SPEECH EMOTION RECOGNITION USING SEMI-SUPERVISED LEARNING WITH EFFICIENT LABELING STRATEGIES

Zhi Zhu, Yoshinao Sato

Fairy Devices Inc., Japan

ABSTRACT

The collection of large amounts of labeled data for speech emotion recognition requires considerable time and effort. As a result, the sizes of existing corpora are limited. One promising solution to this difficulty is semi-supervised learning, i.e., learning from both labeled and unlabeled data. In this study, we applied the noisy student training (NST) method to speech emotion recognition. We experimentally investigate the trade-off between the amount and reliability of labeled data. For this purpose, we prepared labeled and unlabeled data by limiting the available annotations in the CREMA-D dataset. The experimental results showed that a model trained using the NST method with some of the annotations achieved almost the same performance as the one trained using supervised learning with all the annotations if the amount and reliability of the available annotations were appropriate. Our findings are significant in identifying the most efficient labeling strategy when utilizing a large-scale dataset without labels for speech emotion recognition.

Index Terms— speech emotion recognition, semisupervised learning, noisy student training

1. INTRODUCTION

A large-scale corpus for speech emotion recognition (SER) is challenging to build. One of the reasons is that emotion classification is partially subjective. An emotion expressed by a speaker does not always match one perceived by a listener. Furthermore, listeners' perceptual evaluations do not necessarily agree with each other because their sensitivities and biases vary widely. Consequently, we need to ask many annotators to perform the task of classifying each utterance by emotion. In fact, 3 annotators, 6 to 12 annotators, and 20 annotators are assigned to each utterance in [1], [2], and [3, 4], respectively. The majority voting method is commonly used to aggregate the answers from annotators and determine the emotion label. More sophisticated statistical methods proposed in [5, 6, 7, 8] can also be used. These methods may reduce the number of annotators per utterance required to decide the emotion labels with high reliability. However, it still requires much time and effort to collect a large number of labeled data for SER. One promising solution to overcome the scarcity of labeled data is to utilize unlabeled data in addition to labeled data.

In some previous studies, unlabeled speech data was used for SER. For example, an unsupervised autoencoder was combined with a supervised emotion classifier in [9, 10, 11]. In [12], an autoencoder-based model was extended with adversarial learning and multi-task learning that consisted of emotion classification, speaker recognition, and gender recognition. An intrinsic limitation of an unsupervised autoencoder-based model is that irrelevant information is inevitably retained in the autoencoder. Hence, the emotion classifier should be trained to extract relevant features using labeled data.

In this study, we applied the noisy student training (NST) method [13] to SER to investigate how to utilize unlabeled data efficiently. NST is a variant of the self-training method that uses a teacher model trained using labeled data to infer soft labels of unlabeled data. Labels inferred in this manner are referred to as pseudo labels. The original labeled data and the pseudo labeled data are used to train a student model, which replaces the old teacher model. This process is repeated over multiple generations. One of the self-training method's problems is that errors in the inference of pseudo labels are amplified as the generation progresses. Data augmentation with class balancing using oversampling and filtering of pseudo labeled data are introduced in the NST method to ensure data integrity. The self-training method was used for multimodal emotion recognition in [14], where collaborative semi-supervised learning was proposed to correct mislabeled samples. The effectiveness of NST for image classification [13], automatic speech recognition [15], keyword spotting [16], and singing voice separation [17] was investigated in previous studies.

We note that whether or not the NST method is effective for SER because SER is partially subjective. The emotion expressed in speech is often ambiguous and can be perceived differently between listeners. In other words, not all utterances definitely belong to one of the given emotion classes. Therefore, it is inevitable that unlabeled data includes a significant number of utterances whose "correct" emotion class is ambiguous. Such utterances may hinder the effectiveness of the NST method. This characteristic of SER contrasts with other tasks that NST was applied to in previous stud-

ies [13, 15, 16, 17]. In this study, we experimentally demonstrated the effectiveness of the NST method for the SER task regardless of its subjectivity.

Specifically, we address the trade-off between the amount and reliability of labeled data. Previous studies showed that the performance improvement becomes more substantial as the amount of labeled data becomes larger [9, 10]. However, the effectiveness of semi-supervised learning should also rely on the reliability of labeled data. If we increase the number of annotators per utterance, then we can obtain more reliable labels using the majority voting method as studied in [3, 4], whereas much more time and effort are required. Consequently, we are forced to reduce the number of labeled utterances. Therefore, we must consider the balance between the amount and reliability of labeled data if we suppose the total cost for preparing labeled data is fixed.

