagrangian dual form:

$$\begin{aligned} f(\boldsymbol{\gamma}) = \min_{\mathbf{g}(t)} L(\mathbf{g}(t), \boldsymbol{\gamma}) \\ s.t. \quad \boldsymbol{\gamma} \geq 0, \\ where \quad L(\mathbf{g}(t), \boldsymbol{\gamma}) = \frac{1}{2} \mathbf{g}(t)^{\mathrm{T}} \mathbf{g}(t) - \hat{\mathbf{g}}_k(t)^{\mathrm{T}} \mathbf{g}(t) \\ - \boldsymbol{\gamma} \tilde{\mathbf{G}}(t)^{\mathrm{T}} \mathbf{g}(t), \end{aligned} \tag{7}$$

$f(\boldsymbol{\gamma})$ is the lagrangian dual function, $L(\mathbf{g}^*(t), \boldsymbol{\gamma})$ is the lagrangian function and $\boldsymbol{\gamma}$ is the lagrange multiplier.

We can obtain a solution to (7) by the Karush-Kuhn-Tucker conditions, which is also the optimal solution of (6) since (7) satisfies the condition of strong duality [39][1]:

$$\mathbf{g}^*(t) = \hat{\mathbf{g}}_k(t) - \tilde{\mathbf{G}}(t) [\frac{\hat{\mathbf{g}}_k(t)^{\mathrm{T}} \tilde{\mathbf{G}}(t)}{\tilde{\mathbf{G}}(t)^{\mathrm{T}} \tilde{\mathbf{G}}(t)}]_-^{\mathrm{T}}, \tag{8}$$

where $[a]_- = \min(a, 0)$. The shared parameter is updated:

$$\boldsymbol{\theta}_{sh} = \boldsymbol{\theta}_{sh} - \mathbf{g}^*(t). \tag{9}$$

3.3. MAFA for multilingual ASR model training

Combining the section 3.1 and 3.2, we propose a MAFA to make the model pay more attention to learning the complex language without influencing other languages, in which the multilingual ASR model is trained by normalizing the magnitudes of the shared gradients and adjusting their angles. In Fig. 1, we show the proposed MAFA.

At each training step of our method, we first compute the training loss of each language and update the language-specific parameters. The DGN is then employed to normalize the magnitudes of the shared gradients. And we can speed up the algorithm by applying the DGN only to the top shared layer. Finally, we find the complex language based on the current mini-batch, a new mini-batch or the evaluation set and obtain the balancing gradient by the MGB to update the shared parameter. Pseudocode is provided in Algorithm 1.

[1]Here, we mainly consider there exists a feasible region. The detailed mathematical derivations and an example code are given in `https://github.com/JiabinXue/MAFA`

Algorithm 1 MAFA for multilingual ASR model training.

Input: Language set: $S = \{1, 2, \ldots, N\}$, hyperparameter: α.
Output: Finally obtained multilingual ASR model parameter.

