A Tiny Whisper-SER: Unifying Automatic Speech Recognition and Multi-label Speech Emotion Recognition Tasks

Huang-Cheng Chou
National Tsing Hua University, Taiwan
E-mail: huangchengchou@gmail.com

***Abstract*—Speech Emotion Recognition (SER) is critical in human-computer interaction (HCI). Previous SER works often utilize Automatic Speech Recognition (ASR) systems to improve SER performance. Some studies have investigated the possibility of modeling multi-task ASR and SER. However, prior studies and our preliminary experiments show the conflicts between ASR and SER if standard multi-task learning is used. We adopt a two-stage and weighted training strategy to overcome the conflicts between SER and ASR tasks. We utilize the public Whisper model that supports ASR tasks and others and add a small light adapter on the Whisper for the SER task, Whisper-SER. We first fine-tune the model on the ASR task and then fine-tune the model using the weighted loss, controlling loss between the ASR and multi-label SER task during training. The proposed method allows the Whisper-SER to recognize emotions and transcribe without ASR and SER performance degradation within the same encoder and decoder.**

I. Introduction

Speech is the most common modality for daily human interaction, and emotional signals in speech can improve communication. More and more devices support speech-based requests for users to ask machines to achieve tasks. For instance, people can use Amazon Alexa to make restaurant reservations and leave messages to friends. Detecting emotions from human speech can improve HCI and user experiences. Take the same example; the machine can add suitable "emojis" by recognizing emotions from user requests when sending messages to others for users. Therefore, machines need the ability to know the content and emotions of speech. However, ASR and SER tasks are not easy to integrate into one system, and two tasks are in conflict, degrading performance for both tasks [1].

Additionally, humans use the semantic and acoustics to express their emotions and intentions in speech. Therefore, many prior studies [2]–[4] improve performances of the SER systems by integrating linguistic and acoustic information, leading to better recognition of emotions than only considering acoustic information. The common practice is to deploy a pre-trained ASR system to generate transcripts and use a pre-trained language model, such as BERT, to extract linguistic embeddings as inputs for further improved SER performance [5]–[7]. However, prior studies only focus on SER tasks, and very few works jointly train a *"shared-parameter"* model for SER and ASR tasks due to the well-known catastrophic forgetting issues, which means the deep neural networks forget the original tasks when fine-tuning the models for new tasks. Therefore, we aim to avoid the catastrophic forgetting issue by the proposed method to make the pre-trained ASR model recognize emotions without ASR performance degradation.

The recent rapid development of large foundation speech processing systems produces an excellent opportunity to unify SER and ASR tasks into one shared-layer model. One of the state-of-the-art (SOTA) ASR systems, the Whisper [8], trained with 680,000 hours of weakly-supervised audio data, can handle four tasks: speech recognition, speech translation, voice activity detection, and language identification tasks. However, this model can not perform SER tasks. While some previous studies have proposed models for multi-task SER and ASR [6], [9], [10], they define the SER task as a single-label task, allowing each utterance only to have one emotion. Nevertheless, emotion perception can co-occur with emotions, [11]–[13]. Therefore, we follow the emerging perspective to define SER as a multi-label classification task to reflect the realist world. We aim to jointly adopt a Whisper model to validate the proposed method for SER and ASR tasks in one unified architecture. To the best of our knowledge, we are the first paper to integrate ***multi-label*** SER and ASR in the share-layer model.

In conclusion, we developed a model that is based on the Whisper model and added the lightweight component for multi-label SER tasks. Different from the recent SER studies that utilize speech Self-Supervised Learning (SSL) models as encoders [1], [14]–[16] and fine-tune them for SER tasks, we investigate whether extracting embeddings from both encoder and decoder of Whisper can benefit the performance of SER task. The two main contributions of this work are as follows:

- We comprehensively study a SOTA ASR model's word error rate (WER) on emotion-rich corpora and investigate the relationship between ASR and SER performances, considering inter-rater agreement, speaking rate, duration, and emotion category.
- We are the first work to model ASR and ***multi-label*** SER tasks with one unified weakly supervised system that enables parameter sharing between these two tasks for both encoders and decoders with the adapted two-stage training strategy, mitigating cartographic forgetting issues.

