Improving Noisy Student Training for Low-Resource Languages in End-to-End ASR Using CycleGAN and Inter-Domain Losses

Chia-Yu Li and Ngoc Thang Vu
Institute for Natural Language Processing (IMS), University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
licu@ims.uni-stuttgart.de, thang.vu@ims.uni-stuttgart.de

Abstract

Training a semi-supervised end-to-end speech recognition system using noisy student training has significantly improved performance. However, this approach requires a substantial amount of paired speech-text and unlabeled speech, which is costly for low-resource languages. Therefore, this paper considers a more extreme case of semi-supervised end-to-end automatic speech recognition where there are limited paired speech-text, unlabeled speech (less than five hours), and abundant external text. Firstly, we observe improved performance by training the model using our previous work on semi-supervised learning "CycleGAN and inter-domain losses" solely with external text. Secondly, we enhance "CycleGAN and inter-domain losses" by incorporating automatic hyperparameter tuning, calling "enhanced CycleGAN inter-domain losses." Thirdly, we integrate it into the noisy student training approach pipeline for low-resource scenarios. Our experimental results, conducted on six non-English languages from Voxforge and Common Voice, show a 20% word error rate reduction compared to the baseline teacher model and a 10% word error rate reduction compared to the baseline best student model, highlighting the significant improvements achieved through our proposed method.

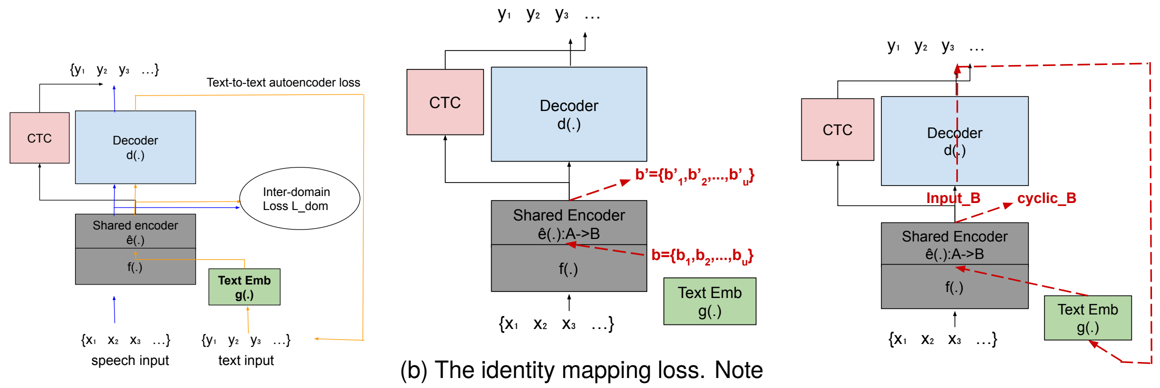
Keywords: speech recognition, low resource, semi-supervised training, CycleGAN, noisy student training

1. Introduction

Over the last decade, there has been a significant improvement in the performance of speech and language processing technologies, with an increasing number of systems being deployed across multiple languages and applications. However, the majority of these efforts have been focused on a limited set of languages. Given that there are over 6,900 languages worldwide, the biggest challenge today is to quickly and cost-effectively transfer speech processing systems to new languages with minimal manual effort. In the field of automatic speech recognition (ASR), semi-supervised end-to-end (E2E) can be applied to reduce the amount of annotated data. Two prominent approaches include consistency-based and iterative self-training-based methods. The consistency-based method focuses on enhancing the model by improving the representation of input through training a separate task (Tjandra et al., 2017; Hayashi et al., 2018; Renduchintala et al., 2018; Karita et al., 2018; Hsu and Glass, 2018; Chung and Glass, 2018; Chorowski et al., 2019; Hori et al., 2019; Schneider et al., 2019; Baevski et al., 2019; Ling et al., 2020). The iterative self-training technique utilizes augmentation to improve the overall network performance (Zavaliagkos et al., 1998; Novotney and Schwartz, 1998; Thomas et al., 2013; Parthasarathi and Strom, 2019; Li et al., 2019; Kahn et al., 2020a; Synnaeve et al., 2020; Hsu et al., 2022). Among the various techniques, a widely recognized approach known as noisy student training (NST) has emerged. NST is an iterative self-training method that leverages unlabeled data to enhance accuracy, particularly in the domains of image classification and machine translation (Xie et al., 2020). Park et al. adapted and improved NST by employing techniques such as SpecAugment (Park et al., 2019a,b) and incorporating shallow fusion with a language model (LM) into the teacher network. Additionally, they introduced a normalized filtering score that aids in generating enhanced transcripts for training the student network (Park et al., 2020). The results demonstrate significant performance on Librispeech (Panayotov et al., 2015) and LibriLight (Kahn et al., 2020b).