To investigate the effect of the amount and reliability of labeled data on the performance improvement by utilizing unlabeled data using the NST method, we limit the number of available annotations in an existing emotional speech corpus, the crowdsourced emotional multimodal actors dataset (CREMA-D) [2], and examine different labeling strategies: quantity-first and quality-first. Furthermore, we compared hard and soft labels to determine the best way to utilize the initial labeled data. It was reported in [18, 19] that the use of soft labels provides better performance of SER by supervised learning. In this study, we aim to investigate whether soft labels are beneficial when using semi-supervised learning. The findings of this study are significant for identifying the most efficient annotation strategy for utilizing a large-scale emotional speech dataset without labels.

2. METHODS

2.1. Noisy Student Training

This section briefly reviews the NST method and explains the data and model used in this study. We use NST as the semi-supervised learning method, which was applied to image recognition in [13] and automatic speech recognition in [15]. The algorithm of NST is summarized as follows:

1. Train a teacher model M_t using the labeled data with data augmentation.
2. Infer pseudo labels of each sample in the unlabeled data using the teacher model. Filter the unlabeled data on the basis of the pseudo labels.
3. Train a student model M_s using both the labeled data and the filtered unlabeled data with data augmentation.
4. Set the student model as the teacher model $M_t := M_s$ at the next generation, and repeat Step 2 to Step 4.

The pseudo labels can be soft or hard. We refer to the model trained using supervised learning with only labeled data as the zeroth generation. The models at the precedent generation are trained using the NST method with unlabeled data.

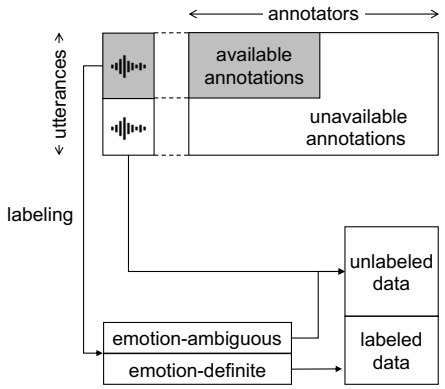

Fig. 1: Data preparation process. For simplicity, it is ignored that not all utterances were classified by each annotator.

2.1.1. *Data augmentation*

We used three data augmentation methods: adding noise, adding reverberation, and SpecAugment [20]. We balance the training data for each emotion class by oversampling with these data augmentation methods.

2.1.2. *Filtering*

At each generation, unlabeled data is filtered on the basis of the pseudo labels inferred by the teacher model. The filtering operation is applied generation by generation gradually so that utterances with higher scores are used for training at an earlier generation. In other words, restrictive and permissive filtering is imposed at early and late generations.

2.2. Data preparation

Assuming that multiple annotators classify each utterance into one of predefined emotion classes, we introduce some terminology. An utterance is referred to as *emotion-definite* if the label decided using majority voting by the annotators agrees with the one intended by the speaker. Otherwise, it is referred to as *emotion-ambiguous*. We note that this is not the only possible criterion to discriminate between emotion-definite and emotion-ambiguous utterances. If the emotion classes intended by the speakers are not available, e.g., in the case of a spontaneous emotion corpus, we may classify a given utterance based only on voting by annotators. Therefore, our method can also be applied to spontaneous emotion corpora.The discrimination between emotion-definite and emotion-ambiguous utterances depends on which annotations are used. If not specified explicitly, the discrimination is supposed to be performed using all annotations in a given corpus.

Table 1: Statistics of the initial data.

Available annotation ratio	r	1.0	0.5	0.5	0.25	0.25	0.25
Annotations per utterance	a	12	12	5	12	5	3
Labeled data size	$\|L_{r,a}\|$	3,085	1,551	2,845	770	1,373	2,272
Unlabeled data size	$\|U_{r,a}\|$	4,356	5,890	4,596	6,671	6,068	5,168
Label quality	$q_{r,a}$	1.0	1.0	0.897	1.0	0.903	0.856

We extract the labeled and unlabeled data from an existing emotional speech corpus, discarding part of the annotations. Fig. 1 illustrates the data preparation process. In this study, we evaluate the effectiveness of different strategies on preparing labeled data for semi-supervised learning. For this purpose, we impose the number of available annotations. Under this condition, there are two opposite directions concerning the annotation strategy. One direction is to prioritize the number of labeled utterances, sacrificing their reliability. The other is to give priority to the reliability of labels, sacrificing the amount. By comparing these strategies, we investigate the trade-off between the amount and reliability of labeled data for semi-supervised learning. We note that labeled and unlabeled data obtained in this manner belong to the same domain. Hence, there is no domain mismatch between them.

