Domain-Aware Data Selection for Speech Classification via Meta-Reweighting

Junghun Kim, Ka Hyun Park, Hoyoung Yoon, U Kang

Seoul National University, South Korea

{bandalg97, kahyunpark, crazy8597, ukang}@snu.ac.kr

Abstract

Given speeches from diverse domains, how can we train an accurate classifier for a specific target domain utilizing the other source domains? The problem commonly arises in real-world scenarios, such as identifying the intents of speeches from individuals with a specific speech disorder using those of other disorders. However, existing data selection methods for utilizing the source instances encounter two main challenges: they cannot consider the diversities of source domains, and their hard selection schemes may ignore helpful source instances if the given information of the target domain is insufficient. In this work, we propose DOREME, a domain-aware data selection method for accurate speech classification on a target domain. The key idea is to softly select source instances by dynamically assigning importance scores to each instance based on two similarities: *instance-scores* and *domain-scores*. Various experiments show that DOREME achieves the best classification accuracy.

Index Terms: speech classification, meta-reweighting

1. Introduction

Given a limited amount of speech instances from a target domain, how can we leverage speech instances from multiple source domains to maximize the classification accuracy of the target speeches? Speech classification finds application across diverse areas such as emotion analysis [1, 2, 3, 4], speaker identification [5, 6, 7], language identification [8, 9], fake speech detection [10, 11], and environmental sound detection [12]. Many real-world speeches belong to various domains. For speeches consisting of multiple languages, each language can be regarded as a distinct domain [13, 14]. In speeches collected from subjects with diverse disorders, those with a specific impairment exhibit distinct features compared to the others [15]. In this context, our problem aims to train an accurate speech classification model for the targeted disorder domain by leveraging information from the other domains.

A naive approach for such multi-domain settings would be to train an individual classifier for the target domain. However, neglecting valuable information in the source domains degrades the classification performance primarily due to the limited number of instances in the target domain. Another possible approach would involve using speech examples from all domains simultaneously while disregarding their domains. In this case, however, the inherent similarities and disparities in properties across domains present a challenge. While source domains that resemble the target domain facilitate the training of the classifier, those with distinct properties impede the process.

Many data selection methods are studied for domain adaptation [16, 17] and noisy training [18]. Previous data selection approaches for utilizing the source speech instances to perform various tasks in a target domain include speech recognition [19, 20] and emotion recognition [21, 22, 23]. However, they cannot adaptively consider the diversities of source domains. Furthermore, their hard selection scheme may ignore beneficial instances if there is insufficient information given for the target domain. To train an accurate speech classifier specialized for a target domain, it is crucial to carefully select advantageous examples from the diverse source domains.

In this work, we propose DOREME (Domain-AwaRe Data Selection via Meta-Reweighting), a domain-aware data selection method for accurate speech classification. DOREME softly selects important source instances for training a classifier on the target domain. This is done by assigning importance scores to the source speeches based on the domain-aware meta-reweighting. The meta-reweighting [24] reduces the influence of noisy examples while enlarging that of clean ones with the goal of improving the robustness of models (see Section 2.2 for details). The main idea of DOREME is to treat the source speeches that have different properties from the target domain as noisy examples. DOREME effectively utilizes the source speeches by focusing on examples that belong to the source domains which have similar properties with the target domain.

In the following, we summarize our contribution. First, we propose a novel method that softly selects important source instances for training an accurate classifier on the target domain. DOREME enlarges the importance of source instances if they belong to domains encompassing similar properties with the target domain. This is done by assigning learnable domain-aware importance scores to the source instances. Second, we conduct various experiments and show that DOREME effectively utilizes the source instances for training the classifier on the target domain, achieving the best performance among the baselines. The code and datasets are available at https://github.com/snudatalab/DoReMe.

2. Related Works

2.1. Data Selection for Speeches

To train an accurate speech classifier on a target domain with limited instances, it is crucial to select suitable instances from multiple source domains. Many data selection methods for speeches are studied to improve many downstream tasks including speech recognition [19, 20] and emotion recognition [21, 22, 23]. Park et al. [25] select noise-free training instances based on the noisy student training strategy [26]. Lu et al. [27] select important data instances from the source corpus that are matched to the target domain. However, those approaches fail to adaptively select beneficial or important source instances for training a speech classifier specialized for the target domain, significantly degrading the classification performance.