Although NST is simple and effective, it depends on a substantial quantity of paired speech-text to train a teacher model, which is used for labeling the unlabeled speech data that the student model could train on. For low-resource languages, the paired speech-text is expensive. There are techniques that can be explored to address this limitation. One approach is to leverage pre-trained models, such as wav2vec (Schneider et al., 2019), where leverages transfer learning to learn contextual representations from a large corpus of unlabeled speech data. The model can then be fine-tuned for the target domain using unlabeled speech data from the same target domain. However, this approach still requires a reasonable quantity of speech data, which is still expensive in low-resource scenario. Besides, this technique requires multi-stage tuning processing which introduces computational cost. How to improve inexpensively the teacher model

SIGUL2024 Workshop, pages 133–142
21-22 May, 2024.



(a) The architecture of semi-supervised E2E ASR.

(b) The identity mapping loss. Note that b is the representations from encode speech or text.

(c) The cycle-consistent inter-domain loss.

Figure 1: The framework of CycleGAN and inter-domain losses (Li and Vu, 2022).

in NST remains a key challenge especially in language with very small data.

Our previous work "cycle-consistent generative adversarial networks (CycleGAN) and inter-domain losses", which is the dissimilarity between the intermediate representations of encoded speech and its hypothesis (Li and Vu, 2022), was proposed for semi-supervised E2E ASR. The architecture is shown in Figure 1a. CycleGAN and inter-domain losses (CID) encourage the model to learn the common representations from the speech and text. With the advantage of this structure allowing speech and text input, we observe that training a model by CID with small paired speech-text and additional external text (without additional speech) can still improve the ASR performance. Therefore, we propose leveraging it into the training pipeline of NST to enhance the teacher model solely using a large amount of external text. Subsequently, the improved teacher model generates better labels for the unlabeled speech, which the student model can train on.

In this paper, we make several contributions in the following aspects: Firstly, we observe that training a model by CID (Li and Vu, 2022) with lots of external text significantly boosts performance (subsection 2.2); Secondly, we enhance CID by incorporating automatic hyperparameter tuning, calling enhanced CID (subsection 2.3); Thirdly, we improve the NST training pipeline for low-resource scenarios by boosting the teacher model using enhanced CID (subsection 2.4); Fourthly, we evaluate our method on six languages on the Voxforge and Common Voice (section 3 and section 4). The results demonstrate that our proposed approach achieves a 20% word error rate reduction (WERR) compared to the baseline (NST) teacher model, and a 10% WERR compared to the baseline student model for most languages. Notably, the improvement of teacher model is accomplished without the need for additional speech data. Lastly, we provide an analysis of the recognition output and cherry-pick hypothesis (section 5).

For the sake of simplicity, throughout the rest of this paper, we use the term "paired data" to refer to "paired speech-text," the term "unpaired data" to refer to "unpaired speech-text," the term "CID" to refer to the "CycleGAN and inter-domain" approach, and our proposed NST pipeline designed for low-resource using CID is denoted as "cNST".

2. Method

2.1. CycleGAN and Inter-Domain Losses (CID)

Figure 1a shows the CID architecture, which is based on semi-supervised E2E speech recognition and joint CTC-attention E2E (Kim et al., 2017; Watanabe et al., 2017; Karita et al., 2018). The encoder is $e = \hat{e} \circ f$ when the input is speech. If the input is text, the encoder is the composition of text embedding $g(.)$ and the share encoder $\hat{e}$. i.e., $\hat{e} \circ g$. The model is trained by jointly CTC-attention objective on paired data $S = \{X, Y\}$ and by CID on unpaired data $U = \{X', Y'\}$ simultaneously. The objective is as follows (Karita et al., 2018; Li and Vu, 2022),

$$\mathcal{L} = \alpha \mathcal{L}_{pair}(e, d, S) + (1 - \alpha)\mathcal{L}_{unpair}(f, g, \hat{e}, d, U) \tag{1}$$

where the supervised ratio α is a tunable parameter.

The supervised objective is negative log likelihood of the ground-truth y given the encoded speech $e(x)$ (Watanabe et al., 2017):

$$\begin{aligned}\mathcal{L}_{pair}(e, d, S) &= -\sum_{(x,y)\in S} \log d(e(x)) \\ &= -\sum_{(x,y)\in S} \log \prod_{t=1}^{|y|} \Pr(y_t | y_{t-1}, e(x))\end{aligned} \tag{2}$$

Model	**paired data**	**unpaired text** (#lines)	**without LM** WER(%)	**with LM** WER(%)
Initial model (M_0)	Voxforge German (5 hrs.)	0	63.6	63.1
CID model (M_1)	Voxforge German (5 hrs.)	10K (Goldhahn et al., 2012)	38.6	36.3
	Voxforge German (5 hrs.)	100K (Goldhahn et al., 2012)	31.2	29.4
	Voxforge German (5 hrs.)	300K (Goldhahn et al., 2012)	**30.8**	**29.1**

Table 1: WERs on the Voxforg German test set. Note that the initial model is trained by supervised objective in Equation 2 with five-hour Voxforg German train data, and the CID model (M_1) is trained with same five-hour Voxforg German train data and external text from Leipzig corpus (Goldhahn et al., 2012) via semi-supervised objective in Equation 1.

