Accented Speech Recognition: A Survey

Arthur Hinsvark[1], *Natalie Delworth*[1], *Miguel Del Rio*[1], *Quinten McNamara*[1], *Joshua Dong*[1], *Ryan Westerman*[1], *Michelle Huang*[1], *Joseph Palakapilly*[1], *Jennifer Drexler*[1], *Ilya Pirkin*[1], *Nishchal Bhandari*[1], *Miguel Jetté*[1]

[1]Rev.com

{ arthur, natalie.delworth, miguel.delrio, quinn, joshua.dong, ryan.westerman, michellehuang, joseph.palakapilly, jennifer.drexler, ilya.pirkin, nishchal, miguel } @rev.com

Abstract

Automatic Speech Recognition (ASR) systems generalize poorly on accented speech. The phonetic and linguistic variability of accents present hard challenges for ASR systems today in both data collection and modeling strategies. The resulting bias in ASR performance across accents comes at a cost to both users and providers of ASR.

We present a survey of current promising approaches to accented speech recognition and highlight the key challenges in the space. Approaches mostly focus on single model generalization and accent feature engineering. Among the challenges, lack of a standard benchmark makes research and comparison especially difficult.

Index Terms: accented speech recognition, accent, dialect

1. Introduction

Speech recognition for in-domain audio has reached high levels of performance[1] and widespread use. However, ASR performance degrades significantly on accented speech[2]. Thus, accent-robustness is needed for speech recognition to be solved in the wild. Generalizing speech recognition across dialects is a hard problem for real-world speech systems.

Dialects are variations within a language that differ in geographical regions and social groups, which can be distinguished by traits of grammar, vocabulary[3] and, of particular interest to ASR, phonology[4]. The terms “accent” and “dialect” are used interchangeably in the context of speech recognition as both represent the primary speaker-specific phonetic variation in a given language[5].

Accurate speech transcript data is hard to source, with accurately labeled and transcribed accented speech even more so. This lack of data means standard approaches to speech recognition perform poorly on accented speech. A lack of common training and test suites in the research community makes it difficult to compare similar approaches.

This paper reviews the current state of the art of accent-robust ASR. We first discuss accent-specific models and how their difficulty in production caused their decline in recent research. The primary approaches we observed attempt to improve the data-efficiency and generalization capability of modeling from little available training data using techniques such as multi-task training or semi-supervised learning. Feature engineering approaches also showed success, such as direct lexicon modification, accent embeddings, and adversarial training for feature extraction. Finally, there is some research into application of existing ASR systems to minimize error on accents through techniques such as accent-error correction systems.

Sections 2 through 5 showcase the techniques applied in current literature, while Section 6 and 7 discuss the open problems and suggest key areas for future research.

2. Accent-Specific Modeling

A straightforward approach to accent-robust ASR design is to use an accent identification system to manage multiple accent-specific ASR systems[6]. However, accent-specific performance depends heavily on the availability of data for the specific accent, which can be alleviated through data augmentation such as speed modification[7]. The naive approach is to create a single model, pooling all the data, but this is shown to under-perform compared to accent-specific models[8]. However, training multiple accent-specific models is cumbersome in production and increases maintenance costs. These costs motivate techniques for adapting a unified model that performs well on all accents.

3. Model Generalization

3.1. Multi-task Learning

Multi-task learning incorporates additional information into neural network training. When models trained to perform two (or more) separate tasks share some of their parameters, those shared layers learn to represent the information necessary for all tasks. For ASR, multi-task learning can be used in the context of neural network acoustic models or end-to-end ASR models. The first use of multi-task learning for accented ASR was when neural network acoustic models first became popular[9]; multi-task learning was a natural way to switch from having separate acoustic models for different accents to having accent-specific top-layers but sharing all other acoustic model parameters across accents[9, 10]. While this is both more efficient and better-performing than separate models, it is still a somewhat indirect way of incorporating accent information.

Prasad and Jyothi[11] presented a variety of in-depth approaches to analyzing DeepSpeech2 neural layers for accented speech. The authors show that the strongest gradients of the final neural ASR layer for a given word align with the timing interval for that word. The effect is especially strong in US and Canadian accented speech, and weaker in Indian accented speech. Prasad and Jyothi also show that the first RNN layer in DeepSpeech2 contains the most accent information as demonstrated by the mutual information between the ASR layers and the accent classification. They posit that the preliminary layers of the network are most useful for accent classification tasks.

