Proceedings Paper:
Hollands, S., Blackburn, D. and Christensen, H. (2022) Evaluating the performance of state-of-the-art ASR systems on non-native English using corpora with extensive language background variation. In: Interspeech 2022: Proceedings of the Annual Conference of the International Speech Communication Association. Interspeech 2022, 18-22 Sep 2022, Incheon, Korea. Interspeech Proceedings. International Speech Communication Association (ISCA), pp. 3958-3962. ISSN: 2308-457X. EISSN: 1990-9772.

https://doi.org/10.21437/interspeech.2022-10433

Interspeech 2022
18-22 September 2022, Incheon, Korea

Evaluating the Performance of State-of-the-Art ASR Systems on Non-Native English using Corpora with Extensive Language Background Variation

Samuel Hollands[1], *Daniel Blackburn*[2], *and Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, UK
[2]Sheffield Institute for Translational Neuroscience, University of Sheffield, UK
{shollands1, d.blackburn, heidi.christensen}@sheffield.ac.uk

Abstract

This investigation is an exploration into the performance of several different ASR systems in dealing with non-native English using corpora with extensive language background variation. This study takes two corpora amounting to 191 different native language (L1) backgrounds and looks at how these systems are able to process non-native English (L2) speech. A transformer based ASR system and a CRDNN architecture are both tested, trained on Librispeech [1] and Commonvoice [2] for a three way cross comparison. In addition Google's Speech-to-Text API and AWS Transcribe were investigated in order to evaluate popular mainstream approaches given their current degree of impact in deployed systems. Experiments reveal deficits in the range of 10%-15% mean WER performance difference between L1 and L2 speech. Results indicate ASR systems trained on particular varieties of L2 speech may be effective in improving WERs with outcomes in this paper demonstrating several Google ASR models trained on varieties of African L2 English outperforming L1 trained ASR for under-represented dialect groups in the United Kingdom. Further research is proposed to explore the plausibility of this approach and to critically approach WER as a metric for ASR evaluation, striving instead towards metrics with greater emphasis on evaluating language for communication.

Index Terms: non-native speech recognition, equality diversity and inclusion

1. Introduction

Automatic speech recognition (ASR) systems permeate many areas of human computer interaction. Poor ASR performance on certain language backgrounds has been attributed to serious issues in equality, diversity, and inclusion (EDI). Poor automatic transcription on YouTube for example has been shown to limit the outreach of non-native speakers due to the site's ranking algorithms for video visibility favouring transcribed content for accessibility [3].

L1 and L2 are terms referring to an individual's native spoken language and second language respectively. Whilst English is the most popular language in the world by number of speakers, approximately only 379 million of the 1.132 billion speakers globally are L1 native speakers. In the United Kingdom (UK) non-native speakers account for approximately 8% [4] of the population and in the United State of America approximately 20% of the population [5]. These figures from two of the largest L1 English speaking communities in the world demonstrate that there is a need for L2 language robustness in ASR not just in non English speaking countries, but also in countries where English is the majority L1 language.

At present the argument can be made that the largest exposure individuals will get in their day-to-day lives using an ASR system will be communication with some form of digital assistant. Currently it is estimated 22% of households in the UK own a digital assistant device; although the number of individuals owning devices that have inbuilt digital assistant technologies will be much higher and the number of people actually utilising these devices is going to lower this statistic somewhat. Digital assistants at present offer fairly prosaic abilities: choosing songs, setting timers, checking the weather, calling people, but seldom provide much more utility beyond this. Therefore, whilst these features can be seen to improve accessibility for some individuals, generally these offer more of an entertainment role or mild quality of life (QoL) improvement. However, it is vital to consider that state-of-the-art experimental approaches into medical diagnosis [6], air traffic control (ATC) [7] and other systems are dependent on highly robust ASR and inaccuracies can propagate with severe consequence that devices largely used for entertainment and QoL improvement do not create when they fail to function correctly.