The unsupervised objective CID consists of the identity mapping loss, the cycle-consistent inter-domain loss, and the text-to-text autoencoder loss with tunable hyperparameter speech-to-text ratio $\beta \in [0, 1]$ (Li and Vu, 2022),

$$\begin{aligned}\mathcal{L}_{unpair}(f, g, \hat{e}, d, U) = \mathcal{L}_{idt}(f, g, \hat{e}, U) \\ + \beta * \mathcal{L}_{cyc,dom}(f, g, \hat{e}, d, U) \\ + (1 - \beta) * \mathcal{L}_{text}(g, \hat{e}, d, U)\end{aligned} \quad (3)$$

The identity loss enhances the shared encoder $\hat{e}(.)$ to preserves important features after translation. The computation of loss in Figure 1b is as follows,

$$L_{idt} = \|\hat{e}(b) - b\|_1 \quad (4)$$

where the representation is coming from speech $b = f(x)$ or text $b = g(y)$.

The cycle-consistent inter-domain loss is the dissimilarity between the representations of encoded speech and its hypothesis, which aims to let networks learn common knowledge from speech and text. The illustration of loss is shown in Figure 1c and the definition is as follows,

$$\begin{aligned}L_{cyc,dom} &= \mathcal{D}(input\_B, cycle\_B) \\ &= \mathcal{D}(e(x), \hat{e}(g(d(e(x)))))\end{aligned} \quad (5)$$

where $\mathcal{D}(.)$ is a distance measure of the distributions. In our previous work, we use Maximum Mean Discrepancy (MMD) because it achieves the best result (Li and Vu, 2022).

The text-to-text autoencoder loss measures a negative log-likelihood that the encoder-decoder network can reconstruct text from unpaired text (Hinton and Salakhutdinov, 2006; Karita et al., 2018), see the orange line in Figure 1a. The loss is defined as follows,

$$L_{text} = -\sum \log \Pr(y|\hat{e}(g(y))) \quad (6)$$

2.2. CID Solely with External Text

In low-resource settings, acquiring paired data or speech data can be costly. Therefore, this section focus on enhancing the model inexpensively. In our previous work (Li and Vu, 2022), we train model by CID with an equal amount of unlabeled speech and text. However, training a model by CID without additional unlabeled speech and with only external text (i.e., $U = \{X, Y'\}$) might still gain performance improvements. To validate this hypothesis, Table 1 presents the evaluation of models on Voxforge German test set. These models are trained by jointly CTC-attention objective on paired data $S = \{X, Y\}$ and by CID on speech from paired data and text from Leipzig German corpus (Goldhahn et al., 2012) $U = \{X, Y'\}$ simultaneously. The results demonstrate that CID models trained with 10K/100K/300K lines of external text improve WERs from 63.6% to 38.6/31.2/30.8% without involving a language model. Moreover, when evaluated with a language model, the CID model improves WERs from 63.1% to 36.3/29.4/29.1%. These findings highlight the effectiveness of incorporating CID with external text to enhance the performance of E2E model. It also indicates that the CID allows text to benefit not only the language model (LM) but also the encoder-decoder model.

2.3. Enhanced CID by Incorporating Automatic Hyperparameter Tuning

Although the CID model achieves a significant reduction in character error rate (CERR) across English datasets, WSJ and Librispeech, as well as low supervision non-English datasets (Voxforge) (Li and Vu, 2022), it requires effort to tune the two hyperparameters, the supervised ratio α and the speech-to-text ratio β, for each dataset. To streamline the training pipeline, we propose using supervised ratio decay and automatic speech-to-text ratio tuning by performing an operation on the unsupervised losses with all the possible values for the speech-to-text ratio during the training. The details are as follows: Firstly, we suggest that the model obtains lots of guidance from the supervision data at the early stages of training. Therefore, α starts at 0.9 for the first three epochs and gradually decays after three epochs until the training is completed, which enables the model to explore the

Model	supervised ratio α	adapted Equation 3 $\mathcal{L}_{unpair}$	CER(%)
Baseline(Li and Vu, 2022)			46.9
MIN-UNPAIR-LOSS	0.5	$\min_{\beta \in \{0,0.1,0.2,\ldots,1.0\}} \mathcal{L}_{unpair}$	30.6
MAX-UNPAIR-LOSS	0.5	$\max_{\beta \in \{0,0.1,0.2,\ldots,1.0\}} \mathcal{L}_{unpair}$	39.5
AVG-UNPAIR-LOSS	0.5	$\overline{\mathcal{L}_{unpair}}$	50.6
MED-UNPAIR-LOSS	0.5	Median($\mathcal{L}_{unpair}$)	50.4
DECAY-MIN-UNPAIR-LOSS	decay	$\min_{\beta \in \{0,0.1,0.2,\ldots,1.0\}} \mathcal{L}_{unpair}$	29.6
DECAY-MAX-UNPAIR-LOSS	decay	$\max_{\beta \in \{0,0.1,0.2,\ldots,1.0\}} \mathcal{L}_{unpair}$	44.1
DECAY-AVG-UNPAIR-LOSS	decay	$\overline{\mathcal{L}_{unpair}}$	46.6
DECAY-MED-UNPAIR-LOSS	decay	Median($\mathcal{L}_{unpair}$)	30.3