II. Background and Related Work

A. SER Systems Using Pre-trained ASR Models

Linguistic information plays an essential role in understanding emotions in speech. Many prior studies integrated semantics from human-annotated transcripts or predicted ones by ASR systems. For instance, Yoon et al. [17] applied the pre-trained ASR system to extract linguistic information to obtain better performances of the SER systems than only using acoustic information for training SER systems. Also, recent papers utilize hidden states from the audio encoder of an ASR system [8] to strengthen the performances of the SER systems [7], [18], [19]. However, the above-mentioned prior literature only optimized for a single SER task and does not jointly optimize ASR and SER tasks. Besides, our preliminary study revealed

that the Whipser Large V3 [1] has a WER of 12.7% on the v1.11 of the MSP-PODCAST, which is higher than most of the ASR benchmark databases, WER 7.7% on the "Open ASR Leaderboard" [2] [20]. Moreover, Li et al. [9] reported the same finding on the well-known emotion corpus, IEMOCAP [21]. Therefore, this work plans to deeply analysis the potential reasons why one of the SOTA ASR systems performs worse on emotional speech than more neutral utterances by considering the inter-agreement between raters, speaking rate, categorical emotions, and duration lengths.

B. Multi-task SER and ASR Systems

For most practical applications, speech-based devices must automatically transcribe speech requests to understand users' intentions, actions, or item information. Therefore, many previous studies have combined SER and ASR tasks in the systems. the studies [22], [23] apply supervised fine tuning on a pre-trained SSL system for ASR and SER tasks. However, some research [10], [24] repeatedly reveals that emotional speech is challenging for ASR systems, leading to higher error rates. Also, Gao et al. [10] showed that ASR and SER tasks often conflict with each other when using common multi-task learning. To overcome this problem, Feng et al. [24] proposed an SER model integrated with an acoustic-to-word ASR model. Gao et al. [10] utilized a two-stage fine-tuning method to separately fine-tune two components of models for SER and ASR tasks, respectively. Nevertheless, the study [10] still split the SER and ASR tasks using two individual layers, which means the encoder part is not shared, but we use the shared encoder and decoder parts in the work.

To make the model more efficient, we aim to make one share-layer model for ASR and SER tasks, so we adopt a two-stage training approach [10] with the weighting loss to jointly optimize ASR and SER tasks without compromising either task. Besides, all the above-mentioned studies regard the SER task as a single-label task, only assigning one emotion for each utterance. They used the majority rule or plurality rule to obtain a consensus label, but they directly excluded the data sample without a consensus label and ignored a minority of emotional ratings, making the test easier to predict. This work defines SER as a multi-label classification task (section II-C) to reflect a realistic scenario using the complete test without discarding any ambiguous data samples, suggested by [13].

C. Disagreement Among Raters and Multi-label SER

Disagreement among raters is often in the public emotion databases [25]. However, most previous studies regard the disagreement as noise and define SER as a single-label task using majority or plurality aggregation rules [3], [26] to obtain a single consensus label. However, the SER task differs from other speech tasks, like ASR or part-of-speech tagging, as it is a genuinely and naturally subjective task, so majority/plurality aggregation rules are unsuitable for SER task [13]. The disagreement is not necessarily to be removed, and the non-consensus samples represent real-world emotion perception. Also, conveyed and perceived emotions are not always discrete, as founded by the psychological studies [11], [12], [27]. In other words, there is a co-occurrence of emotions in one speech. Therefore, to make the SER systems recognize co-occurring emotions, we use the all-inclusive label aggregation rule proposed by [13] to calculate distributional labels as ground truth, and we define SER tasks as multi-label emotion recognition.

III. Methodology

Prior studies observed conflicts between ASR and SER tasks [10], [24], [28], and we also found that direct joint training for ASR and SER can degrade the performance of both tasks. Therefore, we are curious about the feasibility of a one-shared multitask system for ASR and SER based on the Whisper architecture [8]. We introduce an SER and ASR multitask framework, named Whisper-SER, as shown in Fig. 1, by supervised fine-tuning the SOTA ASR models with a two-stage weighting method to minimize ASR and SER performance degradation. To the best of our knowledge, we are the first work to reveal ASR performances on an emotional dataset with the **complete** tests, allowing samples to have co-occurrence of emotions.

A. Distributional-Label Processing for SER

We follow the studies [13] to calculate the distributional labels for training SER systems. For instance, given an emotional audio for four-class emotion recognition (neutral (N), angry (A), sad (S), and happy (H)), the votes are {N, N, A, S, S}. The distributional label can be calculated by the number of votes divided by the total votes for each emotion. In the example, we get a distributional label $\{\text{N, A, S, H}\} = \{2/5, 1/5, 2/5, 0/5\} = \{0.4, 0.2, 0.4, 0.0\}$. Fig. 2 shows the detailed distribution across different data sets. For conventional SER studies, datasets with mixed emotions are removed for simplicity; however, we kept all the data to reveal the model's actual SER performances, as suggested by [13], [29]. During the evaluation of systems, the threshold, $1/C$, is used to convert the distributional predictions into hard decisions. The C means the number of emotion classes. We use one example in Table I to clearly describe the ad-hoc process. Therefore, with the sample example, we assign the data sample has co-occurring emotions, neutral and sadness.

