MINING HARD SAMPLES LOCALLY AND GLOBALLY FOR IMPROVED SPEECH SEPARATION

Kai Wang[1], *Yizhou Peng*[1], *Hao Huang*[1,3*], *Ying Hu*[1], *Sheng Li*[2]

[1]School of Information Science and Engineering, Xinjiang University, Urumqi, China
[2]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[3]Xinjiang Provincial Key Laboratory of Multi-lingual Information Technology, Urumqi, China

terry_wang@stu.xju.edu.cn, hwanghao@gmail.com

ABSTRACT

Speech separation dataset typically consists of hard and non-hard samples, and the former is minority and latter majority. The data imbalance problem biases the model towards non-hard samples and weakens the generalization capability. Given that the average separation performance is sufficiently good, improving hard samples may contribute more to back-end tasks. In this paper, we propose two methods to alleviate data imbalance in speech separation task, based on local and global hard sample mining. For the local, we propose weighted loss to compensate for hard samples by increasing their weights in each batch. For the global, we perform global hard sample mining and re-sample to increase the proportion of hard samples in the training set. Because hard sample mining using objective loss in dynamic mixing leads to local results, we propose an indirect method using speaker-specific parameters, based on the fact that pitch median difference and x-vector cosine distance of two speakers in a mixture are closely correlated with separation SI-SNRi. Experimental results show that both methods decrease the percentage of hard samples in the test set than using dynamic mixing only while keeping the average SI-SNRi comparable, and the global method shows more promising results than the local one.

Index Terms— Speech separation, data imbalance, dynamic mixing, weighted loss, hard sample mining

1. INTRODUCTION

Single-channel speech separation is important and challenging in speech processing. Recently, speech separation based on deep learning has achieved promising performance [1, 2, 3, 4, 5, 6, 7, 8]. However, these methods typically report the average metric between estimates and reference signals. In the training process, the data are sampled uniformly. Therefore, unbiased models are trained with large variances, which leads to more failures, i.e., hard samples, in the generalization. Given that the average separation performance is sufficiently good, improving hard samples may contribute more to back-end tasks (e.g., speech recognition). The following studies improve hard samples from different aspects. Some studies [3, 9] demonstrate that frame-level permutation invariant training (PIT) outperforms utterance-level PIT in reducing separation failures. Tzinis et al. [10] propose a gradient reweighting scheme to bias the model towards bad predictions. Zeghidour et al. [11] generate training samples dynamically to significantly improve separation performance, including hard samples, but it is still sampled uniformly.

According to objective loss, training samples with low performance can be regarded as hard and the others as non-hard. Typically, the former is the minority and latter the majority. Assuming that hard and non-hard samples are two classes with different internal attributes, the model is actually biased towards non-hard samples, which weakens the generalization capability in the test set. On the basis of this observation, we assume that uniform sampling in the training set leads to data imbalance, and balancing between hard and non-hard samples may benefit generalization. Previous studies on data imbalance mainly focus on classification tasks, in which traditional solutions include re-sampling [12, 13, 14] and cost-sensitive weighting [15, 16, 17]. Some researchers propose re-weighting or novel loss function to compensate for the imbalanced class [18, 19, 20, 21, 22, 23]. Some corresponding learning paradigms for data imbalance have been investigated [24, 25, 26]. Data augmentation is another common approach used to address data imbalance [27, 28, 29, 30, 31]. Speech separation is fundamentally a regression task [32]. Though Yang et al. [33] proposed distribution smoothing for both labels and features for deep imbalanced regression, data imbalance in regression tasks has not been as well explored.

