Interspeech 2024
1-5 September 2024, Kos, Greece

Oversampling, Augmentation and Curriculum Learning for Speaking Assessment with Limited Training Data

Tin Mei Lun[1], *Ekaterina Voskoboinik*[1], *Ragheb Al-Ghezi*[1], *Tamas Grosz*[1], *Mikko Kurimo*[1]

[1] Department of Information and Communications Engineering, Aalto University, Finland
first.lastname@aalto.fi

Abstract

Automated assessment systems for spontaneous speech are an increasingly important component in language proficiency tests and learning platforms. These systems have seen remarkable development in recent years, driven by advances in self-supervised learning. Nevertheless, in languages such as Finnish and Finland Swedish, their performance is still limited by the low-resource and imbalance nature of their data. To alleviate these issues, this work evaluates two data-level methods: oversampling and curriculum learning. Our results reveal that combining these methods results in the greatest boost to model performance, achieved without additional data or modification to the model structure.

Index Terms: L2 speaking assessment, data imbalance, low-resource, oversampling, curriculum learning

1. Introduction and Related Work

The increasing amount of second language (L2) learners worldwide has prompted the need for effective automatic speaking assessment (ASA) systems to support formal language tests and computer-assisted language learning. However, the development in this area has focused mainly on L2 English [1, 2, 3] owing to its relatively high demand and abundant resources. For languages such as Finnish and Finland Swedish, building such ASA systems remains a very challenging task due to their low-resource nature. Additionally, in the small datasets [4] that are available, data imbalance emerges as a prominent issue.

More specifically, this work focuses on ASA systems that evaluate spontaneous L2 speech and generate holistic scores as output. Due to high data collection cost and the specificity of the target domain, such data are unlikely to become available on a large scale. Additionally, beginner L2 learners are often reluctant to being recorded while near-native learners tend to be fewer in number, thus naturally results in a proficiency imbalance in the datasets with intermediate level samples being the majority. Given the intrinsic data scarcity and imbalance in these datasets, alongside model development, it is crucial to explore methods to fully utilize available data.

Wav2vec 2.0 is a framework that leverages a large amount of unlabelled speech to learn powerful contextualised representations which can be fine-tuned to perform specific speech tasks using significantly smaller datasets [5]. This architecture has not only dramatically improved performance across a range of speech recognition [6, 7] and classification [3, 8] tasks, but it has also enabled low-recourse tasks that were previously considered unattainable, including the ASA tasks of interest. In [9], the authors developed L2 Finnish and Finland Swedish ASA systems using Wav2vec 2.0 features, which outperformed hand-crafted speech features. However, due to insufficient data, certain classes (lowest and highest proficiency levels) were not evaluated. As a result, their model performance was still limited by the small and imbalanced training data despite leveraging pre-trained Wav2vec 2.0 models.

Inspired by the challenges faced in [9], this work explores strategies to mitigate data scarcity and imbalance in the context of ASA. Common strategies include data-level methods such as resampling, and algorithm-level methods such as cost-sensitive learning [10]. This work favours the former approach not only because of its adaptability and simplicity, but also the intuitive interpretations it offers, which is beneficial in the context of ASA. In this work, two data-level methods are examined: 1) oversampling and 2) curriculum learning.

Random **oversampling** is a resampling method that achieves class balance by replicating instances in minority classes. SMOTE [11], A more robust oversampling method, interpolates between neighbouring instances from minority classes to simultaneously achieve balanced distribution and enhance in-class variation. These techniques have been shown to improve model performance in imbalanced speech and text classification applications [12, 13]. To apply SMOTE on speech classification tasks, one could interpolate between speech representations [13]. However, it suffers from poor interpretability and strong assumptions on the properties of the speech representations. Therefore, we propose combining random oversampling with augmentation techniques to enrich in-class variability on the input level.

Curriculum learning (CL) introduces training data to the model in gradually increasing level of difficulty as opposed to a random order, drawing inspiration from education systems [14]. While CL could be implemented on the data-level, algorithm-level [15] and a combination of both [16], this work focuses on the former. Although tackling class imbalance was not its initial objective, CL is shown in some studies [17, 16, 18] to be effective in classification tasks that suffered from this issue. Indeed, instances in minority classes are typically harder for models to generalise [19] and can thus be considered to be difficult compared to instances from majority classes. In [17], the authors proposed class-wise curriculum learning (CCL), that was shown to mitigate class imbalance in a multi-class image classification task. Training data were sorted in a class-wise manner according to the baseline model's performance on each class. As the training progress, more classes are added from easy to difficult.