B. Objective Functions

We employ the cross-entropy loss as an objective function for ASR and SER tasks.

1) Cross-entropy Loss for ASR: In this work, we choose the Whisper model [8] as a backbone model, an encoder-decoder-based ASR model by sequence-to-sequence leanings. The decoder of the Whisper generates the IDs of each token in the input utterances, and the Whisper tokens can transform the ground truth transcripts into IDs. With the ID vectors of the ground truth and predictions, we can calculate the cross-entropy loss denoted by $\mathcal{L}_{ASR}$ to update the whole model.

2) Cross-entropy Loss for SER Task: Chou et al. [13] reported that soft-label learning using cross-entropy loss is the best approach to obtain higher recognition of SER systems among hard-label learning with the cross-entropy loss, soft-label learning with the cross-entropy loss, and distributional-label learning with the Kullback–Leibler Divergence (KLD) loss. Therefore, we follow the study to use soft-label learning using cross-entropy loss, denoted by $\mathcal{L}_{SER}$, to train the model for the SER task.

3) Standard Multitask Learning for ASR and SER Tasks: The objective function in the standard multitask learning (MTL) becomes

$$\mathcal{L}_{MTL} = \mathcal{L}_{SER} + \mathcal{L}_{ASR}. \tag{1}$$

C. Two Stage Training Method with the Weighting Loss

We aim to enable the Whisper to recognize emotions from speech and keep the same ASR performances. We have two training stages to

TABLE I
OVERVIEW OF ONE SAMPLE FOR 4-CLASS EMOTION RECOGNITION. THE **A**, **S**, AND **N**, ARE ANGER, SADNESS, AND NEUTRAL EMOTIONS, RESPECTIVELY. THE NUMBER MEANS THE COUNT OF EMOTIONS. FOR INSTANCE, **A*1, S*2, AND N*2** MEANS A, S, S, N, N

Raw Annotation	A*1, S*2, N*2
Label for **Training** Stage {N, A, S, H}	{0.4, 0.2, 0.4, 0.0}
Label for **Testing** Stage {N, A, S, H}	{1, 0, 1, 0}

[1] https://huggingface.co/openai/whisper-large-v3

[2] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

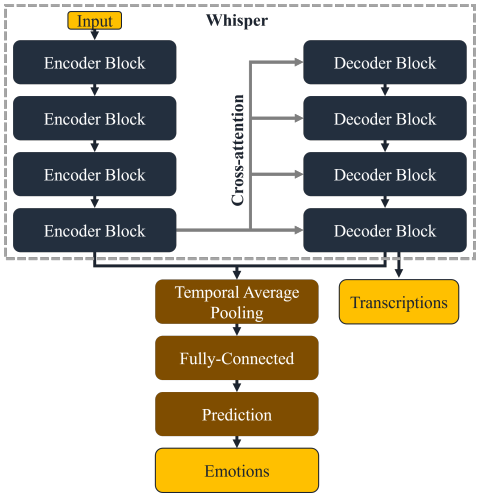


Fig. 1. The figure illustrates the overall structure of Whisper-SER. We added the light models to the Whisper model (brown color) for the SER task, enabling Whisper to do ASR and SER tasks using the proposed training method with weighting loss.

achieve this goal. In the first training stage, we fine-tune the Whisper model on the emotional dataset with the training and development sets on the ASR task only. Then, in the second training stage, we use the weighting loss to fine-tune the model for SER and ASR tasks. In the weighting loss, we introduce an **alpha** to control weights between L_{SER} and L_{ASR} below, and the alpha value ranges from 0.1 to 0.9, and the step size is 0.1. We also reveal the relationship between the alpha and the ASR/SER performances by introducing the weighting factor. Notice that the order of training tasks matters. The first stage is to fine-tune the model for the ASR task and use the proposed weighting loss for multitasking ASR and SER. It won't work based on our experiments if we inverse the tasks (train for the SER task first, then the ASR task).

$$\mathcal{L}_{MTL}^{\mathcal{W}} = \alpha * \mathcal{L}_{ASR} + (1-\alpha) * \mathcal{L}_{SER} \quad (2)$$

IV. Experimental Settings

A. The MSP-PODCAST Corpus and Preprocessing

We validate the proposed method on version 1.11 of the MSP-PODCAST emotion dataset [30]. This version has about 237 hours of segmented audio recordings in English, containing 84,030 sentences (134.34 hours) in the train set, 19,815 utterances (31.72 hours) in the development set, 30,647 utterances in the test1 set, and 14,815 utterances in the test2 partition. Test2 set is overall more challenging

