Automatic Severity Classification of Dysarthric speech by using Self-supervised Model with Multi-task Learning

Eun Jung Yeo*, Kwanghee Choi*, Sunhee Kim, Minhwa Chung

ICASSP 2023

Contents

Motivation

Motivation

What is dysarthric speech?

1. Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.

Main challenge of dysarthric speech

Motivation

What is dysarthric speech?

1. Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.
2. People with dysarthria suffer from degraded speech intelligibility.

Main challenge of dysarthric speech

Motivation

What is dysarthric speech?

1. Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.
2. People with dysarthria suffer from degraded speech intelligibility.
3. Accurate and reliable speech assessment is essential in the clinical field.

Main challenge of dysarthric speech

Motivation

What is dysarthric speech?

1. Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.
2. People with dysarthria suffer from degraded speech intelligibility.
3. Accurate and reliable speech assessment is essential in the clinical field.

Main challenge of dysarthric speech

: Technologies related to dysarthric speech suffers from data scarcity.

Motivation

What is dysarthric speech?

1. Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.
2. People with dysarthria suffer from degraded speech intelligibility.
3. Accurate and reliable speech assessment is essential in the clinical field.

Main challenge of dysarthric speech

: Technologies related to dysarthric speech suffers from data scarcity.

→ Self-supervised pre-trained model

Motivation

What is dysarthric speech?

1. Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.
2. People with dysarthria suffer from degraded speech intelligibility.
3. Accurate and reliable speech assessment is essential in the clinical field.

Main challenge of dysarthric speech

: Technologies related to dysarthric speech suffers from data scarcity.

→ Self-supervised pre-trained model

→ Multi-Task Learning

Our method

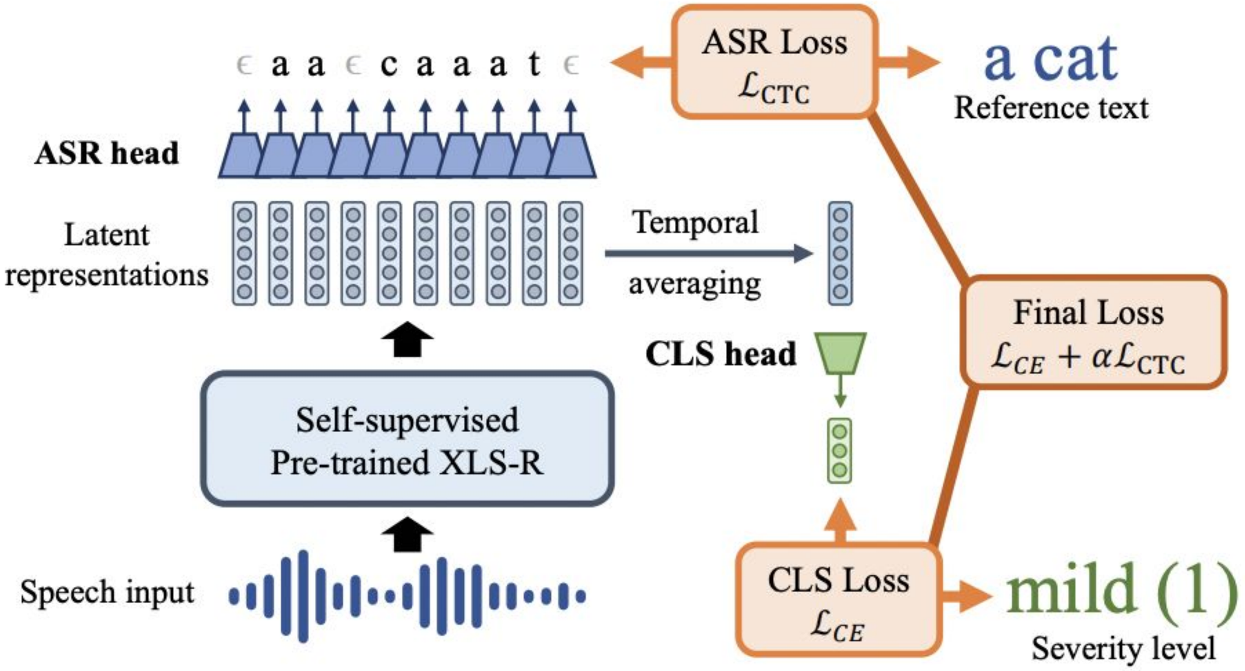


Fig. 1: Illustration of our proposed method.

Our method - Self-supervised model

Fine-tuning self-supervised models are known to solve data-scarce problems effectively.

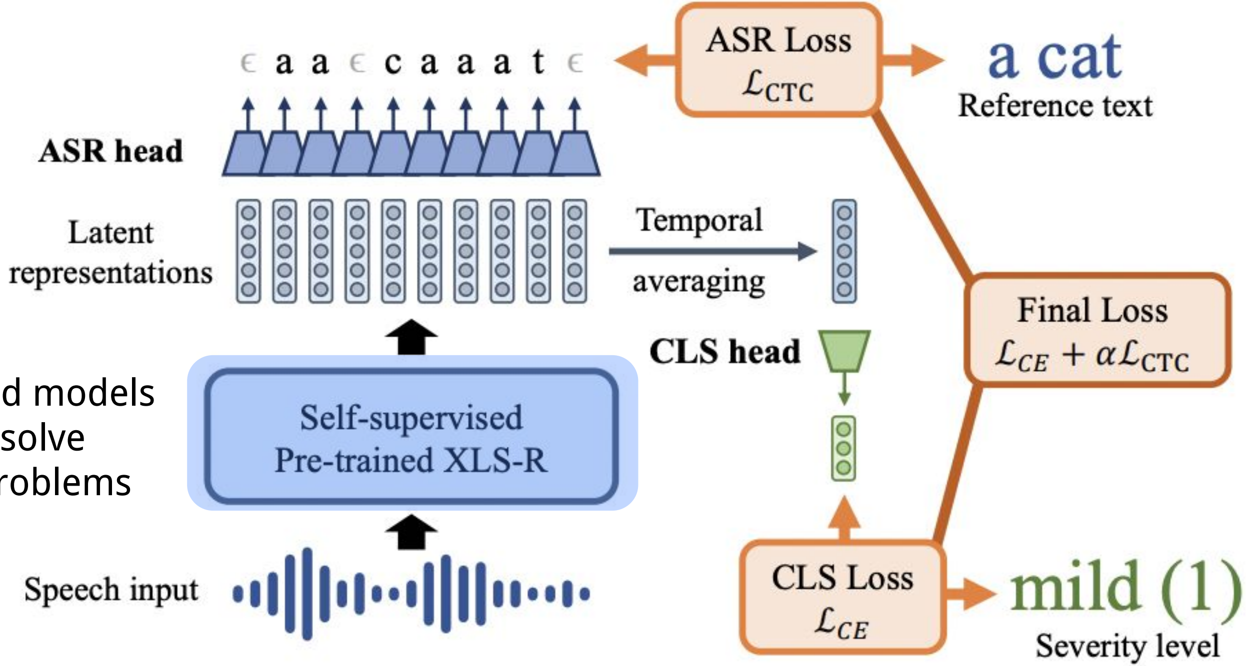


Fig. 1: Illustration of our proposed method.