$\mathbf{W}(0) = [1, 1, \ldots, 1]$;
Pre-training the multilingual ASR model parameter: $\{\boldsymbol{\theta}_{sh}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N\}$;
Initializing $\boldsymbol{\mathcal{L}}(0)$;
for $i = 1 \rightarrow N$ **do**
    Randomly selecting a mini-batch of language i;
    Computing the training loss of language i: $\mathcal{L}_i(0)$; ▷ [standard forward pass]
    $\boldsymbol{\mathcal{L}}(0).insert(\mathcal{L}_i(0))$;
end for
for $t = 1 \rightarrow max\_train\_steps$ **do**
    Randomly selecting a language set from S: S^*;
    Initializing $\mathbf{G}(t)$ and $\boldsymbol{\mathcal{L}}(t)$;
    for $i \in S^*$ **do**
        Computing the language loss :$\mathcal{L}_i(t)$; ▷ [standard forward pass]
        Updating language-specific parameter: $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i - \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{\theta}_i}(t)$; ▷ [standard backward pass]
        Computing the shared gradient: $\mathbf{g}_i(t) = \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{\theta}_{sh}}(t)$;
        $\boldsymbol{\mathcal{L}}(t).insert(\mathcal{L}_i(t))$;
        $\mathbf{G}(t).insert(w_i(t-1)\mathbf{g}_i(t))$;
    end for
    $\mathbf{W}(t), \hat{\mathbf{G}}(t) = \mathrm{DGN}(\mathbf{W}(t-1), \boldsymbol{\mathcal{L}}(0), \boldsymbol{\mathcal{L}}(t), \mathbf{G}(t), \alpha)$;
    $\mathbf{g}^*(t) = \mathrm{MGB}(\hat{\mathbf{G}}(t))$;
    Updating the shared parameter by (9);
end for
function $\mathrm{DGN}(\mathbf{W}(t-1), \boldsymbol{\mathcal{L}}(0), \boldsymbol{\mathcal{L}}(t), \mathbf{G}(t), \alpha)$
    Initializing $\mathbf{s}(t)$;
    for $\mathcal{L}_n(0) \in \boldsymbol{\mathcal{L}}(0)$ and $\mathcal{L}_n(t) \in \boldsymbol{\mathcal{L}}(t)$ **do**
        Computing loss decreasing ratio $l_n(t)$ by (2);
        Computing training speed factor $s_n(t)$ by (3);
        $\mathbf{s}(t).insert(s_n(t))$;
    end for
    Computing the gradient loss $\mathcal{L}_g(t)$ by (4);
    Updating $\mathbf{W}(t)$ by (5);
    Computing the normalized gradient $\hat{\mathbf{G}}(t)$ by (1);
    return $\mathbf{W}(t), \hat{\mathbf{G}}(t)$;
end function
function $\mathrm{MGB}(\hat{\mathbf{G}}(t))$
    Generating a evaluation set, and finding the normalized gradient of the complex language in S^*: $\hat{\mathbf{g}}_k(t)$;
    Computing the balancing gradient $\mathbf{g}^*(t)$ by (8);
    return $\mathbf{g}^*(t)$;
end function

4. Experiments and discussion

4.1. Experimental setting

We evaluated our training method on the Common Voice corpus [40], a massively-multilingual collection of transcribed speech intended for speech technology research. This corpus was widely used in the previous study about multilingual ASR [41–43]. We selected four languages from this corpus, i.e., English (en),

Table 1: *The durations of multiple languages.*

language	train (h)	dev (h)	test (h)	total (h)
en	878.3	110.2	107.5	1096
ca	195.5	24.2	24.3	245
fr	274.0	34.0	34.0	342
de	377.7	47.5	47.8	472

Table 2: *The breakdown effect for each proposed component.*

Method	ca (%)	fr (%)	de(%)	Average(%)
Monolingual	18.2	25.8	21.7	21.9
Multilingual	18.1	26.0	21.3	21.8
DGN	15.7	25.5	**21.0**	20.7
MGB	19.5	**23.1**	21.4	21.3
DGN + MGB	**14.8**	24.1	21.6	**20.1**

Catalan (ca), French (fr), and German (ge). The en was used to pre-train the multilingual ASR model and the other languages were used to build a multilingual corpus. A detailed description of the corpus is given in Table 1.

All the models in this paper were built by ESPent [44]. The model architecture and training strategy were similar to our previous work. The difference was that we generated subword units used as the output units by byte-pair encoding [45].

4.2. Ablation study

We first did an ablation experiment to verify the effectiveness of proposed components, i.e., DGN and MGB. The baseline model of this experiment was the publicly available attention-based multilingual ASR model for the Common Voice corpus [2]. We show the word error rates (WERs) of the monolingual and various multilingual ASR models in Table 2.

We compared the performance of the multilingual baseline model trained by the original training method [17] and the monolingual ASR model before formal ablation study. From the first two rows in Table 2, we can find that the multilingual baseline model achieved a similar or better performance compared with the monolingual ASR model, which proves that our ASR model has the ability to learn multiple languages simultaneously and it conforms to the famous stein's paradox [46]. We next compared the effectiveness of various proposed components. It can be found that the models trained by the proposed DGN and MGB were able to achieve relative reductions of 5.1% and 2.3% in the average WER by comparing the middle rows in this table. And they obtained the best performance on fr and de respectively. Furthermore, the optimal performance was achieved by the model trained by the DGN + MGB, and its average WER was 20.1%. Moreover, for ca, it obtained a relative reduction of 18.2% in the WER. The results showed that the MAFA was able to enhance the multilingual ASR model.

4.3. State-of-the-Art comparisons

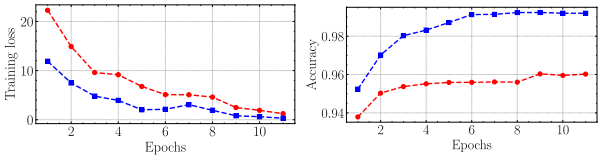
To further verify the effectiveness of the proposed MAFA, we designed a comparison experiment, in which it was compared with other model-agnostic multilingual ASR model training methods [18, 22] and multi-objective balancing methods [34–38].