In this paper, we propose two methods from the perspective of hard sample mining to alleviate data imbalance, both of which are based on dynamic mixing [11]. For local hard sample mining, we search hard samples using objective loss in each batch. We then propose weighted loss to compensate for hard samples by increasing their weights during training. We also apply the weighted loss in the validation stage to select the model biased towards hard samples. However, local hard samples may lead to sub-optimal results. For global hard sample mining, when applying dynamic mixing, discriminating hard samples using objective loss during each training epoch leads to local hard samples. We propose an indirect method for global hard sample mining. Specifically, we first analyze the correlation between speaker-specific parameters and the separation results. The work [34] shows that the pitch median difference of two speakers in a mixture is correlated with separation results. Additionally, we investigate two more parameters, the energy ratio and x-vector cosine distance of two speakers in a mixture. We find that the x-vector cosine distance has a good correlation with the separation results. Then we add data preprocessing before training to search hard samples globally and indirectly, and then re-sample to increase the proportion of hard samples during training. The experimental results demonstrate that both methods effectively decrease the percentage of hard samples in the test set while keeping the average

This work was supported by the Opening Project of Key Laboratory of Xinjiang, China (2020D04047); The National Key R&D Program of China (2020AAA0107902); Xinjiang Uyghur Autonomous Region Graduate Research and Innovation Project, China (XJ2021G066); NSFC (61663044, 61761041).

metrics comparable, and the method based on global hard sample mining shows more promising results than the local method.

The contributions of this work are summarized as follows: 1, To the best of our knowledge, this is the first study in which speech separation is improved by alleviating data imbalance. 2, We propose a novel weighted loss based on local hard sample mining. 3, We discover that the x-vector cosine distance between two speakers in a mixture is correlated with the separation results. 4, We propose a novel indirect method for global hard sample mining and a new data augmentation method using hard re-sampling.

2. DYNAMIC MIXING BASED ON GLOBAL HARD SAMPLE MINING

To separate the speech of two speakers, let $\mathcal{D} = ((x_i^1, x_i^2), y_i)_i$ be the training set, where x_i^1 and x_i^2 are individual audios, y_i is the mixture, i is the sample index, and $x_i^1 + x_i^2 = y_i$. The probability of selecting each sample is typically equal: $P(i|\mathcal{D}) = \frac{1}{|\mathcal{D}|}$. We assume that the training set consists of hard and non-hard samples, and set a threshold to define hard samples as those whose evaluation metrics are lower than the threshold, and the others are non-hard samples. Generally, hard samples account for a small proportion of all samples. Let $\mathcal{D}_h$ and $\mathcal{D}_{nh}$ be the set of hard and non-hard samples, respectively. Then $\mathcal{D} = \mathcal{D}_h \cup \mathcal{D}_{nh}$, and $|\mathcal{D}_h| << |\mathcal{D}_{nh}|$. This imbalanced training set biases the model towards non-hard samples during training and weakens the generalization capability of hard samples in the test set. One possible solution is to increase the proportion of hard samples in the training set using data augmentation so that $|\mathcal{D}_h|$ and $|\mathcal{D}_{nh}|$ are comparable to balance model training.

The proposed method consists of three steps: dynamic mixing, hard sample mining and hard re-sampling, as shown in Fig.1. Specifically, we search hard samples globally in the training set generated using dynamic mixing. Instead of using objective loss, the search is performed using speaker-specific parameters indirectly, based on the correlation between these parameters and the separation results. Then we re-sample to increase the proportion of hard samples.

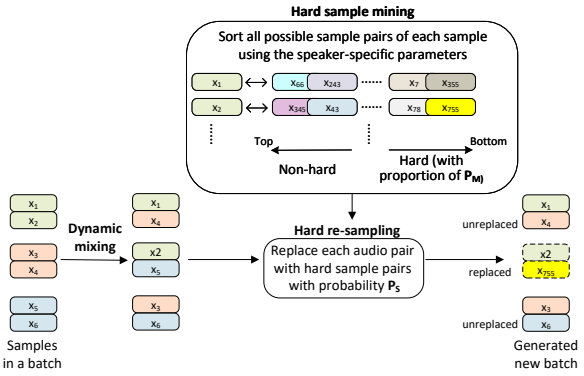

Fig. 1. Process of dynamic mixing with hard sample mining

2.1. Dynamic mixing

The general dataset in speech separation is static and does not change after creation. In this paper, we apply the method of dynamic mixing [11] to the training set for data augmentation. Specifically, the model receives a batch of $B \times 2$ audios, where B is the batch size and 2 is the number of speakers. A new batch is created by shuffling the audios randomly, where the gains are sampled for each audio and each pair of audios is summed to obtain B mixtures.

