Unsupervised Out-of-Distribution Dialect Detection with Mahalanobis Distance

Sourya Dipta Das, Yash Vadi, Abhishek Unnam, Kuldeep Yadav

SHL Labs, India

Sourya.Das@shl.com, yash.vadi@yahoo.com, Abhishek.Unnam@shl.com, Kuldeep.Yadav@shl.com

Abstract

Dialect classification is used in a variety of applications, such as machine translation and speech recognition, to improve the overall performance of the system. In a real-world scenario, a deployed dialect classification model can encounter anomalous inputs that differ from the training data distribution, also called out-of-distribution (OOD) samples. Those OOD samples can lead to unexpected outputs, as dialects of those samples are unseen during model training. Out-of-distribution detection is a new research area that has received little attention in the context of dialect classification. Towards this, we proposed a simple yet effective unsupervised Mahalanobis distance feature-based method to detect out-of-distribution samples. We utilize the latent embeddings from all intermediate layers of a wav2vec 2.0 transformer-based dialect classifier model for multi-task learning. Our proposed approach outperforms other state-of-the-art OOD detection methods significantly.

Index Terms: Out of Distribution Detection, Open Set Classification, Outlier Detection , Dialect Identification, Wav2vec 2.0, Automatic Speech Recognition

1. Introduction

Dialect identification [1] has received a lot of interest in the speech processing community in recent decades. Dialect identification plays an important role in speech processing systems such as automated speech recognition (ASR) [2], multilingual translation systems, targeted advertising, and biometric authentication since it helps to target certain dialects. In recent years, numerous approaches [3, 4] have been proposed with great success, for dialect identification. However, almost all of these latest state-of-the-art approaches only address closed-set dialect identification, where the set of dialects to predict is fixed. For every input speech data, the predicted dialect within that set is returned. But in a real-life scenario, deployed applications rarely receive regulated inputs from known dialects and are vulnerable to an ever-changing set of unlabeled user inputs with unknown dialects. To solve this issue, a system should have the option to 'reject' that prediction and identify when the input speech does not match any known dialects well. It can also be used to identify and learn new dialects for the system. This task is defined as out-of-distribution(OOD) detection for dialect identification and is essential to the design of trustworthy AI applications in real-world use cases [5]. The out-of-distribution (OOD) detection problem, in general, has received a great deal of attention in the literature, with cutting-edge algorithms [6, 7, 8] being supervised in the sense that they require fine-tuning on OOD data to accomplish high performance in OOD detection. Nevertheless, supervised OOD detection algorithms have the problem of requiring expensive training on

OOD data, curating the OOD dataset with diverse samples to make it more distinguishable from the in-distribution data, and additional model hyperparameter tuning.

In this paper, we present a joint framework for both dialect classification that automatically classifies known dialects from input speech and out-of-distribution detection which also detects input audio that does not belong to any of the dialects used to train the model. We used a pre-trained wav2vec 2.0 model and fine-tuned it on the known dialects to adapt the feature embedding. This model is used for dialect identification tasks for input data with known dialects. Further, we retrieved features from several intermediate transformer layers to capture rich micro and macro phonetic feature information that may outperform the last layer of the fine-tuned transformer model. Later, we used those latent representations to estimate the mean, and covariance matrix for each layer by using close-set training data only. During inference, we used previously calculated layer-wise mean and covariance matrices to compute the Mahalanobis distance score for each layer, which was then used as a feature vector. This allows us to use classical outlier rejection methods like KNN [9] to enumerate outlier scores for the OOD detection task. We have evaluated our proposed method by comparing it with state-of-the-art out-of-distribution detection methods [10, 11, 12, 6, 13, 14]. Our contributions can be summarized as follows:

- We present the first joint dialect identification framework with unsupervised OOD dialect detection, which is a plug-and-play technique to identify known dialects or reject input speech samples of unknown dialects in a single forward pass.

- We propose Mahalanobis's distance-based feature to be used by a KNN-based [9] outlier classification model for the OOD dialect identification task. This can be used without modifying the backbone architecture, unlike previous approaches [12, 13] that required us to modify the model architecture.

- We evaluated the performance of our solution on two different language datasets i.e. English and Spanish to see how well it rejects unknown classes while maintaining its performance on close-set dialect classification.