Our method - Multi-Task Learning

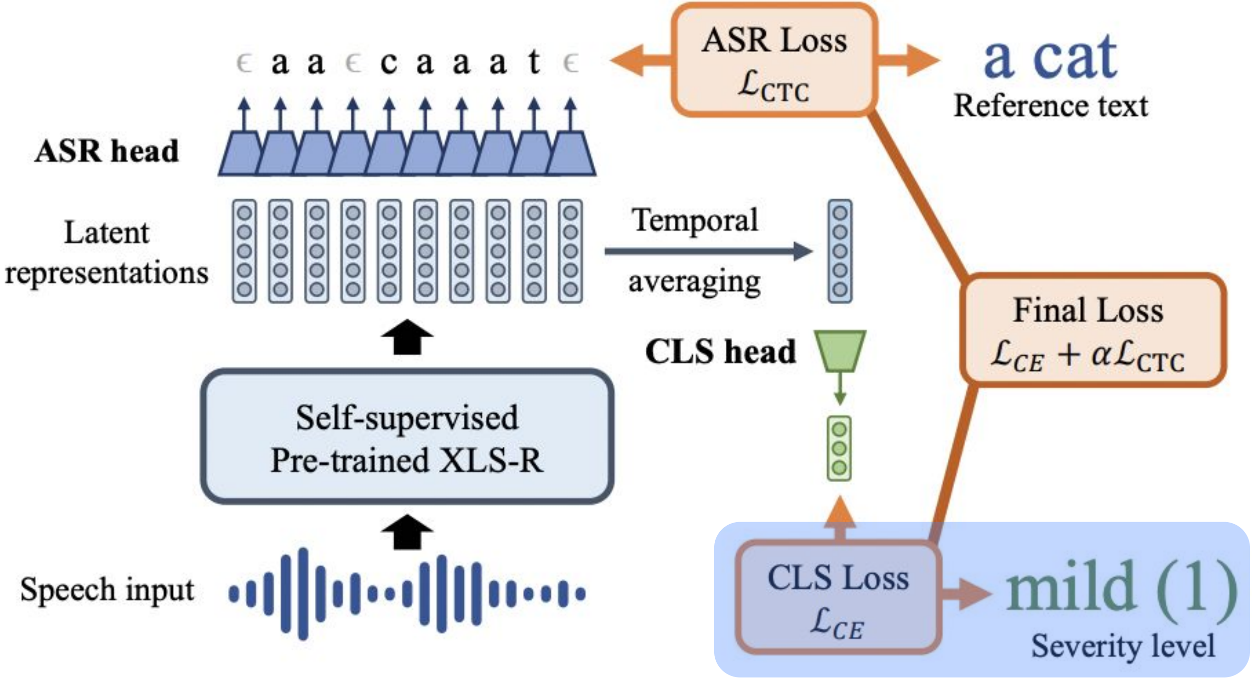


Fig. 1: Illustration of our proposed method.

Our paper's focus!
Severity classification

Our method - Multi-Task Learning

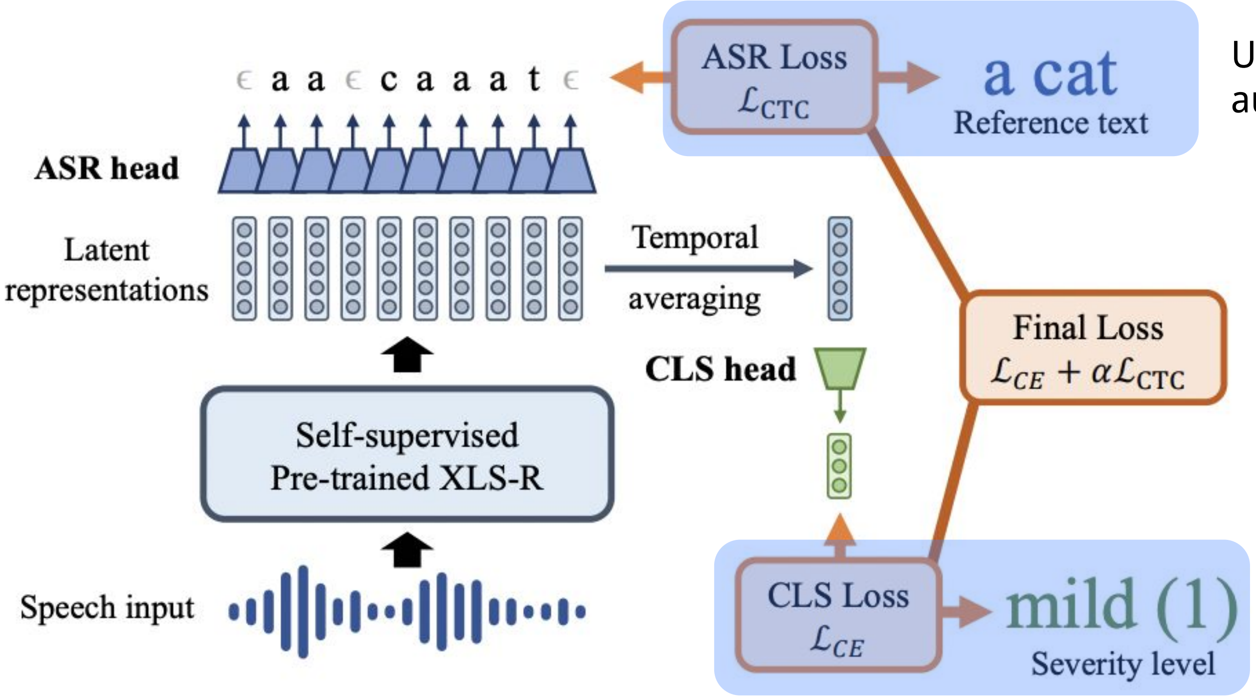


Fig. 1: Illustration of our proposed method.