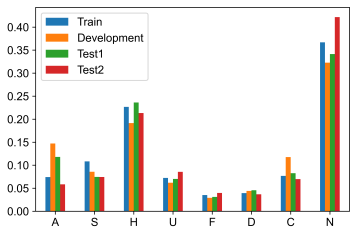


Fig. 2. Distributional-label distribution across different data sets. There are 8 emotions, including anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), and neutral (N).

than test1 and closer to real-life scenarios described in [30]. To the best of our knowledge, we are the first ones to show the ASR and SER performances on the Test2 set since the prior studies ignored the Test2 set. The work uses the primary emotions where every rater can only select one emotion from the given options, including anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), neutral (N), and "other". We exclude "other" to define the SER task as 8-class multi-label emotion recognition. The corpus also provides transcripts suitable for developing multitask ASR and SER systems. We use the Whisper tokenizer Fig. 2 illustrates the averaged distributional label across different data sets. to normalize the transcripts and remove special comments, e.g., "["00:30:33" Cross Talk" or "[inaudible 00:08:44]". More details about the corpus are in [30].

B. Whisper-SER and Implementation Details

Fig. 1 illustrates the proposed Whisper-SER framework. We employ the Whisper model proposed by [8] (marked in dark blue color in Fig. 1) as the primary backbone model since the Whisper architecture is one of the SOTA ASR models. Whisper models have sizes ranging from tiny (39M parameters) to large (1.5B parameters). For computing efficiency, our experiments are based on Whisper Tiny. Whisper Tiny has four transformer-based blocks on the encoder and decoder, and the number of nodes for each layer is 384. We add a module (marked in brown color in Fig. 1) for the SER task, including one average pooling layer, one fully-connected layer with a ReLu activation (the number of nodes is 384), and one prediction layer. The inputs concatenate the hidden states of the last layers of the encoder and decoder. In addition, to study if the encoder encodes more emotion-related information than the decoder, we conduct experiments to take the hidden states from the encoder's last layer, the decoder's, or both.

There are two training stages for the Whisper-SER framework introduced in section III-C. In the first stage, we fine-tune the pre-trained Whisper architecture on the emotion dataset for ASR tasks by setting $\alpha = 1$ in equation (2). In the second stage, we introduce non-zero α and jointly optimize the system with a convex combination of ASR and SER losses (equation (2)). The maximum number of generated tokens is 160. We use the Adam optimizer [31] with an initial learning rate of 1e-7 and a batch size of 32 and a warm-up decay learning rate scheduler with the maximum and minimum learning, 1e-5 and 1e-7, respectively, for 5 epochs. We train the models for 100 epochs with the DeepSeed stage 2 [32] and select the best checkpoint with the lowest loss on the development set. We use PyTorch and HuggingFace library . We conduct experiments on an NVIDIA Tesla v100 GPU with 16 GB of memory. The total of GPU hours is 100 for all results.

C. Evaluation Metric

For the ASR task, we use the Word-Error-Rate (WER) as the evaluation metric, and the WER values can be calculated by $WER = (S + I + D)/N$, where S, I, D, and N are a number of substitution, insertion, deletion errors, and the number of words in the ground-truth transcripts, respectively. In terms of multi-label emotion recognition, the study [13] investigated the advantages and disadvantages of using distribution-based assessment (e.g., KLD) versus hard-decision-based assessment (e.g., F1 scores) on the SER task. Using distribution-based assessment, it is hard to observe differences in performance between the baseline and proposed model, which is not easy to interpret since the scale is very small. Therefore, we decide to utilize weighted-F1 (**weF1**), macro-F1 (**maF1**), and micro-F1 (**miF1**) scores to measure multi-label classification accuracy. This involves selecting target classes by applying thresholds to the ground truth data. A prediction is deemed successful if the proportion for a class exceeds $1/8$ mentioned in section III-A.

TABLE II

ASR AND SER PERFORMANCE RESULTS OF THE MODELS. THE COLUMN **EXPERIMENTS** EXPLAINS THE EXPERIMENTS WE CONDUCT. THE COLUMN **INPUT** INDICATES THE INPUT SOURCE OF THE SER MODELS MENTIONED IN SECTION IV-B. THE COLUMN **FINE-TUNE** INDICATES IF THE PRE-TRAINED MODELS ARE FINE-TUNED. THE COLUMN **#PARA. (M)** INDICATES THE NUMBER OF MODEL PARAMETERS IN MILLIONS (M). ↑ SIGNALS THAT A METRIC IS BETTER IF THE VALUE IS HIGHER. OTHERWISE, WE USE ↓.