The proposed approach outperforms the state-of-the-art method and achieves an overall AUROC of 96% for the *English Dialect Dataset* and 80% AUROC for the *Spanish Dialect Dataset*.

2. Related Work

There has been existing research work on speech dialect identification, but no prior work on out-of-distribution dialect detection. Torres *et al.* [1] had done earlier work on dialect identification where they used the Gaussian mixture model (GMM) with

shifted delta cepstral features (SDC). Zhang *et al.* [3] proposed an unsupervised bottleneck feature extraction approach for phonetic label estimation. They also used an alternate autoencoder and an adversarial autoencoder in the alternative phase of the speech feature extraction process. Kong *et al.* [4] proposes a new model architecture that consists of dynamic kernel convolution, local multi-scale learning, and global multiscale pooling layers. These proposed custom layers are used to capture features in both short-term and long-term contexts, multiscale granular features from wide receptive fields, and aggregated features from different bottleneck layers, respectively. Hamalainen *et al.* [15] worked on a Finnish dialect identification system that used speech recording and transcription data. Similarly, Imaizumi *et al.* [16], Ma *et al.* [17] and Lin *et al.* [18] also worked on Japanese, Chinese and Arabic dialects identification techniques, respectively.

Recently, there have been several studies that have explored the out-of-distribution problem in various domains like vision, text, etc. Liang *et al.* [10] have shown that thresholding onto the softmax output of the predicted class provides a good proxy score for detecting Out of Distribution (OOD) data. Shu *et al.* [11] suggested another approach called DOC (Deep Open Classification). In contrast to conventional classifiers, DOC constructs a multi-class classifier with a 1-vs-rest final layer of sigmoids instead of a softmax to minimize the risk associated with open spaces. By reducing the decision bounds of sigmoid functions with Gaussian fitting, it significantly lowers the open space risk for rejection. Bendale *et al.* [12] then presented a new neural network layer, OpenMax, which estimates the likelihood that an input belongs to an unknown class. They estimate the unknown class rejection probability value by adapting the extreme-value Meta-Recognition-inspired distance normalization process to the activation patterns in the penultimate layer of the network. Lee *et al.* [6] proposed building a Gaussian model from features extracted from the hidden layer and calculating the distance from this multivariate distribution (Mahalanobis Distance) and used this distance for OOD detection. Ren *et al.* [14] modified the Mahalanobis distance by subtracting the distance calculated from the entire training distribution to make it suitable for detecting near OOD samples. Liu *et al.* [13] has developed a robust uncertainty-based methodology that delivers an uncertainty score to each prediction and may be used to discover outliers. Here, we compared recent state-of-the-art machine learning and deep learning-based OOD detection techniques, as stated above, in an audio dialect setting.

3. Problem Statement

The problem is formulated as a variant of conventional multiclass classification which is also referred to as a close-set classification problem. Given dialect classification training data $D_{train} = \{(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\}$ where N is the total number of samples in training data, x_i is the input audio sample, $y_i \in L_{knw} = \{1, \ldots, M\}$ is corresponding target dialect label with M number of dialect classes, contains samples from a fixed set of known dialect classes. During inference, the test set contains samples from both the set of known dialect classes during training and additional unknown dialect classes, i.e., $D_{test} = \{(x'_1, y'_1), (x'_2, y'_2) \ldots (x'_n, y'_n)\}$ where $y'_i \in (L_{knw} \cup L_{unk})$ and L_{unk} includes classes that are not observed during training.

In this paper, we use a more realistic scenario in which we have no prior knowledge of what out-of-distribution inputs look like. It is not possible to train a separate supervised classifier directly in this scenario. So, our task is to train a classifier $F_D(x)$ with D_{train} training data, that correctly predicts the dialect class from a set of known dialect classes, i.e. $F_D(x) = [d_1, d_2, \ldots, d_M]$ where d_m is prediction score of m-th known dialect class and accurately detects audio samples with unknown dialect class by classifying those audio samples as a rejected class, which is denoted as the M + 1 class.

$$\hat{y}_p = \begin{cases} \text{argmax}_m F_D(x) & \text{if } G_D(x) \leq \delta \\ M + 1 & \text{if } G_D(x) > \delta \end{cases} \quad (1)$$

Here, $\hat{y}_p$ is the predicted class and $G_D(x)$ is a class rejection score function that determines if the input corresponds to the unknown dialect class or rejected class and δ is a threshold value. Here, the OOD dialect detection problem is as simple as constructing a class rejection score function, $G_D(x)$ that assigns lower scores to inputs with known dialect class than to out-of-distribution inputs.