Table 2: This table compares the CERs on the Common Voice Finnish test set of models with or without (1) the supervised ratio decay and (2) automatic speech-to-text ratio tuning. We also observe the same conclusion in six languages test sets from Common Voice and Voxforge.

unpaired data with increased flexibility. Secondly, we integrate the speech-to-text ratio into the training process, we propose to use minimal, maximal, average, or median operations on the unsupervised losses with β from 0.0 to 1.0. Table 2 shows our proposed adapted unsupervised losses and the corresponding CERs on the Common Voice Finnish test set. This table reveals that the model using minimal operation outperforms the ones using other operations and baseline. The best model is the model using the supervised ratio decays and minimal operations on the unsupervised losses over β. We observe the same conclusion in six languages from Common Voice and Voxforge. Figure 2 and Figure 3 present the training loss and the accuracy of baseline and models trained by our adapted objective in Table 2. The model using minimal operation on unsupervised loss performs stable and improved accuracy during the training, whereas the baseline and other models using maximum, average, and median operations produce mismatched training loss and validated loss, as well as fluctuating model accuracy during the training. These figures resonated with the result from the Table 2, the model trained by Equation 1 using supervised ratio decay and performing minimal operation on unsupervised loss achieves the best performance.

2.4. Noisy Student Training with CycleGAN and Inter-Domain Losses (cNST) for Low-Resource Languages

NST for speech recognition is effective when sufficient paired data is available. However, the paired data and unlabeled speech are often limited in a low-resource setting. That leads to a low performance teacher model, which generates low-quality labels for unlabeled speech; the training for the student model can be severely affected, resulting in inefficient training.

We aim to improve the teacher model with little effort and less cost regarding time and finances. subsection 2.2 demonstrates that the model can be improved by CID solely with external text. Therefore, we propose to exploit the enhanced CID in subsection 2.3 and external text to improve the teacher model. A LM is also trained with the in-domain and external text $\{Y, Y'\}$. The NST algorithm is revised as follows,

1. Train M_0 on S using SpecAugment.

2. Train M_1 on S and $U = \{Y'\}$ by enhanced CID and using SpecAugment. Set $M = M_1$.

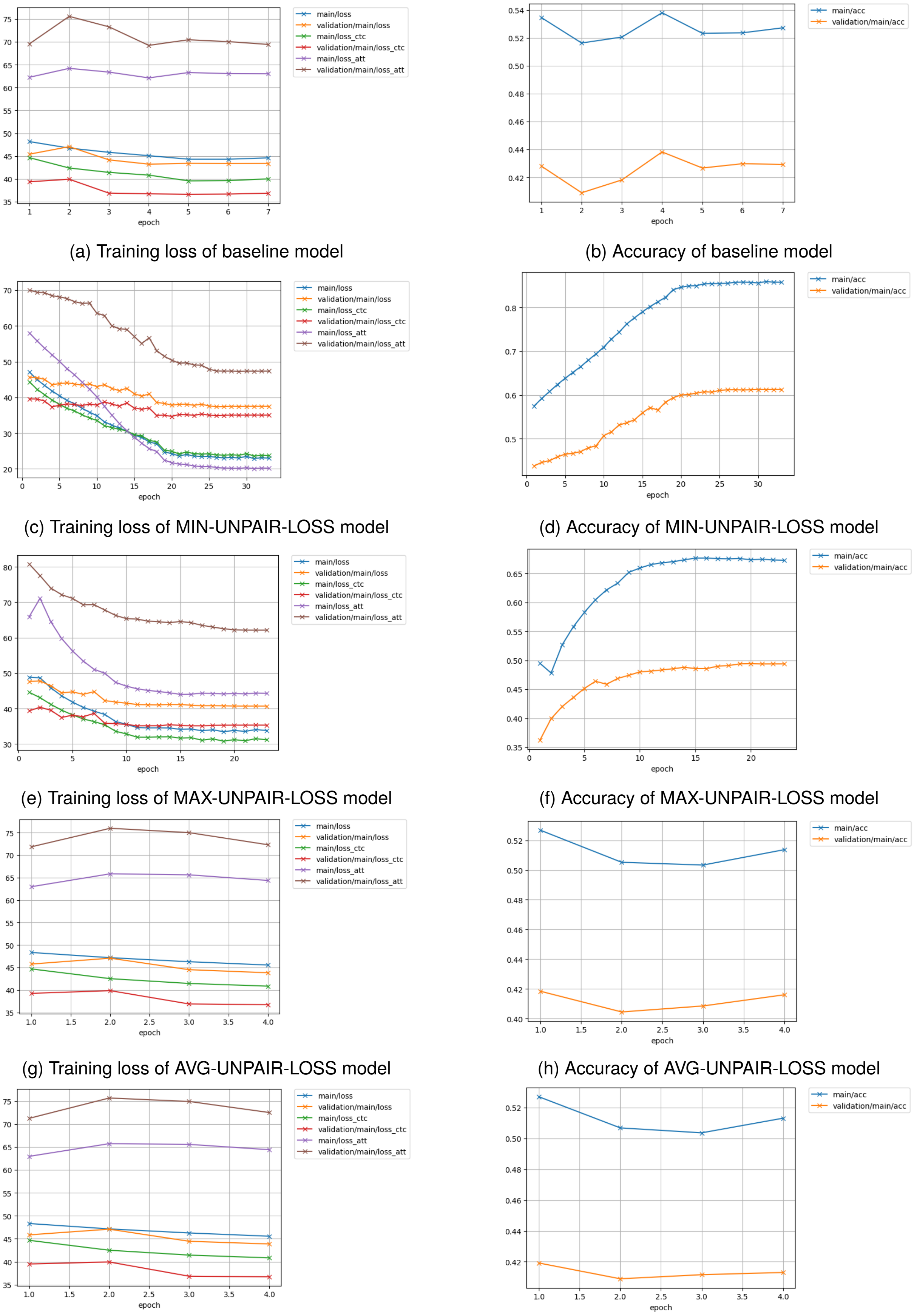
3. Fuse M with LM and measure performance.