Use ASR as an auxiliary task

Our paper's focus!
Severity classification

Our method - Multi-Task Learning

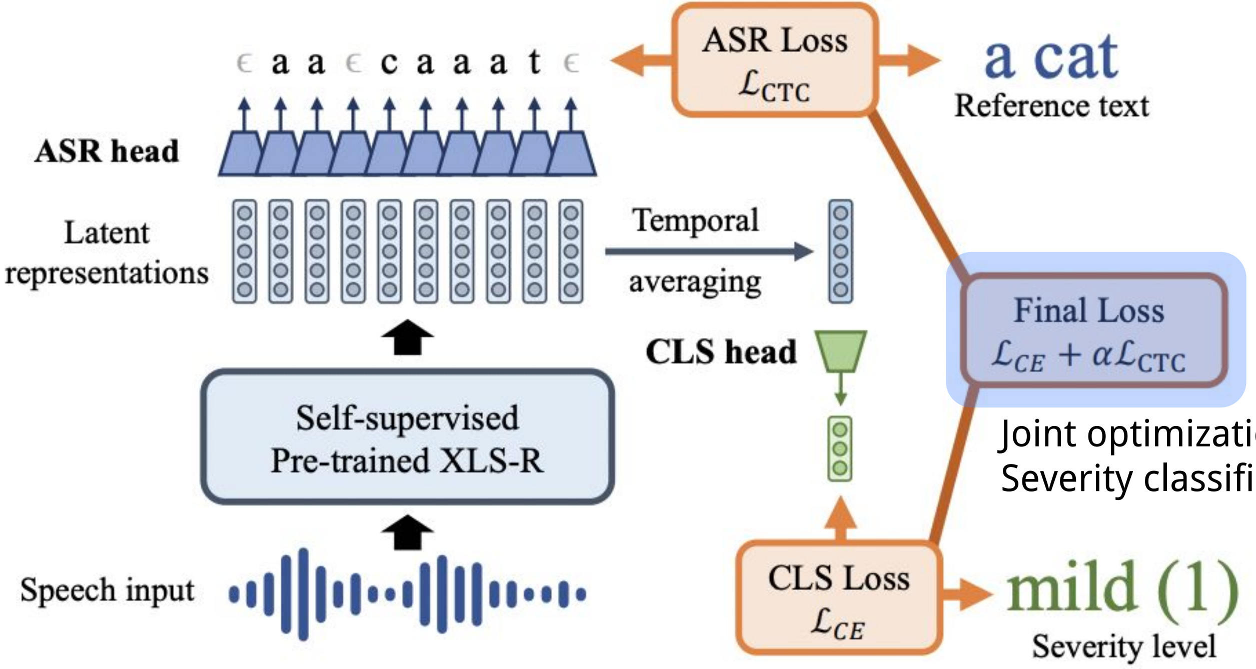


Fig. 1: Illustration of our proposed method.

Joint optimization of classification
Severity classification + ASR

Our method - Multi-Task Learning

Why should this work?

Why should this work?

1. The model is enforced to learn both acoustic and phonetic/pronunciation features for severity classification.

Why should this work?

1. The model is enforced to learn both acoustic and phonetic/pronunciation features for severity classification.
2. The auxiliary ASR task can act as a regularizer, as the model is trained to focus on two different tasks.

Why should this work?

1. The model is enforced to learn both acoustic and phonetic/pronunciation features for severity classification.
2. The auxiliary ASR task can act as a regularizer, as the model is trained to focus on two different tasks.

→ Prevents **overfitting** and yield **better performances**!

Dataset

- QoLT Korean dysarthric speech dataset
  - Speakers
    - 10 healthy speakers (5 males, 5 females)
    - 70 dysarthric speakers (45 males, 25 females)
      - 25 mild, 26 mild-to-moderate, 12 moderate-to-severe, 7 severe

Dataset

- QoLT Korean dysarthric speech dataset
  - Speakers
    - 10 healthy speakers (5 males, 5 females)
    - 70 dysarthric speakers (45 males, 25 females)
      - 25 mild, 26 mild-to-moderate, 12 moderate-to-severe, 7 severe

  - Materials
    - Each speaker recorded five sentences twice → Total of 800 utterances

Dataset

- QoLT Korean dysarthric speech dataset
  - Speakers
    - 10 healthy speakers (5 males, 5 females)
    - 70 dysarthric speakers (45 males, 25 females)
      - 25 mild, 26 mild-to-moderate, 12 moderate-to-severe, 7 severe
  - Materials
    - Each speaker recorded five sentences twice → Total of 800 utterances
  - Experiment
    - 5-way cross-validation in a speaker-independent manner.

Results

Table 1: Classification performance compared to the baselines.

Input	Classifier	Accuracy	Precision	Recall	F1-score
eGeMAPS	SVM	55.01	53.89	53.27	52.28
	MLP	50.79	44.46	48.60	46.58
	XGBoost	52.20	55.07	50.85	50.61
Hand-crafted features	SVM	61.02	64.19	63.19	62.41
	MLP	55.74	60.06	60.34	58.85
	XGBoost	55.72	61.14	56.21	56.16
eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65
	MLP	50.21	48.40	47.31	46.76
	XGBoost	56.29	62.29	56.23	56.68
Raw audio	STL	61.02	64.09	57.93	57.13
	MTL	**65.52**	**66.47**	**64.86**	**63.19**

Results

Feature-based

Table 1: Classification performance compared to the baselines.

Input	Classifier	Accuracy	Precision	Recall	F1-score
eGeMAPS	SVM	55.01	53.89	53.27	52.28
	MLP	50.79	44.46	48.60	46.58
	XGBoost	52.20	55.07	50.85	50.61
Hand-crafted features	SVM	61.02	64.19	63.19	62.41
	MLP	55.74	60.06	60.34	58.85
	XGBoost	55.72	61.14	56.21	56.16
eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65
	MLP	50.21	48.40	47.31	46.76
	XGBoost	56.29	62.29	56.23	56.68
Raw audio	STL	61.02	64.09	57.93	57.13
	MTL	**65.52**	**66.47**	**64.86**	**63.19**

Results

Table 1: Classification performance compared to the baselines.

	Input	Classifier	Accuracy	Precision	Recall	F1-score
Feature-based	eGeMAPS	SVM	55.01	53.89	53.27	52.28
		MLP	50.79	44.46	48.60	46.58
		XGBoost	52.20	55.07	50.85	50.61
	Hand-crafted features	SVM	61.02	64.19	63.19	62.41
		MLP	55.74	60.06	60.34	58.85
		XGBoost	55.72	61.14	56.21	56.16
	eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65
		MLP	50.21	48.40	47.31	46.76
		XGBoost	56.29	62.29	56.23	56.68
Self-supervised model-based	Raw audio	STL	61.02	64.09	57.93	57.13
		MTL	**65.52**	**66.47**	**64.86**	**63.19**

Results

Table 1: Classification performance compared to the baselines.