2.2. Meta-reweighting

Meta-reweighting method [24] aims to enhance the robustness of deep learning models by dynamically adjusting the weights of noisy training examples. The main idea is to decrease the weights of noisy examples while increasing the weights of clean examples. This is done by comparing the gradient of each training instance with that of noise-free validation examples.

Suppose a deep learning model f_θ parameterized by θ, and an instance-wise loss function $l(x_i)$ for each training example x_i are given. Then the meta-reweighting method trains f_θ with the following weighted loss $\mathcal{L}$ with weights ϵ:

$$\mathcal{L} = \sum\nolimits_{i=1}^{N} \epsilon_i l(x_i) \tag{1}$$

where ϵ_i is a weight for an instance x_i and N is the number of training instances. The weight ϵ_i is computed as follows:

$$\epsilon_i \leftarrow \max\left(0, -\eta \frac{\partial}{\partial \epsilon_i} \frac{1}{N_t} \sum_{j=1}^{N_t} l(x_j; \theta(\epsilon))\right)\Bigg|_{\epsilon_i=0} \tag{2}$$

where N_t is the number of validation instances, $\theta(\epsilon) = \theta - \eta\nabla_\theta\mathcal{L}$, and η is the learning rate.

The intuition behind Equation (2) is to define the importance of x_i in training a model f_θ by assessing the magnitude of gradients of validation loss $\sum_{i=1}^{N_t} l(x_i; \theta(\epsilon))$ with respect to the score ϵ_i. Note that the magnitude of gradient in terms of a variable represents the importance of the variable for minimizing the objective function. Thus, if x_i takes a significant role in minimizing the validation loss, the magnitude of gradient with respect to ϵ_i becomes larger, and vice versa.

Meta-reweighting is applied in various fields including knowledge extraction [28], entity recognition [29], and language model [30]. They employ the method to enhance the robustness of model by considering explicit noise present in training instances. To the best of our knowledge, DOREME is the first approach to explicitly reweight the importance of various source domains through meta-reweighting to train an accurate speech classifier on the target domain.

3. Proposed Method

We propose DOREME, a domain-aware data selection approach for accurate speech classification within a target domain. The main challenges and our approaches are as follows:

1. **How can we utilize the speeches from multiple source domains for classifying the speeches of the target domain?** We assign an importance score to each source speech instance (*instance-score*) according to the similarities with the target domain. Then instances with their corresponding scores are utilized to train a classifier on the target domain (Section 3.1).
2. **How can we handle the instance imbalance across domains?** The instance-score may overfit the classifier to the domain with large number of instances. To address this, we propose domain-wise importance score (*domain-score*) which measures the similarity between domains, and normalize each instance-score according to them (Section 3.2).

The overall architecture of DOREME is depicted in Figure 1. We give importance scores to the source instances in two perspectives: *instance-score* which measures the importance of each instance and *domain-score* which quantifies the importance of each domain. We then train a classification model for the target domain with scored source instances.

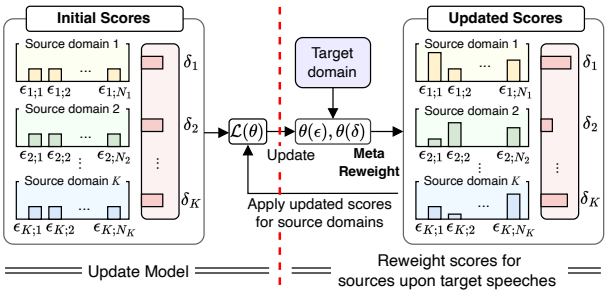

Figure 1: *Overall structure of* DOREME. *The method gives two kinds of importance scores to the source speeches: instance-score $\epsilon_{s;i}$ for each i-th instance in source domain s, and domain-score δ_s for each domain s.* DOREME *updates the model parameters θ using the scores, and reweights the importances based on the meta-reweighting scheme on target domain.*

3.1. Instance-scores for Source Domains

The objective of DOREME is to accurately classify the speeches of a target domain utilizing the speeches from multiple source domains. The challenge is that there are disparities between the source and target domains. Source speech instances which have different properties from those of the target domain degrade the performance of the classifier on the target domain. To address this, we propose to assign an importance score to each source speech instance, and leverage them for training a classification model on the target domain.