4. Generate labelled dataset $M(X')$ with fused model.

5. Mix dataset $M(X')$ and S. Use mixed dataset to train new model M' with SpecAugment.

6. Set $M = M'$ and go to 3.

where the initial model M_0 is trained with the paired data S using SpecAugment (Park et al., 2019a), and we further re-train it at the stage 2 using the enhanced CID with external text with SpecAugment. At stage 3, the teacher model is then fused with a LM to generate labels for the unlabeled speech. Subsequently, the student model is iteratively trained with the paired and newly labeled speech data by the supervised objective. We work with small data, so it is better to utilize the available data wisely rather than removing any of it. Therefore, we simplify the NST training recipe, making it easily applicable to all languages by discarding the sophisticated filtering and balancing stages in (Park et al., 2020).



(a) Training loss of baseline model

(b) Accuracy of baseline model

(c) Training loss of MIN-UNPAIR-LOSS model

(d) Accuracy of MIN-UNPAIR-LOSS model

(e) Training loss of MAX-UNPAIR-LOSS model

(f) Accuracy of MAX-UNPAIR-LOSS model

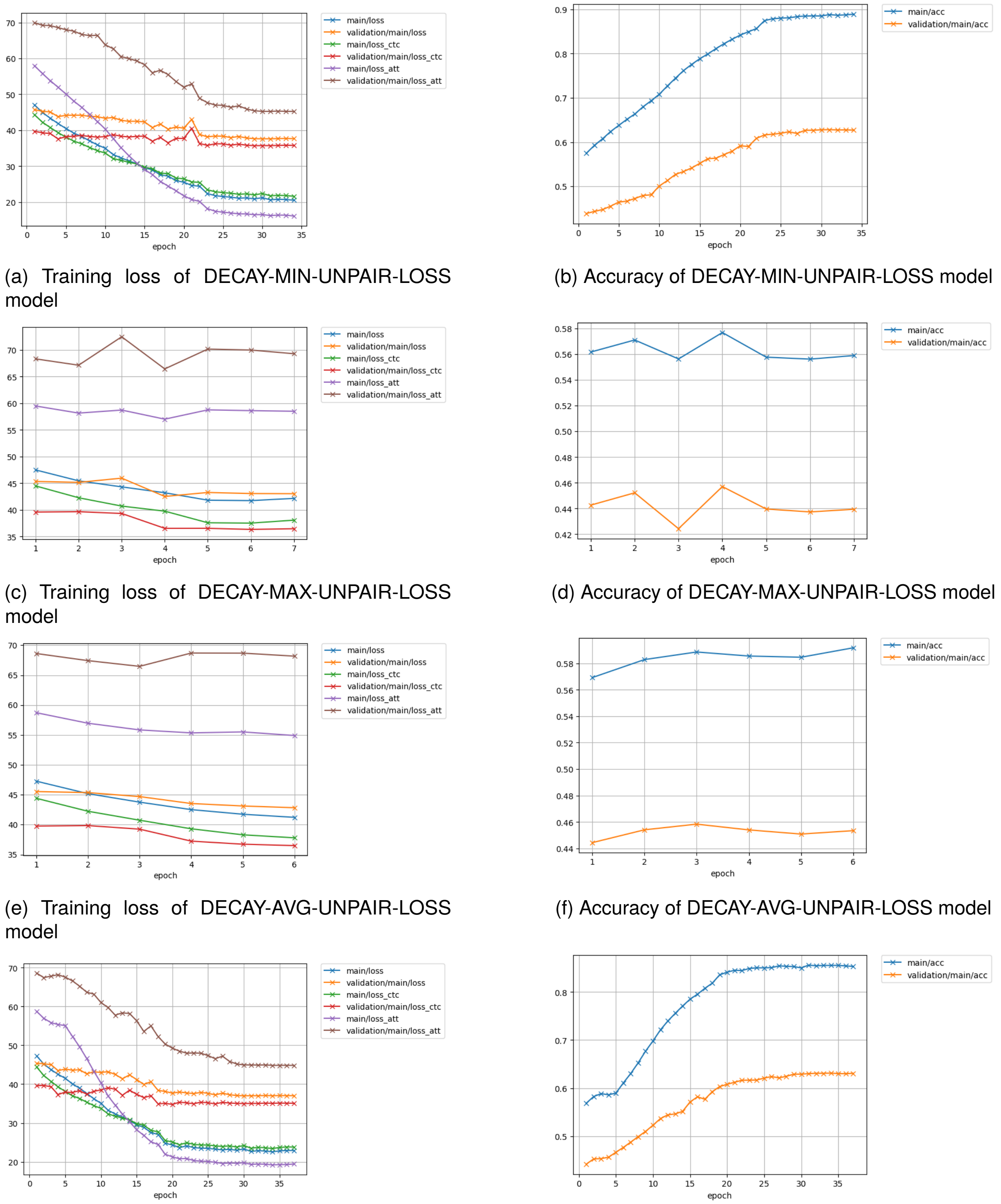
(g) Training loss of AVG-UNPAIR-LOSS model

