OPEN SET DIALECT CLASSIFICATION USING WAV2VEC 2.0 MODEL

Sourya Dipta Das[1], *Yash A. Vadi*[2], *Abhishek Unnam*[3], *Kuldeep Yadav*[4]

SHL Labs[1,2,3,4]

ABSTRACT

Automatic Speech Recognition (ASR) systems perform poorly on dialects which are not well represented in training datasets, so dialect-specific ASR models are selected using automatic dialect classification models. In a real-world scenario, a deployed system can encounter both *known* (i.e. seen during training) and *unknown* dialects (i.e. unseen during training). A system must be able to handle both known and unknown dialects. In this paper, we fine-tuned a wav2vec 2.0 transformer-based architecture for known dialect classification. The fine-tuned model is able to classify *known dialects* that are part of training data. Further, we proposed a class rejection score estimation method that uses Mahalanobis distance to detect unknown dialects during inference. We present an experimental evaluation with two different datasets of English and Spanish languages. The proposed approach achieves an overall AUROC of 96% for the *English Dialect Dataset* and 80% AUROC for the *Spanish Dialect Dataset*.

Index Terms— speech recognition, Open Set Classification, Dialect identification, Wav2vec 2.0, automatic speech recognition, ASR

1. INTRODUCTION

Deep Learning based Automated Speech Recognition (ASR) has emerged as an essential component in many business applications to convert audio signals to text. However, the current ASR system fails to generalize well across the different dialects. Dialects are variants of the same language that are unique to distinct geographical locations or social groupings and may have significant variances, including phonological, grammatical, orthographic, and other levels. A previous study [1] has showcased that the average word error rate (WER) of a cross-dialect system was 30.9%, whereas it was only 23.4% for a dialect-specific system. Therefore, systems that are simultaneously trained on multiple dialects exhibit poor performance on individual dialects. One of the approaches to deal with low accuracy is that classify the dialect of the incoming speech and then select an ASR model specifically trained for that dialect. In open-set [2] dialect classification task [3], there are not only known dialect classes to be recognized but also unknown classes that were not present during the training. This type of difficulty is significantly more common in speech-processing applications because it is not possible to guarantee that all probable dialects of speech that may occur during the deployment of the system have been previously known. Hence, it is essential for an ASR system to have the capability to recognize unknown dialects.

In this paper, we present an approach for open-set dialect classification that automatically classifies known dialects from input speech and also detects input audio that does not belong to any of the dialects used to train the model. We used a pre-trained wav2vec2.0 model and fine-tuned it on the known dialects to adapt the feature embedding. Further, we retrieved features from intermediate transformer layers and computed the Mahalnoobis distance score for each layer, which was then used as a feature vector for KNN [4]-based outlier detector. We have evaluated our proposed method by comparing it with state-of-the-art open set classification methods [5, 6, 7, 8, 9, 10].

Our contributions can be summarized as follows: (1) We proposed the first open-set dialect identification method, which is a low-cost, easy-to-use technique to identify known dialects and/or reject input speech samples of unknown dialects in a single forward pass. (2) We evaluated the performance of open-set dialect classification 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. (3) We curated two open-source datasets for an open set dialect classification problem and plan to make them available for the research community.

2. RELATED WORK

There has been very limited research work on dialect identification and no prior work on open-set dialect identification. Torres *et al.* [11] had done some earlier work on dialect identification where they used the Gaussian mixture model (GMM) with shifted delta cepstral features (SDC). Recently, there have been several studies that have explored general open-set classification problems in various domains like vision, text, etc. Liang *et al.* [5] 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.* [6] 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.* [7] 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.* [8] 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.* [10] 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.* [9] has developed a robust uncertainty-based methodology that delivers an uncertainty score to each prediction and may be used to discover outliers.

3. PROBLEM STATEMENT

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

Now, our task is to train a classifier $F_D(x)$ with only training data, D_{train} that correctly predicts the dialect class from a set of known dialect classes, i.e. $F_D(x) = [d_1, d_2, \dots, d_M]$ 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.

