Interspeech 2024
1-5 September 2024, Kos, Greece

M^{2}ASR: Multilingual Multi-Task Automatic Speech Recognition via Multi-Objective Optimization

A F M Saif[1], *Lisha Chen*[1], *Xiaodong Cui*[2], *Songtao Lu*[2], *Brian Kingsbury*[2], *Tianyi Chen*[1]

[1]Rensselaer Polytechnic Institute, Troy, NY, USA
[2]IBM Research AI, T. J. Watson Research Center, Yorktown Heights, NY, USA
{saifa,chenl21,chent18}@rpi.edu, {cuix, bedk}@us.ibm.com, songtao@ibm.com

Abstract

To enable the capability of speech models across multiple languages, training multilingual, multi-task automatic speech recognition (ASR) models has gained growing interest. However, different languages and tasks result in distinct training objectives, potentially leading to conflicts during training and degrading the model's performance. To overcome this issue, we introduce M^{2}ASR, a multilingual, multi-task ASR framework, which formulates the problem as a constrained multi-objective optimization (MOO), where multilingual multi-task supervised training augmented by speech-to-text translation (S2TT) serve as supervised objectives and are subject to the desired performance of multilingual unsupervised training. We employ MOO techniques to avoid conflicts among multiple linguistic representations and tasks during training. Extensive experiments demonstrate that M^{2}ASR outperforms conventional multilingual ASR models by 28.3% to 38.6% across diverse ASR tasks.
Index Terms: multilingual speech recognition, speech-to-text translation, multi-objective learning, multi-task learning

1. Introduction

Developing a unified speech model that is capable of simultaneously handling multiple tasks across diverse languages has recently gained significant attention. One reason is that utilizing a single speech model can save computational resources compared to using multiple models separately for different tasks. Another reason is that it broadens the applicability of speech processing systems [1, 2]. However, training these models is often challenging because of the diversity of languages, limited data availability for some languages, and the inherent complexity of handling multiple tasks within a single model [3, 4, 5]. For example, conventional multilingual models typically feature a shared backbone acoustic encoder across languages, but this backbone needs to handle different phonetic structures and acoustic characteristics, which may cause potential conflicts among multiple objectives that represent the performance of different tasks and different languages [6]. Consequently, performance improvements in one task/language may lead to performance degradation in another.

In addition to the conflicts among objectives, another challenge is the lack of labeled data. To tackle this, most existing studies in the speech community use self-supervised learning (SSL) techniques [7, 8] on either multilingual ASR or S2TT tasks [9, 10]. However, self-supervised pre-training and task-specific fine-tuning are often executed in separate stages, elevating the risk of negative transfer or catastrophic forgetting [11, 12]. In this paper, we aim to address the above challenges through a multi-objective optimization framework.

1.1. Our contributions

We propose a multi-objective optimization (MOO) framework that we call **M^{2}ASR** to tackle the **M**ultilingual and **M**ulti-task **ASR** problem. Our contributions can be summarized as follows:

- **C1) We formulate the multilingual multi-task ASR within a unified MOO framework.** The multiple objectives in this framework are defined by either the self-supervised pretraining objective or the supervised fine-tuning objectives for multiple languages and multiple tasks such as ASR and S2TT.
- **C2) We propose an algorithm to mitigate multi-objective conflicts during training.** Our algorithm explicitly employs a conflict-avoidant update to mitigate multi-objective gradient conflicts during optimization, aiming for simultaneous improvements across all objectives after each training epoch.
- **C3) We conduct extensive experiments on different datasets.** Experiments conducted on the Librispeech, AISHELL, and CoVoST-v2 datasets demonstrate that M^{2}ASR exhibits superior performance compared to existing multilingual baselines.

2. Related Work

Multilingual ASR. Multilingual ASR has been studied extensively in the literature. Earlier works on multilingual ASR were based on hidden Markov models and multilayer perceptrons [13, 14, 15]. These models were effective but had limitations, especially when dealing with complex speech patterns and multiple languages. Recently, Seq2Seq models and transformer-based models have been utilized, achieving state-of-the-art (SOTA) results thanks to their ability to capture long-range global context [16, 17]. More recently, SSL with transformer-based models has attained SOTA performance [18, 19]. Despite these advancements, these models lack the capability to perform multiple speech-related tasks.

