INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice

Juan Zuluaga-Gomez[⋆,†,‡], *Sara Ahmed*[1], *Danielius Visockas*[§], *Cem Subakan*[♮,♯,♭].

[†]Idiap Research Institute, Switzerland [‡]Ecole Polytechnique Federale de Lausanne, Switzerland
[1]Sketch Recognition Lab, Texas A&M University, USA
[§]Vilnius Gediminas Technical University, Lithuania
[♮]Université Laval, Canada [♯]Concordia University, Canada [♭]Mila-Québec AI Institute, Canada

juan-pablo.zuluaga@idiap.ch

Abstract

Despite the recent advancements in Automatic Speech Recognition (ASR), the recognition of accented speech still remains a dominant problem. In order to create more inclusive ASR systems, research has shown that the integration of accent information, as part of a larger ASR framework, can lead to the mitigation of accented speech errors. We address multilingual accent classification through the ECAPA-TDNN and Wav2Vec 2.0/XLSR architectures which have been proven to perform well on a variety of speech-related downstream tasks. We introduce a simple-to-follow recipe aligned to the SpeechBrain toolkit for accent classification based on Common Voice 7.0 (English) and Common Voice 11.0 (Italian, German, and Spanish). Furthermore, we establish new state-of-the-art for English accent classification with as high as 95% accuracy. We also study the internal categorization of the Wav2Vev 2.0 embeddings through t-SNE, noting that there is a level of clustering based on phonological similarity.[1]

Index Terms: automatic accent classification, ECAPA-TDNN, Wav2Vec 2.0, SpeechBrain, Common Voice dataset,

1. Introduction

Large acoustic models (LAMs) have become the standard choice for a large variety of downstream tasks, such as automatic speech recognition (ASR) or language identification [1].[2] These LAMs are trained using self-supervised learning (SSL) techniques on vast amount of data, often exceeding 50k hours of audio. Their powerful capabilities are evident in the application of datasets such as Librispeech [2]. However, a significant limitation of these models is that they primarily utilize data spoken in native English or a single accent, thereby neglecting the diversity of accents among speakers. Examples of these pretrained LAMs in only-English are Wav2Vec 2.0 (w2v2) [3] or in a multilingual setup, the w2v2-XLSR model [4].

Previous studies have demonstrated that end-to-end ASR models based on pretrained LAMs (e.g. w2v2) exhibit significant performance disparities when applied to non-native English speech. For example, up to $\sim$50% relative increase in word error rate (WER) has been observed when comparing native (US) and non-native (Malaysian) English speech [5]. As a result, it has become crucial to develop and implement accent-aware or accent-invariant ASR systems. However, only a limited number of studies have specifically addressed this issue.

In this paper, we study the accent classification problem, which is a critical building block towards accent-aware ASR. Our aim is to provide insight and guidance for evaluating fine-tuned LAMs in a more inclusive manner, highlighting the importance of considering accent variability not only in ASR, but also in different downstream tasks where accent disparity might degrade performances (e.g., spoken language understanding). Specifically, in this work, we introduce a simple-to-follow recipe on the SpeechBrain [6] toolkit to perform accent classification based on speech recordings. The recipe fine-tunes either ECAPA-TDNN [7] or w2v2 [3] models (also XLSR) in the accent classification task. Our system follows closely the CommonLanguage recipe[3] available in SpeechBrain. Additionally, we open-source fine-tuned models in the HuggingFace Hub [8, 9].

Our recipe utilizes data from Common Voice [10] dataset in four different languages, namely: English, German, Spanish, and Italian. Our contributions are four-fold as follows:

- We open-source ECAPA-TDNN [7] and w2v2 [3] fine-tuned models that recognize 16 different accents in the English language, which to the author's knowledge is the largest open-source accent classification system to date. In addition, we also cover 4 accents in German, 6 in Spanish, and 5 in Italian.
- We set the first baseline for accent classification based on Common Voice dataset [10] which, to the author's knowledge, is the largest open-source and free-access acoustic database that provides accent annotations.
- We introduce CommonAccent, a subset of Common Voice compiled as a benchmark dataset optimized for accent classification in multiple languages, e.g., English, German, Spanish, and Italian.
- Finally, we open-source a recipe, named CommonAccent, in the SpeechBrain tookit [6] for performing accent classification based on speech recordings. CommonAccent can be easily adapted to other languages.