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

where $\hat{y_p}$ is the predicted class, $G_D(x)$ 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.

4. METHODOLOGY

4.1. Model Architecture and Fine-Tuning

Wav2Vec 2.0 speech model [12] is pre-trained on unlabeled speech data using self-supervised learning for learning good representations of speech. It shows promising results when transferred to other tasks [13, 14] like speech classification, speech recognition, speech frame classification, etc. Because of this reason, We have used a pre-trained wav2vec 2.0 model and 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 $k \in [1, 2, \dots, K]$. we further passed those intermediate feature embeddings through a hyperbolic tangent function, $tanh(.)$ to transform the features into same restricted semantic space, i.e., $h_k^t(x) = tanh(F_D^k(x))$ From a previous study [15], different transformer layers of 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., $\phi_h(x) = [h_1^t, h_2^t, \dots, h_K^t]^T \in \mathbb{R}_{d.K}$. From previous work [8], 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 [16]. 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 equation.

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

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

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

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

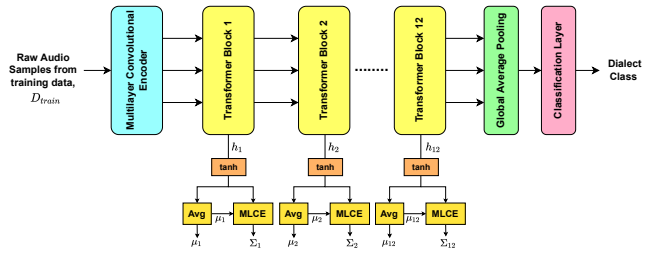

Fig. 1. Illustration of Layer Feature Embedding Mean(μ_k), Covariance Matrix (Σ_k) Estimation for k-th transformer layer from the feature embeddings,$F_D^k(x)$ of training data, D_{train}. Here, *Avg* is component-wise vector average operation and *MLCE* is Maximum Likelihood Covariance Estimator.

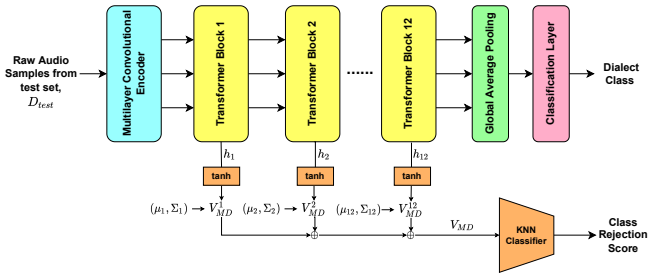

Fig. 2. OpenSet wav2vec 2.0 Dialect Classifier Architecture. Here, $\oplus$ is concatenation operator and $V_{MD}(x) = [V_{MD}^1(x) \oplus V_{MD}^2(x) \oplus \cdots \oplus V_{MD}^K(x)]$, is Mahalanobis Distance Feature Vector.

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 [4] classifier 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.

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. 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 [4] 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 (12GB) with 16 GB RAM. For both training and validation, all experiments use solely close set data with fixed known classes.

Table 1. Details of Open-set Speech Dialect Classification Datasets.

Dialect Dataset	Dataset Split	No of Samples	Time Duration (in Hours)
English	train-set	9738	17.194
	validation-set	550	1
	known_dialect_test-set	5895	2.004
	unknown_dialect_test-set	4800	5.5433
Spanish	train-set	13715	21.29
	validation-set	2010	3.233
	known_dialect_test-set	5973	3.213
	unknown_dialect_test-set	3974	5.817

5.1. Dataset Details

We have used two custom-made datasets, *English Dialect Dataset*, *Spanish Dialect Dataset* for the two most spoken languages, English and Spanish respectively to evaluate our method. We have sampled speech data from *AccentDB* [17], *UK and Ireland English Dialect speech dataset* [18] and *Google Nigerian English speech dataset*[6] for custom *English Dialect Dataset*. We have used whole *Latin American Spanish speech dataset* [19] 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. More 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 open set evaluation. Close set data is a subset of *UK and Ireland English Dialect speech dataset* and outlier 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'; 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 open set evaluation. More details on the respective train, validation, and test set are provided in Table 1.