The data preparation process consists of two steps: selection of annotations and labeling of utterances. In the first step, we fix the number of available annotations to obtain labeled data. The ratio to all annotations in the original corpus is denoted by r. Moreover, we set the upper limit of the number of annotations for each utterance to a. Ignoring that part of the selected utterances may have less than a annotations, a represents the number of annotations per utterance. We can say that a represents the annotation strategy under a given cost constraint. Then, we select a subset of utterances from the original corpus so that the total number of annotations does not exceed the specified value. Here, utterances with more annotations are chosen first. No remaining annotations are used for training. While the label reliability becomes higher as we increase the number of annotations per utterance [9, 10], the amount of labeled data becomes smaller with the fixed number of available annotations. In short, r determines the number of available annotations, whereas a controls the balance between the amount and reliability of the labeled data.

In the second step, we divide the original corpus into labeled and unlabeled datasets using the selected annotations. Each of the selected utterances is classified as labeled data if it is emotion-definite. Here, only the selected annotations are used to determine an utterance is emotion-definite or emotion-ambiguous. The other utterances are classified as unlabeled data. The unlabeled dataset consists of the utterances that were not selected at the first step and the emotion-ambiguous utterances. The labeled and unlabeled datasets obtained in this manner are denoted by $L_{r,a}$, $U_{r,a}$, respectively. L and U denote the labeled and unlabeled datasets when all annotations are used, which are the same as the sets of emotion-definite and emotion-ambiguous utterances, respectively. The quality $q_{r,a}$ of the labeled dataset $L_{r,a}$ can be estimated as the ratio of utterances whose labels are "correct." Here, labels in the emotion-definite dataset L, determined using all annotations and the actors' intention, are supposed to be "correct," whereas this is not the only definition. We note that if an utterance in the labeled dataset $L_{r,a}$ is classified to the emotion-ambiguous dataset U, then it is counted as an incorrect label.

2.3. Model

We use an attention-based convolutional recurrent neural network (ACRNN) model that is equivalent to that investigated in [21]. This network consists of two convolutional layers, one time-distributed fully-connected layer, one bidirectional recurrent layer, one attention layer, one fully-connected layer, and one softmax layer. ACRNN is known to be efficient for SER [22, 23, 24, 25]. A 40-dimensional log mel-spectrogram was used as the input features calculated with a window size of 25 ms and a window shift of 10 ms. We applied z-score normalization to the input features.

3. EXPERIMENTS

3.1. Setup

In our experiments, we used the CREMA-D [2], which consists of 7,442 utterances by 91 actors, as the original corpus. The actors read aloud given sentences expressing one of the six emotions: neutral, happiness, sadness, anger, disgust, and fear. Crowdsourced annotators classified the emotion and rated the intensity of emotion level of presented utterances on the audio-only, visual-only, or audio-visual information. Between these types of annotations, we used all 73,058 categorical annotations based on audio-only information. 6 to 12 annotators evaluated each utterance. Therefore, the maximum value of the number of annotations per utterance is 12.

We trained the ACRNN model up to the fifth generation using the NST method. In other words, the final generation in our experiments was the fifth generation. Specifically, we compared the performance of the model trained on different labeled and unlabeled datasets at the zeroth generation shown in Table 1. We examined two ways of labeling for the initial labeled data: hard and soft labels. A hard label gives a value

of one to the emotion class that receives the most votes by annotators and a value of zero to the other classes. A soft label is obtained as a ratio of the number of votes received by each emotion class to the total number of votes. Conversely, we defined the pseudo labels assigned to the unlabeled utterances in the NST method as soft in our experiments. The utterances with scores in the top 20%, 40%, 60%, 80%, and 100% of unlabeled data were selected at each generation. Therefore, the size of training data grows linearly with each generation. Note that not only emotion-definite utterances that are not annotated but also emotion-ambiguous utterances are inevitably used in the training of student models.