To make full use of accent labels, accent identification is the natural choice for an auxiliary task within a multi-task learning framework. In Yang et al.[4], the authors train an acoustic model with separate output layers for American and British English, as well as an accent classifier that branches off from the first layer of the acoustic model. The authors introduce a tuning parameter, α, that scales the accent ID objective relative to the acoustic model objective and is tuned on the development data.

Having separate output layers for different accents requires making a choice of which output layer to use at inference time, based on either knowledge of the input accent or on the output of an accent classifier. Jain et al.[12] use a similar approach to Yang et al.[4], but instead use a single output layer for all accents. Additionally, Jain et al.[12] explores the use of a standalone accent classifier to produce features that are then used as additional input to the ASR model. Combining accent embedding input and multi-task learning performs best.

Li et al.[13] apply multi-task learning with accent ID to end-to-end ASR. They insert special accent ID tokens to the text labels sequences during training. For example, the output targets for a British accented speech utterance of "hello" would be "<sos> <en-gb> h e l l o <eos>" instead of "<sos> h e l l o <eos>". This technique makes accent ID part of the ASR objective function and thus does not require any tuning of the trade-off between the ASR objective and the accent ID objective. The authors experiment with different positions for these special tokens and find that they work best at the end.

Multi-task learning approaches are also useful without accent labels in training data. Rao et al.[14] train a single DNN acoustic model with both phonemic and graphemic output targets, and then use the graphemic output targets for inference. The idea behind this technique is that the use of graphemic targets may allow the model to compensate for places where the lexicon does not accurately reflect the accents in the training or test data. At the same time, the authors recognize that graphemic output targets alone perform worse than phonemes, and thus want their model to be able to take advantage of the information contained in a pronunciation dictionary. Multi-task learning allows them to combine both sources of information in one model.

Ghorbani et al.[15] explore a special case of multi-task learning to improve recognition of non-native speech. The authors train a multi-lingual end-to-end model with separate output layers for English, Spanish, and Hindi. They then test the English output layer with English speech from native Spanish and Hindi speakers. They show that this model better represents the speech of these non-native English speakers than a model trained on English alone.

3.2. Domain Expansion

Domain expansion is a method for generalizing a model to other domains. In accented ASR, this method takes a baseline model that performs well on an initial set of accents and improves it's performance on a new accent while maintaining performance on the original set. Contrast this to fine tuning, where we train on a new accent to improve performance on that accent, but run into the problem of catastrophic forgetting where performance on the original accents decreases dramatically. Domain expansion is focused on techniques such as regularization and novel architectures to minimize this loss.

Regularization in this context adds a term to the loss that penalizes diverging from the baseline model. This can be done through KL-divergence[9] between the adapted model and the baseline model, elastic weight consolidation (EWC), and weight constraint adaptation (WCA). Ghorbani et al.[16] investigates several of these methods for one transfer step finding a hybrid of soft-KL-divergence and EWC to be the best, while Houston[17] looks at the performance over multiple transfer steps finding that EWC covers 65% of the gap between a fine-tuned and a pooled (best-case) model across all accents. An advantage of these regularization techniques is that the original data need not be present in order to adapt and improve the model. Architectures techniques such as Progressive Neural Networks[1] and Progress and Compress[18] have yet to be applied to this space.

Recently Winata et al.[19] applied model agnostic meta learning (MAML) to accented ASR, showing that a model could generalize to unseen accents. Meta learning, or learning to learn, has also been applied to train models to quickly adapt to new accents in zero or few shots, possibly even at inference[2]. However other methods, such as optimizers and different representations, have yet to be applied.

3.3. Pronunciation Modification

Many traditional ASR techniques require a pronunciation lexicon, but pronunciation dictionaries are constructed for particular canonical dialects (e.g. American English) which can lead to issues with recognizing accented speech[20]. Many pronunciation modification techniques involve learning or discovering a mapping between accented and canonical phones. These mappings may be purely handcrafted and based on domain knowledge[21], learned through training on accented and non-accented speech[22, 23], or learned given some handcrafted constraints[24, 25]. The phone mappings may be used to either augment the pronunciation lexicon[21, 23], applied during inference so that a canonical pronunciation lexicon may still be used[22, 25], or used as part of a modified, dialect-specific grapheme-to-phoneme model[23].