To the best of our knowledge, this is the first study to apply these methods in the training of ASA systems with limited data. Our findings demonstrate that 1) oversampling combined with time-domain augmentation consistently improves model performance, and 2) with appropriate configurations, CCL also has a positive impact on the model performance.

10.21437/Interspeech.2024-760

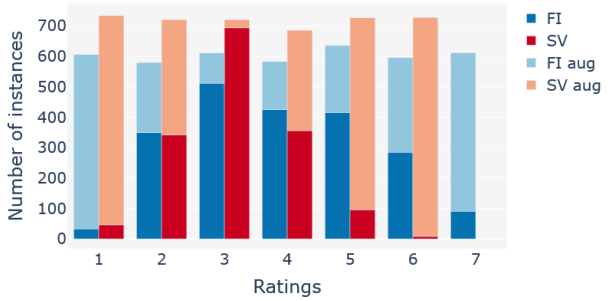

Figure 1: *Class distribution across ratings in the original datasets (FI and SV) and after oversampling (FI + FI aug and SV + SV aug). Class 1 represents pre-A1 and 7 represents C2.*

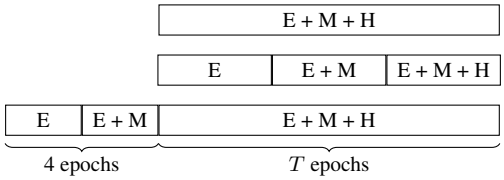

Figure 2: *Pacing strategies for the **baseline** (top), **CCL 1** (middle) and **CCL 2** (bottom). **E** denotes the Easy subset, **M** denotes Medium and **H** denotes Hard.*

2. Methods

2.1. Oversampling (OS) + augmentation

First, the datasets were randomly oversampled to achieve class balance. Then, to introduce diversity, duplicated instances were augmented using *two* randomly chosen time-domain augmentation techniques including 1) time masking, 2) frequency masking, 3) additive noise, 4) reverberation, 5) pitch shift and 6) tempo perturbation. Augmentation was performed using the WavAugment toolkit [20] and the parameters were chosen to ensure the labels (holistic scores) were unchanged. Thus, this method aimed to construct oversampled datasets that:

1. were class-balanced;
2. contained a copy of the original (un-augmented) dataset and oversampled instances augmented with *two* randomly chosen methods from the six listed above;
3. had $\alpha \times N$ instances, where N is the number of instances in the original dataset and α is the oversampling factor.

Objective 2 was introduced to ensure at least one un-augmented copy of each instance was present in the training data, as some of the augmentation methods (e.g. time masking) could introduce disruption to the speech. For objective 3, note that to achieve class balance while including all instances in the original dataset, α had to be higher for datasets with a higher imbalance ratio. Figure 1 illustrates the final class distribution after oversampling (FI + FI aug and SV + SV aug).

2.2. Class-wise curriculum learning (CCL)

To implement CCL, the class (holistic score) subsets were first sorted according to their F_1 scores using a trained baseline model: the lower the F_1 score, the higher the difficulty. The sorting was done before the first epoch and the order remained static throughout training.

Similar to [21], these class subsets were then grouped into three difficult levels: Easy, Medium and Hard. During training, the model first learned from the Easy subset. Subsequent subsets were added to the training data after every τ epochs, where τ was determined by the chosen pacing strategy. Two pacing strategies (illustrated in fig. 2) were examined:

1. **CCL 1**: $\tau = \lfloor T/3 \rfloor$ and total training epoch = T
2. **CCL 2**: $\tau = 2$ and total training epoch = $T + 4$

where T was the number of training epochs for the baseline model. Strategy 1 was similar to the pacing strategies experimented in the original CL paper [14], while strategy 2 used CL as a warm-up technique, similar to the SortaGrad method introduced in [22]. Note that in each epoch, the training examples were fed to the model in a random order.

In the original CCL paper [17], loss estimates were used as the sorting criteria. However, results in [21] showed that using the actual evaluation metrics led to better model performance (using WER in ASR task) which motivated our choice of F_1 scores as the sorting criteria.