(h) Accuracy of AVG-UNPAIR-LOSS model

(i) Training loss of MED-UNPAIR-LOSS model

(j) Accuracy of MED-UNPAIR-LOSS model

Figure 2: The training loss (left) and the accuracy (right) of models using different automatic speech-to-text ratio tuning defined in Table 2.



(a) Training loss of DECAY-MIN-UNPAIR-LOSS model

(b) Accuracy of DECAY-MIN-UNPAIR-LOSS model

(c) Training loss of DECAY-MAX-UNPAIR-LOSS model

(d) Accuracy of DECAY-MAX-UNPAIR-LOSS model

(e) Training loss of DECAY-AVG-UNPAIR-LOSS model

(f) Accuracy of DECAY-AVG-UNPAIR-LOSS model

(g) Training loss of DECAY-MED-UNPAIR-LOSS model

(h) Accuracy of DECAY-MED-UNPAIR-LOSS model

Figure 3: The training loss and accuracy of models using supervised ratio decay and different automatic speech-to-text ratio tuning defined in Table 2.

3. Experimental Setup

3.1. Dataset

Common Voice is a massively multilingual collection of transcribed speech, which is also recorded by user on Mozilla website, and recently it reaches 100 languages (Ardila et al., 2020). We conducted experiments on a subset of European languages which has limited data: Hungarian, Finnish and Greek. Additionally, we ensured that there were

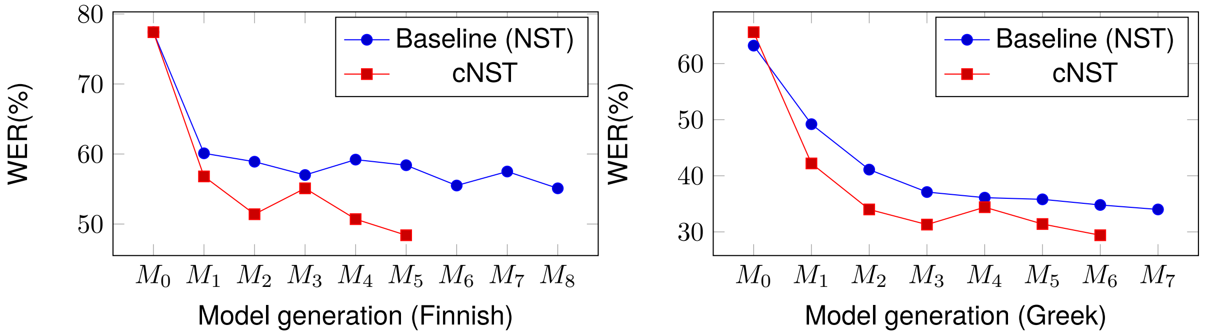


Figure 4: WERs on the Common Voice (Finnish and Greek) test set against model generations.

Model	**Voxforge (WER%)**			**Common Voice (WER%)**		
	German	Italien	Dutch	Hungarian	Finnish	Greek
Initial Model (M_0)	63.1	71.2	63.1	84.8	77.4	63.2
Baseline (NST)	49.7	47.1	58.2	72.0	55.1	34.0
Proposed Method (cNST)	**27.3**	**42.0**	**56.3**	**58.6**	**48.4**	**29.4**
WERR % (NST-cNST)/NST	45.1	10.8	3.26	18.6	12.7	13.5

Table 3: WERs comparison between baseline best student model and our proposed cNST best student model across corpus.

no overlapping sentences or speakers between the train, development and test set. The data size of train/development/test sets are in an 80:10:10 ratio and the test set contains at least two hours speech. The train set is further split to five hours paired data and the remaining portion (around three hours to five hours) is dedicated to the unlabeled speech. Voxforge consists of user submitted audio clips using their own microphone (Voxforge.org) and has eight European languages. Each language has limited size of data, ranging from approximately eight to twenty hours. In this paper, we evaluate our proposed method on German, Italian and Dutch languages. The train set is further divide into five hours paired data, while the remaining portion is dedicated to the unlabeled speech X'. The Leipzig corpus, which consists of annual collections of documents from various sources such as wikis, news, and the web (Goldhahn et al., 2012), is used as external text Y' in the experiment.