	Input	Classifier	Accuracy	Precision	Recall	F1-score	
Feature-based	eGeMAPS	SVM	55.01	53.89	53.27	52.28	
		MLP	50.79	44.46	48.60	46.58	
		XGBoost	52.20	55.07	50.85	50.61	
	Hand-crafted features	SVM	61.02	64.19	63.19	62.41	
		MLP	55.74	60.06	60.34	58.85	
		XGBoost	55.72	61.14	56.21	56.16	
	eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65	
		MLP	50.21	48.40	47.31	46.76	
		XGBoost	56.29	62.29	56.23	56.68	
Self-supervised model-based	Raw audio	STL	61.02	64.09	57.93	57.13	CLS loss only
		MTL	**65.52**	**66.47**	**64.86**	**63.19**	

Results

Table 1: Classification performance compared to the baselines.

	Input	Classifier	Accuracy	Precision	Recall	F1-score	
Feature-based	eGeMAPS	SVM	55.01	53.89	53.27	52.28	
		MLP	50.79	44.46	48.60	46.58	
		XGBoost	52.20	55.07	50.85	50.61	
	Hand-crafted features	SVM	61.02	64.19	63.19	62.41	
		MLP	55.74	60.06	60.34	58.85	
		XGBoost	55.72	61.14	56.21	56.16	
	eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65	
		MLP	50.21	48.40	47.31	46.76	
		XGBoost	56.29	62.29	56.23	56.68	
Self-supervised model-based	Raw audio	STL	61.02	64.09	57.93	57.13	CLS loss only
		MTL	**65.52**	**66.47**	**64.86**	**63.19**	CLS + ASR

Results

1. SSL > Feature-based

Table 1: Classification performance compared to the baselines.

Input	Classifier	Accuracy	Precision	Recall	F1-score
eGeMAPS	SVM	55.01	53.89	53.27	52.28
	MLP	50.79	44.46	48.60	46.58
	XGBoost	52.20	55.07	50.85	50.61
Hand-crafted features	SVM	61.02	64.19	63.19	62.41
	MLP	55.74	60.06	60.34	58.85
	XGBoost	55.72	61.14	56.21	56.16
eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65
	MLP	50.21	48.40	47.31	46.76
	XGBoost	56.29	62.29	56.23	56.68
Raw audio	STL	61.02	64.09	57.93	57.13
	MTL	**65.52**	**66.47**	**64.86**	**63.19**

Feature-based

Self-supervised model-based

CLS loss only

CLS + ASR

Results

1. SSL > Feature-based
2. CLS + ASR > CLS only

Table 1: Classification performance compared to the baselines.

Feature-based

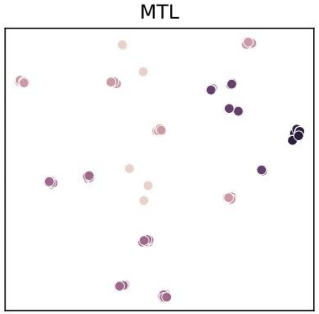
Input	Classifier	Accuracy	Precision	Recall	F1-score
eGeMAPS	SVM	55.01	53.89	53.27	52.28
	MLP	50.79	44.46	48.60	46.58
	XGBoost	52.20	55.07	50.85	50.61
Hand-crafted features	SVM	61.02	64.19	63.19	62.41
	MLP	55.74	60.06	60.34	58.85
	XGBoost	55.72	61.14	56.21	56.16
eGeMaps + Hand-crafted features	SVM	57.83	58.83	57.59	56.65
	MLP	50.21	48.40	47.31	46.76
	XGBoost	56.29	62.29	56.23	56.68
Raw audio	STL	61.02	64.09	57.93	57.13
	MTL	**65.52**	**66.47**	**64.86**	**63.19**

Self-supervised model-based

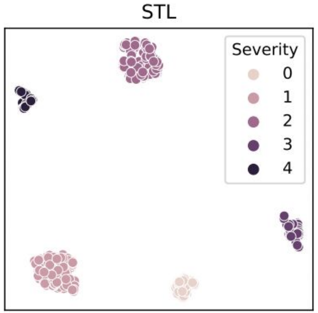
CLS loss only

CLS + ASR

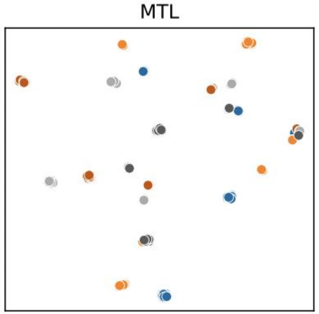
Analysis 1: Representation space visualization



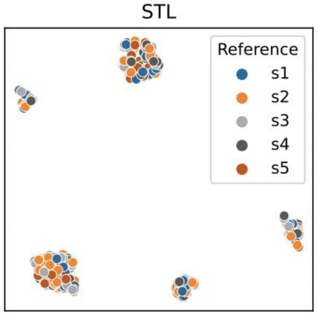
(a) Trained with MTL, color-coded with severity.



(b) Trained with STL, color-coded with severity.

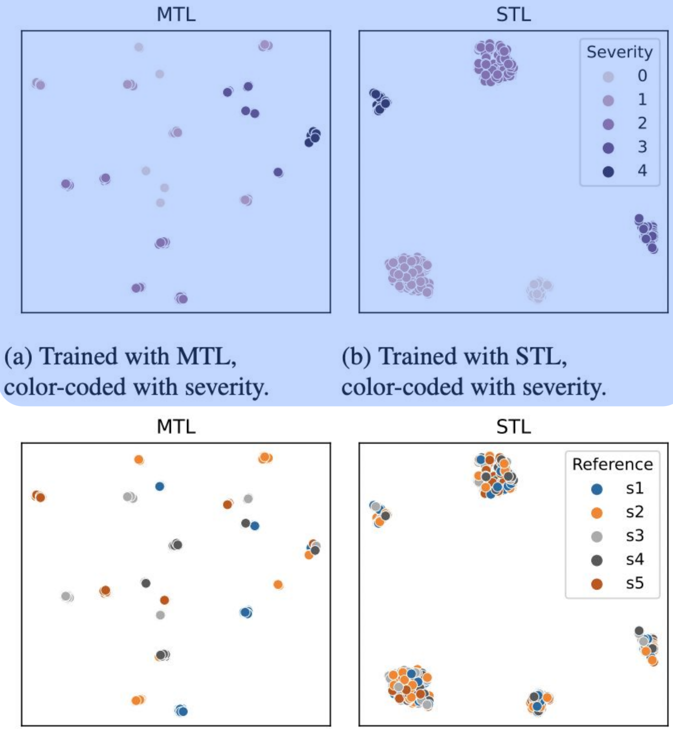


(c) Trained with MTL, color-coded with setence.