For the data augmentation during training, we added noise signals from the DEMAND database [26] to each utterance with a signal-to-ratio (SNR) chosen from the uniform distributed from 0 to 30 dB randomly. Moreover, we randomly chose a room impulse response from the BIRD database [27] and convolved it for each utterance. Furthermore, we set the time wrapping, time masking, and frequency masking parameters of SpecAugment to 20, 15, and 100, respectively.

We evaluated the performance in F_1 score using 10-fold leave-one-speaker-group-out cross-validation. The 91 speakers in the corpus were divided into ten groups, each of which included 9 or 10 speakers. All samples were grouped into three sets based on the speaker groups: eight groups for training, another group for validation, and the last group for testing. From the training set, we chose the initial labeled data based on the limited annotation. The validation set was used to select the best generation and epoch. For testing, the utterances belonging to the emotion-definite dataset determined using all annotations, or namely $L = L_{1.0,12}$, were used.

The emotion-ambiguous utterances were eliminated from testing. Hence, the training and developing sets were dependent on r and a, while the test set was not. In other words, the evaluation was conducted using the "correct" labels.

3.2. Results

The results of our experiments are shown in Table 2. We note that the zeroth-generation model was trained using supervised learning only on the initial labeled data. The final-generation model was trained using the NST method. Therefore, the difference in performance between the zeroth- and final- generation models represents the improvement achieved by utilizing the unlabeled data. According to the results, the hard label outperformed the soft label under all conditions. When all the annotations were available ($r = 1.0$), there was almost no significant performance improvement by the NST method. This result indicates that adding emotion-ambiguous utterances to the training dataset does not bring significant benefits nor drawbacks.

On the other hand, the models trained using the NST method on both the labeled and unlabeled data outperformed those trained using supervised learning only on the same

Table 2: F_1 scores. r denotes the ratio of available annotations; a denotes the number of annotations for each utterance. Column titled label represents labeling type of the initial data.

r	a	**Label**	**Generation**	
			zeroth	final
1.0	12	hard	0.721	**0.727**
1.0	12	soft	0.681	0.717
0.5	12	hard	0.643	0.680
0.5	12	soft	0.641	0.647
0.5	5	hard	0.705	**0.722**
0.5	5	soft	0.673	0.677
0.25	12	hard	0.635	0.650
0.25	12	soft	0.587	0.628
0.25	5	hard	0.659	0.676
0.25	5	soft	0.637	0.647
0.25	3	hard	0.703	**0.705**
0.25	3	soft	0.654	0.677

labeled data for all conditions on a when only part of the annotations were available (i.e., $r = 0.5, 0.25$). For example, the improvement in the F_1 score between the zeroth and final generations with the hard initial label was 0.037 when $(r, a) = (0.5, 12)$. The p-value, effect size, and power were 0.001, 1.38, and 0.97 in a paired t-test based on 10-fold cross-validation with the significance level of 0.05 [28, 29]. These results indicate that semi-supervised learning, or utilizing unlabeled data, contributes to improving the performance of SER regardless of the inclusion of utterances whose emotion is ambiguous in the unlabeled data. In particular, when half the annotations were available ($r = 0.5$), the annotation strategy of five annotators per utterance ($a = 5$) achieved the best performance. The difference in the F_1 score between $a = 5$ and $a = 12$ at the final generation with the hard initial label was 0.043. The p-value, effect size, and power were 0.017, 0.91, and 0.73 in the same t-test as above.When the ratio of available annotations was one quarter ($r = 0.25$), the annotation strategy of three annotators per utterance ($a = 3$) was the best, whereas the performance improvement by NST is tiny. We note that there was no deterioration in the performance by using the NST method under all conditions examined in this study. To summarize, our experiments show that semi-supervised learning is beneficial for improving the performance of SER when we adopt an appropriate labeling strategy for the preparation of initial labeled data, even if the unlabeled data includes utterances with ambiguous emotion.

4. DISCUSSION

4.1. Analysis

In this section, we analyze the results of our experiments further. Below, we let (r, a) denote a condition on the available

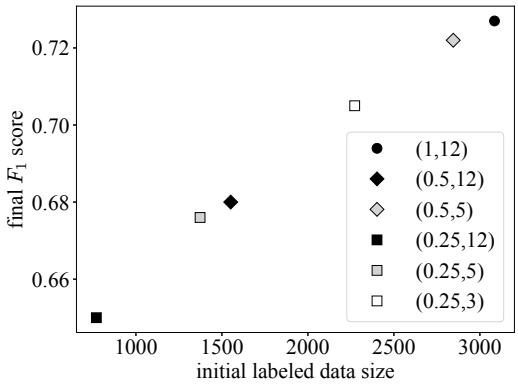

Fig. 2: Relationship between the initial labeled data size $|L_{r,a}|$ and the final F_1 score.