[1] https://huggingface.co
[2] https://huggingface.co/facebook/wav2vec2-base
[3] https://huggingface.co/facebook/wav2vec2-base-10k-voxpopuli-ft-es
[4] https://pyod.readthedocs.io/en/latest/
[5] https://scikit-learn.org/
[6] https://openslr.org/70/

5.2. Evaluation Metric

Known dialect (close set) classification performance is measured using precision, recall, and F1 score. For the detection of open set or unknown dialects, we are using the evaluation metrics that have been previously used in [20] and ([5] [8]) because this can be considered as an out of domain detection. Specifically, TP, TN, FP, FN, TPR, and FPR represent true positive, true negative, false positive, false negative, true positive rate, and false positive rate respectively. We use the following metrics for open-set 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 class as positive class 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)).

$$\mathrm{EER} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{2}$$

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 train CBLOF (cluster based local outlier factor) [21], Isolation Forest [22], KNN [4], local outlier factor [23], and one-class SVM [24] models. Table 3 and Table 4 show that the KNN-based method delivers the best outcomes in both datasets.

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 6 and Table 5 for each dialect dataset. 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 [8] and RMD [10].

Table 3. Ablation study results of English Dialect Dataset

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
CBLOF [21]	0.1625	92.12	76.09	9778
Isolation Forest [22]	0.1351	94.14	82.93	98.43
LOF [23]	0.2146	85.55	59.52	95.36
OC-SVM [24]	0.9762	51.19	60.25	90.89
KNN [4]	**0.0959**	**96**	**86.78**	**98.81**

Table 4. Ablation study results of Spanish Dialect Dataset

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
CBLOF [21]	0.2836	77.8	64	86.02
Isolation Forest [22]	0.2791	78.36	66.47	85.52
LOF [23]	0.4412	58.07	39.88	71.70
OC-SVM [24]	0.6698	64.82	69.12	85.91
KNN [4]	**0.2726**	**80.31**	**71.33**	**86.76**

Table 5. Quantitative Comparison Results of Spanish Dialect Dataset

Methods	EER	AUROC	AUPR (IN)	AUPR (OUT)
Max Thresold [5]	0.4227	63.54	57.28	74.52
DOC [6]	0.3250	55.34	66.98	79.67
Openmax [7]	0.3585	55.81	37.84	71.56
MD [8]	0.3116	74.94	63.79	84.6
SNGP [9]	**0.2496**	62.39	55.31	78.65
RMD [10]	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 [5]	0.1342	63.81	56.98	90.22
DOC [6]	0.3376	53.94	52.88	94.93
Openmax [7]	0.3468	78.34	56.51	91.54
MD [8]	0.3004	78.35	51.62	93.19
SNGP [9]	0.1716	86.38	57.37	95.13
RMD [10]	0.2876	79.97	49.63	94.17
Our method	**0.0959**	**96**	**86.78**	**98.81**

5.6. Conclusion

This paper showcased the open-set dialect classification problem in open-world scenarios and proposed 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. In this paper, we tested 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. In future work, we would like to investigate how adversarial training and contrastive learning can be helpful for open-set dialect classification.

6. REFERENCES

[1] Eiman Alsharhan and Allan Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 975–998, 2020.

[2] Scheirer et al., "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[3] Wang et al., "An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model," *Proc. Interspeech 2021*, pp. 3266–3270, 2021.

[4] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim, "Efficient algorithms for mining outliers from large data sets," *SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, may 2000.

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

[6] Lei Shu, Hu Xu, and Bing 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.

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

[8] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[9] Liu 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.

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

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

[12] Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[13] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

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

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

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

[17] Ahamad 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.

[18] Demirsahin 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.

[19] Guevara-Rukoz 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.

[20] Ryu 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.

[21] Zengyou He, Xiaofei Xu, and Shengchun Deng, "Discovering cluster-based local outliers," *Pattern Recogn. Lett.*, p. 1641–1650, 2003.

[22] Fei Tony Liu, Kai Ming Ting, and Zhi 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.

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

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