In Section 2 we cover early work on accent classification. Section 3 formalizes the data preparation phase for CommonAccent, while also describing the datasets partition. We discuss the main results in Section 4 and 5 and conclude the paper in Section 6.

[⋆] Corresponding author.

This work was partially supported by *ROXANNE*, a European Union's Horizon 2020 research project leveraging real-time network, text, and speaker analytics for combating organized crime (grant 833635).

[1] Our recipe is open-source on the SpeechBrain toolkit, see: `https://github.com/speechbrain/speechbrain/tree/develop/recipes`

[2] See more downstream applications in the SUPERB [1] website: `https://superbbenchmark.org/`

[3] `https://github.com/speechbrain/speechbrain/tree/develop/recipes/CommonLanguage`.

10.21437/Interspeech.2023-2419

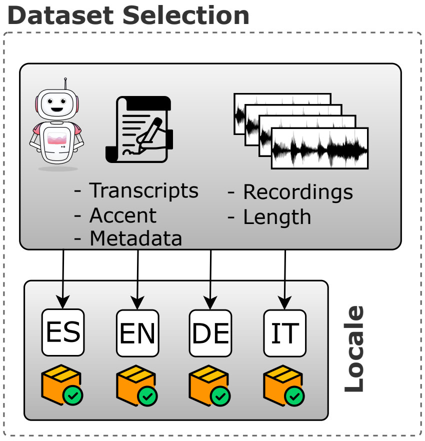


Figure 1: *Audio data selection and packaging for English, German, Spanish, and Italian locales of Common Voice dataset.*

2. Related Work

Accents are considered one of the main sources of speech variability. Differences between accents are reflected primarily in three aspects: stress, tone, and length [11]. Accent classification is similar to language identification [12, 13] and speaker verification [14, 15, 16] as it classifies sequences of speech at the full-length utterance level. Prior research has extensively explored ways to improve the classification of accented speech and its effect on ASR. Early works explored contextual Hidden Markov Model (HMM)-based units [17] and the use of formant frequency features into GMM models [18]. More recent research has leveraged techniques from a variety of speech technology domains which has lead to promising results. In a recent Interspeech (2020) competition [11],[4] the highest performing model used a TDNN based classification network with phonetic posteriorgram (PPG) features as input and TTS (text-to-speech) to augment the training data [11, 19]. Another proposed accent classification network, mined elements from a deep speaker identification framework to make it applicable for accent classification. In detail, they implemented a Convolutional Recurrent Neural Network as a front-end encoder, integrated local features using a Recurrent Neural Network, included a Connectionist Temporal Classification (CTC) based speech recognition auxiliary task, and introduced some strong discriminative loss functions [20]. This work expands upon this area of research by studying more recent deep neural networks, e.g., XLSR, on accent classification for different languages.

Accent information, when integrated into an ASR framework, yields promising results. Multitask learning provides a way to transition from maintaining separate acoustic models for each accent to sharing all acoustic model parameters except for accent-specific top-layers [21, 22]. The inclusion of the speaker's native language data in multitask models has also led to notable decreases in the error rate as well [23, 24]. Other areas such as domain expansion [25] and pronunciation modification have also been explored [26]. A CTC model initialized by a w2v2 encoder with LAS rescoring achieved the highest results when compared to other models at the Interspeech accent recognition challenge [11]. In addition to accent, end-to-end systems that account for dialect variations in grammar and vocabulary have also led to error reduction [27].