	Test Set				Test1				Test2			
Experiments	Input	Backbone	Fine-Tune	#Para. (M)	WER (%)↓	MaF1↑	MiF1↑	WeF1↑	WER (%)↓	MaF1↑	MiF1↑	WeF1↑
SER SOTA [13]		WavLM Large	V	317	-	0.4850	0.6210	0.6060	-	0.4264	0.6127	0.5842
Whisper Large V3 [8]		Large V3		1550	12.57				13.63			
Whisper Tiny [8]		Tiny		39	29.75				54.15			
ASR		Tiny	V	39	20.26				19.98			
SER^{DO}	Decoder	Tiny	V	39	769.2	0.4242	0.6183	0.5679	621.2	0.3591	0.5986	0.5303
SER^{EO}	Encoder	Tiny	V	39	1614	0.3691	0.5911	0.5216	1974	0.3338	0.5924	0.5126
SER^{BO}	Both	Tiny	V	39	348.6	0.4295	0.6191	0.5724	239.8	0.3624	0.5989	0.5332
MTL^{DO}	Decoder	Tiny	V	39	24.86	0.4209	0.6175	0.5646	22.68	0.3509	0.5998	0.5244
MTL^{EO}	Encoder	Tiny	V	39	23.80	0.3186	0.5843	0.4775	21.72	0.2858	0.5908	0.4750
MTL^{BO}	Both	Tiny	V	39	25.30	0.4239	0.6133	0.5642	22.57	0.3568	0.5936	0.5263
Whisper-SER	Both	Tiny	V	39	20.90	0.4111	0.6172	0.5611	19.72	0.3563	0.6009	0.5295

D. Experiments

1) Single-task Baseline: We fine-tune the Whisper Tiny model on the MSP-PODCAST as the ASR baseline using $\mathcal{L}_{ASR}$, denoted by **ASR**. In terms of SER, we add light models, including the temporal average pooling layer, Fully-Connected (FC) layer, and prediction layers. To investigate whether the encoder block has more emotional information than the decoder one, we conduct the experiences to extract the embeddings from the last layer of the encoder block (denoted by **EO**), decoder block (denoted by **DO**), or both blocks (denoted by **BO**). In fact, we fine-tune the Whisper Tiny model on the MSP-PODCAST for the SER task using $\mathcal{L}_{SER}$ as the SER baseline, denoted by **SER**. For instance, the $\textbf{SER}^{BO}$ means the input of the light models for the SER task comes from both of the last layers of encoder and decoder blocks.

2) Multi-task Baseline: To show the conflicts between SER and ASR tasks, we conduct experiments to show the performances using standard loss, $\mathcal{L}_{MTL}$, denoted by **MTL**. We also changed the source of inputs for the light model for the SER task to determine which blocks encode more emotional cues (e.g., $\textbf{MTL}^{BO}$).

3) SER and ASR SOTA Models: To understand the gap between the SOTA and the proposed method, we employ pre-trained Whisper Large V3 model[3] as the ASR SOTA model, denoted by **Whisper Large V3**. In terms of SER SOTA, we use the model[4] proposed by [15], and we fine-tune the model on the MSP-PODCAST using the distributional labels by following the study [13]. We denoted the model by **SER SOTA**. The more details about model structure and training approaches are in [15] and [13].

4) Finding the Best α Value in $\mathcal{L}_{MTL}^{\mathcal{W}}$: To find the best α value in $\mathcal{L}_{MTL}^{\mathcal{W}}$ (Equation 2), we conducted the experiments by assigning the α value as from 0.1 to 0.9, as shown in Fig. 3. The 0.8 is the best α value in the work. Besides, when the α is 0.1 or 0.2, the performances for the SER task is higher than the model $\textbf{SER}^{BO}$, and the finding aligns with the previous studies [1], [9].

5) Whisper-SER: Based on the preliminary experiments, we found that using embeddings from both of encoder and decoder blocks ($\textbf{SER}^{BO}$) has better recognition performance. Therefore, the proposed Whisper-SER uses this same structure. In the first stage, we only fine-tune the Whisper components where in the dotted line in Fig. 1 for the ASR task. Then, in the second stage, we fine-tune the whole Whisper-SER model using the $\mathcal{L}_{MTL}^{\mathcal{W}}$ loss. To find the best α value in Equation 2. We conducted the experiments by assigning the α value as from 0.1 to 0.9, as shown in Fig. 3. The 0.8 is the best α value in the work.

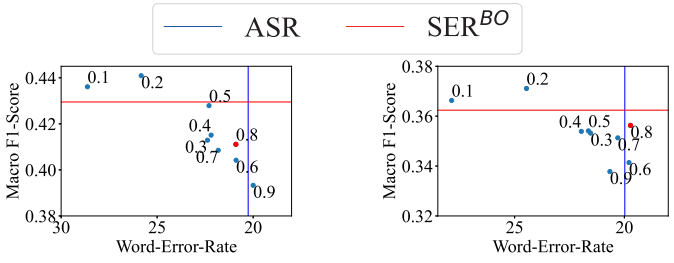
V. RESULTS AND ANALYSIS

Table II summarizes all results on the MSP-PODCAST in various F1 scores across two different test sets. We answer the research questions one by one in detail below.