The main idea is to enlarge the influence of the instances which have similar properties with the target speeches by giving higher scores ϵ to them. We merge all the speeches from the source and target domains, and train a classifier using the following weighted cross-entropy loss function:

$$\mathcal{L}(\theta) = \sum_{s\in\mathcal{S}} \sum_{i=1}^{N_s} \epsilon_{s;i} l(x_{s;i}; \theta) + \sum_{i=1}^{N_t} l(x_{t;i}; \theta) \tag{3}$$

where $\mathcal{S}$ is a set of source domains, t is the target domain, θ is the parameter of the classifier, N_d is the number of instances in domain d, $x_{s;i}$ is the i-th speech instance in domain s, and $\epsilon_{s;i}$ is the instance-score for each speech instance $x_{s;i}$. For the loss $l(x_{d;i}; \theta)$ of instance $x_{d;i}$, any classification loss can be used. We use cross-entropy loss for the loss function l due to its robust and accurate performance [31]. Note that the instance-score $\epsilon_{s;i}$ adjusts the influence of $x_{s;i}$ in the loss function $\mathcal{L}(\theta)$. For simplicity, we denote the first and second terms of $\mathcal{L}(\theta)$ as $\mathcal{L}_\mathcal{S}(\theta)$ and $\mathcal{L}_t(\theta)$, respectively in the rest of this paper.

The instance-scores in Equation (3) are computed based on the meta-reweighting scheme (see Section 2.2). To enhance the robustness and accuracy of deep learning models, the meta-reweighting assigns a learnable weight to each training example and reduces the weights of noisy examples. The intuition is to give a small weight to an example if the gradient of it does not align with that of validation instances.

We apply this meta-reweighting concept within the context of multi-domain scenarios. The main idea is to treat source instances with distinct properties from those of the target domain as noisy examples. Following Equation (2), the importance score $\epsilon_{s;i}$ for each $x_{s;i}$ in Equation (3) is computed as

$$\epsilon_{s;i} \leftarrow \max\left(0, -\eta \frac{\partial}{\partial \epsilon_{s;i}} \frac{1}{N_t} \sum_{j=1}^{N_t} l(x_{t;j}; \theta(\epsilon))\right)\Bigg|_{\epsilon_{s;i}=0} \tag{4}$$

where $\theta(\epsilon) = \theta - \alpha\nabla_\theta\mathcal{L}_\mathcal{S}(\theta)$, $x_{t;j}$ is the j-th instance in the target domain t, and η and α are the learning rates. The intuition is to give a larger weight $\epsilon_{s;i}$ to $x_{s;i}$ if the instance significantly affects the loss of the target domain. This is done by measuring the magnitude of gradient in terms of $\epsilon_{s;i}$. However, there are challenges for directly using the instance-scores in Equation (4) in our problem. In the following, we list up two sub-problems and their solutions for implementing the instance-scores.

Subproblem 1. In the objective function $\mathcal{L}(\theta)$ of Equation (3), we incorporate training instances from the target domain (target-train instances) alongside those from diverse source domains. Since the losses of target training instances are uniformly weighted with 1, the instance-scores ϵ should not exceed 1. This is based on the assumption that every source instance is less important than the target instances. Thus, we rescale the instance-scores in the range of $(0, 1)$ by dividing each $\epsilon_{s;i}$ with $\max(\epsilon_s)$ where $\max(\epsilon_s)$ is the maximum $\epsilon_{s;i}$ in a source domain s.