2.3. CCL + mixing

This method extended CCL with an additional step. Upon obtaining the three difficulty groups, a small portion (e.g. 20%) of the instances from the Easy group were swapped with instances from the Medium and Hard group (e.g. 10% from each group). This method was shown in [21] to enhance the performance of metric-based CL in ASR task. This method can be interpreted as a "softer" grouping strategy where instances are sampled without replacement based on sampling weights determined based on their difficulty scores.

3. Data

This study was repeated on two L2 speech datasets: Finnish and Finland Swedish. Unfortunately, only the Finnish one is openly shared[1]. The datasets contained recorded responses to speaking tasks, their corresponding transcripts and ratings. The collection process was described in detail in [4]. The L2 Finnish dataset contained 2112 spoken responses ranging from 2 to 91 seconds for 29 different speaking tasks collected from 308 unique participants (secondary school and university students). The L2 Finland Swedish data contained 1542 responses ranging from 1 to 31 seconds for 22 tasks collected from 178 participants (secondary school students). Their total duration was 14.1 hours and 5.6 hours respectively.

Among the six criteria covered by the speech ratings, this work focused primarily on the *holistic* scores which is a seven-level scale: pre-A1, A1, A2, B1, B2, C1 and C2, corresponding to the Common European Framework Reference (CEFR) [23] for language proficiency. Majority of the responses were assessed by at least two raters. To determine their reference classes, we first averaged the scores given to each spoken response by different raters. For averaged scores that fell at the midpoints between two integers (e.g. 2.5), we randomly assigned the reference classes from adjacent integers. For other cases, we rounded the averaged scores to their closest integers. The final class distribution is illustrated in fig. 1 (FI and SV). Unlike the experiments in [9], the underrepresented classes were not omitted, except class 7 of the L2 Finland Swedish dataset which did not contain any instances.

[1] https://www.kielipankki.fi/corpora/digitala/

4. Experiments and Results

Similar to [9], four-fold cross-validation was adopted to make use of all instances in the small datasets for training and evaluation. The data splitting was performed such that all subsets contained non-overlapping speakers. Training was performed separately for each fold and the evaluation metrics were averaged across folds to obtain the final results.

For training, we set $T = 10$, hence the total number of epochs was 14 for CCL 2 models and 10 for others. After experimenting with different T values, we concluded that $T > 10$ did not yield improved results. The total training procedure took between 4 and 9 hours per fold for the Finnish dataset and between 1 and 3 hours for the Finland Swedish dataset using Nvidia Tesla V100 GPU cards. The training script and parameters can be found on our Github repository[2].

4.1. Scoring models

The scoring models followed a Wav2vec 2.0-based architecture as described in [9]. The Finnish model was initialised using a Wav2vec2-Large model pre-trained on 42.5k hours of Uralic speech and fine-tuned on 100 hours of transcribed colloquial Finnish speech. For Finland Swedish, the model was initialised using a monolingual Wav2vec2-Large model pre-trained on 11.5k hours of Swedish speech and fine-tuned on transcribed speech in the same language[3]. It is worth noting that up to this point, the models were trained using purely native (L1) speech from various domains. To adapt to L2 speech and the desired domain (i.e. the speaking tasks), the two models were further ASR-fine-tuned using the target datasets. Such continually fine-tuned models were shown in [9, 24] to achieve improved performance. Finally, using these models as initialisation, the scoring models were trained with classification heads and with all layers configured as trainable.

4.2. Evaluation metrics

To evaluate the model performance across all classes equally, *macro* precision, recall and F_1 scores were used as metrics. In addition, we also report quadratic-weighted kappa (QWK), a common metric used to evaluate inter-rater agreement for scoring tasks. The evaluation was performed on the models from the *best* epochs determined by their macro F_1 scores. In addition, we computed the confidence intervals for the macro F_1 scores using the toolkit [25] with speakers as sampling condition.

4.3. Results

After training the baseline model (**BASE** in table 2), the difficulty grouping illustrated in table 1 was obtained and used for CCL training. For the OS experiments, we used an oversampling factor α of 2 for the Finnish models and 2.8 for the Finland Swedish models, the latter being higher due to the greater class imbalance in the dataset.

Table 2 presents the results of our experiments. For the Finnish dataset, **OS_CCL_2_M** achieved the best precision and F_1 scores, although fell slightly behind with recall and QWK. As for the Finland Swedish dataset, this model outperformed others across all metrics. More specifically, they brought a 12% and 26% relative improvement to the F_1 scores, respectively, compared to **BASE**.