State-of-the-art systems developed by Google and Microsoft have been able to achieve an English ASR word error rate (WER) of 4.9% [8] and 5.1% [9] respectively; favourably comparable to human transcription WERs which are about 4%. However, whilst these are impressive results the leap to discussion of having achieved human parity [9] without evaluating on a non-native English corpus is unfavourable as particularly in the context of English, L1 speech represents a minority of all English speakers. This paper offers an evaluation into Google Speech to Text, AWS Transcribe, and transformer and CRDNN CTC/Attention ASR models trained on Commonvoice [2] and Librispeech [1]. The key aim is to explore over 190 different language backgrounds; evaluating and emphasising the need for more L2 robustness in ASR whilst actively contributing to the literature by highlighting which clusters of languages are particularly underperformant; helping to optimise research focuses to tackle these most substantial limitations.

2. Literature Overview

Non-native speech introduces a wealth of challenges for ASR development. Whilst L1 and L2 speech are often discussed as classes of data, it is imperative to recognise that neither is a homogeneous dataset. Mispronounced segments [10], longer pause duration [11], abnormal pause location within clauses [12], and non-reduction of function words [13] are some key features of L2 speech hindering ASR performance [14]. However the reality is a far more extensive list of issues, naturally the result of the heterogeneity of L2 speech [14].

Whilst mainstream applications of ASR rest largely in the entertainment or QoL domains, state-of-the-art experimental technologies utilising ASR reach into far more critical applications such as ATC systems [7] and medical diagnostic tools

10.21437/Interspeech.2022-10433

[6]. Taking the scope of dementia diagnosis, a key focus of our wider research, enormous strides have been made in developing modular classification systems that utilise ASR as a component of a diagnostic pipeline which can be used either in a clinical setting [15] or a home environment [16] as a long-term monitoring tool. Linguistic features have been found to be highly successful for the detection of dementia [17, 18] and indeed most classifiers use linguistic information as an feature set for cognitive impairment detection, some systems exclusively. Whilst early work exploring the potential for using machine learning to detect dementia through language analysis used human transcriptions of speech, scalable automatic approaches which are the inevitable successor to these early experiments demand the automation of the transcription process, leaning firmly on ASR technologies. It is important to note current diagnostic approaches for diagnosing dementia are not perfect, indeed there is strong evidence that existing metrics are problematic from the context of linguistic diversity [19]. Inaccurate ASR in this context will only seek to increase these levels of disparity and worsen EDI outcomes for diagnosis. In the cases of both ATC, medical diagnosis, and many other critical domains it becomes clear to see how under performing ASR as the first module in a downstream system, that is wholly or mostly dependent on accurate transcription, has a risk of either creating highly undesirable outcomes or has to be targeted at a sub category of users who will receive accurate results; challenges for EDI we need to overcome.

ASR performance on L1 speakers has frequently been reported to have surpassed the accuracies of human perception [20, 21, 22]. Studies into the impacts of L2 speech thus far typically focus on a single language [22, 23] as the oculus for investigation. Studies demonstrate around a 20% difference in performance between L1 and L2 English on ASR [22]. Although it is important to note that L2 performance can be substantially better, or substantially worse than this. Existing solutions to non-native ASR have yielded positive results. Methods have been attempted have ranged from transfer learning [24, 25] and language background specific corpora [26] to automatic manipulations of the speech signal and level so-called 'accented' speech to more similarly reflect the ASR training data [27].