Table 1: *Partition of train, dev and test sets.* [†]*also includes South Atlantic and Bermuda accented English.* [‡]*includes accents only from Italy.*

Language	**Accents**	**# Utterances [k]/dur [hrs]**		
(Nb. accents)		**Train**	**Dev**	**Test**
English (16)[†]	en-MY, en-SG, zh-HK, fil-PH, af-ZA, en-NZ, ga-IE, gd-GB, en-AU, en-CA, en-GB, en-IN, en-US, cy-CB,	93.5/154	1.4/2.4	1.4/2.3
German (4)	de-IT, de-CH, de-AT, de-DE	39.8/70	0.5/0.9	0.5/0.9
Spanish (6)	Mexico, Chile, Caribe, Rioplatense, Andino, Spain	51.4/78	0.6/0.9	0.6/0.9
Italian (5)[‡]	Romagna, it-Meridional, Veneto, Sicilia, Trentino	2.6/3.7	0.37/0.6	0.37/0.6

There are still many challenges facing accented speech and accented ASR. For example, the lack of a standard benchmark, dialectal difference encompassing grammar, vocabulary, and spelling, generalizing to a larger selection of accents, and whether accent-tuning approaches in English are applicable to other languages as well, particularly those that show more dialectal variance [28]. We address these problems in regard to accent classification by introducing a standard, open-source benchmark derived from Common Voice. We take the approach that implementing LAMs, without further modification to the architecture, yields promising results.

3. CommonAccent

CommonAccent is a recipe that is derived from the Common Voice dataset [10]. Common Voice is a massive multilingual corpus that contains annotations of speech recordings in over 100 languages (as of May 2023). The data is constantly updated with new recordings, therefore, the authors label each release with a number, e.g., *Common Voice 11.0*. In addition to transcripts and language ID locales, some recordings contain information such as speaker's gender, accent, and age. Train, dev and test splits intended primarily for ASR are also provided.

This work employs the Common Voice dataset to perform accent classification on different languages. The CommonAccent recipe uses two versions of Common Voice. For English (EN), we use Common Voice 7.0 (CV7), while for German (DE), Spanish (ES), and Italian (IT) we use Common Voice 11.0 (CV11).[5]

Data selection process: To prepare each dataset (e.g., EN) for analysis, all samples that are devoid of accent annotations are removed. Next, the train, dev, and test sets are combined into a single dataset since the original dev and test sets have an insufficient number of samples for some accents; in some cases, none at all. The number of samples per accent are then tallied to create new train, dev, and test sets, ensuring that no more than 100 samples per accent are included in the dev and test sets. If a given accent has fewer than 300 samples, the remaining samples are split using a 60/20/20 split for train, dev, and test sets, respectively. The data selection process is summarized in Figure 1. CV7 is sourced directly from the CommonVoice website, while CV11 from the HuggingFace Hub [8, 9]. The resulting train, dev, and test set proportions for each language in consideration are presented in Table 1.[6]

[4] Interspeech 2020 challenge: Accented English Speech Recognition Challenge.

[5] This recipe can be used with other datasets such as L2 Artic [29].

[6] We open-source a data preparation script to parse the

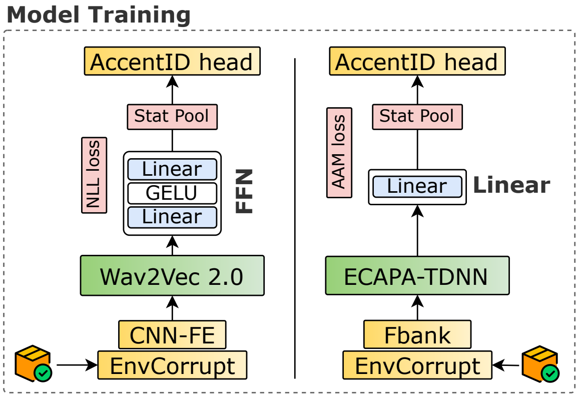


Figure 2: *AccentID classification system. Model selection and training. We either fine-tune a pre-trained w2v2 (XLSR) or a ECAPA-TDNN model. The former uses NLL loss, while the later AAM loss. The w2v2 model is also interchangeable by other acoustic models. CNN-FE stands for Convolutional Neural Network Front-end.*

4. Experimental setup

The proposed experimental setup is split in two parts: 1) fine-tuning w2v2-XLSR and ECAPA-TDNN models in order to obtain baseline results in English and 2) expanding the CommonAccent recipe for three additional languages using the w2v2-XLSR model [4]. This serves the purpose of establishing that CommonAccent generalizes well regardless of different training scenarios with different pre-trained models and architectures. During experimentation, data augmentation by speed perturbation and additive noise from the OpenRIR database was applied [30].