4. Methodology

4.1. Model Architecture and Fine-Tuning

Wav2Vec 2.0 speech model [19] is pre-trained on unlabeled speech data using self-supervised learning for learning high-quality representations of speech. It shows promising results when transferred to other tasks [20, 21] like speech classification, speech recognition, speech frame classification, etc. Therefore, we have used a pre-trained wav2vec 2.0 model and have fine-tuned it on the D_{train} dataset for closed set known dialect classification on M classes for learning the feature embeddings on the dataset. After fine-tuning, we obtain a wav2vec 2.0 architecture-based dialect identification model, F_D with K transformer layers.

4.2. Class Rejection Score Estimation Method

We denote $F_D^k(x) \in \mathbb{R}_d$ as the d-dimensional feature embeddings corresponding to the k-th transformer layer for input x where $k \in [1, 2, \ldots, K]$. we further passed those intermediate feature embeddings through a hyperbolic tangent function, $tanh(.)$ to transform the features into the same restricted semantic space, i.e., $h_k^t(x) = tanh(F_D^k(x))$. We used this function to limit the value of each embedding vector element to between +1 and -1, similar to feature value scaling. From a previous study [22], different transformer layers of the wav2vec 2.0 model capture distinct semantic properties from the input speech. Thus, we use latent representations from all transformer layers of fine-tuned wav2vec 2.0 dialect classifier model by concatenating feature embeddings from all transformer layers, i.e., $\phi_h(x) = [h_1^t, h_2^t, \ldots, h_K^t]^T \in \mathbb{R}_{d.K}$. From previous work [6], we use Mahalanobis distance to calculate the distance between test audio samples and training data distribution, D_{train} for detecting unknown classes. Here, we defined the Mahalanobis distance score by using a simple and computationally efficient approximation method in a prior work [23]. We achieved that by decomposing the feature space into several subspaces and solving a low-dimensional constrained convex optimization. We illustrate this estimation process in Figure 1. Thus, we define Mahalanobis distance score, $V_{MD}^k(x_i)$ in the following equa-

tion.

$$V_{MD}^k(x_i) = \left(h_k^t(x_i)) - \mu_k\right)^T \Sigma_k^{-1} \left(h_k^t(x_i)) - \mu_k\right)$$

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}\left[h_k^t(x_i))\right]$$

$$\Sigma_k = \frac{1}{(n-1)w_k}\sum_{i=1}^{n}\left(h_k^t(x_i) - \mu_k\right)\left(h_k^t(x_i) - \mu_k\right)^T$$

where μ_k, Σ_k are mean and covariance for k-th transformer layer from the feature embeddings of training data, D_{train} respectively, w_k is a layer-dependent constant from that optimization process for k-th transformer layer and the square root of $V_{MD}^k(x_i)$ is the Mahalanobis distance of the transformer layer embedding of data x_i from the k-th layer. We enumerate the value of w_k during that optimization process to extract relevant hidden state features from transformer layer embeddings. we further define Mahalanobis distance feature vector, $V_{MD}(x) = [V_{MD}^1(x) \oplus V_{MD}^2(x) \oplus \cdots \oplus V_{MD}^K(x)]$ by concatenating Mahalanobis distance scores, $V_{MD}^k(x)$ for all transformer layers. Then, we train a KNN [9] based outlier detection model with Mahalanobis distance feature vectors extracted from training samples to estimate Class Rejection Score, $G_D(x)$ for detection of unknown class with a threshold value, δ. We illustrate the Inference pipeline of the proposed method in Figure 2.
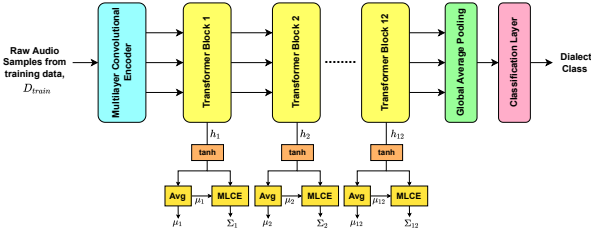