However, we were unable to find any work evaluating ASR on highly language background diverse corpora. Current studies developed to tackle poor L2 ASR performance tend to focus on at most a small handful of L2 language backgrounds, typically highly related and generally geographically similar in location. This provides excellent highly detailed insight and innovation for the languages covered, often providing substantial improvements for these language backgrounds but at the cost of broader scope. One reason for the lack of L2 background diversity studies is data scarcity. Corpora for L2 English are often language background specific which naturally encourages and serves this type of research well. However several databases do exist to allow for an investigation into a greater variety language backgrounds. For instance the Speech Accent Archive [28], CSLU Foreign Accented English Release 1.2 [29], as well as other corpora such as the International Dialects of English Archive (IDEA) [30], and even the Commonvoice [2] English corpus which provides an impressive variety of L2 language backgrounds. These databases are however imperfect. A major issue is the absence of migration information which is vital for providing the necessary metadata to greater predict the dialect and accent of an individual which could be used to optimise model selection. Similarly an individual who lived in Kenya for 6 months, moved to London for 40 years, and then returned would likely skew WER results due to a high probability of dialect and accent assimilation to some degree. However this can be countered by using metadata at some level by exploring years of English language exposure which can help to provide a low resolution gauge for proficiency. It would benefit the domain of L2 ASR evaluation substantially for a corpus to be developed with the background diversity of the Speech Accent Archive [28] but migration information and spontaneous speech included.

This paper will evaluate the performance of state-of-the-art and mainstream approaches to ASR highlighting existing strengths and weaknesses of current approaches and revealing several routes forward for investigative research to improve L2 robustness. Section 3 provides a methodology overview of the models developed and used. Section 4 explores the two corpora used in this investigation, focusing on a breakdown of the scope and demographics of each corpus. Moreover, Section 5 breaks down our results, uncovering some key findings not just in L2 evaluation and robustness, but also potentially avenues for improving ASR accuracy on under-represented dialects. Section 6 proceeds onto a discussion of the results, highlighting the enormous potential in investigating ASR performance on L2 English trained on different language backgrounds to potentially indicate languages that can produce better performing solutions for L2 speech than L1 trained systems. Finally Section 7 concludes by laying out future research goals and indicating a strong need to move beyond WER as a evaluation metrics for ASR.

3. Methodology

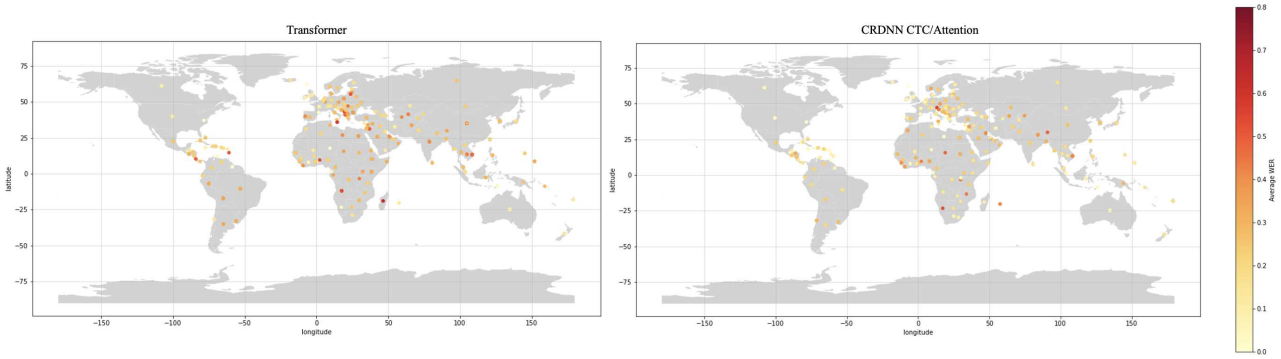
A total of four ASR models were developed. A transformer and CRDNN based model was used for each of the Commonvoice [2] and Librispeech [1] corpora. The justification for this that Librispeech [1] is a corpus of read speech, which may result in a potential bias in performance when evaluated against the Speech Accent Archive [28] in particular where there are key speech characteristics present in read speech that are not present in extemporaneous and spontaneous discourse. Both Librispeech [1] models were imported from the SpeechBrain [31] Hugging Face library of pretrained models. The Commonvoice [2] systems were developed using modified prebuilt recipes in the SpeechBrain [31] Python package using an Nvidia Tesla V100. In addition, AWS and Google speech-to-text solutions were investigated. As the aim of this paper is to determine not just the effectiveness of experimental ASR solutions but also solutions that are currently widely deployed and utilised as these systems will act as modular components to a large amount of software depending on speech-to-text technology. Google Speech-to-text contains a very impressive 16 language backgrounds for English; 10 of which are from countries where English is not the primary language spoken. Conversely, AWS provides only 3 English backgrounds to choose from, all of which are from native English speaking countries. All models were evaluated based on their WER performance on the Speech Accent Archive [28] (Section 4.1) and CSLU Foreign Accented English Release 1.2 [29] (Section 4.2).