Subproblem 2. In our scenario, the number N_t of target-train instances is significantly small. Under such circumstances, the target-train instances do not adequately represent general properties of the target domain due to the scarcity of data. The challenge arises as the instance-scores of source domains are computed based on the insufficient target instances; valuable source instances, which have general properties of the target domain, may be neglected with 0 scores since they are not reflected in the target-train instances. To tackle this, we rescale the instance-scores in the range of $(0.5, 1)$:

$$\epsilon_{s;i} \leftarrow 0.5\left(1 + \epsilon_{s;i}/\max(\epsilon_s)\right) \tag{5}$$

This mitigates the risk of disregarding the valuable source instances during training. Furthermore, this effectively prevents the oscillation of instance-scores, leading to a more stable learning of the classifier on the target domain.

3.2. Domain-scores for Source Domains

The loss function $\mathcal{L}(\theta)$ in Equation (3) assigns an importance score $\epsilon_{s;i}$ to each instance $x_{s;i}$. However, this leads to overfitting of the classification model to a domain with a large number of instances. For example, let a source domain s has different properties from those of the target domain, but has an extremely large number of instances. Although the instance-scores $\epsilon_{s;i}$ are small, the overall effects of domain s in the loss function $\mathcal{L}_\theta$ would be large due to the excessively large N_s. Thus, it is crucial to reweight the effect of each domain in the loss function.

To address this, we propose another importance score δ_s for each source domain $s \in \mathcal{S}$. The *domain-score* δ_s measures the domain-wise similarity between the source domain s and the target domain. The key idea is to use the domain-score δ_s as a weight for the loss of domain s while normalizing the instance-scores $\epsilon_{s;i}$. This effectively reweights the influence of each domain on the loss function, ensuring that data from important domains have a greater impact on model training. With the domain-scores, the loss function $\mathcal{L}(\theta)$ in Equation (3) is rewritten as follows:

$$\mathcal{L}'(\theta) = \sum_{s\in\mathcal{S}}\sum_{i=1}^{N_s}\delta_s\frac{\epsilon_{s;i}}{\sum_{i=1}^{N_s}\epsilon_{s;i}}l(x_{s;i};\theta) + \sum_{i=1}^{N_t}l(x_{t;i};\theta) \tag{6}$$

where $\epsilon_{s;i}$ is computed as in Section 3.1. We denote the first term of $\mathcal{L}'(\theta)$ as $\mathcal{L}'_\mathcal{S}(\theta)$ for brevity.

Table 1: *Summary of datasets.*

Dataset	Hours	# of speakers	# of domains	# of classes	# of speeches per domain		
					Min	Max	Mean
Skit-S2I[1]	13.80	14	5	14	1,239	3,834	2669.0
ITALIC[2]	15.46	70	13	18	108	6,645	1180.1

[1] https://github.com/skit-ai/speech-to-intent-dataset
[2] https://github.com/RiTA-nlp/ITALIC/

Similar to the instance-scores ϵ, we compute the domain-score δ_s for each source domain s as follows:

$$\delta_s \leftarrow \max\left(0, -\eta\frac{\partial}{\partial\delta_s}\frac{1}{N_t}\sum_{j=1}^{N_t}l(x_{t;j};\theta(\delta))\right)\Bigg|_{\delta_s=0} \tag{7}$$

where $\theta(\delta) = \theta - \beta\nabla_\theta\mathcal{L}'_\mathcal{S}(\theta)$, and β is a learning rate.

4. Experiments

Through experiments, we answer the following questions:

Q1. Speech classification performance (Section 4.2). How accurate is DOREME in classifying speeches of the target domain while utilizing those from source domains?

Q2. Ablation study (Section 4.3). Does each module of DOREME contribute to the classification performance?

Q3. Domain-score analysis (Section 4.4). How do the domain-scores of the source domains change as the training proceeds?

4.1. Experimental Settings

Datasets. We use Skit-S2I [32] and ITALIC [33] to evaluate the performance of DOREME. We summarize the statistics of datasets in Table 1. Skit-S2I is a collection of speeches from Indian-English speakers using banking services via telephony. Speeches are categorized into 14 topics as classes such as card issuance and changing limits. Speakers come from five native languages: Bengali, Hindi, Kannada, Malayalam, and Punjabi. We treat the native languages of speakers as domains because linguistic traits vary depending on their native languages. ITALIC is a speech dataset recorded by speakers from 13 Italian regions. We treat regions as domains given that the linguistic features of individuals are significantly influenced by their respective native regions. There are 18 classes which are the topics of speeches such as recommendation and cooking recipe.