4.1. ECAPA-TDNN

The ECAPA-TDNN model has shown state-of-the-art results in speaker verification tasks. It builds on the original x-vector architecture [31] through an increased focus on channel attention, propagation, and aggregation. The architecture includes an incorporation of Squeeze-Excitation blocks, multi-scale Res2Net features, extra skip connections, and channel dependent attentive statistics pooling as output [7].

We examine the implementation of this architecture in accent classification through two models: one trained with SpecAugmentation [32] and speed perturbation and a baseline Accent Identification model without data augmentation. Both models are fine-tuned from the checkpoints on HuggingFace[7] and we use additive angular margin loss [33].

4.2. Wav2Vec 2.0 & XLSR

The w2v2-XLSR model is designed to acquire cross-lingual speech representations for 53 languages, utilizing the raw waveform of speech to train on 56K hours of unlabeled data [4]. Based on the w2v2 architecture [3], it learns contextualized speech representations and multilingual quantized latent speech representations simultaneously. These shared representations facilitate cross-lingual knowledge transfer from high-resource languages to low-resource languages, thereby improving the performance of the latter. The w2v2-XLSR model has proven superior to prior approaches in both language identification and speaker identification tasks [4].[8] To merge the information across embeddings from the same utterance, we added a `StatPooling()` layer for both w2v2-XLSR models. Finally, we trained end-to-end with the NLL loss function. A global overview of the architecture can be found in the left panel of Figure 2.

4.3. Training

We perform two training strategies. The first one fine-tunes a pre-trained Wav2Vev 2.0/XLSR [3, 4] model with NLL loss. The second one performs training with a ECAPA-TDNN network [7]. A workflow of the proposed architectures is on Figure 2. Adam [34] optimizer is used across all experiments, with a learning rate scheduler that anneals the learning rate ($\alpha = 1e^{-3}$) after the end of each epoch ($\beta = 0.95$). All models are trained for 30 epochs. Additionally, a dynamic batching strategy is used while training each model. Dynamic batching aims at reducing the amount of zero-padding in the input batches to the model. This in turn, reduces the overall training time. We use effective $max\_batch_{len} = 600$ and $num_{buckets} = 200$. At decoding time we use a fixed batch size of 16. All experiments run in a GeForce RTX 3090.

5. Results

The results section is divided into three main parts. Firstly, we assess the accent classification performance of ECAPA-TDNN and w2v2-XLSR models on the English CommonAccent dataset (CV7), and conclude that the w2v2 pre-trained models are better suited for accent classification purposes. Secondly, we fine-tune these models for accent recognition in German, Spanish, and Italian. At the end, we conduct a clustering-based analysis of the embeddings derived from the fine-tuned w2v2-XLSR models.