2.2. Correlation analysis

We investigate three speaker-specific parameters that are expected to be correlated with the separation results.

Difference of fundamental frequency medians: The fundamental frequency, i.e., pitch, denoted by f_0, is an intrinsic property of periodic signals. f_0 is a time-varying parameter for a fixed speaker. We use the median of the f_0 sequence of an utterance as the considered parameter [34]. The f_0 median difference is defined as the absolute value of the difference between the median of two f_0 sequences.

Energy ratio: We define the energy ratio as:

$$ER = 10 \left| \log_{10} \frac{E_1}{E_2} \right|, \tag{1}$$

where E_1 and E_2 are the energies of two utterances in a mixture.

Cosine distance of x-vectors: The x-vectors [35], which map variable-length utterances to fixed-dimensional embeddings, capture speaker characteristics. The cosine distance of x-vectors is typically used to discriminate between speakers.

Pearson's correlation coefficient is used to measure the correlation between speaker-specific parameters and separation results:

$$\rho = \frac{Cov\,(S, P)}{\sqrt{SD\,(S)}\sqrt{SD\,(P)}}, \tag{2}$$

where Cov and SD are the covariance and standard deviation, and S and P are sequences of separation metrics and speaker-specific parameters, respectively. Results of correlation analysis regarding the speaker-specific parameters will be presented in the later experiments.

2.3. Hard sample mining

Generating samples dynamically makes it challenging to search hard samples globally during training. We apply an indirect method using speaker-specific parameters for hard sample mining during preprocessing. Specifically, for each audio sample, we sort all its possible sample pairs generated by dynamic mixing using the speaker-specific parameters and select the bottom M pairs with low performance as hard samples. We traverse all samples to complete the corresponding hard sample mining. The value of M determines the constrained strength of hard samples during training. The smaller M is, the stronger the constraint, and the larger M is, the more relaxed the constraint. We can replace M with the percentage of M in all possible sample pairs, which is recorded as P_M, where P_M is the proportion of hard samples in the original training set of dynamic mixing.

2.4. Hard re-sampling

Based on hard sample mining results, we re-sample in the training set generated by dynamic mixing, to increase the proportion of hard samples during training, i.e., hard re-sampling. For an input batch that has $B \times 2$ audios, we replace each audio pair with hard samples with probability P_S. Specifically, we randomly select an audio in a pair as a pivot and then sample a new audio pair in all its corresponding hard sample pairs randomly. Typically, P_M is much smaller than P_S. Therefore, P_S approximately determines the proportion of hard samples in the new training set after hard re-sampling. P_S also needs to be set appropriately. A high P_S makes the model too biased towards hard samples and easy to overfit, whereas a low P_S makes the model insufficient for improving hard samples.

3. WEIGHTED LOSS BASED ON LOCAL HARD SAMPLE MINING

We propose an alternative method based on local hard sample mining. Existing speech separation methods typically use the average scale-invariant signal-to-noise ratio (SI-SNR) [6] in a batch as the objective training loss, which biases the trained model towards non-hard samples in the majority. We propose the following weighted SI-SNR(wSI-SNR):

$$\text{wSI-SNR} = \sum_{i=1}^{B} w_i l_i, \tag{3}$$

where B is the batch size, l_i is the SI-SNR of the ith audio pair, and w_i is the new weight. We use negative wSI-SNR as the training objective loss. We sort l_i in descending order in a batch and determine w_i using its index:

$$w_i = \frac{i}{\sum_{i=1}^{B} i}, \tag{4}$$

where i is the index after sorting. We aim to increase the weights of hard samples in a batch. It is worth noting that the weighted loss compensates for hard samples by increasing their weights in the objective loss, while dynamic mixing with hard sample mining compensates for hard samples by increasing their proportion in the training set. Therefore they achieve the similar goal from different aspects. To avoid repeated compensation, we do not apply the weighted loss in dynamic mixing with hard sample mining during training.

We also apply wSI-SNR during validation. Instead of considering a batch in training, we calculate all validation samples for wSI-SNR, by sorting all validation samples according to SI-SNR. The weighted validation loss helps to select models that are more biased towards hard samples.