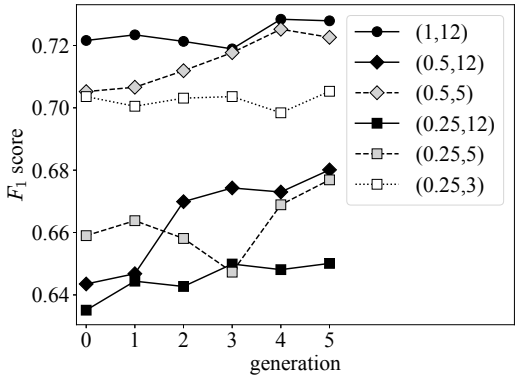

Fig. 3: F_1 scores at each generation.

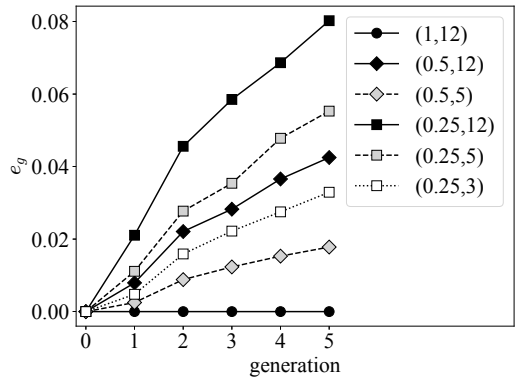

Fig. 4: Ratio of emotion-definite utterances with incorrect pseudo labels in the training data e_g at each generation.

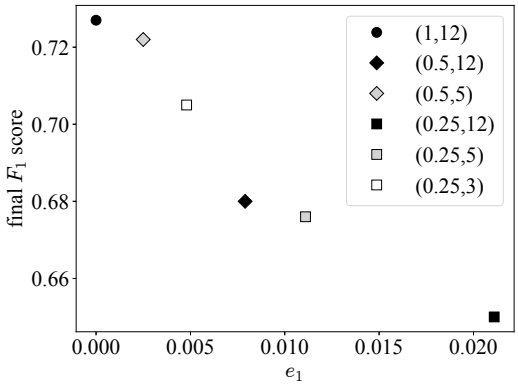

Fig. 5: Relationship between e_1 and the final F_1 score.

annotations ratio and the number of annotations per utterance. Moreover, as the initial data with hard labels outperformed that with soft labels as shown in Table 2, we focus on the former case.

Two major factors affect the effectiveness of NST for SER: the amount and reliability of the initial labeled data. With the fixed number of available annotations, there is a trade-off between these two factors, which is controlled by the number of annotations per utterance, namely a. In Table 1, we can see that increasing the number of annotators per utterance a results in a large drop in the initial labeled data size $|L_{r,a}|$, whereas it resulted in only a small improvement in the initial label quality q. This observation suggests that we do not significantly benefit from increasing a. This is supported by the results in Fig. 2 that show the relationship between the initial labeled data size $|L_{r,a}|$ and the final F_1 score. The results shown in this figure indicate that the overall performance of SER obtained using the NST method is mainly affected by the amount of initial labeled data, as long as its reliability remains at an acceptable level. Note that we did not examine an extremely small value of a, such as a single annotator per utterance, because the initial data obtained under such a condition tends to be inconsistent and lacks integrity. In the case of the CREMA-D dataset, five annotators per utterance were sufficient to resolve most of the subjectivity of emotion classification and benefit from the NST method.

The effect of the initial labeled data size can be understood in the following way. Fig. 3 shows the F_1 scores up to each generation. When we increase a with a fixed value of r, we obtain a small number of utterances with highly reliable labels. Owing to an insufficient amount of training data, a model trained using supervised learning on such small-size initial labeled data results in low generalization performance. On the other hand, if the value a is small, many utterances with moderately reliable labels are obtained. With the benefit of a large amount of training data, we can train a model with high generalization performance using supervised learning. Furthermore, the high or low F_1 scores at the zeroth generation cascaded through the generations during the NST process, consequently resulting in high or low final F_1 scores, as shown in Fig. 3. In other words, the annotation strategy represented by a that achieved the best performance at the zeroth generation also achieved the best performance at the final generation, compared with the same number of available annotations, i.e., the same value of r. Therefore, a large amount of initial labeled data with moderate reliability results in high performance after the NST process.