4. Data

4.1. Speech Accent Archive

The Speech Accent Archive [28] is a corpus of L1 and L2 English speakers reading out the same utterance. It provides key metadata such as L1 language, place of birth, place of residence,

Figure 1: *Average WER of Librispeech [1] Transformer and CRDNN CTC/Attention ASR systems plotted based on country of origin*



and the amount of time (in years) an individual has been exposed to and actively learning English. 25% of speakers in this corpus are L1 English speakers providing a rather substantial dataset of just over 1,500 L2 speakers with 192 different L1 backgrounds. The population sex balance is 48.3% female to 51.7% male, an extremely well balanced dataset given the size. However, it should be noted that there are cross cultural difficulties in comparing these two figures: men were present in 57 more languages and are over represented in African languages whilst women are over represented in European languages. It is worth noting the metadata does not include migration information which would be extremely useful for trying to study the effectiveness of ASR systems on second generation plus immigrants.

4.2. CSLU: Foreign Accented English Release 1.2

The Foreign Accented English Release 1.2 [29] is a corpus of spontaneous phone-call L2 speech. It is worth noting enhanced models used in the Google Speech-to-Text segment of this study are documented [32] to have been trained and tested to work well on phone call speech. An important aim of this study was to both train and test ASR systems on spontaneous speech as well as read speech. This allows for a three way comparison on the two corpora looking at the performance differences between Librispeech [1] and Commonvoice [2] which are read and spontaneous data respectively (see Section 3). The Foreign Accented English Release 1.2 [29] has fewer language backgrounds than the Speech Accent Archive [28] with only 22 L2 language backgrounds as opposed to 191. Transcriptions are not included in this corpus therefore an open-sourced series of transcriptions was used [33] which were evaluated for quality and deemed to be accurate.

5. Results

Whilst the development of CRDNN based ASR systems was not motivated by L2 speech performance, it performs substantially better than the previously dominant transformer based approach in all of the experiments conducted, although interestingly not unanimously for every language background. Figure 1 is a visualisation of WER performance averages for each country of birth for each individual in the Speech Accent Archive [28] corpus, this allows us to see not just ASR performance on a large scale, but also to see clusters of under performing areas that may benefit from additional L2 ASR research. This visualisation demonstrates excellent improvements made globally comparing

Table 1: *Performance of Librispeech and Commonvoice based Transformer and CRDNN CTC/Attention based ASR systems alongside AWS Transcribe and Google Speech-to-Text (Both using American English model). WER performance evaluated on Speech Accent Archive [28] (Acc-L#) and Foreign Accented English Release 1.2 (FAE-L#)*

	Acc-L1	Acc-L2	FAE-L2
Lib-Trans	11.23%	29.55%	31.60%
Lib-CRDNN	**5.48%**	21.57%	24.44%
Com-Trans	11.86%	20.13%	29.46%
Com-CRDNN	6.01%	20.39%	**22.68%**
AWS Transcribe	10.19%	**19.72%**	23.90%
Google STT	11.34%.	24.62%	24.91%