Does the encoder block of the Whisper have more emotion-related cues? To our surprise, the SER^{DO} and SER^{BO} models have better performance than the SER^{EO}, which might imply that the decoder encodes more emotion-related information than the encoder, which has never been investigated in the previous. We are the first work to report this finding, suggesting we should consider a decoder as well.

Are there conflicts between SER and ASR tasks? We find that the joint training with ASR and SER using vanilla multi-task loss, $\mathcal{L}_{MTL}$, degrades ASR and SER task performance. For instance, the model, **ASR**, has 20.26% WER and $\textbf{SER}^{BO}$ has 42.95% in macro-F1 score on the Test1. However, the $\textbf{MTL}^{BO}$ has 25.30% WER and 42.39% macro-F1 score ASR and SER tasks, respectively. All performances decreased.

Why should we use two stages to train the models for SER and ASR tasks? The results of $\textbf{MTL}^{DO}$, $\textbf{MTL}^{EO}$, and $\textbf{MTL}^{BO}$



(a) Test1 Set (b) Test2 Set

Fig. 3. Relationship between ASR/SER performances and α values in $\mathcal{L}_{MTL}^{\mathcal{W}}$ (Equation 2). We mark the point in red to show the best α considering both ASR and SER performances.

[3]https://huggingface.co/openai/whisper-large-v3

[4]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

TABLE III

OVERVIEW OF ASR AND SER PERFORMANCES WITH THE THREE FACTORS (INTER-RATER AGREEMENT, SPEAKING RATE, AND DURATION LENGTH). WE USE THE WER (%)↓ AND MAF1↑ TO SHOW PERFORMANCES ON ASR AND SER TASKS, RESPECTIVELY. THE BOLD NUMBERS MEAN THE WORST PERFORMANCES ON ASR OR SER TASKS.

Factor	**Inter-rater Agreement**						**Speaking Rate**						**Duration Length**					
Task	**ASR**			**SER**			**ASR**			**SER**			**ASR**			**SER**		
Experiments	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Slow**	**Medium**	**Quick**	**Slow**	**Medium**	**Quick**	**Short**	**Medium**	**Long**	**Short**	**Medium**	**Long**
Whisper Large V3 [8]	12.64	12.50	**14.15**				**24.74**	12.15	11.47				**16.92**	12.69	11.38			
Whisper Tiny [8]	39.54	38.73	**40.95**				**61.83**	39.69	37.41				**50.59**	40.66	35.43			
ASR	21.16	21.14	**24.05**				**36.68**	21.67	21.60				**28.51**	22.95	18.53			
SER^{BO}				0.411	0.423	**0.296**				0.404	0.410	0.405				0.381	0.407	0.423
MTL^{BO}	22.64	24.77	**31.75**	0.393	0.416	**0.306**	**51.23**	25.94	26.18	0.403	0.404	0.400	**37.52**	27.73	21.84	0.381	0.402	0.413
Whisper-SER^{BO}	23.72	22.87	**27.05**	0.403	0.408	**0.296**	**42.15**	23.88	22.77	0.393	0.397	0.399	**31.42**	24.94	20.35	0.378	0.398	0.398

show that the models use the standard multi-task learning without two-stage training. The WER is higher than the results of the **ASR** model, so the standard multi-task learning can not benefit the ASR and SER tasks. However, in the practical application, the ASR task is very important, so our main goal is to maintain the same level of performance on the ASR task and have the model recognize emotions. Based on the results of the proposed framework, Whisper-SER, the two stages are critical to achieving the goal.

Can we unify ASR and multi-label SER multitasks in one shared-layer system without ASR performance degradation? Yes. To overcome the conflicts between SER and ASR, the adapted two-stage training strategy using the weighting loss ($\mathcal{L}_{MTL}^{\mathcal{W}}$) can mitigate this issue by keeping ASR performance the same as **ASR**. Fig. 3 shows the relationship between ASR/SER performances and α values in Equation 2. Using both encoder and decoder features, Whisper-SER, when α is 0.8, achieves the best overall performance for both ASR and SER tasks. The best-performing model, Whisper-SER, has on-par ASR and SER accuracy with the baseline models (**ASR** and **SER**) on Test1 and Test2.