[2]https://github.com/lunsanna/asa-with-limited-data

[3]https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish

Table 1: *Difficulty grouping based on the F_1 of **BASE**.*

	Group	Easy		Medium		Hard		
FI	Class	3	2	6	5	4	1	7
	F_1 %	54	52	50	44	43	0	0
SV	Class	3	2	4	5	1	6	7
	F_1 %	62	57	51	42	0	0	–

Comparing the results of each method more closely, for both datasets, **OS** models outperformed their **BASE** counterparts across precision, recall and F_1 with only one exception for precision. As for QWK, the same pattern was observed in the Finland Swedish models, although absent in the Finnish models. Overall, there was enough evidence that combining oversampling and time-domain augmentation improved model performance in our ASA tasks.

As for the CCL models, neither **CCL_1** nor **CCL_2** showed a consistent advantage over their non-CCL counterparts (**BASE** and **OS**). For the Finnish dataset, all CCL models achieved lower F_1 scores than their non-CCL counterparts, except the best performing model **OS_CCL_2_M**. A similar observation was also found for the Finland Swedish dataset. As for mixing, it consistently led to improved results in the Finland Swedish CCL models, although its efficacy on the Finnish CCL models was less conclusive. All in all, even though these methods alone did not show definitive improvement, their combination coupled with appropriate configuration led to the best models in both datasets.

Table 2: *Experimental results. **P** denotes precision and **R** denotes recall. **OS** refers to the experiments conducted on the oversampled datasets while **BASE** refers to those conducted on the original datasets. Models ending with **M** refer to CCL models performed with mixing strategy applied.*

	P %	R %	F_1 %	QWK
Finnish				
BASE	33.56	36.36	34.59	0.801
BASE_CCL_1	32.63	34.38	33.02	0.790
BASE_CCL_2	33.38	36.15	34.39	0.800
BASE_CCL_1_M	32.59	35.55	33.11	**0.805**
BASE_CCL_2_M	33.32	35.80	34.00	0.801
OS	39.06	**38.86**	38.61	0.797
OS_CCL_1	37.39	38.28	37.27	0.791
OS_CCL_2	37.80	38.06	37.79	0.780
OS_CCL_1_M	36.14	37.66	36.33	0.803
OS_CCL_2_M	**41.62**	38.26	**38.80**	0.797
Finland Swedish				
BASE	38.00	34.87	35.49	0.658
BASE_CCL_1	37.43	29.53	28.33	0.597
BASE_CCL_2	37.75	32.94	33.37	0.637
BASE_CCL_1_M	38.17	30.92	29.32	0.618
BASE_CCL_2_M	37.96	33.88	34.60	0.629
OS	41.94	38.02	39.19	0.668
OS_CCL_1	36.15	35.29	35.53	0.658
OS_CCL_2	39.10	38.76	38.74	0.641
OS_CCL_1_M	42.73	40.29	41.08	0.643
OS_CCL_2_M	**61.23**	**41.28**	**44.83**	**0.685**

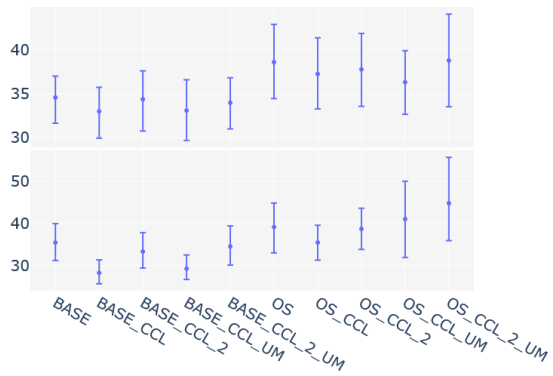

Figure 3: *Confidence intervals of the F_1 scores of all **Finnish (top)** and **Finland Swedish (bottom)** models.*

5. Discussion

By applying methods to mitigate data scarcity and imbalance, we achieved moderate improvement in the overall ASA model performance without additional data or modification to the model structure. Oversampling combined with augmentation techniques, despite all its simplicity, was found to be the most reliable method that led to consistent improvement to the models. However, this came with the cost of higher model variance, revealed by their wider confidence intervals, compared to their BASE counterparts (illustrated in fig. 3). Indeed, overfitting on minority classes and thus higher model variance is an inherent issue associated with oversampling methods [10]. Therefore, one must consider such a trade-off when adopting this method.