Multitask ASR. The first multitask framework dedicated to joint ASR and S2TT was introduced by [20]. Subsequent works explore word embedding, and transformer-based models for multitask ASR and S2TT [21, 22, 23]. Notably, studies on multitask learning in this context have been limited to a specific language. And there is a clear need for more extensive research across a broader range of languages.

3. A Unified Framework - M^{2}ASR

In this section, we will formulate the problem of M^{2}ASR and introduce a MOO training algorithm.

This work was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration.

10.21437/Interspeech.2024-1492

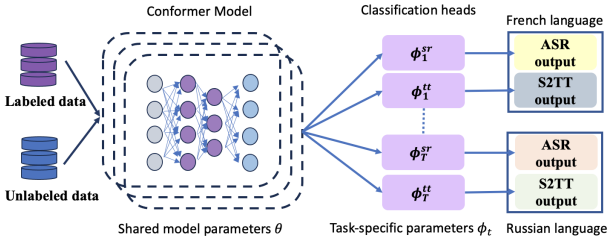

Figure 1: *An overview of the multi-head conformer model used for the multilingual, multitask ASR problem.*

3.1. MOO formulation of M^{2}ASR

In M^{2}ASR, we deal with numerous training objectives, which are elaborated below, in a unified framework based on MOO.

Objectives of unsupervised and supervised training In multilingual acoustic modeling, similar to the monolingual case, there is a large amount of unlabeled data and a small amount of labeled data from each language. The unlabeled and labeled data across languages are pooled together for unsupervised and supervised training respectively. We use the contrastive predictive coding (CPC) loss, $L_{\text{cpc}}(\theta)$ [7], for the former and the connectionist temporal classification (CTC) loss, $L_{\text{ctc}}(\theta, \phi)$ [24], for the latter. Here, θ represents the backbone parameters learned during unsupervised training, and ϕ denotes the task-specific parameters learned during supervised training. The unsupervised and supervised training are conducted jointly as a MOO problem.

Objectives of language-specific outputs We use separate classification heads per language, each associated with its own CTC loss, $L_{\text{ctc}}(\theta, \phi_t)$, where $t \in [T]$ represents a specific language.

Objectives of ASR and S2TT In the case of ASR augmented by S2TT, we use two classification heads with separate objectives, denoted as $L_{\text{ctc}}(\theta, \phi_t^{\text{sr}})$ and $L_{\text{ctc}}(\theta, \phi_t^{\text{tt}})$, for each language $t \in [T]$. Here, ϕ_t^{sr} and ϕ_t^{tt} represent the parameters of the classification heads for the ASR and S2TT tasks, respectively. Note that all ASR and S2TT objectives can be represented as a vector, $L_{\text{ctc}}(\Theta)$, where $\Theta := [\theta, \phi_1^{\text{sr}}, \phi_1^{\text{tt}}, \ldots, \phi_T^{\text{sr}}, \phi_T^{\text{tt}}]$ and $L_{\text{ctc}}(\Theta) := [L_{\text{ctc}}(\theta, \phi_1^{\text{sr}}), L_{\text{ctc}}(\theta, \phi_1^{\text{tt}}), \ldots, L_{\text{ctc}}(\theta, \phi_T^{\text{sr}}), L_{\text{ctc}}(\theta, \phi_T^{\text{tt}})]$.

Figure 1 illustrates the network architecture for the proposed M^{2}ASR with a conformer backbone and task/language-specific heads. The shared backbone is parameterized by θ and each of the task/language-specific heads is parameterized by ϕ_t^{sr} for the ASR task or ϕ_t^{tt} for the S2TT task. Given this architecture, we formulate M^{2}ASR as a constrained MOO problem, given by

$$\min_{\Theta} \; [L_{\text{ctc}}(\theta, \phi_1^{\text{sr}}), L_{\text{ctc}}(\theta, \phi_1^{\text{tt}}), \ldots, L_{\text{ctc}}(\theta, \phi_T^{\text{sr}}), L_{\text{ctc}}(\theta, \phi_T^{\text{tt}})]$$
$$\text{s.t.} \;\; L_{\text{cpc}}(\theta) - \min_{\theta} L_{\text{cpc}}(\theta) \leq \epsilon \tag{1}$$

where the CPC loss $L_{\text{cpc}}(\theta)$ needs to be optimized below a specified threshold ϵ, which serves as a constraint imposed on the supervised CTC losses in various languages and tasks (ASR and S2TT). The specified ϵ ensures a feasible region of the optimization landscape and guarantees a sufficiently good acoustic representation with a small CPC loss.