Table 3 shows the performance of the models trained by various publicly available methods. As shown in the top part of this table, the best performance of the model-agnostic training methods was achieved by the model trained by the language-specific loss weight [22], and its WER was 20.9%. However, this performance was still worse than that of the model trained by the proposed MAFA. Furthermore, it can be found that the performance of all models trained by the commonly used multi-objective balancing methods was significantly worse than that of the multilingual baseline model from the middle part of Table 3. The model trained by the frank-wolfe-based method [38] performed poorly since it frequently failed to converge within the specified iterations number. This experimentally confirmed that MAFA outperformed the other methods.

[2]The baseline code is available at `https://github.com/espnet/espnet/blob/master/egs/commonvoice/asr1/conf/tuning/train_rnn.yaml`

Table 3: *WERs [%] of various commonly used methods.*

Method	ca (%)	fr (%)	de(%)	Average(%)
A. Kannan [18]	17.4	25.6	**20.3**	21.1
T. Alumäe [22]	17.5	24.7	20.4	20.9
Z. Chen [34]	20.0	30.6	25.0	25.2
S. Liu [35]	21.7	26.0	21.7	23.2
M. Guo [36]	24.1	32.7	30.1	28.9
A. Kendall [37]	18.9	27.5	29.9	25.4
O. Sener [38]	27.8	32.4	30.1	30.1
Our	**14.8**	**24.1**	**21.6**	**20.1**



(a) *Training loss curves.* (b) *Accuracy curves.*

Figure 2: *An illustration of the convergence. The blue and red dotted lines indicate that the MAFA and the baseline model.*

4.4. Convergence analysis

For a more comprehensive evaluation of the proposed method, we compared the convergence of our method with the original training method [17]. We first show the training loss curves of the two training methods in Fig. 2a. We find that the proposed MAFA was able to converge toward values significantly lower than that of the original training method, and it had a faster convergence speed. The accuracy curves are next illustrated in Fig. 2b. We can find that the MAFA can achieve higher accuracy on the evaluation set. The results further verified that the proposed MAFA was fit for the training of the multilingual ASR model.

5. Conclusion

We propose a model-agnostic multilingual ASR model training method, in which a novel multi-objective balancing algorithm, i.e., MAFA, is designed to balance multiple languages. In the method, we proposed a DGN to balance the learning speed of the languages by normalizing the magnitudes of the shared gradients. The pareto solution to the objectives of the languages is obtained by the MGB, which lets the model learn the complex language without influence on the rest of the ones. We evaluated the MAFA on the Common Voice corpus, and for ca, a relative reduction of 18.2% in WER was achieved.

6. Acknowledgements

This research was supported by the National Key Research and Development Plan of China under Grant 2017YFB1002102 and the National Natural Science Foundation of China under Grant U1736210.

7. References

[1] K. Lee, H. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 1, pp. 35–45, 1990.

[2] G. E. Dahl, D. Yu, L. Deng *et al.*, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[3] O. Abdel-Hamid, A. Mohamed, H. Jiang *et al.*, "Convolutional neural networks for speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[4] S. R. Madikeri, B. K. Khonglah, S. Tong *et al.*, "Lattice-free maximum mutual information training of multilingual speech recognition systems," in *Proc. INTERSPEECH*, 2020, pp. 4746–4750.

[5] A. Graves, S. Fernández, F. J. Gomez *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[6] J. Xue, T. Zheng, and J. Han, "Convolutional grid long short-term memory recurrent neural network for automatic speech recognition," in *ICONIP*, 2019, pp. 718–726.

[7] J. Chorowski, D. Bahdanau, D. Serdyuk *et al.*, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015, pp. 577–585.

[8] W. Chan, N. Jaitly, Q. V. Le *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.

[9] J. Xue, T. Zheng, and J. Han, "Structured sparse attention for end-to-end automatic speech recognition," in *Proc. ICASSP*, 2020, pp. 7044–7048.

[10] S. Watanabe, T. Hori, S. Karita *et al.*, "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[11] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. ACL*, 2017, pp. 1073–1083.

[12] D. Bahdanau, J. Chorowski, D. Serdyuk *et al.*, "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP*, 2016, pp. 4945–4949.