Despite the potential benefits of CL, it was evident that not all configurations yielded improved performance. Comparing the two CCL strategies, **CCL 2** consistently achieved higher F_1 scores than **CCL 1**. The latter is a traditional pacing strategy that divides the training into equal-length phases, which was shown to be effective in the original CL [14] and CCL [17] paper. However, their datasets contained 10,000 - 50,000 training samples for image classification tasks and over 600 million for language modelling tasks, which were significantly larger than ours. Compared to this traditional pacing method (**CCL 1**), we hypothesise that CL is more effective as a warm-up strategy (**CCL 2**) for datasets that are *both imbalanced and small*. We speculate that with small datasets, the scarcity of unique samples in the difficult (underrepresented) classes limits their ability to contribute significantly once the models are trained extensive on well-represented classes. As a result, it is more beneficial to use easier samples only for initialisation while performing the majority of the training on the full dataset. The best **CCL 2** models achieved F_1 scores of 25.16 (FI) and 33.17 (SV) after the warm-up epochs, demonstrating the effectiveness of CL initialization.

Nonetheless, in most settings, **CCL 2** still yielded lower F_1 scores compared to their non-CCL counterparts, except when combined with the mixing strategy. Consider the OS experiments, although **CCL 2** alone (**OS_CCL_2**) led to lower performance compared to **OS**, pairing it with the mixing strategy (**OS_CCL_2_M**) achieved the best models for both datasets. This finding suggests that mixing a small portion of samples across difficult groups enhances the effectiveness of CCL on our small and imbalanced datasets. This phenomenon can be explained by the enhanced class diversity, especially in the early stage of training. This interpretation is consistent with the results in [26, 27], which demonstrated that exposing the model to a diverse set of samples throughout training can lead to better performance compared to strict difficulty ordering.

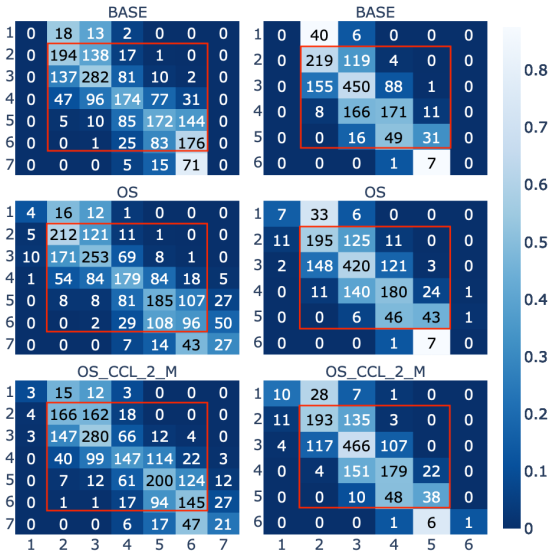

Figure 4: *Confusion matrices of the classification results of three chosen models for the **Finnish (left)** and **Finland Swedish (right)** dataset. Rows represent labels and columns represent predictions. Grid colours represent values normalised over rows, and annotations are the number of samples. The red boxes highlight classes that are not considered under-represented.*

As we try to enhance the classifier's performance for underrepresented classes, it is also worth evaluating whether the performance in other classes is compromised. From fig. 4, for both datasets, **OS** showed a performance gain in underrepresented classes (1 and 7 for FI, 1 and 6 for SV) and lost in other classes compared to **BASE**. Interestingly, for the Finnish dataset, the results of **OS_CCL_2_M** demonstrated a balanced performance trade-off between these two groups, while for the Finland Swedish dataset, an enhanced performance for both groups. A possible explanation is that through CCL warm-up on easier classes while maintaining class diversity, **OS_CCL_2_M** delays the overfitting issue caused by oversampling.

Finally, the QWK scores between human raters were **0.745** (FI) and **0.578** (SV), lower than those between our classifiers and the average human ratings. This could be explained by the reduced human bias as we used average scores for model training. However, QWK varied only slightly among classifiers without showing a consistent trend across datasets, indicating no strong evidence that these data-level methods improved model performance in terms of agreement between the human raters and our classifiers.