Baselines. We compare DOREME with previous data selection approaches. *wav2vec-T* [34] is a simple method that generates speech embeddings using a pretrained wav2vec2.0 model followed by a single-layered linear classifier. wav2vec-T minimizes the cross-entropy loss of the target training instances to optimize the classification model. *wav2vec-S*, a variant of wav2vec-T, utilizes training instances from both the target and source domains. *Meta-reweighting* [24] uses the same model structure as wav2vec-S while giving importance weights to each source instance based on the meta-reweighting scheme. *Contrastive-selection* [27] selects acoustically similar source speeches to the target domain. For Contrastive-selection, we use a pretrained wav2vec2.0 embedding model followed by two-layered linear classifier.

Evaluation and Settings. We use the F1-score as the main evaluation metric along with the classification accuracy to provide more comprehensive assessment. We generate embeddings of speech instances with wav2vec2.0 [34], where the initial parameters are set using the pretrained wav2vec2-large-960h downloaded from the transformer library [35]. In the Skit-S2I dataset, we set the domain with the least number of instances

Table 2: *Speech classification performance in terms of F1-score and accuracy. The best performance is in bold. Note that* DoReMe *shows the best results in all cases.*

Model	Skit-S2I		ITALIC	
	F1-score	Accuracy	F1-score	Accuracy
wav2vec-T	43.42 ± 2.59	45.18 ± 2.32	18.76 ± 0.78	24.66 ± 0.64
wav2vec-S	63.54 ± 1.18	63.90 ± 1.17	29.29 ± 0.46	33.73 ± 0.44
Meta-reweighting	63.76 ± 1.14	64.08 ± 1.11	30.41 ± 0.51	34.92 ± 0.45
Contrastive-selection	62.14 ± 1.42	62.79 ± 1.42	28.29 ± 0.75	33.41 ± 0.51
DoReMe (proposed)	**65.55 ± 1.72**	**65.92 ± 1.68**	**31.56 ± 0.55**	**36.21 ± 0.48**

Table 3: *Ablation Study. The best F1-score and accuracy are in bold. Note that each module of* DoReMe *effectively improves the classification performance.*

Model	Skit-S2I		ITALIC	
	F1-score	Accuracy	F1-score	Accuracy
DoReMe-w.o.-ϵ	62.28 ± 2.53	62.87 ± 2.41	31.41 ± 0.47	36.01 ± 0.48
DoReMe-w.o.-δ	63.76 ± 1.14	64.08 ± 1.11	30.41 ± 0.51	34.92 ± 0.45
DoReMe-w.o.-rescale	56.40 ± 2.60	56.39 ± 2.50	25.47 ± 0.39	27.82 ± 0.46
DoReMe (proposed)	**65.55 ± 1.72**	**65.92 ± 1.68**	**31.56 ± 0.55**	**36.21 ± 0.48**

as the target domain. In ITALIC, we set the domain with the largest number of instances as the target one due to the excessively smaller numbers of instances in the other domains. The smallest domain consists of 108 instances, which is inadequate for accurately evaluating the model. We split the datasets into target-training and target-test sets with ratio 0.1:0.9. We train each model using the Adam optimizer [36] for 1,000 epochs. We report the mean and the standard deviation of the results after conducting the experiment 100 times with random seeds.

4.2. Speech Classification Performance (Q1)

We present the speech classification performance of DoReMe and the baselines in Table 2. Note that DoReMe shows the best accuracy in every case. This highlights the importance of adaptively utilizing source instances based on the domain-aware meta-reweighting scheme. In Skit-S2I, the improvement of DoReMe compared to Meta-reweighting is marginal due to the limited number of source domains (only 4), making the domain-scores of DoReMe less effective. However, DoReMe significantly outperforms all baselines including Meta-reweighting in ITALIC, which has a sufficient number of source domains. This indicates that simply giving importance scores to the source instances without considering the diversity of source domains limits the classification performance.