[13] H. Miao, G. Cheng, P. Zhang *et al.*, "Online hybrid ctc/attention architecture for end-to-end speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2623–2627.

[14] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP*, 2018, pp. 5884–5888.

[15] X. Chang, W. Zhang, Y. Qian *et al.*, "End-to-end multi-speaker speech recognition with transformer," in *Proc. ICASSP*, 2020, pp. 6134–6138.

[16] Q. Zhang, H. Lu, H. Sak *et al.*, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *Proc. ICASSP*, 2020, pp. 7829–7833.

[17] S. Toshniwal, T. N. Sainath, R. J. Weiss *et al.*, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*, 2018, pp. 4904–4908.

[18] A. Kannan, A. Datta, T. N. Sainath *et al.*, "Large-scale multilingual speech recognition with a streaming end-to-end model," in *Proc. INTERSPEECH*, 2019, pp. 2130–2134.

[19] S. Vandenhende, S. Georgoulis, W. Van Gansbeke *et al.*, "Multi-task learning for dense prediction tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[20] A. I. García-Moral, R. Solera-Ureña, C. Peláez-Moreno *et al.*, "Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems," *IEEE Trans. Speech Audio Process.*, vol. 19, no. 3, pp. 468–481, 2011.

[21] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR," in *Proc. ICASSP*, 2016, pp. 4955–4959.

[22] T. Alumäe, S. Tsakalidis, and R. M. Schwartz, "Improved multilingual training of stacked neural network acoustic models for low resource languages," in *Proc. INTERSPEECH*, N. Morgan, Ed., 2016, pp. 3883–3887.

[23] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 459–468, 2016.

[24] S. Toshniwal, T. N. Sainath, R. J. Weiss *et al.*, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*, 2018, pp. 4904–4908.

[25] B. Li, T. N. Sainath, K. C. Sim *et al.*, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *Proc. ICASSP*, 2018, pp. 4749–4753.

[26] M. Grace, M. Bastani, and E. Weinstein, "Occam's adaptation: A comparison of interpolation of bases adaptation methods for multi-dialect acoustic modeling with LSTMS," in *Proc. SLT*, 2018, pp. 174–181.

[27] S. Tong, P. N. Garner, and H. Bourlard, "Multilingual training and cross-lingual adaptation on ctc-based acoustic model," *arXiv preprint arXiv:1711.10025*, 2017.

[28] The Foreign Service Institute, "Language Difficulty Ranking," Website, 2020, https://www.state.gov/foreign-language-training/.

[29] A. Ericsson and R. Pool, *Peak: Secrets from the new science of expertise*. Houghton Mifflin Harcourt, 2016.

[30] J. Waitzkin, *The art of learning: A journey in the pursuit of excellence*. Simon and Schuster, 2007.

[31] D. Coyle, *The culture code: The secrets of highly successful groups*. Bantam, 2018.

[32] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning*, ser. Adaptive computation and machine learning. MIT Press, 2016.

[33] D. T. Luc, *Pareto Optimality*. Springer New York, 2008, pp. 481–515.

[34] Z. Chen, V. Badrinarayanan, C. Lee *et al.*, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. ICML*, 2018, pp. 793–802.

[35] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. CVPR*, 2019, pp. 1871–1880.

[36] M. Guo, A. Haque, D. Huang *et al.*, "Dynamic task prioritization for multitask learning," in *Proc. ECCV*, 2018, pp. 282–299.

[37] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR*, 2018, pp. 7482–7491.

[38] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. NIPS*, 2018, pp. 525–536.

[39] W. S. Dorn, "Duality in quadratic programming," *Quarterly of applied mathematics*, vol. 18, no. 2, pp. 155–162, 1960.

[40] R. Ardila, M. Branson, K. Davis *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020, pp. 4218–4222.

[41] G. I. Winata, S. Cahyawijaya, Z. Lin *et al.*, "Meta-transfer learning for code-switched speech recognition," in *Proc. ACL*, 2020, pp. 3770–3776.

[42] G. I. Winata, S. Cahyawijaya, Z. Liu *et al.*, "Learning fast adaptation on cross-accented speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 1276–1280.

[43] A. Wu, C. Wang, J. Pino *et al.*, "Self-sproc. INTERSPEECH," 2020, pp. 1491–1495.

[44] S. Watanabe, T. Hori, S. Karita *et al.*, "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[45] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, 2016.

[46] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," Stanford University Stanford United States, Tech. Rep., 1956.