What are the relationships between ASR and SER performances and inter-rater agreement, speaking rate, and duration lengths? We first combine the test1 and test2 as one test set. Table IV summarizes the distributions considering inter-rater agreement, speaking rate, and duration length across Train, Development, and Test1+Test2 sets. We calculate the mean and standard deviation (std.) from the Train set and define two thresholds by (mean - std.) and (mean + std.) to split the data into three groups (high/slow/short, medium, and low/quick/long) based on the values of inter-rater agreement, speaking rate, and duration length, respectively.

Then, we measure the inter-rater agreement of each sentence by Cohen's Kappa [33], speaking rate, and duration length on the MSP-PODCAST. We calculate the mean and standard deviation (std.) from the train set and define two thresholds by (mean - std.) and (mean + std.) to split the data into three groups (high/slow/short, medium, and low/quick/long) based on the values of inter-rater agreement, speaking rate, and duration length, respectively. The detailed distributions are summarized in Table IV.

TABLE IV

DATA DISTRIBUTIONS CONSIDERING INTER-RATER AGREEMENT, SPEAKING RATE, AND DURATION LENGTH ACROSS TRAIN, DEVELOPMENT, AND TEST1+TEST2 SETS.

Data Set		**Train**	**Development**	**Test1+Test2**
Total Number		**84,030**	**19,815**	**45,462**
Inter-rater Agreement	**Low**	5.33%	4.11%	4.96%
	Medium	79.06%	81.14%	76.79%
	High	15.61%	14.76%	18.25%
Speaking Rate	**Low**	14.97%	3.52%	2.24%
	Medium	72.10%	85.01%	79.48%
	High	12.94%	11.47%	18.28%
Duration Length	**Low**	17.59%	16.40%	21.30%
	Medium	62.28%	64.65%	61.09%
	High	20.13%	18.95%	17.60%

Table III summarizes the performances on the ASR and SER tasks considering the three factors. To our surprise, all models in Table III performed worst on the high-agreement samples for ASR and SER tasks. With the finding on the ASR task, we align with prior studies [1], [6], [9], [34] that ASR systems have worse accuracy on emotion-rich utterances than emotion-neutral ones. The potential reason might be because the pronunciation tends to be wrong in strong emotions (e.g., anger) [35]. Also, Chou et al. [13] also reported that adding more ambiguous (lower inter-rater agreement) samples can decrease performance on the samples that have a high inter-rater agreement, but adding the ambiguous samples can have better effectiveness and robustness on the ambiguous samples, which is a real-world scenario that we can not know which sample has a high or low inter-rater agreement. This perspective is very important since the previous studies removed the ambiguous sample from the test set. For instance, the 19.85% samples in the test set were removed by the plurality rule on the MSP-PODCAST.

Besides, for the speaking rate and duration length, we found that the models performed worse on the utterances with a slow speaking rate (less than two words per second) and the ones with a short duration (less than 3.39 seconds) on the ASR task. However, there are no apparent performance differences on the SER task. This is the first work to observe and reveal the relationship between the performance of ASR and SER and the three factors.

VI. CONCLUSIONS AND FUTURE WORK

This paper uses the modified two-stage training strategy with the weighting loss to jointly model ASR and **multi-label** SER tasks based on the Whisper architecture and evaluated on the **complete test set** of the MSP-PODCAST corpus without discarding any data samples, such as the samples without consensus label. Our findings indicate that simultaneous training for ASR and SER with equal weights for each loss can cause catastrophic forgetting of pre-trained ASR tasks. To circumvent this issue, we use the two-stage weighting training method to keep the same ASR and SER performance levels as the pre-trained ASR and SER models, respectively. The proposed model, **Whisper-SER**, gets 29.75% and 63.58% relative reduction in WER for the ASR task on Test1 and Test2 sets, respectively, compared to the **Whisper Tiny** baseline model. However, the **Whisper-SER** only gets 4.28% and 1.68% relative reduction in macro F1 on Test1 and Test2 sets, respectively, on the SER task, compared to the **SER**^{BO}. In the further analysis, we are the first paper to reveal

that the modern models perform worse on ASR tasks when the utterances have a higher inter-rater agreement on emotions, slower speaking rater, and shorter duration. Also, the study first showed the possibility of multitasking ASR and multi-label SER models without ASR performance degradation. In future works, we aim to use LoRa [36] to use the bigger Whisper model to train the Whisper-SER on the limited computational GPU resource and deal with imbalanced emotion distribution by class-balanced loss [37].

REFERENCES

[1] Y. Gao *et al.*, "Enhancing Two-Stage Finetuning for Speech Emotion Recognition Using Adapters," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 316–11 320. DOI: 10.1109/ICASSP48485.2024.10446645.

[2] M. Kang *et al.*, "ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models," in *Proc. INTERSPEECH*, 2023, pp. 4339–4343. DOI: 10.21437/Interspeech.2023-754.