Figure 1: *Illustration of Layer Feature Embedding Mean(μ_k), Covariance Matrix (Σ_k) Estimation for k-th transformer layer from the feature embeddings,$F_D^k(x)$ of training data, D_{train}. Here, Avg is component-wise vector average operation and MLCE is Maximum Likelihood Covariance Estimator.*

5. Experimentation and Results

We implemented our dialect classifier model using PyTorch on top of the Wav2Vec 2.0 model from Hugging Face transformer library [1]. We have used two different pre-trained wav2vec 2.0 model [2][3] for English and Spanish dialect datasets. Furthermore, we have used all 12 transformer layers of our Wav2Vec 2.0 base model for defining the Mahalanobis distance score in all of our experiments. For class rejection score estimation, we have used PyOD [4] package to implement KNN [9] with 0.01 outlier fraction value and scikit-learn [5] for implementing Maximum likelihood covariance estimator(MLCE). We train each dialect classification model with 6 epochs using 1 NVIDIA GTX 1080 GPU (12 GB) with 16 GB RAM. For both training and validation, all
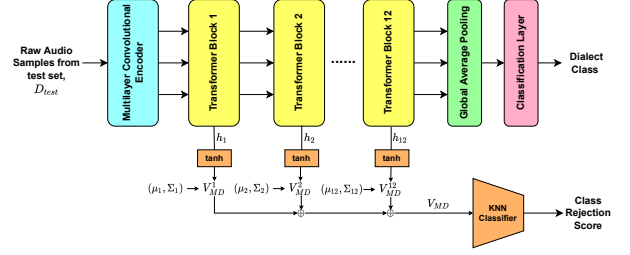
Figure 2: *OpenSet wav2vec 2.0 Dialect Classifier Architecture. Here, $\oplus$ is concatenation operator and $V_{MD}(x) = [V_{MD}^1(x) \oplus V_{MD}^2(x) \oplus \cdots \oplus V_{MD}^K(x)]$, is Mahalanobis Distance Feature Vector.*

experiments use only close-set training data with fixed known classes.

5.1. Dataset Details

We have used two custom-made datasets, *English Dialect Dataset*, and *Spanish Dialect Dataset* for the two most spoken languages, English and Spanish respectively to evaluate our method. We have sampled speech data from *AccentDB* [24], *UK and Ireland English Dialect speech dataset* [25] and *Google Nigerian English speech dataset*[6] for custom *English Dialect Dataset*. We have used whole *Latin American Spanish speech dataset* [26] for our custom *Spanish Dialect Dataset*. During training, we hide a few classes and used those hidden classes as unknown classes in the test set for efficacious open-world evaluation. The details on these custom speech datasets are following.

English Dialect Dataset consists of 11383 audio samples (spoken by 80 speakers) with 4 classes which are 'Southern', 'Northern', 'Welsh', and 'Scottish', used as fixed known classes, and 4800 audio samples (spoken by 12 speakers) with 4 classes which are 'Indian', 'American', 'Nigerian', and 'Australian' used as unknown class samples in the test set for OOD dialect detection evaluation. Close-set data is a subset of *UK and Ireland English Dialect speech dataset* and outlier or out-of-distribution samples are from both *AccentDB* and *Google Nigerian English speech dataset*. More details on the respective train, validation, and test set are provided in Table 1.

Spanish Dialect Dataset consists of 17724 audio samples spoken by 79 speakers with known 4 classes: 'Argentinian', 'Peruvian', 'Colombian', and 'Chilean', used as fixed known classes, and 3674 audio samples (spoken by 23 speakers) with 2 classes which are 'Venezuelan' and 'Puerto rico', which are used as unknown class test samples in the test set for OOD dialect detection evaluation. More details on the respective train, validation, and test set are provided in Table 1.