Transformers to CRDNN CTC/Attention architecture, particularly in Africa and South America. Whilst no country speaking L1 English saw an improvement greater than 3%, countries such as Portugal and Brazil saw improvements nearing 20% in absolute terms scaling down from WERs between 35-45% to between 15-25%. There are notable performance drops in countries such as Finnish which performed 7% worse. However, on average the vast majority of language backgrounds of L2 speakers saw substantial improvements. Figure 1 also demonstrates rather starkly the reality of L1 performance in comparison to L2 globally. Despite English as L1 representing only 25% of the Speech Accent Archive [28] corpus it is responsible for 85% of all 0% WERs and is the 6th most accurately transcribed language despite containing over 500 speakers, all of the other 5 languages have 1-3 speakers. No other language with over 20 speakers came in the top 20.

Performance on eastern European languages is poor compared to the rest of Europe as well as languages within the Indian subcontinent, China, and large swathes of sub-Saharan Africa. A particularly interesting aspect of this find is that both Chinese and many Indian languages would be considered fairly high-resource with many corpora available in L2 English for system development and training. Therefore it is interesting that both language backgrounds perform poorly in a one-size-fits-all ASR system given the amount of data available for evaluation and training. This highlights an initial key issue. When an ASR system is developed and evaluated often it will be tested using a testing partition of the same corpus. For a corpus with a lack of diverse data the system is going to perform better; rewarding homogeneous datasets. If a system is evaluated and advertised

based on its performance on global English, metrics need to be demonstrated for global English, not simply optimal performance for a handful of English dialects and accents. From an EDI perspective an individual deploying a modular ASR system without intricate knowledge of the technology should be provided with information on how this system is going to perform across a wide range of individual backgrounds.

Sex was not found to meaningfully correlate with ASR performance on native speech. Differences of $<2\%$ WER were found in all L1 experiments with the female class often out performing the male class. A 5% WER difference was discovered between women and men for the L2 speech, with women out performing men. Despite being over-represented in WER accuracy women were under-represented in language diversity, the asymmetric distribution within European countries explains the slight differences in WER with western European countries being the highest performing cluster of L2 English varieties.

5.1. Clustering L2 ASR Systems

Google Speech-to-Text, unlike AWS, provides non-native English ASR systems. Whilst the 3 systems provided by AWS cover only native varieties of English, 10/16 of Google's ASR systems are built and trained for L2 variations of English. However it is entirely possible both AWS and Google possess in-house ASR solutions that exceed the performance and diversity of these publicly available APIs. However, as the scope of the study is to investigate contemporary utilisation of these systems, evaluation of these publicly available APIs remains important for discussing ASR robustness on L2 speech for the large number of systems that will utilise these APIs. Given the substantial linguistic differences between the languages chosen this provides an optimal opportunity to compare each model's performance globally. The reason for doing such is to tackle issues both of data scarcity, and of efficiency.

Out of 191 languages in the Speech Accent Archive, the American English ASR system (en-US) was either the best performing or joint best performing model for 43 language backgrounds. Moreover, 94 language backgrounds saw less than a 3% absolute reduction in WER through using a model that was more accurate than en-US. Conversely, 61 languages saw an absolute increase in performance of at least 10% WER with 25 languages seeing an absolute increase of greater than 20% WER. The Ghanaian ASR system however dominates L2 performance achieving lowest or joint lowest WER in 122 language backgrounds. The top five performing ASR systems by number of top performing languages are Ghanaian (122), South African (120), Kenyan (119), Nigerian (116), and Tanzanian (112). The sixth top performing is an L1 Canadian English ASR system with top performance in 45 languages. Whilst L1 model enormously under performs compared to the top five systems, there are several ASR models which performed poorly across the whole data with less than five top or joint top performances including Hong Kong, Philippines, India, and Pakistan. There were a few instances of models performing poorly for countries they were built for including the Indian system scoring a 44% WER in India, and the United Kingdom scoring 24% compared to Nigerian, Tanzanian, and South African scoring 18% WER. After manually inspecting the data it becomes clear in the case of the United Kingdom that the Nigerian, Tanzanian, and South African ASR systems are generally outperforming the native L1 system on northern, Scottish, and Welsh accents. This could suggest an inherent bias in the data used to train the United Kingdom model.

