Interspeech 2022
18-22 September 2022, Incheon, Korea

Knowledge of accent differences can predict speech recognition errors

Tünde Szalay[1], *Mostafa Shahin*[2], *Beena Ahmed*[2], *Kirrie Ballard*[1]

[1]The University of Sydney, Sydney, New South Wales, Australia
[2]The University of New South Wales, Sydney, New South Wales, Australia

tuende.szalay@sydney.edu.au

Abstract

If accent differences can predict the type of speech recognition errors, a smaller dataset systematically representing accent differences might be sufficient and less resource intensive for adapting an automatic speech recognition (ASR) to a novel variety compared to training the ASR on a large, unsystematic dataset. However, it is not known whether ASR errors pattern according to accent differences. Therefore, we tested the performance of Google's General American (GenAm) and Standard Australian English (SAusE) ASR on both dialects using words systematically representing accent differences. Accent differences were quantified using the different number of vowel phonemes, the different phonetic quality of vowels, and differences in rhoticity (i.e., presence/absence of postvocalic /ɹ/). Our results confirm that word recognition is significantly more accurate when ASR dialect matches the speaker dialect compared to the mismatched condition. Our results reveal that GenAm ASR is less accurate on SAusE speakers due to the higher number of vowel phonemes and the lack of postvocalic /ɹ/ in SAusE. Thus, the data need of adapting ASR from GenAm to SAusE might be reduced by using a small dataset focusing on differences in the size of vowel inventory and in rhoticity.

Index Terms: automatic speech recognition, accent differences, adapting ASR to novel varieties

1. Introduction

An accessible ASR must recognise speakers' intended meaning irrespective of their idiosyncrasies [1, 2]. Speaker-independence is achieved through training ASR on vast amounts of speech; however, recognition accuracy is reduced when the domain is mismatched between training and test data [1, 3]. A pervasive source of mismatch is difference in dialect between training and test speakers [3, 4, 5, 6]. ASR performs better on General American English compared to Californian American English [5]. Word recognition is consistently more accurate for five Arabic dialects when ASR is trained and tested on the same dialect compared to training ASR on all dialects [6]. Recognition accuracy of 14 Swiss dialects varies from 40% to 80% depending on dialect, despite training ASR on all 14 [7].

Although both the overall negative effect of accent differences on ASR, and the details of accent differences are known, it is still not clear whether ASR errors can be predicted from linguistic differences between training and speaker data. Therefore, we tested Google's commercially available ASR to investigate whether accent differences could predict ASR performance. We selected the accents General American (GenAm) and Standard Australian English (SAusE), because the differences in the number of phonemes, in the acoustic-phonetic quality of vowels, and in rhoticity (i.e., presence vs. absence of postvocalic /ɹ/) are documented in great detail for these varieties [8, 9, 10]. We hypothesised that (1) both ASRs would be more accurate when ASR-dialect and speaker-dialect are congruent, i.e., GenAm ASR would be more accurate on GenAm speakers than SAusE speakers and vice versa. We hypothesised that (2) in the incongruent conditions, both ASRs would be less accurate due to (2.1) phonemic differences (2.2) acoustic-phonetic vowel differences and (2.3) differences in rhoticity. If our hypotheses are born out, showing that accent differences can predict ASR errors, then GenAm ASR might be adapted for SAusE using a smaller training set that represents accent differences systematically. Thus, the data need for adapting ASR for a novel variety might be reduced.

1.1. Accent differences between GenAm and SAusE

Differences between accents of English can be captured using standard lexical sets that match vowels of a particular set of words and are identified by a key exemplar (Table 1) [8]. For instance, almost all the words that have /aɪ/ in GenAm have /ɑɛ/ in SAusE (e.g., *price, dice, vice*). This correspondence between GenAm /aɪ/ and SAusE /ɑɛ/ is identified by the key word PRICE, making the PRICE-vowel /aɪ/ in GenAm and /ɑɛ/ in SAusE.

Lexical sets capture three key differences between GenAm and SAusE, namely (1) phonemic differences (2) acoustic-phonetic differences (3) and difference in rhoticity [8]. Phonemic differences between GenAm and SAusE are shown in the different number of vowel phonemes: GenAm has 15 stressed vowels and schwa, while SAusE has 18 stressed vowels and schwa (Table 1) [9, 10]. For instance, the GenAm sets TRAP and BATH contain /æ/, whereas in SAusE, TRAP is pronounced with /æ/ and BATH with /ɐː/. That is, the lexical sets TRAP and BATH contain the same vowel in GenAm, but different vowels in SAusE. GenAm and SAusE vowels differ in their acoustic-phonetic characteristics [11, 10]. For example, FACE is more closed in GenAm than in SAusE, and GOOSE is back in GenAm, but central in SAusE. GenAm is a rhotic variety of English, i.e., /ɹ/ can appear before a vowel, a consonant, or a pause. That is, there is an /ɹ/ in *red, cart,* and *car* [9]. SAusE is a non-rhotic accent, i.e., /ɹ/ can only appear before a vowel, but not before consonant, or a pause. That is, there is an /ɹ/ in *red*, but not in *cart* or *car* [10]. The lexical sets NEAR, SQUARE, CURE, START, NORTH, and NURSE contain a postvocalic /ɹ/ in GenAm, but not in SAusE.