5.2. Evaluation Metric

Known dialect classification performance is measured using precision, recall, and F1 scores. For the detection of OOD dialects or unknown dialects, we are using the evaluation metrics that have been previously used in [27, 10, 6] because this can be considered as an out-of-distribution detection. Specifically, TP, TN, FP, FN, TPR, and FPR represent true positive, true negative, false positive, false negative, true positive rate,

Table 1: *Details of Speech Dialect Classification Datasets for Out of Distribution Detection.*

Dialect Dataset	Dataset Split	No of Samples	Time Duration (in Hours)
English	Train-set	9738	17.194
	Validation-set	550	1
	Test-set (known dialects)	5895	2.004
	Test-set (unknown dialects)	4800	5.5433
Spanish	Train-set	13715	21.29
	Validation-set	2010	3.233
	Test-set (known dialects)	5973	3.213
	Test-set (unknown dialects)	3974	5.817

and false positive rate respectively. We use the following metrics for OOD dialect detection evaluation:

AUROC (Higher is better) is the area under the Receiver Operating Characteristic (RoC) Curve. The RoC is plotted TPR against FPR by varying the threshold.

AUPR (Higher is better) is the area under the curve plotted precision against the recall by varying the threshold value. AUPR(IN) and AUPR(OUT) represent the fixed known classes and the outlier unknown classes as positive classes respectively.

EER (Lower is better) is the error rate of the classifier when the confidence threshold is set where the FPR (FPR = FP/(FP+TN)) is equal to FNR (FNR = FN/(FN+TP)).

$$EER = \frac{FP + FN}{TP + TN + FN + FP} \quad (2)$$

5.3. Close-set Performance Results

Table 2 shows the performance of the proposed model on known dialect categorization tasks for both datasets. These results show that the proposed method does not compromise the accuracy of the model in dialect classification tasks for known dialect classes.

Table 2: *Close-set Dialect Classification Performance Results*

Dataset	Recall	Precision	F1
English Dialect	90.1	89.7	89.23
Spanish Dialect	97.77	97.51	97.57

5.4. Ablation Study

Here, we experiment with different outlier detection models to show the effectiveness of our proposed model for class rejection score estimation. For this study, we have used Mahalanobis distance features to train CBLOF (cluster-based local outlier factor) [28], Isolation Forest [29], KNN [9], local outlier factor [30], and one-class SVM [31] models with the same setup as our own model for OOD task. Table 3 and Table 4 show that our proposed method delivers the best outcomes in both datasets.

Table 3: *Ablation study results of English Dialect Dataset*

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
CBLOF [28]	0.1625	92.12	76.09	9778
Isolation Forest [29]	0.1351	94.14	82.93	98.43
LOF [30]	0.2146	85.55	59.52	95.36
OC-SVM [31]	0.9762	51.19	60.25	90.89
Our Method [9]	**0.0959**	**96**	**86.78**	**98.81**

Table 4: *Ablation study results of Spanish Dialect Dataset*

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
CBLOF [28]	0.2836	77.8	64	86.02
Isolation Forest [29]	0.2791	78.36	66.47	85.52
LOF [30]	0.4412	58.07	39.88	71.70
OC-SVM [31]	0.6698	64.82	69.12	85.91
Our Method [9]	**0.2726**	**80.31**	**71.33**	**86.76**

5.5. Quantitative Comparison

We compare our method to other state-of-the-art methods discussed in recent literature and have reported their performance in Table 5 and Table 6 for each dialect dataset. We use the same Wave2vec2.0 model as the backbone throughout all these comparison methods to validate the comparison setup. From these results, it is very prominent that our proposed method outperforms other methods by considerable margins. Since our method makes use of multiple hidden layer embeddings and the KNN classifier model, it outperforms the closest method MD [6] and RMD [14].

Table 5: *Quantitative Comparison Results of Spanish Dialect Dataset*

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
Max Thresold [10]	0.4227	63.54	57.28	74.52
DOC [11]	0.3250	55.34	66.98	79.67
Openmax [12]	0.3585	55.81	37.84	71.56
MD [6]	0.3116	74.94	63.79	84.6
SNGP [13]	**0.2496**	62.39	55.31	78.65
RMD [14]	0.3106	75.02	63.07	84.64
Our method	0.2726	**80.31**	**71.33**	**86.76**