3.2. Algorithm

Leveraging recent advances in unconstrained MOO [25], we employ a penalty-based approach [26, 27] to convert the constrained MOO problem in (1) to an unconstrained MOO problem:

$$\min_{\Theta} L_\eta(\Theta) := [L_{\text{ctc}}(\theta, \phi_1^{\text{sr}}) + \eta L_{\text{cpc}}(\theta), L_{\text{ctc}}(\theta, \phi_1^{\text{tt}}) + \eta L_{\text{cpc}}(\theta),$$
$$\ldots, L_{\text{ctc}}(\theta, \phi_T^{\text{sr}}) + \eta L_{\text{cpc}}(\theta), L_{\text{ctc}}(\theta, \phi_T^{\text{tt}}) + \eta L_{\text{cpc}}(\theta)] \tag{2}$$

where η is a penalty constant that depends on ϵ.

It is generally difficult, if not impossible, for a model to simultaneously minimize all the objectives. For general non-convex objectives, we seek algorithms that are guaranteed to find models satisfying Pareto stationarity conditions. A model $\Theta \in \mathbb{R}^d$ is Pareto stationary if there exists $\lambda \in \Delta^{2T} := \{\lambda \in \mathbb{R}^{2T} \mid \mathbf{1}^\top \lambda = 1, \; \lambda \geq 0\}$ such that $\nabla L(\Theta)\lambda = 0$, i.e., $\min_{\lambda \in \Delta^{2T}} ||\nabla L(\Theta)\lambda|| = 0$, where $\nabla L(\Theta) \in \mathbb{R}^{d \times 2T}$.

Limitation of static weighting. To guarantee Pareto stationarity for ASR and S2TT objectives, we can employ either static or dynamic weighting MOO methods. In static weighting, we optimize the (weighted) average of the multiple objectives [28]. This method is simple but it may suffer from conflicting objective gradients. For instance, considering $l_t(\Theta) = L_{\text{ctc}}(\theta, \phi_t) + \eta L_{\text{cpc}}(\theta)$ and $l_{t'}(\Theta) = L_{\text{ctc}}(\theta, \phi_{t'}) + \eta L_{\text{cpc}}(\theta)$, if the two objectives having conflicting directions, $(t, t') \in T$, then $\langle \nabla l_t(\Theta), \nabla l_{t'}(\Theta) \rangle < 0$. This indicates that updating along one gradient will degrade the performance of the other objective.

Proposed dynamic weighting. To mitigate gradient conflicts, we use the dynamic weighting method MGDA [29] which finds a conflict-avoidant (CA) direction. Specifically, a CA direction d is the steepest common descent direction that maximizes the worst descent, given by

$$d(\Theta) = \arg\max_{d} \min_{\lambda \in \Delta^{2T}} -\langle \nabla L_\eta(\Theta)\lambda, d \rangle - \frac{1}{2}||d||^2. \tag{3}$$

By reformulation, such a direction $d(\Theta)$ is equal to dynamically weighted gradients of different objectives [25], given by $d(\Theta) = -\nabla L_\eta(\Theta)\lambda^*(\Theta)$ with the weight $\lambda^*(\Theta)$ computed by

$$\lambda^*(\Theta) = \arg\min_{\lambda \in \Delta^{2T}} ||\nabla L_\eta(\Theta)\lambda||^2. \tag{4}$$

However, computing the full-batch gradients $\nabla L_\eta(\Theta)$ during optimization is computationally expensive. Hence, in our problem, we employ a stochastic variant of MGDA, the MoDo algorithm [25], which obtains an unbiased stochastic gradient estimate for (4) via a double sampling technique.

At each iteration k, denote ξ_k' and ξ_k'' as two independent samples from labeled dataset D, and $\nabla L_{\text{ctc}}(\xi_k'; \Theta_k)$ and $\nabla L_{\text{ctc}}(\xi_k''; \Theta_k)$ as the stochastic gradients. We leverage the MoDo update in [25] by

$$\lambda_{k+1} = \Pi_{\Delta^{2T}}\Big(\lambda_k - \gamma_k\big(\nabla L_\eta(\xi_k'; \Theta_k)^\top \nabla L_\eta(\xi_k''; \Theta_k)\big)\lambda_k\Big) \tag{5}$$

where γ_k is a step size and $\Pi_{\Delta^{2T}}(\cdot)$ denotes projection to Δ^{2T}.