Pronunciation modification techniques are particularly useful for low-resource or no-resource data. [22] achieves lower WER for a majority of accented speakers using a phone mapping technique and no accented training data. African American Vernacular English (AAVE) may not necessarily be a low resource dialect, but with an off-the-shelf ASR model trained on broadcast news data, Lehr et al.[21] uses an augmented lexicon along with LM parameter estimation for a 2.1% absolute WER gain on AAVE. Humphries et al.[20] achieves 9.7% absolute WER gain on American English using pronunciation modification techniques with an acoustic model trained on British English. When more data is available, however, these techniques may be less useful compared to an accent-specific model - an acoustic model trained on American English achieves 19.3% absolute WER gain compared to the original British English acoustic model in [20].

4. Accent Features

4.1. Adaptation

4.1.1. Auxiliary Features

Inspired from successes in speaker adaptation research, a common approach to adapting ASR to accented speech is incorporating accent information into a single generic model. In this section we focus on approaches that augment the input acoustic features with accent information.

[1] arXiv:1606.04671

[2] arXiv:2004.05439

After testing out several complex interpolation-of-bases adaptation techniques, Grace et al.[26] found that simply adapting the bias of the first layer (mathematically equivalent to concatenating a one-hot encoding of the dialect to the feature vector) performed best and even outperformed dialect-specific models. The conclusion contradicts prior work but the authors claim that the architecture progression (from DNN to LSTM) combined with the large dataset of training audio nullify the need for complex adaptation techniques. Li et al.[13] and Zhu et al.[27] reached similar conclusions, finding improvement when feeding in a one-hot dialect vector into each component of an LAS architecture or a single hidden layer of a hybrid ASR acoustic model using a gate mechanism, respectively. [13] was also able to surpass dialect-specific model accuracy, while [27] did not, and saw equally-promising results using a multi-task network with a gate mechanism (a more practical approach without need for accent information a priori).

Yoo et al.[28] build upon this one-hot approach by further integrating dialect information using Feature-wise Linear Modulation (FiLM) to apply feature-wise affine transformations between intermediate LSTM layers conditioned on dialect. The authors find the best results when conditioning on the dialect and an utterance-level feature extractor built on previous layer outputs. [28] sets itself apart through measuring modeling unseen dialects by training an ”unknown” accent label from random samples of other dialects. [29] echoes a similar philosophy to FiLM, using the Mixture of Experts (MoE) approach to modulate acoustic model input with affine transformations of ”experts”, covering phonetic variability and accent variability in subsequent expert groups.

4.1.2. Accent Embeddings

In recent years, it has become common (inspired by x-vectors in speaker identification literature) to achieve adaptation by augmenting input features with accent embeddings, or vectors extracted from a late hidden layer of an accent identification neural network. Providing embeddings seems to improve both hybrid architecture acoustic models[12, 30] as well as end-to-end models[31, 32]. Furthermore, the accent embeddings can also be combined with semi-supervised lattice-free maximum mutual information (LF-MMI) fine-tuning[30] and multi-task training with accent identification[12, 32] to improve upon baseline modest gains from using the embeddings alone. Using accent embeddings combined with any of the other approaches cited here seems to be one of the most promising areas for accent adaptation research.

4.2. Adversarial Training

Adversarial networks are a popular approach to generative models in computer vision, but they are also useful in speech recognition to improve model robustness to noise[33] and other perturbations. Two papers have recently used adversarial training techniques to improve recognition of accented speech. Sun et al.[34] take inspiration from Domain Adversarial Training (DAT)[35] to learn domain-invariant features and minimize the difference between standard and accented speech. Chen et al.[36] devise AIPNet, a GAN architecture to separate accent-invariant information from audio before passing it along to a typical ASR network.

A DAT[34] architecture consists of three components: the feature generation network; the domain classification network, which discriminates between source and target domains during training; and the senone classification network, which predicts senones during training and inference. In order to learn domain-invariant features, the feature generation network should produce an output feature that makes the domain classification network fail to determine the correct domain, while also remaining discriminative enough for the senone classifier to identify the correct senone. The loss function for the DAT network contains a hyperparameter λ, which can be used to switch the operation of the network between DAT and multi-task learning. The authors report that multi-task learning shows inconsistent improvement compared to DAT, and suggest that this is due to the former incorporating domain-discriminative features. DAT has a relative Character Error Rate improvement of 3.8% on average compared to the baseline trained only on unaccented Mandarin.