5.1. ECAPA-TDNN vs Wav2Vec 2.0

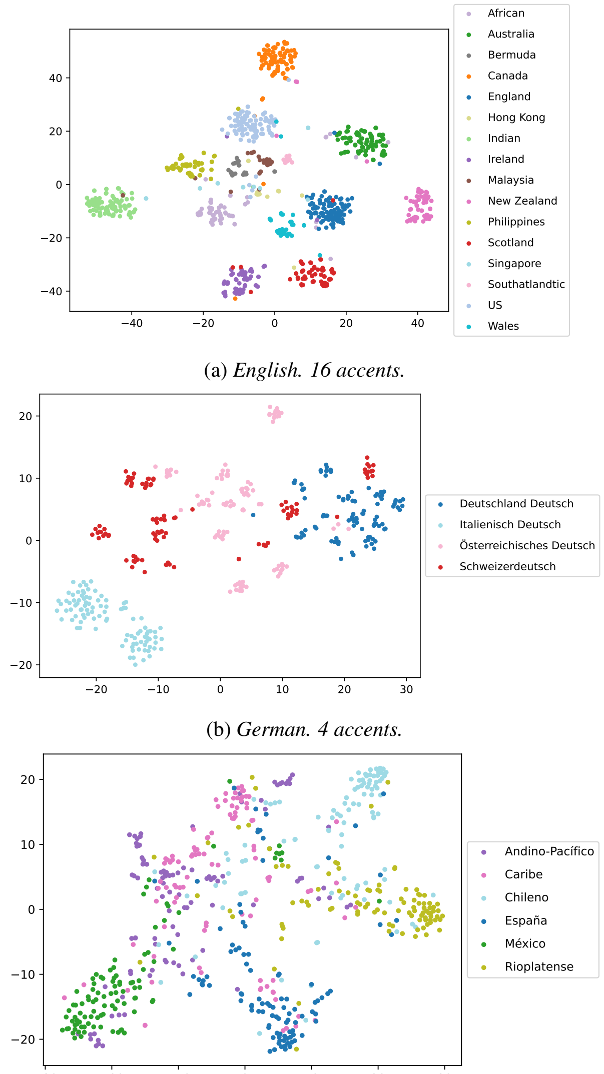
Table 2 presents the accuracy scores for both ECAPA-TDNN and w2v2 models fine-tuned on the English CommonAccent dataset. The table clearly shows that the w2v2 models consistently outperform ECAPA-TDNN on the dev and test sets, with an accuracy improvement 79.0% → 95.1%. We also observe a significant improvement in accuracy for both models when applying SpecAugment technique. For instance, ECAPA-TDNN improves: 79.0% → 89.7% and w2v2-XLSR: 95.1% → 97.1%.

These results are not unexpected and demonstrate the effectiveness of data augmentation in addressing imbalanced data distributions, especially for low-resource accents within the same language. Additionally, the w2v2 model is less sensitive to data augmentation and consistently maintains a high level of accuracy. Perhaps this is due to the large-scale pretraining stage of XLSR.

CommonVoice dataset from HuggingFace into a CSV file.

[7] The model we fine-tuned was based on Language Identification and was trained on the CommonLanguage dataset, see `https://github.com/speechbrain/speechbrain/tree/develop/recipes/CommonLanguage/lang_id` Checkpoint at HuggingFace: `https://huggingface.co/speechbrain/lang-id-commonlanguage_ecapa`

[8] Access to pre-trained w2v2-XLSR model checkpoint: `huggingface.co/facebook/wav2vec2-large-xlsr-53`



(a) *English. 16 accents.*

(b) *German. 4 accents.*

(c) *Spanish. 6 accents.*

Figure 3: *Analysis of the internal categorization of Accent Classifiers by t-SNE plots. The plots list the internal categorization of (a) English, (b) German, and (c) Spanish. The embeddings are obtained from the last layer of the fine-tuned w2v2-XLSR model on each independent language (only test set). The embeddings have a dimension of 1024.*

5.2. Baselines In Other Languages

To further analyze these results, a series of experiments were conducted by fine-tuning the w2v2-XLSR models on the German, Spanish, and Italian CommonAccent datasets. Based on the superior results of the w2v2-XLSR model from § 5.1 we only continue studies with this model on the other 3 languages and summarize the results in Table 3. Although the w2v2-XLSR model yielded above 95% accuracy for English, substantial degradation for other languages can be seen. Overfitting was observed in the Italian model (76.1% dev → 99% test) and underfitting in the Spanish model (below 69% accuracy on dev and test sets). One possible explanation is the high degree of phonological similarity between the accents, which makes the distinction more subtle. For example, five out of the six Spanish accents are from closely located regions (Latin America), which could be a contributing factor to the model's low performance. However, the German model performed relatively well, achieving an accuracy of over 75%. We also used t-SNE plots (see Figure 3) to visualize the clustering of the accents, which clearly show that the models are learning meaningful representations of the accents within the same language.