(d) Trained with STL, color-coded with setence.

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

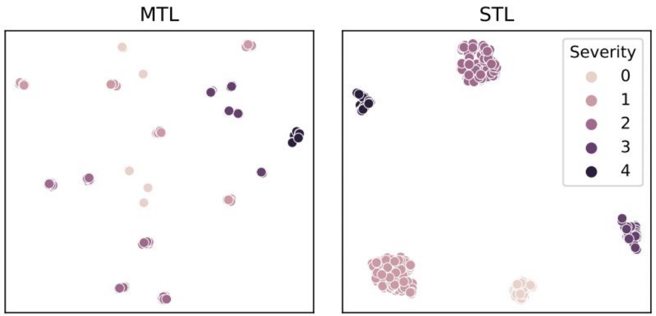
(b) Trained with STL, color-coded with severity.

(c) Trained with MTL, color-coded with setence.

(d) Trained with STL, color-coded with setence.

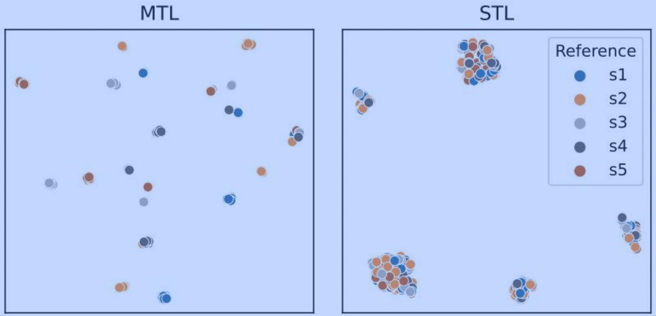
Severity

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

(b) Trained with STL, color-coded with severity.

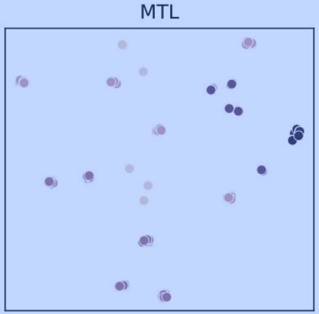


(c) Trained with MTL, color-coded with setence.

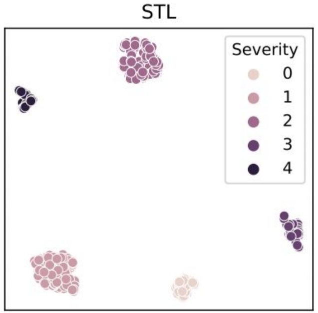
(d) Trained with STL, color-coded with setence.

Sentence

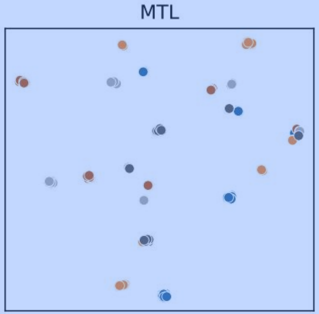
Analysis 1: Representation space visualization



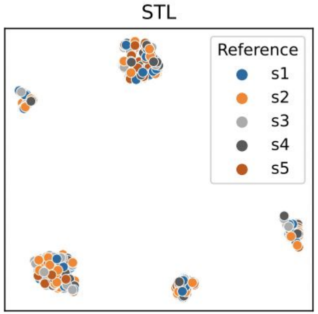
(a) Trained with MTL, color-coded with severity.



(b) Trained with STL, color-coded with severity.



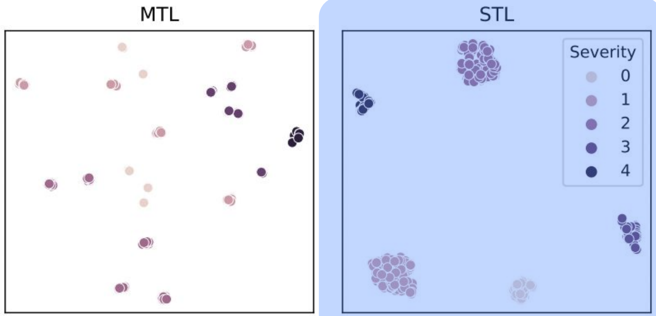
(c) Trained with MTL, color-coded with setence.



(d) Trained with STL, color-coded with setence.

CLS + ASR

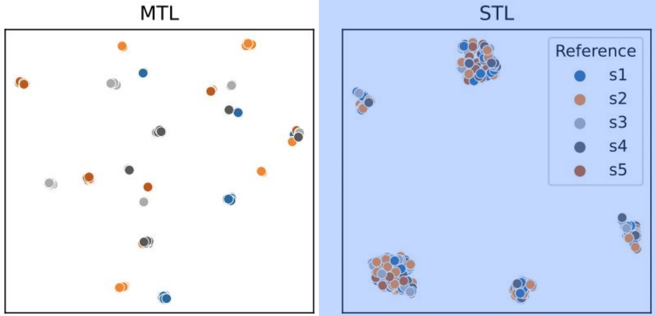
Analysis 1: Representation space visualization



(a) Trained with MTL,
color-coded with severity.

(b) Trained with STL,
color-coded with severity.

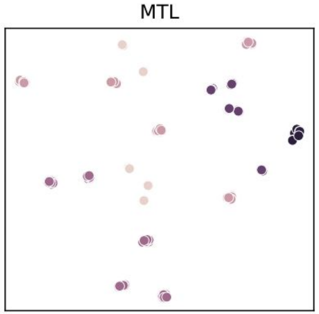
CLS Only



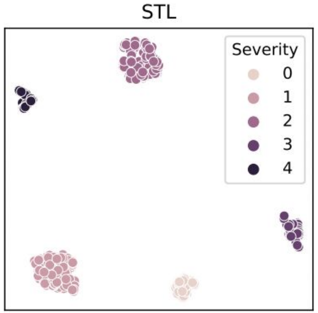
(c) Trained with MTL,
color-coded with setence.

(d) Trained with STL,
color-coded with setence.