Next, we consider the effect of the pseudo label reliability. The error in the inference of pseudo labels is amplified as the generation progresses. Therefore, we need to prevent the occurrence of "incorrect" pseudo labels to obtain a high performance using the NST method. Here, we define the correctness of a pseudo label in the following way: First, we classify an emotion-definite utterance into the emotion class with the highest score on its pseudo label. If the class determined in this manner is the same as the "correct" class determined using all annotations in the original corpus, then the pseudo label is judged as correct; otherwise, it is incorrect. In addition, we introduce an important quantity denoted by e_g: the ratio of emotion-definite utterances with incorrect pseudo labels in the training data at the generation g. As only the initial labeled data was used at the zeroth generation, e_0 is 0.0. Fig. 4 and 5 show the e_g at each generation and the relationship between e_1 and the final F_1 score, respectively. In Fig. 4, we can observe that if the degree of error in the inference of pseudo labels was low in the first generation, then it remained low as the generation progressed. Consequently, the annotation strategies that started with lower e_g values achieved higher performance, as measured by the F_1 score at the final generation; this is shown in Fig. 5. Therefore, we can say that a low e_g (i.e., reliable pseudo labels) is a good indicator of how well the NST method is working.

Moreover, we consider the effect of emotion-ambiguous utterances. When $r = 1.0$, i.e., all annotations in the original corpus were used, the unlabeled data consisted only of emotion-ambiguous utterances. Under this condition, we found no degradation in the performance, as measured by the final F_1 score, as shown in Table 2. This indicates that the NST method contributes to improving performance, even though emotion-ambiguous utterances are inevitably included in a dataset for SER.

4.2. Application on new data

Finally, we discuss how a new dataset for SER should be annotated when the NST method is assumed to be used. The optimal number of annotations per utterance with a fixed number of total annotations is not necessarily the same among different datasets. It may be influenced by various characteristics of the dataset, such as style (acted or spontaneous), situation (script reading or improvisation), recording environment, language, and culture. Therefore, we need to find the optimal annotation strategy depending on each dataset.

In the following, we describe a method for annotating a new dataset for SER based on the findings from our experiments: First, we randomly choose a small subset of speakers and their utterances, which is referred to as an evaluation set. Each utterance in the evaluation set should be classified by a sufficient number of annotators such that we can determine their "correct" labels. According to previous studies [3, 4], 20 annotators per utterance are sufficient in most cases. It is important to note that we cannot know the "correct" labels of all utterances in a new dataset without a vast number of annotations, which required considerable time and effort. Furthermore, we choose an indicator of how well the current annotation data is. There are several candidate indicators: the F_1 score of the model trained using supervised learning, the F_1 score of the final-generation model trained using NST, and the ratio of emotion-definite utterances with incorrect pseudo labels in the training data e_g. The value of the indicator I should be calculated solely on the evaluation set because it requires "correct" labels. Finally, we perform the main annotation process in the following way:

1. Set the ratio of the utterances to be annotated s and the number of annotators per utterance a to initial values s_0 and a_0, respectively.
2. Randomly choose utterances and annotate them under the condition of $s = s_0$ and $a = a_0$.
3. Increase a and add annotations until I saturates.
4. Increase s and add annotations until I saturates.
5. Repeat Step 3 to Step 4 until the total number of annotations reaches the upper limit.

The evaluation set should be eliminated from the target of the main annotation process and used solely to evaluate the value of the indicator.

5. SUMMARY

We investigated the utilization of unlabeled data for SER to reveal efficient labeling strategies in preparing initial labeled data, considering the trade-off between its amount and reliability. For this purpose, we limited the available annotations in the CREMA-D dataset and trained the ACRNN model using the NST method with different conditions on the available annotation ratio and the upper limit of annotations per utterance. Furthermore, we examined hard and soft labels concerning the initial labeled data. Our experiments showed that the semi-supervised models trained on both the labeled and unlabeled data outperformed the supervised model trained solely on the same labeled data. However, the degree of performance improvement is greatly influenced by the manner in which the initial labeled data is prepared. Our experiments indicate that we should increase the amount of labeled data with moderate reliability to making the most of given unlabeled data rather than increasing the reliability of a small amount of labeled data.