[3] T. Feng and S. Narayanan, "Foundation Model Assisted Automatic Speech Emotion Recognition: Transcribing, Annotating, and Augmenting," *arXiv preprint arXiv:2309.08108*, 2023.

[4] N. Abhishek and P. Bhattacharyya, "" We care": Improving Code Mixed Speech Emotion Recognition in Customer-Care Conversations," *arXiv preprint arXiv:2308.03150*, 2023.

[5] Y. Li *et al.*, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech*, Graz, Austria, Sep. 2019, pp. 2803–2807.

[6] Y. Li *et al.*, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7362–7366. DOI: 10.1109/ICASSP43922.2022.9746289.

[7] T. Feng and S. Narayanan, "PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches For Speech Emotion Recognition Using Pre-trained Speech Models," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, Sep. 2023.

[8] A. Radford *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 28 492–28 518.

[9] Y. Li *et al.*, "ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition," in *Proc. INTERSPEECH*, 2023, pp. 1449–1453. DOI: 10.21437/Interspeech.2023-2078.

[10] Y. Gao *et al.*, "Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining," in *Proc. INTERSPEECH*, 2023, pp. 3637–3641. DOI: 10.21437/Interspeech.2023-756.

[11] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, E7900–E7909, 2017.

[12] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021, ISSN: 1364-6613.

[13] H.-C. Chou *et al.*, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024. DOI: 10.1109/TAFFC.2024.3411290.

[14] L. Pepino *et al.*, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[15] J. Wagner *et al.*, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023. DOI: 10.1109/TPAMI.2023.3263585.

[16] H. Wu *et al.*, *EMO-SUPERB: An In-depth Look at Speech Emotion Recognition*, 2024. arXiv: 2402.13018 [eess.AS].

[17] S. Yoon *et al.*, "Multimodal Speech Emotion Recognition Using Audio and Text," in *IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 112–118. DOI: 10.1109/SLT.2018.8639583.

[18] Y. Gong *et al.*, "Joint Audio and Speech Understanding," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023)*, Taipei, Taiwan, Dec. 2023.

[19] G. Ioannides *et al.*, "Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1853–1857. DOI: 10.21437/Interspeech.2023-497.

[20] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, Oct. 2019. DOI: 10.48550/arXiv.1910.03771. arXiv: 1910.03771 [cs.CL].

[21] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[22] W. Wu *et al.*, "Integrating Emotion Recognition with Speech Recognition and Speaker Diarisation for Conversations," in *Proc. INTERSPEECH*, 2023, pp. 3607–3611. DOI: 10.21437/Interspeech.2023-293.

[23] Y. Li *et al.*, "Evaluating Parameter-Efficient Transfer Learning Approaches on SURE Benchmark for Speech Understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, Rhodes island, Greece, 2023, pp. 1–5.

[24] H. Feng *et al.*, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," in *Proc. Interspeech 2020*, 2020, pp. 501–505. DOI: 10.21437/Interspeech.2020-1180.

[25] C. Busso *et al.*, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[26] C.-C. Lee *et al.*, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011, Sensing Emotion and Affect - Facing Realism in Speech Processing, ISSN: 0167-6393.

[27] K. Vansteelandt *et al.*, "The co-occurrence of emotions in daily life: A multilevel approach," *Journal of Research in Personality*, vol. 39, no. 3, pp. 325–335, Jun. 2005. DOI: 10.1016/j.jrp.2004.05.006.

[28] M. Tjalve and M. Huckvale, "Pronunciation variation modelling using accent features," in *Proc. Interspeech 2005*, 2005, pp. 1341–1344. DOI: 10.21437/Interspeech.2005-487.

[29] P. Riera *et al.*, "No sample left behind: Towards a comprehensive evaluation of speech emotion recognition system," in *Proc. Workshop on Speech, Music and Mind 2019*, 2019.

[30] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019. DOI: 10.1109/TAFFC.2017.2736999.

[31] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[32] J. Rasley *et al.*, "DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20, Virtual Event, CA, USA: Association for Computing Machinery, 2020, pp. 3505–3506, ISBN: 9781450379984.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] S. Sahu *et al.*, "Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription," in *Proc. Interspeech*, 2019, pp. 3302–3306. DOI: 10.21437/Interspeech.2019-1149.

[35] Z. Aldeneh and E. M. Provost, "You're Not You When You're Angry: Robust Emotion Features Emerge by Recognizing Speakers," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1351–1362, 2023. DOI: 10.1109/TAFFC.2021.3086050.

[36] E. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR 2022)*, 2022.

[37] Y. Cui *et al.*, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, California, USA, Jun. 2019.