6. Discussion

If an adaptive system were to be used that selected the appropriate ASR model for each language background, performance across all L2 backgrounds based off of the Speech Accent Archive we could see an average L2 ASR performance of 17.8% down from 24.62% or 21.17% down from 24.91% from the Foreign Accented English Release. Simply using a Ghanaian ASR system still yields an average WER of 19.57% in L2 speech although naturally this will change depending on the distribution of the data. Evaluating each system on such a broad range of language backgrounds has uncovered the potential to use L2 ASR systems trained on other languages as an alternative to depending on generic L1 trained models. Similarly there may be a benefit for under-represented dialects even in L1 English countries to take advantage of L2 systems that may allow for better accuracy. Developing several L2 ASR systems utilising these L2 language backgrounds and then creating informed methods for choosing the most appropriate system would help both to reduce the environment impact of building an excessive number of language background specific ASR systems, whilst also providing a more viable alternative to the anglocentric L1 exclusive models that exist within a lot of modern industry scale speech technology APIs. Whilst such a solution is unlikely to beat out language specific innovations in the long term, it could help alleviate the disparities for accessing speech technologies in the short to mid term, especially for under-represent groups.

7. Conclusions

This investigation has demonstrated the disparities between L1 and L2 WER performance on global English looking at 191 different language backgrounds. Experiments have demonstrated that ASR systems trained on certain L2 language backgrounds may have the ability to act as wide scale universal alternatives to L1 systems for automatically transcribing L2 speech. A middle alternative between one-size-fits-all and language specific ASR allows for the greater efficiency of not building many thousands of ASR systems combined with lower WERs over single model solutions.

Further research should aim to find key languages that would be most effective in building generalisable L2 ASR systems for reducing disparities between L2 and L1 performance. In addition the use of WER within this study is problematic as it has been repeatedly demonstrated that it fails to accurately correlate with levels of human intelligibility [34]. A large contributor to this is that WER is devoid of any evaluation related to semantics, pragmatics, grammar, and really every fundamental functional aspect of language aside from orthography. Future studies should consider breaking the long established WER tradition in ASR evaluation for evaluative measures that better reward the objective of the ASR system. This could be a semantic approach [35] for human robot interaction for the purposes of information transmission, or in the context of a downstream classifier a metric used to maximise the accuracy of the most salient features analysed within the pipeline.

8. Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

9. References

[1] 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.

[2] 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," *arXiv preprint arXiv:1912.06670*, 2019.

[3] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[4] P. Stokes, "Census: Detailed analysis-English language proficiency in England and Wales: Main language and general health characteristics," *Mode of access: https://www. ons. gov. uk/peoplepopulationandcommunity/culturalidentity/language/articles/detailedanalysisenglishlanguageproficiencyinenglandandwales/2013-08-30)*, 2011.

[5] C. Ryan, "Language use in the united states: 2011," Dec 2021. [Online]. Available: https://www.census.gov/library/publications/2013/acs/acs-22.html

[6] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[7] Z.-G. Juan, P. Motlicek, Q. Zhan, R. Braun, and K. Vesely, "Automatic speech recognition benchmark for air-traffic communications," ISCA, Tech. Rep., 2020.

[8] E. Protalinski, "Google's speech recognition technology now has a 4.9% word error rate," May 2017. [Online]. Available: https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/

[9] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[10] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.

[11] J. Anderson-Hsieh and H. Venkatagiri, "Syllable duration and pausing in the speech of Chinese ESL speakers," *TESOL quarterly*, vol. 28, no. 4, pp. 807–812, 1994.

[12] O. Kang, D. Rubin, and L. Pickering, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English," *The Modern Language Journal*, vol. 94, no. 4, pp. 554–566, 2010.