4.3. Ablation Study (Q2)

We provide an ablation study in Table 3 to demonstrate the impact of each module in DoReMe. DoReMe-w.o.-ϵ and DoReMe-w.o.-δ are DoReMe without the instance-score ϵ and the domain-score δ, respectively. DoReMe-w.o.-rescale is DoReMe without the rescaling of instance-scores from $(0, 1)$ to $(0.5, 1)$. Note that in Skit-S2I, instance-scores exhibit a significant influence relative to domain-scores, whereas in ITALIC, domain-scores have a greater impact. This indicates that the two scores effectively complement each other, optimizing their impacts to perform better on the target domain. Furthermore, DoReMe presents higher accuracy than DoReMe-w.o.-rescale. This shows the significance of avoiding the risk of overlooking valuable source instances with 0 scores.

4.4. Domain-score Analysis (Q3)

We showed that the domain score effectively improves the classification performance through the experiments in Sections 4.2 and 4.3. Our further question is, how closely is the domain score δ_s related to the similarity of a source domain s with a target domain? To answer this, we analyze the domain-scores in our synthetic data which are generated by applying random noises of various levels to the target domain of the Skit-S2I dataset. We treat the noisy domains as source domains, and the original one as the target domain. For each i-th speech instance $x_{s;i}$ of a source domain s, we add Gaussian noises sampled from $\mathcal{N}(0, (\sigma_{x_{s;i}} r_s)^2)$ where $\sigma_{x_{s;i}}$ is the standard deviation of $x_{s;i}$ and r_s is the noise level of domain s. We vary r_s in $\{0.1, 0.2, ..., 1.0\}$ to generate ten source domains. We present the trends of domain-scores as the epoch proceeds in Figure 2. We also show the average domain-scores for each source domain after training in Figure 3. In these figures, we normalize the domain-scores by dividing them with the sum of all scores.

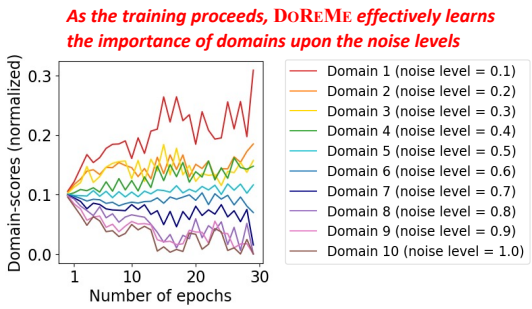

Figure 2: *Change of domain-scores over the epochs. As the training progresses, the scores for source domains with small level of noises gradually increase.*

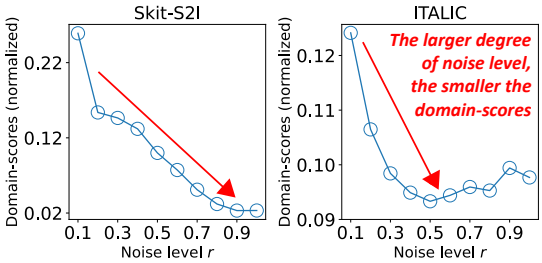

Figure 3: *Average domain-scores over each source domain after training.* DoReMe *assigns high domain-scores to important source domains with small noise, while assigning lower scores to those with higher levels of noise throughout training.*

Figure 2 shows that the domain-scores are uniformly distributed in earlier epochs. As the training proceeds, domain-scores inversely correlate with noise levels, assigning larger weights to domains with less noise. In Figure 3, we observe that the domain-scores decrease as we increase the noise level. This indicates that DoReMe successfully identifies source domains similar to the target domain through the domain-scores.

5. Conclusion

We propose DoReMe, a domain-aware data selection method for training an accurate classifier on a target domain by utilizing data from various source domains. DoReMe softly selects important source instances according to the similarity with the target instances. Furthermore, DoReMe measures the importance of source domains and balances the influence of them for training the classifier on the target domain. Extensive experiments on real-world datasets show that DoReMe achieves the best classification performance on the target domain.

6. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No.2022-0-00641, XVoice: Multi-Modal Voice Meta Learning], [No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and [No.RS- 2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)]. U Kang is the corresponding author.