Table 2: *Accuracy score on two type of models trained on the English CommonAccent dataset with and without data augmentation (speed and noise perturbation).*

Model	Aug.	Dev	Test
ECAPA-TDNN	✗	78.5	79.0
↪	✓	91.5	89.7
w2v2-XLSR	✗	95.2	95.1
↪	✓	96.5	97.1

Table 3: *Accuracy score in German, Spanish, and Italian CommonAccent datasets. Results are with a w2v2-XLSR model fine-tuned on each language independently. †includes speed perturbation and noise perturbation during fine-tuning.*

Locale	Aug.†	Dev	Test	Test loss
German	✓	66.2	75.5	0.937
Spanish	✓	64.2	68.5	1.22
Italian	✓	76.1	99.0	0.392

5.3. Clustering

As shown in Figure 3, the embeddings for each language in the w2v2-XLSR architecture show a level of clustering based on phonological similarity and geographical proximity. For example, the English plot displays England, Wales, and Scotland close to each other. US and Canada are also next to each other as is Australia and New Zealand. While the t-SNE plot shows that the model has recognized similarities in accents, the relational differences between them is not so clear. For example, the Australian accent is equidistantly placed between US and New Zealand, although there are considerable difference between the IPA phonemes for US English, while Australia and New Zealand are the same.[9]

6. Conclusion

In this work, we showed that pre-trained acoustic models like XLSR can be adapted for accent classification systems, particularly for English (§ 5). We also introduced CommonAccent (§ 3), a benchmark dataset for accent classification tasks in four languages. We hope it becomes the default benchmark, like VoxCeleb for speaker recognition, as there is currently a lack of standardized benchmarks within the domain targeted in this paper. Finally, we open-sourced our models in English, Spanish, German, and Italian, along with the accent classification recipe in SpeechBrain that can be easily adapted to other datasets and languages.

[9] As given in the IPA charts for Amazon Polly: `https://docs.aws.amazon.com/polly/latest/dg/ref-phoneme-tables-shell.html`

7. References

[1] S. wen Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "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.

[4] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[5] G. Cámbara, A. Peiró-Lilja, M. Farrús, and J. Luque, "English accent accuracy analysis in a state-of-the-art automatic speech recognition system," *arXiv preprint arXiv:2105.05041*, 2021.

[6] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[7] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[8] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45.

[9] Q. Lhoest *et al.*, "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 175–184.

[10] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[11] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6918–6922.

[12] P. Rangan, S. Teki, and H. Misra, "Exploiting spectral augmentation for code-switched spoken language identification," *arXiv preprint arXiv:2010.07130*, 2020.

[13] N. E. Safitri, A. Zahra, and M. Adriani, "Spoken language identification with phonotactics methods on minangkabau, sundanese, and javanese languages," *Procedia Computer Science*, vol. 81, pp. 182–187, 2016.

[14] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[15] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[17] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1784–1787.

[18] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*. IEEE, 2005, pp. 139–143.

[19] H. Huang, X. Xiang *et al.*, "Aispeech-sjtu accent identification system for the accented english speech recognition challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6254–6258.

[20] W. Wang, C. Zhang, and X. Wu, "Deep discriminative feature learning for accent recognition," *arXiv preprint arXiv:2011.12461*, 2020.

[21] Y. Huang, D. Yu, C. Liu, and Y. Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[22] M. Chen, Z. Yang, J. Liang, Y. Li, and W. Liu, "Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] S. Ghorbani and J. H. Hansen, "Leveraging native language information for improved accented speech recognition," in *Proc. Interspeech 2018*, 2018, pp. 2449–2453. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1378

[24] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Proc. Interspeech 2018*, 2018, pp. 2454–2458. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1864

[25] S. Ghorbani, S. Khorram, and J. H. Hansen, "Domain expansion in dnn-based acoustic models for robust speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 107–113.

[26] K. Radzikowski, L. Wang, O. Yoshie, and R. Nowak, "Accent modification for speech recognition of non-native speakers using neural style transfer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–10, 2021.

[27] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-aware modeling for end-to-end japanese dialect speech recognition," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 297–301.

[28] A. Hinsvark *et al.*, "Accented speech recognition: A survey," *arXiv preprint arXiv:2104.10747*, 2021.

[29] G. Zhao, S. Sonsaat *et al.*, "L2-arctic: A non-native english speech corpus." in *Interspeech*, 2018, pp. 2783–2787.

[30] T. Ko *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[31] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[32] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[33] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.