[13] T.-Y. Jang, "Automatic assessment of non-native prosody using rhythm metrics: Focusing on korean speakers' english pronunciation," in *Proc. of the 2nd International Conference on East Asian Linguistics*, 2009.

[14] S. Park and J. Culnan, "A comparison between native and non-native speech for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1827–1827, 2019.

[15] R. P. D. O'Malley, B. Mirheidari, K. Harkness, M. Reuber, A. Venneri, T. Walker, H. Christensen, and D. Blackburn, "Fully automated cognitive screening tool based on assessment of speech and language," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 92, no. 1, pp. 12–15, 2021.

[16] X. Liang, J. A. Batsis, Y. Zhu, T. M. Driesse, R. M. Roth, D. Kotz, and B. MacWhinney, "Evaluating voice-assistant commands for dementia detection," *Computer Speech & Language*, vol. 72, p. 101297, 2022.

[17] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[18] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech," in *Proc. Interspeech*, 2021, pp. 3810–3814.

[19] M. Goudsmit, J. van Campen, T. Schilt, C. Hinnen, S. Franzen, and B. Schmand, "One size does not fit all: comparative diagnostic accuracy of the Rowland universal dementia assessment scale and the mini mental state examination in a memory clinic population with very low education," *Dementia and Geriatric Cognitive Disorders Extra*, vol. 8, no. 2, pp. 290–305, 2018.

[20] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[21] E. Edwards, W. Salloum, G. P. Finley, J. Fone, G. Cardiff, M. Miller, and D. Suendermann-Oeft, "Medical speech recognition: reaching parity with humans," in *International Conference on Speech and Computer*. Springer, 2017, pp. 512–524.

[22] R. Cumbal, B. Moell, J. Lopes, and O. Engwall, ""you don't understand me!": Comparing ASR results for L1 and L2 speakers of Swedish," *Proc. Interspeech 2021*, pp. 4463–4467, 2021.

[23] Y. Wang, H. Luan, J. Yuan, B. Wang, and H. Lin, "LAIX corpus of chinese learner english: Towards a benchmark for L2 english ASR." in *INTERSPEECH*, 2020, pp. 414–418.

[24] P. Sullivan, T. Shibano, and M. Abdul-Mageed, "Improving automatic speech recognition for non-native English with transfer learning and language model decoding," *arXiv preprint arXiv:2202.05209*, 2022.

[25] T. Shibano, X. Zhang, M. T. Li, H. Cho, P. Sullivan, and M. Abdul-Mageed, "Speech technology for everyone: Automatic speech recognition for non-native english with transfer learning," *arXiv preprint arXiv:2110.00678*, 2021.

[26] K. Kulkarni, S. Sengupta, V. Ramasubramanian, J. G. Bauer, and G. Stemmer, "Accented Indian English ASR: some early results," in *2008 IEEE Spoken Language Technology Workshop*. IEEE, 2008, pp. 225–228.

[27] 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.

[28] S. Weinberger, "Speech accent archive. George Mason University," *Online:¡ http://accent. gmu. edu*, 2015.

[29] T. Lander, *Foreign Accent English Release 1.2*. Linguistic Data Consortium, 2007.

[30] P. Meier and D. Paul, "International dialects of English archive," *IDEA-The International Dialects of English Archive*, 1997.

[31] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[32] "Recognize speech by using enhanced models." [Online]. Available: https://cloud.google.com/speech-to-text/docs/enhanced-models

[33] Wellesleynlp, "Wellesleynlp/emilythesis." [Online]. Available: https://github.com/wellesleynlp/emilythesis

[34] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz *et al.*, "Automatic human utility evaluation of ASR systems: Does WER really predict performance?" in *INTERSPEECH*, 2013, pp. 3463–3467.

[35] S. Roy, "Semantic-wer: A unified metric for the evaluation of ASR transcript for end usability," *arXiv preprint arXiv:2106.02016*, 2021.