4. SETUP

4.1. Setup of correlation analysis

We first perform correlation analysis to discriminate between the speaker-specific parameters that are expected to be correlated with the separation results. We use the 8 kHz sampling of the WSJ0-2mix [1] dataset. The mixtures are generated by mixing two random speakers in the Wall Street Journal dataset (WSJ0) training set s_tr_s with a random SNR of -5 to 5dB, to obtain about 30 hours of training and 10 hours of validation speech data. We mix any two speakers from the WSJ0 development set si_dt_05 and evaluation set si_et_05 in the same way to generate 5 hours of evaluation set, which contains 3000 recordings.

The utilized model of speech separation is Conv-TasNet [6], which contains three modules: the encoder, separation, and decoder. The separation module comprises R convolutional blocks consisting of X 1-D dilated convolutional layers with exponentially increasing dilation factors. In order to reduce the training time, we modify the original $R = 3$ and $X = 8$ to $R = 2$ and $X = 6$ to obtain a light model. We follow the method in [35] to train an x-vector extractor, in which we use the WSJ0 and WSJ1 datasets, which contain 206 speakers. After applying speed perturbation, we obtain 618 speakers with 508 hours of speech in total. We extract the x-vectors at the layer before the nonlinearity of the model, and the output dimension is 512. We employ SI-SNR improvement (SI-SNRi) as the separation evaluation metric.

4.2. Setup of speech separation

Experiments of speech separation involves using the weighted loss and dynamic mixing with hard sample mining. The dataset is consistent with that in section 4.1, and the sample gains during dynamic mixing are between -5 to 5 dB. We use light Conv-TasNet to adjust the hyperparameters and show the final results with the full model. SI-SNRi is used as the evaluation metric. In order to evaluate the performance of hard samples, we report hard sample rate (HSR) [9, 11]. We set a threshold in the test set, and those below the threshold are regarded as hard. It is worth noting that the hard sample threshold in the test set is independent of that in the training set. We set two thresholds: 5 dB and 10 dB, corresponding to HSR5 and HSR10. We add the constraint that audios in the generated sample pair do not come from the same speaker. Adam [36] is used as the optimizer, and the initial learning rate is set to 0.001. All the models are trained for 100 epochs on 4 second long segments. Gradient clipping with a maximum L2 norm of 5 is applied during training.

5. RESULTS

5.1. Results of correlation analysis

Fig. 2 shows the results of correlation analysis, including the correlation coefficient, SI-SNRi line, and density distribution of the speaker-specific parameters. All the samples in the test set are arranged in ascending order according to SI-SNRi. We can find that the SI-SNRi is non-linear, with a flat right part, while it drops rapidly in the left part. For the correlation between the pitch median difference and SI-SNRi, the correlation coefficient is 0.443, which is a promising value; then, the density distribution of pitch median difference is approximately consistent with SI-SNRi, that consists of a flat right part and a left part dropping rapidly. The energy ratio has no correlation with the SI-SNRi based on the observation of the correlation coefficient and density distribution. For the x-vector cosine distance, we observe a slightly higher correlation coefficient (0.447), and a more obvious consistent trend between the density distribution and SI-SNRi. We can conclude that the pitch median difference and x-vector cosine distance are well correlated with the separation SI-SNRi. The well-correlated parameters provide a possible approach to use the indirect method for global hard sample mining.

5.2. Results of speech separation

Results of local hard sample mining: We first examine separation results of the weighted loss based on local hard sample mining, as shown in Table1. The baseline is light Conv-TasNet, to which we add dynamic mixing (DM), weighted loss for the training set (WTL) and weighted loss for the validation set (WVL), respectively. We can find that compared with the baseline, adding DM effectively improve the SI-SNRi and HSRs. After applying the combination of WTL and WVL, the HSRs can be improved further based on DM. Moreover, the individual application of WTL and WVL also results

Table 1. Results of weighted loss

Model	DM	WTL	WVL	SI-SNRi(dB)	HSR5(%)	HSR10(%)
Conv-TasNet				14.41	8.00	15.00
	✓			**15.75**	3.30	7.20
	✓	✓		15.44	2.86	7.77
	✓		✓	15.72	3.15	7.20
	✓	✓	✓	15.72	**2.40**	**6.37**