AIPNet[36] consists of a pre-training stage and a fine-tuning stage. During pre-training, the authors use two generators: one accent-specific and one accent-invariant, along with their respective accent discriminators, plus a decoder that is trained to reconstruct the input audio. The authors utilize three loss functions for the pre-training stage: an adversarial loss to train the discriminators, a reconstruction loss to ensure acoustic information is encoded in both generators, and a consistency loss to encourage linguistic information to appear in the accent-invariant generator. Once pre-training is complete, the accent-invariant generator can be connected as an input to any downstream speech task. One major advantage of this technique is that pre-training only needs to be performed once, and can then be used as the base for any number of traditional ASR architectures.

In a supervised setting where transcriptions are available for all accents, AIPNet gives a relative WER improvement of 3.33% on accented speech compared to the baseline. When transcripts are available only for US accented speech, the relative improvement is 6.29%. Qualitative analysis of AIPNet shows that the accent-independent generator succeeds in disentangling accent content from linguistic content.

5. System Combination

Khandelwal et al.[37] propose a novel coupling of an open-source accent-tuned local model with a commercial ASR model trained on non-accented speech. The authors introduce the “FineMerge” algorithm to enable a commercial ASR system to guide inference of an accent-tuned ASR system at a character level. “FineMerge” aligns characters of the commercial transcript with the original audio frames using per-frame character confidence distributions calculated by the open-source decoder. The algorithm then adjusts the confidence of mismatching characters. The authors report that the combined system reduces the commercial system’s WER by 17-28% relative.

6. Open Problems and Challenges

Standard benchmark ASR research in accented speech lacks a standard benchmark, such as ImageNet for image classification. The lack of standardization makes it hard to track progress and compare findings across papers, even when similar techniques are used.

The Artie Bias dataset[41], based off of the Common Voice Corpus, is a good candidate for such a standard. Artie Bias is a manually verified subset of Common Voice with 17 English accents. However, the sparsity of Chinese, Middle Eastern, European, South and Central American, African, and Australian accents makes it difficult to use as a global benchmark. The dataset is dominated by American, Indian, and British English.

Datasets in Papers Surveyed

Paper	Accents	Total Hours	Source
[13, 26]	6	40000	Google
[25]	4	40000	Amazon
[14]	4	13600	Google
[27]	4	7000	IoA CAS
[36]	9	3800	Facebook
[9]	3	2000	Microsoft
[17]	4	1900	Amazon
[34]	6	960	NPU, china
[29]	8	736	Speech Ocean*
[28]	8	500	Speech Ocean*
[8]	6	450	News
[10]	6	400	Intel
[15, 16]	4	200	UT-CRSS-4
[37]	3	152	Common Voice*
[30]	5	125	Verbmobil/Voxforge*
[6]	14	70	ABI*
[7]	2	43	Google
[38]	2	37	ALS
[12, 31]	7	34	Common Voice*
[5, 39]	4	16	Microsoft

Table 1: *Overview of datasets used in accented speech benchmarking. Public datasets denoted by * include Phondat Verbmobil, Voxforge, Mozilla Common voice v4 (MCV-v4)[40], and Accents of the British Isles (ABI)*

Additionally, the test set only uses country labels and lacks finer regional dialect classifications.

Language modeling A majority of the works surveyed focus on pronunciation differences between dialects, when they can vary in grammar, vocabulary, and spelling as well [3]. End-to-end models may capture this implicitly, but systems that use language models for decoding could benefit from research into adapting them to accents.

Zero-shot accent robustness More research needs to be done on how ASR systems behave and can adapt to unseen accents. The difficulty of collecting accented English, sparsity of accented speech data, and dynamic nature of language is a major obstacle for general ASR systems in the wild. Effectively handling unseen accents could entail moving away from discrete accent categorization and towards more continuous representations of accent attributes. Accent-invariant representations presented in Section 4.2 are promising, though none of the papers in that section discussed unseen accents. Meta-learning may also present a solution for quickly adapting to unseen accents or learning accent-invariant adaptations.

Large numbers of accents The papers surveyed focus on relatively few accents compared to the entire range seen in the wild, as demonstrated in Table 1. Notably, the ABI dataset shows how a single national accent includes many regional dialects. It is unclear if techniques presented in current literature are able to generalize to the hundreds of dialects and subclassifications of accents worldwide.

Accents in other Languages English is the most studied language in ASR literature, with much sparser literature for other languages. It is unclear if accent-tuning approaches appropriate for English are applicable to other languages, especially in languages with greater dialect variation such as Hindi, Chinese, or Arabic.