To train a model using the proposed M^{2}ASR algorithm, the backbone parameters θ are updated using the λ in (5) by

$$\theta_{k+1} = \theta_k - \alpha \sum_{t=1}^{T} \lambda_{k,t} \nabla_\theta L_{\text{ctc}}(\theta_k, \phi_{t,k}^{\text{sr}})$$
$$- \alpha \sum_{t=1}^{T} \lambda_{k,t} \nabla_\theta L_{\text{ctc}}(\theta_k, \phi_{t,k}^{\text{tt}}) - \alpha\eta \nabla_\theta L_{\text{cpc}}(\theta_k) \tag{6}$$

where $\lambda_{k,t}$ is the t-th entry of λ_k, and $\alpha > 0$ is the learning rate. Similarly, taking the gradients of ASR and S2TT objective functions with respect to task-specific output heads, task-specific parameters are updated via

$$\phi_{t,k+1}^{\text{sr}} = \phi_{t,k}^{sr} - \beta \nabla_\phi L_{\text{ctc}}(\phi_{t,k}^{\text{sr}}, \theta_k) \tag{7a}$$
$$\phi_{t,k+1}^{\text{tt}} = \phi_{t,k}^{\text{tt}} - \beta \nabla_\phi L_{\text{ctc}}(\phi_{t,k}^{\text{tt}}, \theta_k) \tag{7b}$$

where $\beta > 0$ is the learning rate; see Algorithm 1.

Algorithm 1 M^{2}ASR for multilingual multi-task ASR

Input: Labeled data (x, y), unlabeled data $X^u := \{x_1^u, x_2^u, \cdots, x_N^u\}$, learning rates α, β, and penalty η;
for $k = 1$ **to** K **do**
  sample $\xi_k^u = X_k^u$, $\xi_k' = (x_k', y_k')$ and $\xi_k'' = (x_k'', y_k'')$
  compute $\nabla L_{\text{cpc}}(\xi_k^u; \theta_k)$
  compute $\nabla L_{\text{ctc}}(\xi_k'; \theta_k, \phi_k)$, $\nabla L_{\text{ctc}}(\xi_k''; \theta_k, \phi_k)$
  update λ_{k+1} by (5)
  update θ_{k+1} by (6)
  update $\phi_{t,k+1}^{sr}$ by (7a) and $\phi_{t,k+1}^{tt}$ by (7b) for all $t \in [T]$
end for
Output: $\theta_K, \phi_{t,K}^{sr}, \phi_{t,K}^{tt}$

Table 1: *WERs on combined Librispeech and AISHELL v1. The percentage in parentheses denotes the WER reduction over MST.*

Language	MST	PT+FT	M^{2}ASR (Static)	M^{2}ASR (Dynamic)
Chinese	10.2%	9.1%	8.0%	**6.2%** (-39.2%)
English	11.8%	9.7%	8.4%	**7.3%** (-38.1%)
Average	11.0%	9.4%	8.2%	**6.7%** (-38.6%)

4. Experiments

In this section, we evaluate M^{2}ASR on various datasets and compare it with existing multilingual training strategies. The code can be accessed at https://github.com/afmsaif/M2ASR.

4.1. Dataset

We evaluate M^{2}ASR on two datasets. First, we consider only the ASR task and create a dataset by combining Librispeech [30], a popular English dataset consisting of 960 hours of speech and AISHELL v1 [31], a popular Chinese dataset consisting of 178 hours of speech. Second, we consider both ASR and S2TT tasks and use CoVoST v2 [5], which is a large-scale multilingual ASR and S2TT corpus covering translations from 21 languages into English and from English into 15 languages.

4.2. Models

We use a conformer backbone with 8 conformer blocks and 512 hidden dimensions [32]. Each conformer block has 8 attention heads of 64 dimensions. The convolutional kernel size is 31. The encoder of the conformer has 58.4 M parameters. The softmax output units depend on different languages, mostly around 1000 units. For unsupervised training, we use the raw data as input; for supervised training, we convert the raw audio files into 80-dimensional logmel features which are then normalized to have zero mean and unit variance. We use SpecAug for data augmentation [33], and SentencePiece for text tokenization and detokenization [34]. For Chinese and Japanese, we use character-based tokens and for other languages, we use word-based tokens. Each task/language-specific head includes a linear layer followed by a softmax layer with SentencePiece units.