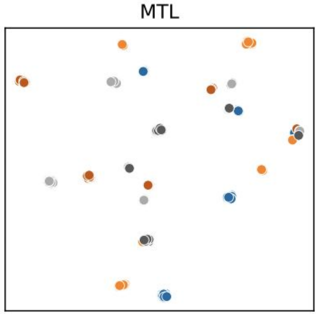
Analysis 1: Representation space visualization



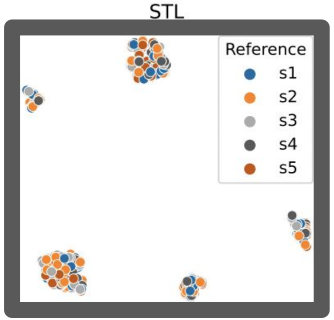
(a) Trained with MTL, color-coded with severity.



(b) Trained with STL, color-coded with severity.



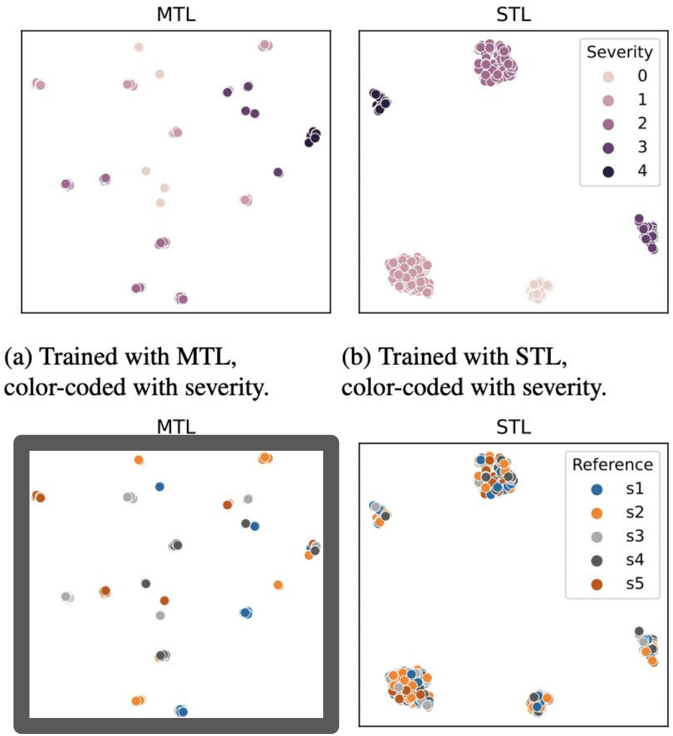
(c) Trained with MTL, color-coded with setence.



(d) Trained with STL, color-coded with setence.

1. STL cannot distinguish **different sentences**

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

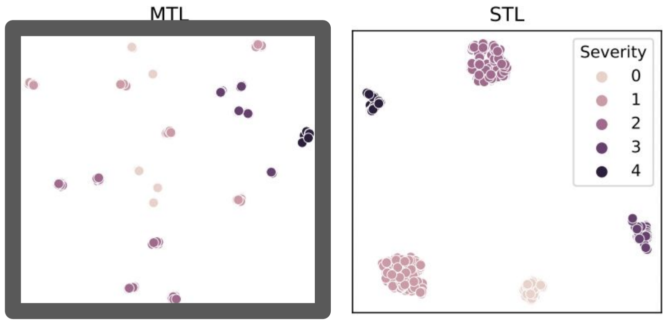
(b) Trained with STL, color-coded with severity.

(c) Trained with MTL, color-coded with setence.

(d) Trained with STL, color-coded with setence.

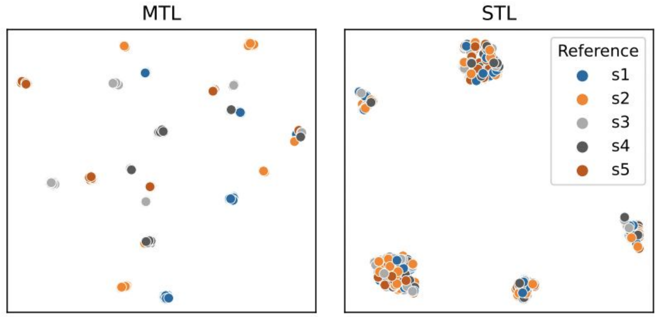
1. STL cannot distinguish different sentences, while MTL's representations are clustered in terms of both **sentences**

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

(b) Trained with STL, color-coded with severity.

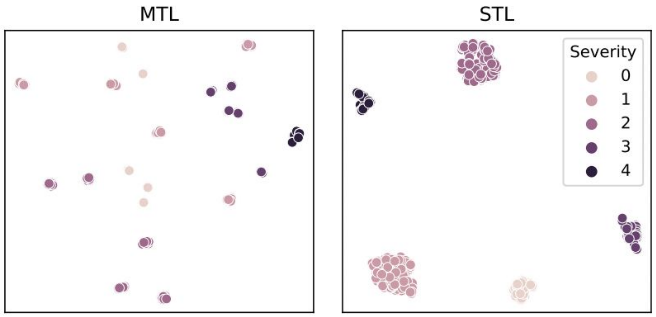


(c) Trained with MTL, color-coded with setence.

(d) Trained with STL, color-coded with setence.

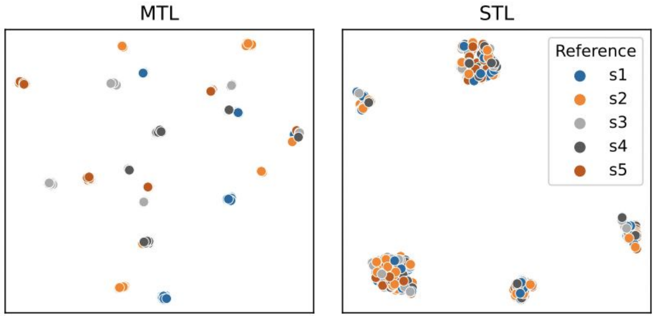
1. STL cannot distinguish different sentences, while MTL's representations are clustered in terms of both sentences and **severity levels**.

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

(b) Trained with STL, color-coded with severity.

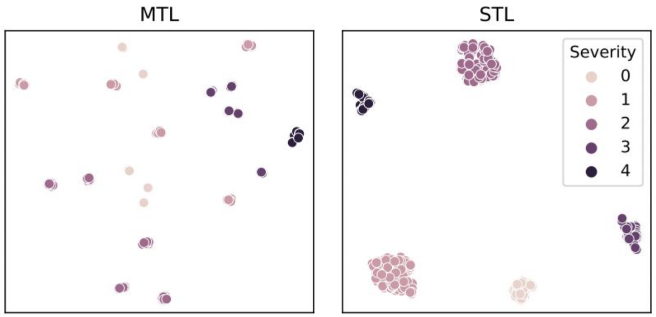


(c) Trained with MTL, color-coded with setence.