7. Conclusions

Accent-robust ASR is increasingly important in practical applications. Significant challenges such as data sparsity and a lack of a standard benchmark impede research progress, and many open questions exist. The research here also shows great progress in the field and a general trend towards successful single-model generalization. We hope this survey serves as a primer for accent-robust ASR research and initiates future research.

8. Acknowledgements

We'd like to thank Petr Schwarz of the Brno University of Technology for his help reviewing this survey, and the rest of the team at Rev.com for their feedback on this paper.

9. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, 10 2016.

[2] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[3] J. Holmes, *An introduction to sociolinguistics*. Routledge, 2013.

[4] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[5] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[6] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," in *Fifteenth annual conference of the international speech communication association*, 2014.

[7] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data augmentation improves recognition of foreign accented speech." in *Interspeech*, 2018, pp. 2409–2413.

[8] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic speech recognition of multiple accented english data," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] 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.

[10] 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.

[11] A. Prasad and P. Jyothi, "How accents confound: Probing for accent information in end-to-end speech recognition systems," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3739–3753.

[12] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in *Interspeech*, 2018, pp. 2454–2458.

[13] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4749–4753.

[14] K. Rao and H. Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4815–4819.

[15] 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

[16] 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.

[17] B. Houston and K. Kirchhoff, "Continual learning for multi-dialect acoustic models," *Proc. Interspeech 2020*, pp. 576–580, 2020.

[18] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4528–4537.

[19] G. I. Winata, Z. Lin, and P. Fung, "Learning multilingual meta-embeddings for code-switching named entity recognition," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 2019, pp. 181–186.

[20] J. J. Humphries and P. C. Woodland, "The use of accent-specific pronunciation dictionaries in acoustic model training," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 1. IEEE, 1998, pp. 317–320.

[21] M. Lehr, K. Gorman, and I. Shafran, "Discriminative pronunciation modeling for dialectal speech recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[22] S. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," *Speech Communication*, vol. 42, no. 1, pp. 109–123, 2004.

[23] L. Loots and T. Niesler, "Automatic conversion between pronunciations of different english accents," *Speech Communication*, vol. 53, no. 1, pp. 75–84, 2011.

[24] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modelling for robust speech recognition," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 4. IEEE, 1996, pp. 2324–2327.

[25] H. Arsikere, A. Sapru, S. Garimella, and A. Learning, "Multi-dialect acoustic modeling using phone mapping and online i-vectors." in *INTERSPEECH*, 2019, pp. 2125–2129.

[26] M. Grace, M. Bastani, and E. Weinstein, "Occam's adaptation: A comparison of interpolation of bases adaptation methods for multi-dialect acoustic modeling with lstms," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 174–181.

[27] H. Zhu, L. Wang, P. Zhang, and Y. Yan, "Multi-Accent Adaptation Based on Gate Mechanism," in *Proc. Interspeech 2019*, 2019, pp. 744–748. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3155

[28] S. Yoo, I. Song, and Y. Bengio, "A highly adaptive acoustic model for accurate multi-dialect speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5716–5720.

[29] A. Jain, V. P. Singh, and S. P. Rath, "A multi-accent acoustic model using mixture of experts for speech recognition." in *INTERSPEECH*, 2019, pp. 779–783.

[30] M. T. Turan, E. Vincent, and D. Jouvet, "Achieving multi-accent asr via unsupervised acoustic model adaptation," in *INTERSPEECH 2020*, 2020.

[31] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *INTERSPEECH*, 2019, pp. 2140–2144.

[32] W. Rao, J. Zhang, and J. Wu, "Improved blstm rnn based accent speech recognition using multi-task learning and accent embeddings," in *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, 2020, pp. 1–6.

[33] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech 2016*, 2016, pp. 2369–2372. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-879

[34] S. Sun, C.-F. Yeh, M. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4854–4858, 2018.

[35] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, Jan. 2016.

[36] Y. Chen, Z. Yang, C. feng Yeh, M. Jain, and M. L. Seltzer, "Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979–6983, 2020.

[37] K. Khandelwal, P. Jyothi, A. Awasthi, and S. Sarawagi, "Black-Box Adaptation of ASR for Accented Speech," in *Proc. Interspeech 2020*, 2020, pp. 1281–1285. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3162

[38] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, and Y. Matias, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Proc. Interspeech 2019*, 2019, pp. 784–788. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1427

[39] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2, pp. 141–153, 2004.

[40] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[41] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6462–6468.