4.3. Baselines of comparison

We consider the following training baselines.
Multilingual Supervised Training (MST): Use supervised training with a shared backbone and language-dependent ASR heads. This is a natural approach to multilingual training [35].
Translation Augmented Multilingual Supervised Training (TAMST): Follow the same training step as MST but include the S2TT head in addition to the ASR head for each language. The ASR head provides token output units from the same language as the speech input (e.g., English input and English output). The S2TT head provides token output units in the translated language (e.g., French input and English output).[1]
Pre-training and Fine-tuning (PT+FT): First follow unsupervised training using multilingual unlabeled data under the CPC loss and then supervised fine-tuning using multilingual labeled data under the CTC loss. In the fine-tuning stage, each language has its own head (ASR or ASR+S2TT).
Wav2Vec2: Follow same steps as PT+FT but using the wav2vec2 architecture. We also adopt the hyperparameters from [8] for consistency. To ensure a fair comparison, we only use 12 conformer blocks in the encoder with each block having 512 hidden units and 8 attention heads to make the size similar to ours.
M^{2}ASR (Static): Use the same conformer model and objectives (CPC and CTC losses) as the TAMST but solve MOO in (2) using static weights λ.
M^{2}ASR (Dynamic): Use the same model as M^{2}ASR (Static) but solve MOO in (2) using dynamic weights in Algorithm 1.

We optimize hyperparameters using grid search. To find optimal learning rates, we experiment with various combinations, consistently setting a higher rate for constrained objective training compared to ASR and S2TT training. We determine the best values to be $\alpha = 5 \times 10^{-4}$ for constrained-objective learning rate, $\beta = 5 \times 10^{-5}$ for ASR and S2TT learning rate, and $\gamma = 0.01$ for MoDo step size.

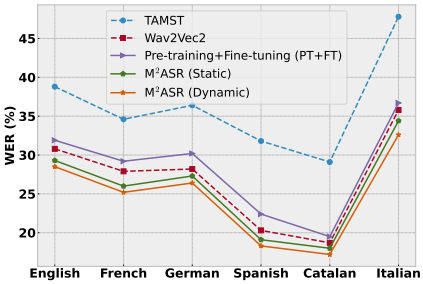

Figure 2: *WERs of high-resource data in ASR.*

4.4. Results

In Table 1, we present the word error rates (WERs) on the merged Librispeech and AISHELL datasets. For MST, we use 'train-clean-100' from Librispeech and the full AISHELL dataset. For PT+FT, M^{2}ASR (Static), and M^{2}ASR (Dynamic) training methods, we combine 960 hours of Librispeech data with the AISHELL v1 dataset for unsupervised training and use 'train-clean-100' from Librispeech and the full AISHELL dataset for supervised training. From Table 1, the M^{2}ASR (Dynamic) method outperforms all other methods by a significant margin. The gain of M^{2}ASR (Dynamic) over PT+FT is 31.8% on AISHELL (Chinese) and 24.7% on Librispeech (English). Compared with M^{2}ASR (Static), it outperforms by 22.5% on AISHELL and 13.0% on Librispeech. Compared with the MST, it reduces the WER by 39.2% on AISHELL and 38.1% on Librispeech, respectively. The higher WERs of MST and PT+FT compared to M^{2}ASR suggest the presence of conflicting gradient updates, which degrade model performance.

We leverage the CoVoST-v2 dataset to explore both ASR and S2TT tasks across 22 languages which leads to 22 classification heads for ASR and 21 classification heads for S2TT. For S2TT augmentation, we translate 21 languages into English (X $\rightarrow$

[1]Due to the absence of translated transcripts in Librispeech and AISHELL, TAMST training is not applied to this combined dataset.

Table 2: *WERs under various training strategies on CoVoST 2 dataset. The percentage in parentheses denotes the relative WER reduction over MST. S2TT heads are applied in all strategies except MST.*

Language	Hours of training Data	MST	TAMST	PT+FT	M^{2}ASR (Static)	M^{2}ASR (Dynamic)	Wav2Vec2
English	1489	39.4%	38.8%	31.9%	29.3%	**28.5%** (−27.6%)	30.8%
French	413	35.2%	34.6%	29.2%	26.1%	**25.2%** (−28.4%)	27.9%
German	539	36.9%	36.4%	30.2%	27.3%	**26.4%** (−28.4%)	28.2%
Spanish	222	32.7%	31.8%	22.4%	19.2%	**18.3%** (−44.0%)	20.3%
Catalan	296	30.5%	29.1%	19.5%	18.0%	**17.2%** (−43.6%)	18.7%
Italian	123	48.2%	47.8%	36.7%	34.5%	**32.6%** (−32.3%)	35.8%
Russian	77	51.9%	51.6%	40.3%	38.7%	**37.1%** (−28.5%)	39.5%
Average	451.2	39.3%	38.5%	30.0%	27.5%	**26.4%** (−32.8%)	28.7%

Table 3: *WERs of English after translation from different languages under various training strategies on CoVoST 2 dataset. The percentage in parentheses denotes the relative WER reduction over TAMST.*

Language → English	Hours of training Data	TAMST	PT+FT	M^{2}ASR (Static)	M^{2}ASR (Dynamic)	Wav2Vec2
French → English	180	39.8%	33.1%	31.5%	**30.4%** (−23.6%)	32.4%
German → English	119	39.5%	35.6%	33.8%	**33.1%** (−16.2%)	34.8%
Spanish → English	97	38.4%	25.2%	23.9%	**22.9%** (−40.3%)	24.3%
Catalan → English	81	36.2%	26.8%	24.7%	**24.2%** (−33.1%)	25.6%
Italian → English	28	55.9%	40.3%	38.0%	**37.2%** (−33.4%)	38.6%
Russian → English	16	57.3%	48.9%	44.7%	**43.6%** (−23.9%)	45.7%
Average	86.8	44.5%	34.9%	32.7%	**31.9%** (−28.3%)	33.5%

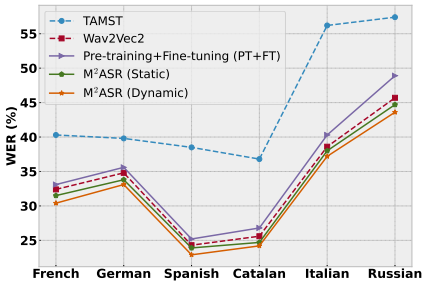

Figure 3: *WERs of high-resource data in S2TT task.*

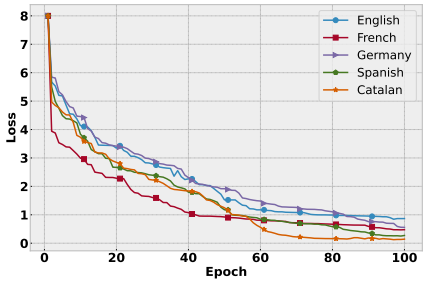

Figure 4: *CTC training curve of ASR task for 5 languages.*

English). We only explore this translation direction in this paper to conduct controlled experiments, and will leave more cross-language translation in the future. We categorize languages into high-resource and low-resource languages based on the data size. If a language has more than 10 hours of training data, we classify it as a high-resource language. Even though we carry out experiments using all language data, given the limited space, we only report WERs on high-resource languages using ASR and S2TT heads in Table 2 and Table 3, respectively.

In Table 2, we compare the WERs of the ASR task among multiple training baselines. MST shows an average WER of 39.3%, while incorporating S2TT as an augmentation technique improves the average WER by 0.8% absolute. The improvement is consistent across all languages. This shows that cross-language translation can help to learn more robust and universal acoustic representations. The PT+FT method demonstrates further performance improvements over TAMST, with an overall 22.0% reduction in WER compared to TAMST. Compared to the conventional PT+FT method, our proposed M^{2}ASR (Static) method exhibits a noteworthy 8.3% reduction in WER. Additionally, employing dynamic weights in M^{2}ASR further improves the WER. M^{2}ASR (Dynamic) achieves an impressive 2.4% reduction in WER compared to the M^{2}ASR (Static) method. We also compare M^{2}ASR (Dynamic) with Wav2Vec2, where M^{2}ASR (Dynamic) shows 8.0% WER reduction over Wav2Vec2.