3.2. Network Architecture

The semi-supervised E2E model using CycleGAN-inter-domain losses is implemented under Espnet1 (Watanabe et al., 2018) and (Li and Vu, 2022). The model consists of three layers of Vgg (Simonyan and Zisserman, 2015) bidirectional long short-term memory with projection (Vggblstmp) encoder and attention based decoder, which is one layer long short-term memory (LSTM) with 320 units. The text embedding $g(.)$ encodes the labels over $\{Y, Y'\}$ to an one-hot vector and process it by one layer bidirectional long short-term memory (BLSTM). Byte pair encoding (BPE) (Gage, 1994; Sennrich et al., 2016) is used for some languages, some have better performance without using BPE. The input acoustic feature is 80-bin log-Mel filterbank with three pitch coefficients. For decoding, we use a beam search algorithm with beam size of 20. Our training recipe and code[1]

4. Result

4.1. WERs against Model Generation

Figure 4 shows WERs on the Common Voice (Finnish and Greek) test sets against model generations . We trained the models using our proposed algorithm cNST in subsection 2.4 and evaluated the teacher model and all the student models at different stages. Based on the observed trend in model performance, it is evident that the red line (cNST) demonstrates a steeper progression compared to the blue line (NST) from M_0 to M_1. This suggests that the enhanced CID plays a crucial role in accelerating the iterative training process and achieving better results compared to the baseline for all the model generations. Besides, red and blue lines fluctuate over the generations, which might be because the models are over-fitting on the train set, but it does not hurt the subsequent student model performance.

[1] https://github.com/chiayuli/Improved-NST-for-low-resource-language.git

Models	Hypothesis
Ground-Truth	es ist sehr beständig gegen witterungseinflüsse und insektenbefall
Initial Model	es ist sehr BESTÄNDIGEN ***** WEITEREN SPÄTEREN SECKER
Baseline(NST)	es ist sehr BESTÄNDIGEN ***** WEITEREN EINFLÜSSE *** **************
CID	es ist sehr BESTÄNDE gegen WEITERUNGSFLÜSSE und IN SEKTEN BEFALL
cNST	es ist sehr BESTÄNDE gegen WEITERUNGSEINFLÜSSE und INSEKTEN BEFALL
Ground-Truth	der anspruch ist von der Frau auf den Mann Übergegangen
Initial Model	der SPRUCH ist *** *** **** *** *** VOLLKOMMEN REGELT
Baseline(NST)	der anspruch ist *** *** **** *** *** **** FREI
CID	ER EINE SPRUCH ist von der frau auf DIE LANDEN Übergegangen
cNST	der anspruch ist von der frau auf DIE LANDEN Übergegangen
Ground-Truth	der Traffic des ersten anbieters wird zum zweiten anbieter weitergeleitet
Initial Model	der ******* *** ****** DRITTES SPÄTER NETZwerK KANN NETZwerK GELEITET
Baseline(NST)	der TRITTE IST ALS anbieters **** *** ZWEI LIETER GELEITET
CID	der TRÄFT IST ES ANBIETS werT ZU zweiten anbieter WEITER GELEITET
cNST	der TRÄFT IST ES anbieters wird ZU zweiten anbieter WEITER GELEITET

Table 4: The hypothesis of all the models on the unlabeled speech from Voxforge German. Note that the words in uppercase are incorrect compared to the ground-truth and the words in yellow means insertion.

Table 5: The WER, insertion, deletion, and substitution at word level on the Voxforge German test set. Note that all the results are with the same LM.

Models	WER(%)	INS	DEL	SUB
Initial Model	63.1	1.8	20.6	40.7
Baseline	49.7	1.0	21.0	27.9
CID	**29.4**	3.3	**4.0**	22.0
cNST	**27.3**	3.2	3.6	**20.5**

4.2. cNST Effectiveness across Corpus

Table 3 presents the performance of our proposed method, cNST, across various corpora. We examine the baseline best student model and our proposed cNST best student model on Voxforge German, Italien, Dutch and Common Voice Hungarian, Finnish Greek datasets. The result shows that cNST outperforms the baseline by achieving at least 10% WERR for most languages. Moreover, when the initial model performs poorly (above 70% WER), our proposed cNST successfully reduces the WERs to 40$\sim$50%, indicating the effectiveness of our proposed method.