This study suggests that it is essential to consider the trade-off between the amount and reliability of labeled data when using a semi-supervised learning method. The efficient labeling strategy revealed in this study is significant when utilizing a large-scale dataset without emotion labels, including public domain video archives.

6. REFERENCES

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335, 2008.

[2] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affecting Computation*, vol. 5, no. 4, pp. 377–390, 2014.

[3] Alec Burmania, Mohammed Abdelwahab, and Carlos Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *ICASSP 2016 – 42th International Conference on Acoustics, Speech, and Signal Processing, March 20-25, Shanghai, China, Proceedings*, 2016, pp. 5190–5194.

[4] Alec Burmania and Carlos Busso, "A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20–24, Stockholm, Sweden, Proceedings*, 2017, pp. 152–156.

[5] Alexander Philip Dawid and Allan M Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[6] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in neural information processing systems*, 2009, pp. 2035–2043.

[7] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie, "The multidimensional wisdom of crowds," in *Advances in neural information processing systems*, 2010, pp. 2424–2432.

[8] Yoshinao Sato and Kouki Miyazawa, "Quality estimation for partially subjective classification tasks via crowdsourcing," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 229–235.

[9] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Fruhholz, and Bjorn Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 31–43, 2018.

[10] Jianhua Tao, Jian Huang, Ya Li, Zheng Lian, and Mingyue Niu, "Semi-supervised ladder networks for speech emotion recognition," *International Journal of Automation and Computing*, vol. 16, pp. 437–448, 2019.

[11] Srinivas Parth and Carlos Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Acousics, Speech and Signal Processing*, vol. 28, pp. 2697–2709, 2020.

[12] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Bjórn Wolfgang Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective computing*, 2020.

[13] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR 2020 – Conference on Computer Vision and Pattern Recognition, June 14–19, Virtual*, 2020, pp. 10687–10698.

[14] Zixing Zhang, Jing Han, Jun Deng, Xinzhou Xu, Fabien Ringeval, and Björn Schuller, "Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning," *IEEE Access*, vol. 6, pp. 22196–22209, 2018.

[15] Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le, "Improved noisy student training for automatic speech recognition," in *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, October 25–29, Shanghai, China*, 2020, pp. 2817–2821.

[16] Hyun-Jin Park, Pai Zhu, Ignacio Lopez Moreno, and Niranjan Subrahmanya, "Noisy student-teacher training for robust keyword spotting," *arXiv preprint arXiv:2106.01604*, 2021.

[17] Zhepei Wang, Ritwik Giri, Umut Isik, Jean-Marc Valin, and Arvindh Krishnaswamy, "Semi-supervised singing voice separation with noisy self-training," in *ICASSP 2021 – 47th International Conference on Acoustics, Speech, and Signal Processing, April 15-20, Calgary, Canada, Proceedings*, 2021, pp. 31–35.

[18] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *IJCNN 2016 – International Joint Conference on Neural Networks, Junly 24–29, Vancouver, Canada*, 2016, pp. 566–570.

[19] Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and

Yushi Aono, "Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification," in *ICASSP 2018 – 44th International Conference on Acoustics, Speech, and Signal Processing, April 15-20, Calgary, Canada, Proceedings*, 2018, pp. 4964–4968.

[20] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15–19, Graz, Austria*, 2019, pp. 2613–2617.

[21] Zhi Zhu and Yoshinao Sato, "Reconciliation of multiple corpora for speech emotion recognition by multiple classifiers with an adversarial corpus discriminator," in *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, October 25–29, Shanghai, China*, 2020, pp. 2342–2346.

[22] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20–24, Stockholm, Sweden, Proceedings*, 2017, pp. 1098–1102.

[23] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[24] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15–19, Graz, Austria*, 2019, pp. 2083–2087.

[25] Zhichao Peng, Xingfeng Li, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi, "Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020.

[26] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591, 2013.

[27] Francois Grondin, Jean-Samuel Lauzon, Simon Michaud, Mirco Ravanelli, and Francois Michaud, "BIRD: Big impulse response dataset," *arXiv preprint arXiv:2010.09930*, 2020.

[28] Thomas G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[29] Jacob Cohen, *Statistical power analysis for the behavioral sciences*, Academic press, 2013.