2. Methods

2.1. Corpora

Sixty-four (male = 30, female = 34) GenAm speakers from the LibriSpeech, and 64 (male = 30, female = 34) SAusE speakers from the AusTalk corpus were selected [13, 14]. The speakers were verified to have a GenAm and SAusE accents by phonetically trained native listeners of their respective dialects.

10.21437/Interspeech.2022-10162

Table 1: *Lexical sets with GenAm [11] and SAusE [12] IPA.*

Phonemic differences			Phonetic differences		
Lexical set	GenAm	SAusE	Lexical set	GenAm	SAusE
KIT	ɪ	ɪ	FLEECE	iː	i̞ː
NEAR	ɪ	ɪə	GOOSE	uː	ʉː
DRESS	ɛ	e	STRUT	ʌ	ɐ
SQUARE	ɛ	eː	PRICE	aɪ	ɑe
FOOT	ʊ	ʊ	MOUTH	aʊ	æɔ
CURE	ʊ	oː	CHOICE	oɪ	oɪ
TRAP	æ	æ	FACE	eɪ	æɪ
BATH	æ	ɐː	GOAT	oʊ	əʉ
START	ɑː	ɐː	THOUGHT	ɔ [ɑː]	oː
LOT	ɑː	ɔ	NORTH	ɔ	oː
			NURSE	ɚ [ɜˠ]	ɜː

2.2. Material

Three (target words) × 21 (lexical sets) = 63 monosyllabic words were selected to systematically represent accent differences. Each lexical set was evaluated based on phonemic and phonetic vowel differences, and rhoticity differences (Table 2).

Table 2: *Lexical sets and scoring accent differences*

Phonemic differences				Phonetic differences			
Lexical Set	CS	WVS	/ɹ/	Lexical Set	CS	WVS	/ɹ/
KIT	2	1	0	FLEECE	1	1	0
NEAR	2	0.5	1	GOOSE	1	0.8	0
DRESS	2	0.8	0	STRUT	1	0.7	0
SQUARE	2	0.7	1	PRICE	1	0.7	0
FOOT	2	1	0	MOUTH	1	0.8	0
CURE	2	0.7	1	CHOICE	1	1	0
TRAP	2	1	0	FACE	1	0.85	0
BATH	2	0.6	0	GOAT	1	0.7	0
START	2	0.7	1	THOUGHT	1	0.8	0
LOT	2	0.6	0	NORTH	1	0.8	1
				NURSE	1	1	1

Phonemic vowel differences were quantified with a Correspondence Score (CS). Lexical sets containing a vowel with a one-to-one correspondence between GenAm and SAusE received a CS of one, indicating higher phonemic similarity. Lexical sets with a GenAm vowel corresponding to two SAusE vowels received a CS of two, indicating lower phonemic similarity.

Phonetic vowel differences were quantified with Weighted Vowel Similarity (WVS) score, adapted from [15]. Vowels were aligned, and each slot was scored on a scale of zero (no agreement) to one (maximal agreement). The score is the sum of agreement scores for height (close, close-lax, close-mid, open-mid, open) ranging from 0.0 to 0.4; for frontess (front, front lax, central, back lax, back), ranging from 0.0 to 0.4; for length (long, short), ranging from 0.0 to 0.1; and for rounding (rounded, unrounded), ranging from 0.0 to 0.1. Having calculated WVS for every slot in a vowel, average WVS was calculated (Table 3, Equation 1). Slots were counted maximally, i.e., lexical sets that are diphthongs in SAusE but not in GenAm (e.g., NEAR) were assigned two slots.

$$WVS = \frac{Height + Frontness + Length + Rounding}{NumberOfSlots} \quad (1)$$

Rhoticity was coded as a binary variable: lexical sets containing a postvocalic /ɹ/ in GenAm but not in SAusE were marked as 1 for R-dropping. Lexical sets not containing a postvocalic /ɹ/ in either of the dialects were marked with 0.

Table 3: *WVS for the lexical set* FACE *calculated as mean WVS for GenAm /e/ and SAusE /æ/ (first segment) and GenAm /ɹ/ and SAusE /ɹ/ (second segment).*

Slot	GenAm	SAusE	Height	Front.	Length	Round.	Sum	Mean
1	e	æ	0.1	0.4.	0.1	0.1	0.7	0.85
2	ɪ	ɪ	0.4	0.4	0.1	0.1	1	

For each speaker, recordings of a semantically neutral carrier sentence (e.g., *Who says_, I said_*) were extracted from the corpora. The carrier sentence was inserted before the target word; sentence and target were separated by 10 ms of silence.

2.3. Procedure

The sentences were transcribed using Google's commercially available ASR for GenAm and SAusE, creating two congruent and two incongruent conditions. In the congruent conditions, ASR dialect matched speaker dialect. That is, target words produced by GenAm speakers were submitted to GenAm ASR whereas target words produced by SAusE speakers to SAusE ASR. In the incongruent conditions, ASR dialect did not match speaker dialect. That is, target words by GenAm speakers were submitted to SAusE ASR and vice versa.

2.4. