Application of ASR to a Sociolinguistic Corpus of Australian English

Maya Weiss[1], Ksenia Gnevsheva[1], Catherine Travis[1], Gerard Docherty[2]

[1] Australian National University, [2] Griffith University

maya.weiss@anu.edu.au, ksenia.gnevsheva@anu.edu.au, catherine.travis@anu.edu.au, gerry.docherty@griffith.edu.au

Abstract

This study applies Automatic Speech Recognition (ASR) to a sociolinguistic corpus of Australian English. We compare a human transcription of excerpts from 20 urban and regional speakers with a transcription generated by Microsoft's Azure AI Speech. The Word Error Rate is comparable to previous studies, and is not impacted by the sociolinguistic variables of speaker region and gender, nor the phonetic variable of vowel formants. Despite the overall low rate of transcription errors, our findings suggest that the quality of certain vowel categories that are particularly characteristic of Australian English can impact on the accuracy of the ASR-generated transcription.

Index Terms: Automatic Speech Recognition, corpus-building, transcription, sociolinguistics, Australian English

1. Introduction

Corpus building is central to much linguistic work, not least for sociolinguistic studies which are routinely centred around large samples of spontaneous speech. The creation of such corpora, however, is a notoriously slow process, due to the time required for both collecting speech data and producing usable transcriptions. Recent advances in technology have revolutionised transcription with the incorporation of ASR into the linguistics toolkit facilitating the processing of hours of speech in just minutes. There has been some work supporting the usability of ASR for sociophonetic analysis of American English [1], but there has been much less done with Australian English.

While the proprietary nature of ASR systems means that the nature of the material on which they are trained is opaque, a recurrent question in recent years has been how well ASR systems deal with cross-accent variation [2-5]. In the case of Australian English, with its well-documented differences from other varieties [6, 7] and its relatively small population of speakers, it is not a given that ASR-generated transcription would work as effectively as it does for other varieties (e.g. from the USA) which are likely to be more highly represented in training materials. For example, vowels with a realisation that is particularly distinctive for Australian English speakers may be prone to transcription errors, and the effectiveness of automatic transcription may be impacted by factors such as region and gender which are known to correlate with accent variability in Australian English varieties.

In this paper, we test Microsoft's Azure AI Speech [8] on a substantial sample of spontaneously spoken Australian English. We use the measure of Word Error Rate to assess accuracy, and we consider this for the data overall, across genders, and across speakers from an urban vs. regional setting. We further consider the impact of the acoustic signal, to compare accuracy across different vowel categories, including those that are more characteristically Australian and those that are not.

2. Automatic Speech Recognition

Current ASR models use combinations of deep learning systems. Typical components include signal processing and feature extraction, acoustic and language models, and a hypothesis search [9]. Together, these convert a speech signal into a frequency-domain representation; extract and match salient features with acoustic and phonetic knowledge; make grammatical and lexical-semantic predictions; and estimate the probabilities of hypothesised word sequences, outputting a word sequence with the highest probability. One salient feature of ASR systems is that they do not rely on segment-by-segment analysis of the acoustic phonetic parameters conventionally used in phonetic analysis, such as formant estimates.

A widely used method for testing ASR accuracy is *Word Error Rate* (WER), which is the sum of three types of errors (substitution, deletion, and insertion) divided by the total number of words in the corrected transcript [3]. Accuracy is impacted by the ASR model used, and the data fed into the system, including audio quality [10]. For example, OpenAI's Whisper achieved an average 12.8% WER across multiple English datasets, with the lowest WER (2.7%) being achieved on read speech [11]. Microsoft's 2017 system achieved an overall 5.1% WER with data from telephone conversations [12]. This is indicative of the fact that ASR models are better at transcribing what can be considered 'clean' data, reinforced by Meta (then Facebook) AI's wav2vec 2.0 model achieving a 1.8% WER on test-clean speech in 2020 [13].

ASR performance differs across varieties of a language. For example, Whisper ASR performed better with American and Canadian English over Australian and British English [14], and YouTube's automatically generated captioning performed better with American than Scottish English [4]. Within American English, several ASR systems were found to be more accurate for the speech of white Americans than of African Americans [3], and of white Americans than Native Americans, African Americans and ChicanX [5]. Similarly, ASR performed better with Standard Dutch than Southern Dutch Dialects [15]. Results for gender vary, with better performance observed for men [4], for women [14], and for neither [16]. Sociophonetic variation seems to play a role in accuracy, and one study of ASR for speakers of the Pacific North-West (USA) pointed to sociolinguistic features of that variety accounting for 20% of word errors [5].

One of the factors impacting ASR performance is the imbalanced nature of training data sets, which are believed to over-represent men and university-educated speakers [4] and to under-represent minoritised speakers [5]. It has been demonstrated that accuracy improves with dialect-specific training sets [17], and with larger training sets [11]. The fact that "stratified sampling of speakers has historically not been the priority during corpus construction for computational applications" [4] is thus potentially problematic for ASR systems, including in their use for sociolinguistic analysis.

3. Method

3.1. Data

The data in this study come from two corpora of Australian English, one recorded in the large metropolis of Sydney (pop. 5.2 million) [18], and the other in a regional district around the town of Braidwood, New South Wales (pop. 1720; 300km from Sydney and 100km from Canberra) [19]. The recordings are from casual interviews – sociolinguistic interviews for the urban data and oral history interviews for the regional data. Participants include five men and five women from each region, who are generally comparable in other social characteristics: they are of Anglo-Celtic background, aged between 35 and 65, and have a range of occupations (from lawyers to bar workers).

The analysis is based on 20,000 words, 1,000 words from each speaker (as counted in the manually corrected transcripts). We endeavoured to extract continuous, uninterrupted stretches of speech as much as possible, though due to the interactive nature of the recordings, it was necessary to extract multiple strings to reach 1,000 words. The sociolinguistic interviews are more conversational than the oral histories, thus for the former, the 1,000 words came from an average of 7.2 strings, and for the latter, from an average of 4 strings.

3.2. ASR Process

For this study, we used the ASR operated via Microsoft Azure AI Speech. The most recent detailed description of its architecture is from 2017 [12]. This does not specify the training data, although corpora such as Switchboard [20], LibriSpeech [21], and GigaSpeech Middle [22] are likely to be included [12, 23]. There is an option to set the language variety to Australian English (among others), implying that there is some Australia-specific information in the system.

Audio files were transcribed by Azure AI Speech via a dedicated speech-to-text transcription platform developed at Griffith University using Microsoft technologies and run locally. The language was set to Australian English, and the number of speakers was specified (two for all files but one with three speakers, including one interviewer, whose speech is not analysed). The recordings were processed, outputting .txt and .json transcription files, as well as a numerical rating (on a 0-1 scale) of the confidence with which the ASR had reached the corresponding output for each individual word and each phrase.

3.3. Data Coding

The ASR output was manually corrected, and these corrected transcripts were then compared word-by-word with the original ASR output to identify errors. Errors were coded for type (substitution, deletion, insertion) and linguistic element affected (vowel, consonant, filler, lexical, or other; see Table 1). We excluded proper nouns that were specific to the recording sites (e.g. *Araluen, Mongarlowe*).

Example (1) provides an illustration of ASR output for a segment of text and (2) the corresponding corrected transcription. Substitutions are underlined (e.g. *points, pawns* and *palms* for *pines*) and deletions are bolded (e.g. *all,* repetition of *the*, *um*). The overall WER in this sample is 17% (9 errors divided by 52 words in the corrected transcription). A third type of error not illustrated here is insertion, where the ASR output contains an addition to what was produced. Multiple errors are sometimes coded for the one word (e.g. ASR *go on* for *gone* has an insertion and a substitution; ASR *annoying* for *and I went* has two deletions and a substitution).

(1) ASR output
*So we had points result **Ø** here that **Ø** the house, we had two hectares all up for pawns. So right here at the house, there's some right up against the house and then the hill **Ø** just at the back of the paddock there, that was full of palm trees as well.*

(2) Corrected transcription
*So we had pine trees **all** here at **the** the house, we had two hectares all up of pines. So right here at the house, there's some right up against the house and then the hill **um** just at the back of the paddock there, that was full of pine trees as well.*

To assess the effect of vowel quality on transcription accuracy, the corrected transcripts were force-aligned to the corresponding speech signal (with LaBB-CAT [24]). F1 and F2 values for the primary/stressed vowel in the target word were estimated (with Praat [25]), at 20% into a diphthong interval and 50% for monophthongs. The F1/F2 estimates were subsequently mapped to vowels transcribed correctly and incorrectly (coded as 'vowel errors'). A total of 22,580 vowel tokens were processed by LaBB-CAT, including 360 vowel error tokens, of which 299 are analysed below (17% being set aside, due to processing issues in Praat or LaBB-CAT).

4. Results

4.1. Error Types and Word Error Rate

As a first measure of performance, we consider the Word Error Rate, and types of errors. Table 1 give the numbers of words impacted for both error types and affected linguistic elements, along with examples of each. SUB+ represents instances where multiple errors are involved (all of which involve a substitution and one or two other errors). As can be seen, substitutions and deletions are equally frequent, while insertions are relatively rare. The most common type of deletion is fillers such as *um, uh*. Though the ASR system used here did not categorically delete all fillers, many ASR models intentionally delete these [26], and thus it may be misleading to consider them errors. If we set fillers aside, then substitutions are overwhelmingly the most common type of error, of which vowel substitutions are particularly frequent.

The total number of errors is 1,420, giving an overall WER of 7%. Individual speakers exhibit a wide range, from 1.6% to 14.7%, with a per speaker mean of 7.2%. An Interquartile Range (IQR) calculation reveals no outliers. Excluding fillers, the overall WER drops to 5.3%, and the range for individual speakers drops to 0.6% to 9.9%, with a mean of 5.3% (again, no outliers, according to an IQR calculation).

Table 1: *Error types and affected linguistic elements.*

Error Type	Linguistic Element	Target word	ASR output	N
SUB	Vowel	*known*	*nine*	269
(N=520)	Consonant	*embers*	*members*	216
	Other	*threw*	*through*	35
DEL	Filler	*um*	*Ø*	372
(N=570)	Lexical	*then*	*Ø*	198
INS	Lexical	*Ø*	*it's*	8
(N=26)	Other	*rem~ ('remember')*	*mean*	18
SUB+	Vowel	*a firey*	*Afari*	91
(N=304)	Other	*Australians*	*this train*	213

4.2. Word Confidence Scores

The confidence with which the ASR has reached a corresponding output is potentially a useful tool for researchers – if the incorrectly transcribed words are given a lower confidence rating than the correctly transcribed words, these could help to locate transcription errors. Table 2 provides these confidence scores and shows that correctly transcribed words do have a higher average score than incorrect words (0.808 vs. 0.595). The range, however, is similar across the two: for correct words, from 0.004 to 0.998, and for incorrect words from 0.012 to 0.984. Thus, in some cases, ASR outputs (whether correct or incorrect) are returned with low levels of confidence, and some instances of incorrect outputs are returned with high levels of confidence.

Table 2. *Confidence scores of correctly and incorrectly transcribed words.*

Error	Average Conf.	StdDev	Min. Conf.	Max. Conf.
Correct	0.808	0.179	0.004	0.998
Incorrect	0.595	0.245	0.012	0.984
Grand Total	0.798	0.189	0.004	0.998

Note: Confidence values come from uncorrected transcripts (total N words = 20,371; N correct = 19,718, N incorrect = 653.

4.3. Social Effects: Region and Gender

We found no significant difference in WERs for the two social factors considered. The WER for the urban speakers is slightly higher than for the regional speakers (6.1% vs. 4.5%; urban Mean = 6.1%, SD = 0.030; regional Mean = 4.5%, SD = 0.022; fillers excluded), but this difference is not significant (based on a two-sample t-test performed in Python, p = 0.200). The regional IQR is 3.8%, and the urban IQR is 4.9%, signifying a somewhat higher level of variation within urban speakers.

For gender, though men have a higher overall WER than women (6.7 vs. 4.4%; Men Mean = 6.2%, SD = 0.022; Women Mean = 4.4%, SD = 0.029; fillers excluded), this difference is not significant (p = 0.122). The female IQR is 4.5%, and the male IQR is 4.3%, thus little difference between the groups.

4.4. Acoustic Effects

We now turn to consider what impact vowel quality has on the ASR performance, for which we focus on the 10 most frequent vowel categories to occur in the data.

We first compare the acoustic properties of the correctly vs. incorrectly transcribed vowels, shown in Figure 1 with diphthongs in the top panel and monophthongs in the bottom panel (8 points with F1>1,100Hz were removed to aid with figure readability). The ellipses represent the correctly transcribed instances, restricted to the central 1SD, and the individual points of the same colour the incorrectly transcribed instances of the same vowel categories. While the incorrect vowels appear to be widely distributed, the majority fall within the ellipse of their respective category. For example, for PRICE 60% of the incorrectly transcribed tokens fall inside the ellipse (comparable to the 67% of correctly transcribed tokens which fall within 1 SD). Furthermore, acoustic characteristics of vowels that were incorrectly transcribed are not necessarily predictive of what they would be mistranscribed as. For example, the incorrectly transcribed instance of a PRICE vowel with the lowest F1 value (the top-most purple point in the diphthong chart, in the word *I*) was incorrectly transcribed as

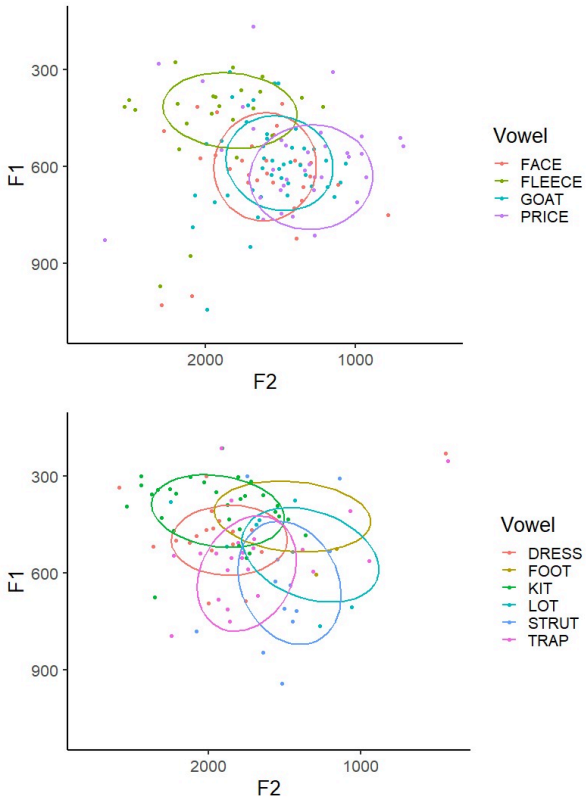

Figure 1. *Diphthongs and monophthongs correctly (ellipses) and incorrectly (points) transcribed by ASR.*

TRAP (*and*), whereas one would expect vowels with relatively high F1 and F2 values to be transcribed as TRAP. The two incorrectly transcribed instances of the FACE vowel with the highest F1 and F2 values (the bottom- and left-most red points, in the words *they* and *mates*) were incorrectly transcribed as GOAT and STRUT (*though* and *mums* respectively).

The fact that relative location within F1/F2 space does not emerge as a compelling predictor of the vowel substitution errors is of course not unexpected given that in ASR multiple channels of information cutting across different temporal domains are processed in parallel, well beyond the segment-based F1/F2 spectral specifications typically applied within acoustic phonetic analysis [2]. Nevertheless, it is interesting to note that the diphthongs tend to have a lower rate of accuracy than the monophthongs, as seen in Table 3, which gives the proportion of instances that are incorrect for these 10 vowels. The highest proportion of errors is with the diphthongs (in particular, GOAT, e.g. *so* as *sorry, inside, a*), while LOT has the lowest proportion of errors (e.g. *was* as *is*).

Table 3. *Proportion of errors per vowel category.*

Vowel category	% incorrect	N
GOAT	3.5%	1238
FACE	2.7%	1131
PRICE	2.6%	1405
FLEECE	1.5%	1681
TRAP	1.4%	1730
DRESS	1.4%	1358
STRUT	0.9%	1497
KIT	0.8%	4024
FOOT	0.8%	1444
LOT	0.6%	1424

A final point is that there is a cluster of vowel substitution transcription errors (77 in total; 21% of the vowel substitution errors) that *prima facie* do point to the ASR analysis being influenced by specific distinctive segmental characteristics of the Australian English vowel system. Table 4 presents some examples of where the ASR output has conceivably been triggered by the typical realisation of a different vowel category by a speaker within our sample. This cluster was not distributed evenly across vowel categories with the majority of instances being found in lexical sets where there is not a good match in phonetic realisation between Australian English and those other varieties that likely constitute the larger component of the ASR training materials. Likewise, this cluster of apparently Australian English vowel-triggered errors was made up of tokens from some but not all speakers, reflecting the fact that within our speaker sample there will inevitably be significant variability in the way in which some of the same vowel categories are realised. Though it is not known the extent to which segmental characteristics are influential in the decisions made by ASR transcription algorithms, examples such as these do suggest that they are relevant to differences in ASR performance across varieties of the same language, albeit only at the margins.

Table 4. *Examples of ASR errors with vowels characteristic of Australian English.*

Vowel category	Target word	ASR Output
FACE > PRICE	*trades*	*tried*
	wave	*why*
FLEECE > FACE	*seen*	*saying*
	green	*grain*
	meal	*mail*
GOAT > STRUT	*ropes*	*rubs*
PRICE > CHOICE	*aisles*	*soils*
STRUT > TRAP	*tugs*	*tags*
(pre-nasal) TRAP > FACE	*plans*	*planes*
(non-rhotic) BATH > TRAP	*bark*	*back*

5. Discussion

This study has explored ASR transcriptions produced by Microsoft's Azure AI Speech for Australian English spontaneous speech, considering WERs, error types, and confidence scores, and testing whether accuracy varies according to social (gender and region) and phonetic factors (formant values).

First, overall WER was 7% with fillers included, and 5.3% with fillers removed. This WER is very close to the 5.1% reported by Microsoft for conversational American English [12]. Though this might suggest that the ASR does not perform worse with Australian English, this study is now seven years old and, given rapid advances in technology, the comparison may not be a valid one. We considered the confidence scores at the word level as a potential tool for researchers to pinpoint areas for correction, but these proved not to be a reliable measure of correct ASR transcription. The wide range of confidence values not only in incorrect words but also in correct words indicates that the ASR does not always recognise when it is making an error.

Looking more specifically at the kinds of errors that occurred, we found that one of the most frequent errors was filler deletions. Although ASR models often purposefully exclude fillers [26], this was not the case here where, despite filler deletion being the most frequent error, *um* and *uh* were nevertheless transcribed in the ASR output multiple times, particularly when said in relative isolation, separated by pauses. It may be that ASR models are trained on clean speech that contains few fillers [27], and thus they cannot reliably be detected, particularly if not said in isolation. Whether fillers are regularly deleted or not, however, is more reliant on the specific ASR model being used. While filler deletions are less problematic for sociophonetic analysis, the other most frequent error is problematic, namely vowel substitutions, which we hypothesised may be related to a lack of adequate representation of Australian varieties of English in datasets used to train ASR models.

Our sociolinguistic comparisons revealed no significant differences, although women had a lower WER than men, and regional speakers had a numerically lower WER than urban speakers. The lack of a significant effect for gender may not be surprising, given the varied results from previous studies [4, 14, 16]. The lack of a significant effect for region, however, is contrary to what we might predict, if the regional speakers accord with the stereotype of using more characteristically Australian English vowels [28] or follow a general trend of lying behind language changes taking place in urban centres [29]. More work is required across varieties of Australian English, including regional varieties to help shed light on this.

Finally, we explored whether the phonetic qualities of vowel segments influenced the accuracy of the ASR output. We found that this was not the case, with segments that were incorrectly transcribed showing a similar spread of vowel formants to those that were correctly transcribed, likely attributable to the fact that ASR does not rely on formant frequencies to identify speech. However, there was an indication that Australian English vowel realisations may have been problematic for the ASR, seen in poorer performance with diphthongs and other vowel categories that are particularly distinctive for Australian English.

6. Conclusion

While the scope of this study is constrained by a relatively small dataset (20,000 words from 20 speakers), our findings provide clear evidence that using an ASR approach to generating automatic orthographic transcription can work effectively with unscripted speech produced by a diverse sample of Australian speakers of English. Furthermore, the accuracy of the transcription does not appear to be influenced by factors that are known to impact on the characteristics of the speakers' spoken performance such as gender or whether they are based in an urban or regional location. The fact that the ASR model is deploying information of a form and complexity that is some way removed from conventional acoustic phonetic parameters explains why vowel location in F1/F2 space is not a strong predictor of whether a transcription instance is correct or incorrect. There is however a suggestion that some incorrect transcriptions are arising as a result of the ASR model mis-interpreting vowel qualities that are known to be particularly distinctive in the speech of (some) Australian speakers of English. Testing a wider range of speakers and accents will allow us to better understand the factors that are most closely associated with lower levels transcription accuracy, thus facilitating more informed use of ASR for sociolinguistic work.

7. Acknowledgements

We thank Clare Young and the Braidwood Museum for making the interviews from the “Heart of the Storm” podcast series available for this research; Gan Qiao for assistance with data management for the ASR process; and an anonymous reviewer for comments on an earlier version of this paper. This work was supported by ARC DP230100464, and was initiated as a Summer Scholar project conducted by the first author, with support from the Commonwealth Government of Australia.

8. References

[1] Coto-Solano, R., Stanford, J. N., and Reddy, S. K., “Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems with DARLA,” Frontiers in Artificial Intelligence, 4, 2021-September-24, 2021. https://doi.org/10.3389/frai.2021.662097.

[2] Coto-Solano, R., “Computational sociophonetics using Automatic Speech Recognition,” Language and Linguistics Compass, 16(9):e12474, 2022. https://doi.org/10.1111/lnc3.12474.

[3] Koenecke, A. et al., “Racial disparities in automated speech recognition,” Proceedings of the National Academy of Sciences, 117(14):7684-7689, 2020. https://doi.org/10.1073/pnas.191576811.

[4] Tatman, R., “Gender and dialect bias in YouTube’s automatic captions,” Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, 53-59, 2017. DOI: 10.18653/v1/W17-1606.

[5] Wassink, A. B., Gansen, C., and Bartholomew, I., “Uneven success: Automatic Speech Recognition and ethnicity-related dialects,” Speech Communication, 140:50-70, 2022/05/01/, 2022. https://doi.org/10.1016/j.specom.2022.03.009.

[6] Cox, F., and Fletcher, J., Australian English pronunciation and transcription, 2nd ed., Cambridge: Cambridge University Press, 2017

[7] Purser, B., Grama, J., and Travis, C. E., “Australian English over time: Using sociolinguistic analysis to inform dialect coaching,” Voice and Speech Review, 14(3):269-291, 2020. https://doi.org/10.1080/23268263.2020.1750791.

[8] Microsoft. "Azure AI Speech," Accessed: 23 September 2024. https://azure.microsoft.com/en-us/products/ai-services/ai-speech/.

[9] Yu, D., and Deng, L., Automatic Speech Recognition: A Deep Learning Approach, London: Springer, 2014. https://doi.org/10.1007/978-1-4471-5779-3.

[10] Loakes, D., “Does Automatic Speech Recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes?,” Frontiers in Communication, 7, 2022-June-14, 2022. https://doi.org/10.3389/fcomm.2022.803452.

[11] Radford, A. et al., “Robust speech recognition via large-scale weak supervision,” International Conference on Machine Learning, 28492-28518, 2023. https://doi.org/10.48550/arXiv.2212.04356.

[12] Xiong, W. et al., “The Microsoft 2017 conversational speech recognition system,” Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5934-5938, 2018. https://doi.org/10.1109/ICASSP.2018.8461870.

[13] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, 33:12449-12460, 2020. https://doi.org/10.48550/arXiv.2006.11477.

[14] Graham, C., and Roll, N., “Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits,” JASA Express Letters, 4(2), 2024. https://doi.org/10.1121/10.0024876.

[15] Ghyselen, A.-S. et al., “Clearing the transcription hurdle in dialect corpus building: The corpus of Southern Dutch dialects as case study,” Frontiers in Artificial Intelligence, 3, 2020-April-15, 2020. https://doi.org/10.3389/frai.2020.00010.

[16] Tatman, R., and Kasten, C., “Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube Automatic Captions,” Interspeech, 934-938, 2017. https://10.21437/Interspeech.2017-1746.

[17] Dorn, R., “Dialect-specific models for Automatic Speech Recognition of African American Vernacular English,” Proceedings of the Student Research Workshop Associated with RANLP 2019, Varna, Bulgaria, 16-20, 2019. DOI: 10.26615/issn.2603-2821.2019_003.

[18] Travis, C. E. et al., "*Sydney Speaks Corpus*," 2023. https://dx.doi.org/10.25911/m03c-yz22.

[19] Travis, C. E., Gnevsheva, K., and Docherty, G., “Voices of Regional Australia Corpus,” In Progress

[20] Godfrey, J. J., Holliman, E. C., and McDaniel, J., “SWITCHBOARD: Telephone speech corpus for research and development,” Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 517-520, 1992. DOI: 10.1109/ICASSP.1992.225858.

[21] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., “Librispeech: An ASR corpus based on public domain audio books,” Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206-5210, 2015. DOI: 10.1109/ICASSP.2015.7178964.

[22] Chen, G. et al., “Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio,” Interspeech, 3670-3674, 2021. http://dx.doi.org/10.21437/Interspeech.2021-1965.

[23] Gong, X. et al., “Advanced long-content speech recognition with factorized neural transducer,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32:1803-1815, 2024. https://doi.org/10.1109/TASLP.2024.3350893.

[24] Fromont, R., and Hay, J., “LaBB-CAT: An annotation store,” Proceedings of the Australasian Language Technology Workshop:113-117, 2012. https://aclanthology.org/U12-1015.

[25] Boersma, F., J., and Weenink, D., Praat: Doing phonetics by computer [Computer program] Version 6.4.13: retrieved 10 June 2024 from http://www.praat.org/, 2024

[26] Lease, M., and Johnson, M., “Early deletion of fillers in processing conversational speech,” Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York City, USA, 73–76, 2006.

[27] Zhu, G., Caceres, J.-P., and Salamon, J., “Filler word detection and classification: A dataset and benchmark,” Interspeech, 3769-3773, 2022. https://doi.org/10.48550/arXiv.2203.15135.

[28] Bradley, D., "Regional characteristics of Australian English: Phonology," Varieties of English: The Pacific and Australasia, Burridge, K. and Kortmann, B., eds., 111-123, Berlin/New York: Mouton de Gruyter, 2008. https://doi.org/10.1515/9783110208412.1.111.

[29] Britain, D., "Space and spatial diffusion," The Handbook of Language Variation and Change, Chambers, J. K. et al., eds., 603-637, Malden, MA: Blackwell, 2004. https://doi.org/10.1002/9780470756591.ch24.