(d) Trained with STL, color-coded with setence.

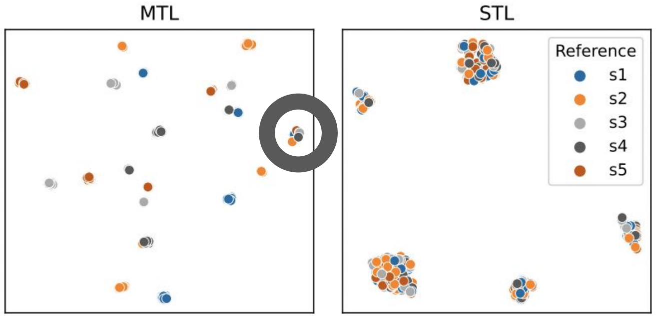
1. STL cannot distinguish different sentences, while MTL's representations are clustered in terms of both sentences and severity levels.
   → Indicates that the MTL model also encodes phonetic/pronunciation information.

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

(b) Trained with STL, color-coded with severity.

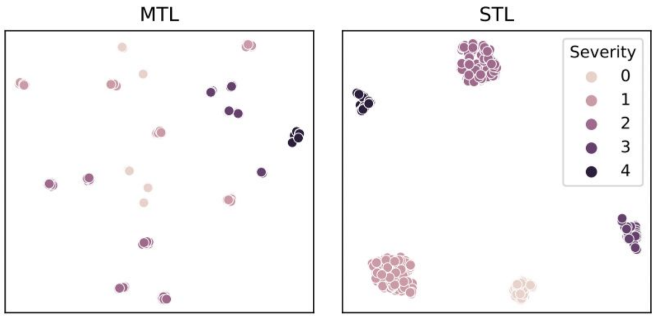


(c) Trained with MTL, color-coded with setence.

(d) Trained with STL, color-coded with setence.

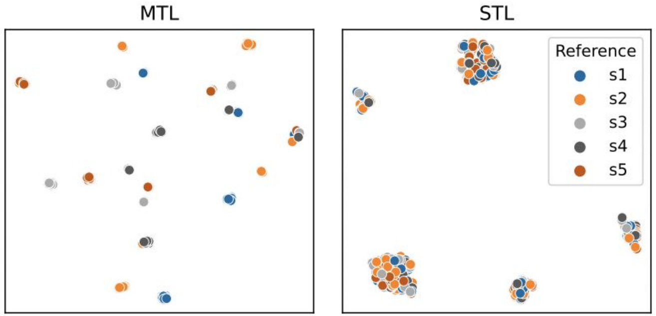
1. STL cannot distinguish different sentences, while MTL's representations are clustered in terms of both sentences and severity levels.
   → Indicates that the MTL model also encodes phonetic/pronunciation information.
2. Unlike others, severe samples are strongly clustered.

Analysis 1: Representation space visualization



(a) Trained with MTL, color-coded with severity.

(b) Trained with STL, color-coded with severity.



(c) Trained with MTL, color-coded with setence.

(d) Trained with STL, color-coded with setence.

1. STL cannot distinguish different sentences, while MTL's representations are clustered in terms of both sentences and severity levels.
   → Indicates that the MTL model also encodes phonetic/pronunciation information.
2. Unlike others, severe samples are strongly clustered.
   → May be due to significantly distorted speech, difficult for ASR

Analysis 2. Regularization effect of ASR

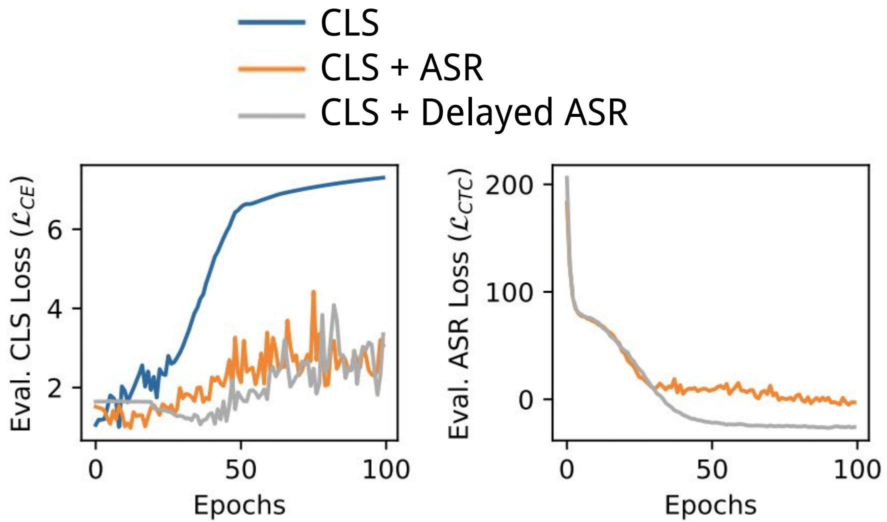


Fig. 3: Classification loss $\mathcal{L}_{\mathrm{CE}}$ and ASR loss $\mathcal{L}_{\mathrm{CTC}}$ on validation set. $\alpha = 0$ denotes the STL case when we use the $\mathcal{L}_{\mathrm{CE}}$ only.

Analysis 2. Regularization effect of ASR

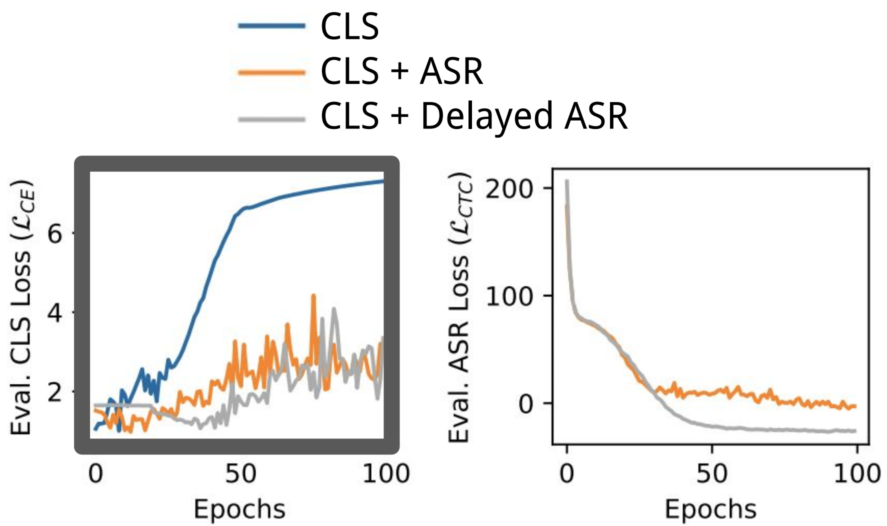


Fig. 3: Classification loss $\mathcal{L}_{\mathrm{CE}}$ and ASR loss $\mathcal{L}_{\mathrm{CTC}}$ on validation set. $\alpha = 0$ denotes the STL case when we use the $\mathcal{L}_{\mathrm{CE}}$ only.