Table 6: *Quantitative Comparison Results of English Dialect Dataset*

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
Max Thresold [10]	0.1342	63.81	56.98	90.22
DOC [11]	0.3376	53.94	52.88	94.93
Openmax [12]	0.3468	78.34	56.51	91.54
MD [6]	0.3004	78.35	51.62	93.19
SNGP [13]	0.1716	86.38	57.37	95.13
RMD [14]	0.2876	79.97	49.63	94.17
Our method	**0.0959**	**96**	**86.78**	**98.81**

6. Conclusion

This paper presents the dialect classification problem in open-world scenarios and proposes a wav2vec 2.0 transformer model-based method to not only recognize dialects known during the training process but also detect unknown dialects as rejected classes at the inference time. We test our approach on two large-scale open-source dialect speech datasets and also present its performance comparison with other methods that are widely used in vision and language processing. Furthermore, our quantitative comparison experiment also indicates that integrating several intermediate layer output vectors to compute Mahalanobis distance-based feature vectors delivers higher performance than other prior Mahalanobis distance-based OOD detection methods [6, 14]. In future work, we would like to investigate how adversarial training and contrastive learning can be helpful for out-of-distribution dialect classification.

7. References

[1] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using gaussian mixture models," in *ODYSSEY04-The speaker and language recognition workshop*, 2004.

[2] A. Das, K. Kumar, and J. Wu, "Multi-dialect speech recognition in english using attention on ensemble of experts," in *ICASSP*. IEEE, 2021, pp. 6244–6248.

[3] Q. Zhang and J. H. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 873–882, 2018.

[4] T. Kong, S. Yin, D. Zhang, W. Geng, X. Wang, D. Song, J. Huang, H. Shi, and X. Wang, "Dynamic multi-scale convolution for dialect identification," *arXiv preprint arXiv:2108.07787*, 2021.

[5] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *5th International Conference on Learning Representations, ICLR 2017*, 2017. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl

[6] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[7] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *7th International Conference on Learning Representations, ICLR 2019*, 2019. [Online]. Available: https://openreview.net/forum?id=HyxCxhRcY7

[8] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pp. 2744–2751, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.244

[9] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *SIGMOD Rec.*, vol. 29, no. 2, p. 427–438, may 2000. [Online]. Available: https://doi.org/10.1145/335191.335437

[10] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations*, 2018.

[11] L. Shu, H. Xu, and B. Liu, "DOC: Deep open classification of text documents," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 2911–2916.

[12] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[13] L. et al., "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7498–7512, 2020.

[14] R. et al., "A simple fix to mahalanobis distance for improving near-ood detection," *arXiv preprint arXiv:2106.09022*, 2021.

[15] M. Hämäläinen, K. Alnajjar, N. Partanen, and J. Rueter, "Finnish dialect identification: The effect of audio and text," *arXiv preprint arXiv:2111.03800*, 2021.

[16] R. Imaizumi, R. Masumura, S. Shiota, H. Kiya *et al.*, "End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[17] B. Ma, D. Zhu, and R. Tong, "Chinese dialect identification using tone features based on pitch flux," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[18] W. Lin, M. Madhavi, R. K. Das, and H. Li, "Transformer-based arabic dialect identification," in *2020 International Conference on Asian Language Processing (IALP)*. IEEE, 2020, pp. 192–196.

[19] B. et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[21] L. et al., "Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings," *arXiv preprint arXiv:2110.03520*, 2021.

[22] S. et al., "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[23] X. et al., "Unsupervised out-of-domain detection via pre-trained transformers," *arXiv preprint arXiv:2106.00948*, 2021.

[24] A. et al., "Accentdb: A database of non-native english accents to assist neural speech recognition," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 5353–5360.

[25] D. et al., "Open-source multi-speaker corpora of the english accents in the british isles," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6532–6541.

[26] G.-R. et al., "Crowdsourcing latin american spanish for low-resource text-to-speech," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6504–6513.

[27] R. et al., "Out-of-domain detection based on generative adversarial network," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 714–718.

[28] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recogn. Lett.*, p. 1641–1650, 2003.

[29] F. T. Liu, K. M. Ting, and Z. hua Zhou, "Isolation forest," in *In ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society*, pp. 413–422.

[30] B. et al., "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, p. 93–104, 2000.

[31] S. et al., "Estimating support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, 07 2001.