6. Conclusion

In this paper, we examine data-level strategies to mitigate data scarcity and class imbalance when training assessment systems for L2 spontaneous speech. Our results reveal that combining oversampling, augmentation, curriculum learning as a warm-up strategy and mixing technique yield the best model performance. Future work could consider integrating these methods with algorithm-level strategies for potential further model enhancement in low-resource and class-imbalanced conditions.

7. Acknowledgements

We are grateful for the Academy of Finland project funding "Digital support for training and assessing second language speaking" and "Automatic assessment of spoken interaction in second languag" (grant number 322625 and 355587 respectively). The computational resources were provided by Aalto ScienceIT.

8. References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009, spoken Language Technology for Education.

[2] J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," in *Proc. SLT*, 2014, pp. 294–299.

[3] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken english using wav2vec 2.0," in *Proc. SLT*, 2023, pp. 1088–1095.

[4] M. Kurimo, Y. Getman, E. Voskoboinik, R. Al-Ghezi *et al.*, "New data, benchmark and baseline for l2 speaking assessment for low-resource languages," in *Proc. SLaTE*, 2023, pp. 166–170.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[6] R. Al-Ghezi, Y. Getman, A. Rouhe, R. Hildén, and M. Kurimo, "Self-Supervised End-to-End ASR for Low Resource L2 Swedish," in *Proc. Interspeech*, 2021, pp. 1429–1433.

[7] K. Masuda, J. Ogata, M. Nishida, and M. Nishimura, "Throat microphone speech recognition using wav2vec 2.0 and feature mapping," in *Proc. Global Conference on Consumer Electronics (GCCE)*, 2022, pp. 395–397.

[8] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. ICASSP*, 2023, pp. 1–5.

[9] R. Al-Ghezi, Y. Getman, E. Voskoboinik, M. Singh, and M. Kurimo, "Automatic rating of spontaneous speech for low-resource languages," in *Proc. SLT*, 2023, pp. 339–345.

[10] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *The Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] L. B. Letaifa and M. I. Torres, "Perceptual borderline for balancing multi-class spontaneous emotional data," *IEEE Access*, vol. 9, pp. 55 939–55 954, 2021.

[13] H. Rathpisey and T. B. Adji, "Handling imbalance issue in hate speech classification using sampling-based methods," in *Proc. 5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 193–198.

[14] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, p. 41–48.

[15] T. Castells, P. Weinzaepfel, and J. Revaud, "Superloss: A generic loss for robust curriculum learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4308–4319.

[16] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan, "Dynamic curriculum learning for imbalanced data classification," in *Proc. ICCV*, 2019, pp. 5016–5025.

[17] M. Escudero-Viñolo and A. López-Cifuentes, "CCL: Class-wise curriculum learning for class imbalance problems," in *Proc. ICIP*, 2022, pp. 1476–1480.

[18] X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Improving imbalanced text classification with dynamic curriculum learning," in *Proc. 18th International Conference on Mobility, Sensing and Networking (MSN)*, 2022, pp. 1031–1036.

[19] S. Yadav and G. P. Bhole, "Handling imbalanced dataset classification in machine learning," in *Proc. Pune Section International Conference (PuneCon)*, 2020, pp. 38–43.

[20] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P.-E. Mazaré, M. Douze, and E. Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," in *Proc. SLT*, 2021, pp. 215–222.

[21] G. Karakasidis, T. Grósz, and M. Kurimo, "Comparison and Analysis of New Curriculum Criteria for End-to-End ASR," in *Proc. Interspeech 2022*, 2022, pp. 66–70.

[22] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai *et al.*, "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proc. 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 173–182.

[23] Council of Europe, *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Strasbourg, 2020.

[24] S. Jie, Z.-H. Deng, and Z. Li, "Alleviating representational shift for continual fine-tuning," in *Proc. CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 3810–3819.

[25] L. Ferrer and P. Riera, "Confidence Intervals for evaluation in machine learning," https://github.com/luferrer/ConfidenceIntervals, 2023, [Computer software].

[26] M. Sachan and E. Xing, "Easy questions first? a case study on curriculum learning for question answering," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 453–463.

[27] P. Soviany, "Curriculum learning with diversity for supervised computer vision tasks," in *Proc. 11th International Workshop on Modelling and Reasoning in Context (MRC)*, J. Cassens, R. Wegener, and A. Kofod-Petersen, Eds. Galicia, Spain: Digital ECAI 2020, Aug. 2020, pp. 37–44.