1. With joint optimization, model overfits much slower.

Analysis 2. Regularization effect of ASR

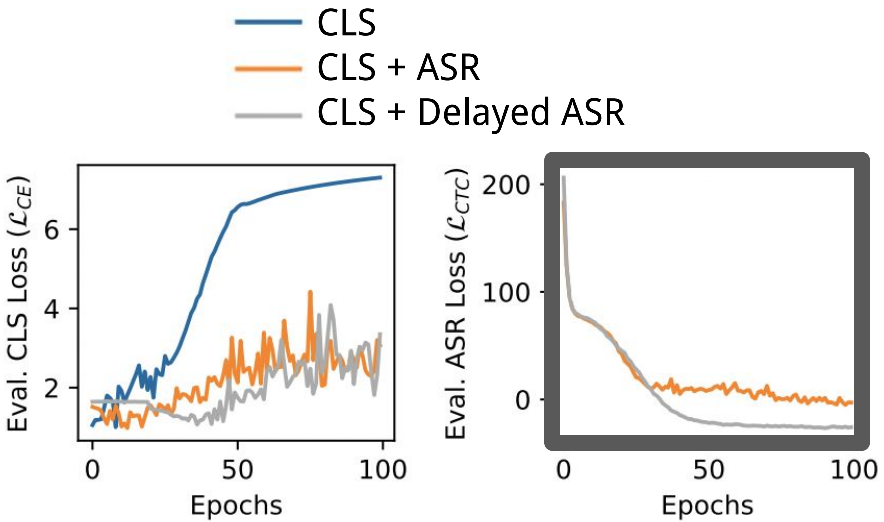


Fig. 3: Classification loss $\mathcal{L}_{\texttt{CE}}$ and ASR loss $\mathcal{L}_{\texttt{CTC}}$ on validation set. $\alpha = 0$ denotes the STL case when we use the $\mathcal{L}_{\texttt{CE}}$ only.

1. With joint optimization, model overfits much slower.
2. Delaying the optimization of ASR loss = stable optimization and better performances

Analysis 2. Regularization effect of ASR

Accuracy	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	60.51	60.69	56.21	54.94
$e = 10$	61.82	63.12	57.14	57.00
$e = 20$	54.77	64.69	59.84	61.27
$e = 30$	57.74	**65.52**	60.10	62.72
$e = 40$	55.47	60.11	62.00	57.96

PER	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	17.50	21.86	88.49	96.91
$e = 10$	14.83	22.37	82.59	96.49
$e = 20$	16.66	18.10	31.12	90.08
$e = 30$	15.87	17.72	23.10	74.54
$e = 40$	15.41	15.95	20.45	56.24

Analysis 2. Regularization effect of ASR

Accuracy	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	60.51	60.69	56.21	54.94
$e = 10$	61.82	63.12	57.14	57.00
$e = 20$	54.77	64.69	59.84	61.27
$e = 30$	57.74	**65.52**	60.10	62.72
$e = 40$	55.47	60.11	62.00	57.96

PER	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	17.50	21.86	88.49	96.91
$e = 10$	14.83	22.37	82.59	96.49
$e = 20$	16.66	18.10	31.12	90.08
$e = 30$	15.87	17.72	23.10	74.54
$e = 40$	15.41	15.95	20.45	56.24

1. Bigger the α and later the e, the Phone Error Rate consistently drops.

Analysis 2. Regularization effect of ASR

Accuracy	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	60.51	60.69	56.21	54.94
$e = 10$	61.82	63.12	57.14	57.00
$e = 20$	54.77	64.69	59.84	61.27
$e = 30$	57.74	**65.52**	60.10	62.72
$e = 40$	55.47	60.11	62.00	57.96

PER	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	17.50	21.86	88.49	96.91
$e = 10$	14.83	22.37	82.59	96.49
$e = 20$	16.66	18.10	31.12	90.08
$e = 30$	15.87	17.72	23.10	74.54
$e = 40$	15.41	15.95	20.45	56.24

1. Bigger the α and later the e, the Phone Error Rate consistently drops.
2. Best accuracy found in the mid-point of the hyper-parameter grid.

Analysis 2. Regularization effect of ASR

Accuracy	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	60.51	60.69	56.21	54.94
$e = 10$	61.82	63.12	57.14	57.00
$e = 20$	54.77	64.69	59.84	61.27
$e = 30$	57.74	**65.52**	60.10	62.72
$e = 40$	55.47	60.11	62.00	57.96
PER	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
$e = 0$	17.50	21.86	88.49	96.91
$e = 10$	14.83	22.37	82.59	96.49
$e = 20$	16.66	18.10	31.12	90.08
$e = 30$	15.87	17.72	23.10	74.54
$e = 40$	15.41	15.95	20.45	56.24

1. Bigger the α and later the e, the Phone Error Rate consistently drops.
2. Best accuracy found in the mid-point of the hyper-parameter grid.
   → **Premature training of CLS** leads to the model being under-trained with the ASR task, which fails to inject enough information.

Takeaways

1. Data scarcity

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning,

   : jointly learns the five-way severity classification task & ASR task.

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning,

   : jointly learns the five-way severity classification task & ASR task.

3. Further analyses

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning,

   : jointly learns the five-way severity classification task & ASR task.

3. Further analyses

   : latent representation

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning,

   : jointly learns the five-way severity classification task & ASR task.

3. Further analyses

   : latent representation → complementary information

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning,

   : jointly learns the five-way severity classification task & ASR task.

3. Further analyses

   : latent representation → complementary information

   : regularization effect

Takeaways

1. Data scarcity
2. Automatic dysarthria severity classification method

   : a self-supervised model fine-tuned with multi-task learning,

   : jointly learns the five-way severity classification task & ASR task

3. Further analyses

   : latent representation → complementary information

   : regularization effect → prevents overfitting

Thank you for your attention!