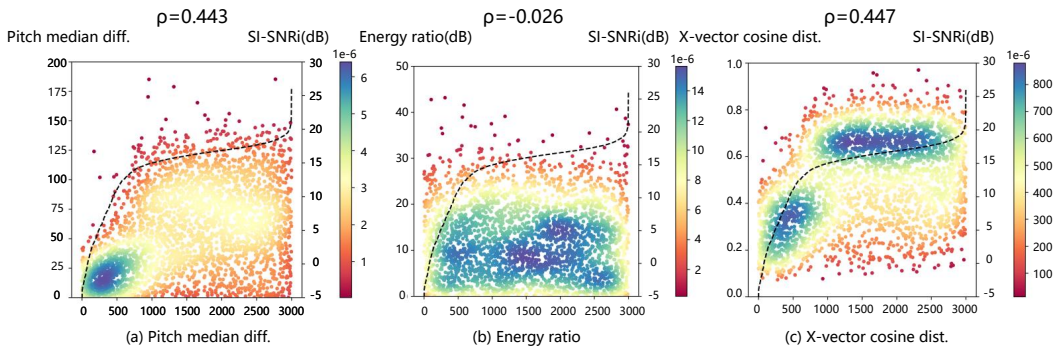

Fig. 2. Diagrams of SI-SNRi line and density distribution of speaker-specific parameters

Table 2. Results of hard sample mining with different P_M

Model	Para. for hard sample mining	P_M(%)	SI-SNRi(dB)	HSR5(%)	HSR10(%)
Conv-TasNet +DM +WVL	pitch median difference	1	15.71	2.63	6.50
		2	**15.81**	**2.40**	**6.00**
		3	15.78	2.53	6.30
		4	15.63	3.17	6.60
		5	15.50	3.10	7.13
	x-vector cosine distance	1	15.32	3.00	7.00
		2	15.46	2.77	6.93
		3	15.50	2.57	**6.17**
		4	**15.68**	**2.00**	6.50
		5	15.62	2.63	6.53

P_S is fixed to 30%

Table 3. Results of hard re-sampling with different P_S

Model	Para. for hard sample mining	P_S(%)	SI-SNRi(dB)	HSR5(%)	HSR10(%)
Conv-TasNet +DM +WVL	pitch median difference	20	15.63	2.90	6.97
		30	**15.81**	**2.40**	**6.00**
		40	15.49	2.77	7.13
	x-vector cosine distance	20	15.57	2.73	6.93
		30	**15.68**	**2.00**	6.50
		40	15.43	2.27	**6.33**

P_M=%2 for pitch median and %4 for x-vector

in some improvements to the HSRs. The results demonstrate the effectiveness of WTL and WVL for improving hard samples. And we apply WVL in the other experiments.

Results of global hard samples mining: We first fix the probability of hard re-sampling, i.e., P_S, as 30%. For different P_M in hard sample mining, we investigate five settings: 1%, 2%, 3%, 4%, and 5%. The results are shown in Table 2. We can find that when applying the pitch median difference as the speaker-specific parameter, the setting of P_M = 2% achieves the best results. For the x-vector cosine distance, the best SI-SNRi and HSR5 are achieved with the setting of P_M = 4%. We compare the results of the pitch median difference and x-vector cosine distance, and find that the former achieves better SI-SNRi and HSR10, whereas the latter wins on HSR5. Then we investigate different settings of P_S, including 20%, 30% and 40%, as shown in Table 3. We find that for both speaker-specific parameters, the setting of P_S=30% provides most of the best results.

Results of full model: Table 4 reports the results of full Conv-TasNet using the P_M and P_S with the best results in the light model as a summary. Similar to results of the light model, applying DM significantly outperforms the baseline Conv-TasNet on SI-SNRi and

Table 4. Results of full Conv-TasNet

Model	DM	WTL	WVL	Para. for hard sample mining	SI-SNRi(dB)	HSR5(%)	HSR10(%)
Conv-TasNet					15.73	5.70	10.63
	✓				**17.47**	2.17	3.67
	✓	✓	✓		17.24	1.13	2.83
	✓		✓	pitch median	17.25	1.13	2.70
	✓		✓	x-vector	17.18	**0.93**	**2.17**

HSRs. Both methods proposed show comparable SI-SNRi compared with DM, while making HSRs decreased obviously. Particularly, the method using the x-vectors achieves the best HSRs of 0.93% on HSR5 and 2.17% on HSR10. It can be seen that the x-vectors can more effectively represent the internal attributes of hard samples, that is, the more similar the characteristics of each audio, the more difficult it is to separate the mixture.