We also report WERs on English after translation from various languages in Table 3 since the S2TT heads represent English token outputs with various input languages. In the S2TT task, TAMST exhibits a WER of 44.5% on average. Furthermore, the PT+FT method significantly improves over TAMST, achieving a 21.5% WER reduction. Our proposed M^{2}ASR (Static) method also shows promising results, with a notable 6.3% reduction in WER compared to the conventional PT+FT method. Additionally, utilizing dynamic weights in M^{2}ASR leads to further improvements in WER. Specifically, M^{2}ASR (Dynamic) achieves an impressive 2.5% reduction in WER compared to M^{2}ASR (Static). Moreover, when comparing M^{2}ASR (Dynamic) with Wav2Vec2, M^{2}ASR (Dynamic) outperforms Wav2Vec2, achieving a 4.8% reduction in WER. This can also be observed from Figures 2 which visually demonstrates that joint training with MOO substantially improves WERs. We also show the training CTC loss curve for the high-resource languages in Figure 4.

5. Conclusions

In conclusion, we introduced M^{2}ASR as a novel and comprehensive approach that integrates unsupervised, supervised learning, multilingual, and multitask ASR via a single multi-objective optimization framework. Experimental results across multiple datasets consistently demonstrate the superior performance of M^{2}ASR in multilingual, multitask learning scenarios, which underscore the utility of explicitly using multi-objective learning techniques in multilingual and multitask ASR.

6. References

[1] T. Schultz and K. Kirchhoff, *Multilingual speech processing*. Elsevier, 2006.

[2] H. Bourlard, J. Dines, M. Magimai-Doss, P. N. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, and F. Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, pp. 885–915, 2011.

[3] Y. J. Kim, A. A. Awan, A. Muzio, A. F. C. Salinas, L. Lu, A. Hendy, S. Rajbhandari, Y. He, and H. H. Awadalla, "Scalable and efficient moe training for multitask multilingual models," *arXiv preprint arXiv:2109.10465*, 2021.

[4] J. Fu, S.-K. Ng, and P. Liu, "Polyglot prompt: Multilingual multitask promptraining," *arXiv preprint arXiv:2204.14264*, 2022.

[5] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv preprint arXiv:2007.10310*, 2020.

[6] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[7] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, and D. Povey, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *2010 IEEE international conference on acoustics, speech and signal processing*, 2010, pp. 4334–4337.

[10] M. A. Di Gangi, M. Negri, R. Cattoni, R. Dessi, and M. Turchi, "Enhancing transformer for end-to-end speech-to-text translation," in *Proceedings of Machine Translation Summit XVII: Research Track*, 2019, pp. 21–31.

[11] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 293–11 302.

[12] X. Chen, S. Wang, B. Fu, M. Long, and J. Wang, "Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[13] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource lvcsr systems," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[14] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7349–7353.

[15] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 7319–7323.

[16] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 4904–4908.

[17] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[18] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[19] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "mslam: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[20] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," *arXiv preprint arXiv:1802.06655*, 2018.

[21] S.-P. Chuang, T.-W. Sung, A. H. Liu, and H.-y. Lee, "Worse wer, but better bleu? leveraging word embedding as intermediate in multitask end-to-end speech translation," *arXiv preprint arXiv:2005.10678*, 2020.

[22] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Attention-passing models for robust and data-efficient end-to-end speech translation," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 313–325, 2019.

[23] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation," *arXiv preprint arXiv:2011.00747*, 2020.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[25] L. Chen, H. Fernando, Y. Ying, and T. Chen, "Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance," *arXiv preprint arXiv:2305.20057*, 2023.

[26] D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)*. Athena Scientific, 1996.

[27] A. Saif, X. Cui, H. Shen, S. Lu, B. Kingsbury, and T. Chen, "Joint unsupervised and supervised training for automatic speech recognition via bilevel optimization," *arXiv preprint arXiv:2401.06980*, 2024.

[28] V. Kurin, A. De Palma, I. Kostrikov, S. Whiteson, and P. K. Mudigonda, "In defense of the unitary scalarization for deep multi-task learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 169–12 183, 2022.

[29] J.-A. Désidéri, "Multiple-gradient descent algorithm (mgda) for multiobjective optimization," *Comptes Rendus Mathematique*, vol. 350, no. 5-6, pp. 313–318, 2012.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[31] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, 2017, pp. 1–5.

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, 2019.

[34] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[35] H. Yadav and S. Sitaram, "A survey of multilingual models for automatic speech recognition," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 5071–5079.