5. Analysis

5.1. Recognition Output

We want to gain insights and the reasons for the improvements brought about by enhanced CID. Table Table 5 presents the WER, insertion, deletion, and substitution on the test set of Voxforge German. The initial model experiences a high number of deletion errors, which are propagated to the subsequent student models in the baseline (NST). However, with enhanced CID, the deletion errors decrease from 20.6 to 4.0. On the other hand, there is a side-effect as the insertion errors increase from 1.8 to 3.3. Overall, the subsequent student model of our proposed cNST achieve the best WER and better substitution and deletion.

5.2. Cherry-Pick Hypothesis

Some cherry-pick examples in Table 4 demonstrate that the initial model and baseline experience high deletion errors. However, the baseline exhibits a further worsening of these errors as the student model undergoes iterative training using labels that contain such errors. This observation resonates with the findings presented in Table 5. The enhanced CID model and our proposed cNST successfully reduce deletion errors. However, there is still room for improvement in terms of substitution and insertion errors. Interestingly, In the last example, if we combine both insertion words “WEITER GELEITET” to “WEITERGELEITET”, it aligns with the correct word in the reference. The issue with insertions can be attributed to inaccurate word boundary predictions from our proposed models.

6. Conclusion

We enhance the CID by incorporating automatic hyperparameter tuning and propose an improved noisy student training that leverages the enhanced CID for low-resource languages. The enhanced CID accelerates the iterative self-training process by sorely utilizing external text. The results demonstrate the effectiveness of our proposed method cNST across six non-English languages from two datasets, surpassing the baseline by 10% WER.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proc. of LREC*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. Vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv:1910.05453*.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. 2019. Unsupervised speech representation learning using WaveNet autoencoders. In *Proc. of IEEE TASLP*.

Yu-An Chung and James R. Glass. 2018. Speech2vec: A sequenceto-sequence framework for learning word embeddings from speech. In *Proc. of Interspeech*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proc. of LREC*.

Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. 2018. Back-Translation-Style Data Augmentation for End-to-End ASR. In *Proc. of SLT*.

G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. In *Science*, volume 313, page 504–507.

Takaaki Hori, Ramón Fernandez Astudillo, Tomoki Hayashi, Yu Zhang, Shinji Watanabe, and Jonathan Le Roux. 2019. Cycle-consistency training for end-to-end speech recognition. In *Proc. of ICASSP*.

Wei-Ning Hsu and James R. Glass. 2018. Extracting domain invariant features by unsupervised learning for robust automatic speech recognition. In *Proc. of ICASSP*.

Wei-Ning Hsu, Ann Lee, Gabriel Synnaeve, and Awni Y. Hannun. 2022. Self-supervised speech recognition via local prior matching. *arXiv:2002.10336*.

Jacob Kahn, Ann Lee, and Awni Y. Hannun. 2020a. Self-training for End-to-End speech recognition. In *Proc. of ICASSP*.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. 2020b. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *Proc. of ICASSP*.

Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2018. Semi-Supervised End-to-End Speech Recognition. In *Proc. of Interspeech*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *Proc. of ICASSP*.

Bo Li, Tara N. Sainath, Ruoming Pang, and Zelin Wu. 2019. Semi-supervised training for End-to-End models via weak distillation. In *Proc. of ICASSP*.

Chia-Yu Li and Thang Vu. 2022. Improving Semi-supervised End-to-end Automatic Speech Recognition using CycleGAN and Inter-domain Losses. In *Proc. of SLT*.

Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. 2020. Deep contextualized acoustic representations for semi-supervised speech recognition. In *Proc. of ICASSP*.

Scott Novotney and Richard Schwartz. 1998. Analysis of low-resource acoustic model self-training. In *Proc. of BNTUW*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proc. of ICASSP*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019a. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. of Interspeech*.

Daniel S. Park, Yu Zhang, Chung-Cheng Chiu, Youzheng Chen, Bo Li, William Chan, Quoc V. Le, and Yonghui Wu. 2019b. SpecAugment on large scale datasets. *arXiv:1912.05533*.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved Noisy Student Training for Automatic Speech Recognition. In *Proc. of Interspeech*.

Sree Hari Krishnan Parthasarathi and Nikko Strom. 2019. Lessons from building acoustic models with a million hours of speech. In *Proc. of ICASSP*.

Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe. 2018. Multimodal data augmentation for End-to-End ASR. In *Proc. of Interspeech*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised pre-training for speech recognition. In *Proc. of Interspeech*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2020. End-to-end ASR: from supervised to semi-supervised learning with modern architectures. In *Proc. of ICML*.

Samuel Thomas, Michael L. Seltzer, Kenneth Church, and Hynek Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proc. of ICASSP*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Listening while speaking: Speech chain by deep learning. In *Proc. of ASRU*.

Voxforge.org. Free speech recognition: voxforge.org. http://www.voxforge.org/. Accessed 06/25/2014.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proc. of Interspeech*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for End-to-End speech recognition. *IEEE Journal of Selected Topics in Signal Processing*.

Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proc. of CVPR*.

George Zavaliagkos, Man-Hung Siu, Thomas Colthurst, and Jayadev Billa. 1998. Using untranscribed training data to improve performance. In *Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)*.