For global hard sample mining, we investigated three speaker-specific parameters and discovered that the x-vector cosine distance of two speakers in a mixture has the best correlation with the separation results. Meanwhile, the correlation between the speaker-specific parameters and separation results is not perfect, e.g. the correlation coefficient of the x-vector cosine distance and SI-SNRi is 0.447, which is still some distance from 1.0, and there are some outliers in the density distribution of the x-vectors whose trend is not perfectly correlated with SI-SNRi. These deficiencies may introduce errors in indirect hard sample mining. Despite this, the method with global hard sample mining still achieves the best hard sample rate in the test set, thereby demonstrating the superiority of global hard sample mining over its local counterpart for speech separation.

6. CONCLUSIONS

We explore an improved separation model for hard samples instead of training a model using average metrics. We assume that sampling uniformly in the training set leads to data imbalance. We propose two methods for improving separation on hard samples: local hard sample mining based, i.e., weighted loss, and global hard sample mining based, i.e., dynamic mixing with hard sample mining. The experimental results demonstrate that both methods outperform the baseline of using dynamic mixing only on hard sample rate while keeping the SI-SNRi comparable. The method of weighted loss is simple and easy to apply, whereas the method of dynamic mixing with hard sample mining shows more promising results.

7. REFERENCES

[1] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[3] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[5] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[8] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[9] Xiaoyu Liu and Jordi Pons, "On permutation invariant training for speech source separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6–10.

[10] Efthymios Tzinis, Dimitrios Bralios, and Paris Smaragdis, "Unified gradient reweighting for model biasing with applications to source separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 531–535.

[11] Neil Zeghidour and David Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] Haibo He and Edwardo A Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[14] Tomasz Maciejewski and Jerzy Stefanowski, "Local neighbourhood extension of smote for mining imbalanced data," in *2011 IEEE symposium on computational intelligence and data mining (CIDM)*. IEEE, 2011, pp. 104–111.

[15] Kai Ming Ting, "A comparative study of cost-sensitive boosting algorithms," in *In Proceedings of the 17th International Conference on Machine Learning*. Citeseer, 2000.

[16] Bianca Zadrozny, John Langford, and Naoki Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Third IEEE international conference on data mining*. IEEE, 2003, pp. 435–442.

[17] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser, "Svms modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, 2008.

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[19] Qi Dong, Shaogang Gong, and Xiatian Zhu, "Class rectification hard mining for imbalanced deep learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1851–1860.

[20] Qi Dong, Shaogang Gong, and Xiatian Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 6, pp. 1367–1381, 2018.

[21] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019.

[22] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2781–2794, 2019.

[23] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[24] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5704–5713.

[25] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," *arXiv preprint arXiv:1902.07379*, 2019.

[26] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," in *NeurIPS*, 2020.

[27] Jan Schlüter and Thomas Grill, "Exploring data augmentation for improved singing voice detection with neural networks.," in *ISMIR*, 2015, pp. 121–126.

[28] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell, "Understanding data augmentation for classification: when to warp?," in *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2016, pp. 1–6.

[29] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[30] Rafael L Aguiar, Yandre MG Costa, and Carlos N Silla, "Exploring data augmentation to improve music genre classification with convnets," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[31] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan, "Data augmentation using gans for speech emotion recognition.," in *Interspeech*, 2019, pp. 171–175.

[32] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[33] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi, "Delving into deep imbalanced regression," *arXiv preprint*, 2021.

[34] David Ditter and Timo Gerkmann, "Influence of speaker-specific parameters on speech separation systems.," in *INTERSPEECH*, 2019, pp. 4584–4588.

[35] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of ICASSP 2018*, 2018, pp. 5329–5333.

[36] Di. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR (Poster)*, 2015.