7. References

[1] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. W. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2019.

[2] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Interspeech*, B. Yegnanarayana, Ed., 2018.

[3] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control.*, 2020.

[4] S. Zhang, A. Chen, W. Guo, Y. Cui, X. Zhao, and L. Liu, "Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition," *IEEE Access*, 2020.

[5] V. Tran and W. Tsai, "Speaker identification in multi-talker overlapping speech using neural networks," *IEEE Access*, 2020.

[6] Z. A. Abbood, B. T. Yasen, M. R. Ahmed, A. D. Duru *et al.*, "Speaker identification model based on deep neural networks," *Iraqi Journal For Computer Science and Mathematics*, pp. 108–114, 2022.

[7] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences*, 2021.

[8] D. Deshwal, P. Sangwan, and D. Kumar, "A language identification system using hybrid features and back-propagation neural network," *Applied Acoustics*, 2020.

[9] H. Mukherjee, A. Dhar, S. Md. Obaidullah, K. Santosh, S. Phadikar, and K. Roy, "A recurrent neural network-based approach to automatic language identification from speech," in *Proceedings of the 2nd International Conference on Communication, Devices and Computing: ICCDC*, 2019.

[10] D. M. B. L., Y. Rodríguez-Ortega, D. Renza, and G. R. Arce, "Deep4snet: deep learning for fake speech classification," *Expert Syst. Appl.*, 2021.

[11] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[12] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, 2021.

[13] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP*, 2022.

[14] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," *arXiv preprint arXiv:1509.02409*, 2015.

[15] C. Wachinger, M. Reuter, A. D. N. Initiative *et al.*, "Domain adaptation for alzheimer's disease diagnostics," *Neuroimage*, 2016.

[16] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *EMNLP*. ACL, 2011, pp. 355–362.

[17] M. Liu, Y. Song, H. Zou, and T. Zhang, "Reinforced training data selection for domain adaptation," in *ACL (1)*, 2019.

[18] Y. Wei, X. Mei, X. Liu, and P. Xu, "DST: data selection and joint training for learning with noisy labels," *CoRR*, 2021.

[19] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve perfomance." in *LREC*, 1998.

[20] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *ICASSP*, 2004.

[21] J.-B. Kim and J.-S. Park, "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition," *Engineering applications of artificial intelligence*, 2016.

[22] D. Le and E. M. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *ACII*, 2015.

[23] C. E. Erdem, E. Bozkurt, E. Erzin, and A. T. Erdem, "Ransac-based training data selection for emotion recognition from spontaneous speech," in *AFFINE*, 2010.

[24] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *ICML*, 2018.

[25] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *Interspeech*, 2020.

[26] Q. Xie, M. Luong, E. H. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR*, 2020.

[27] Z. Lu, Y. Wang, Y. Zhang, W. Han, Z. Chen, and P. Haghani, "Unsupervised data selection via discrete speech representation for ASR," in *Interspeech*, 2022.

[28] Z. Li, J. Nie, Y. Song, P. Du, and D. Li, "Learning to classify relations between entities from noisy data - A meta instance reweighting approach," *Expert Syst. Appl.*, 2022.

[29] L. Wu, P. Xie, J. Zhou, M. Zhang, C. Ma, G. Xu, and M. Zhang, "Robust self-augmentation for named entity recognition with meta reweighting," 2022.

[30] S. Chi, B. Dong, Y. Xu, Z. Shi, and Z. Du, "APAM: adaptive pre-training and adaptive meta learning in language model for noisy labels and long-tailed learning," *CoRR*, 2023.

[31] J. Yoo, J. Kim, H. Yoon, G. Kim, C. Jang, and U. Kang, "Accurate graph-based PU learning without class prior," in *ICDM*, 2021.

[32] S. Rajaa, S. Dalmia, and K. Nethil, "Skit-s2i: An indian accented speech to intent dataset," *CoRR*, 2022.

[33] A. Koudounas, M. L. Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, and E. Baralis, "ITALIC: an italian intent classification dataset," *CoRR*, 2023.

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR, year = 2015,*.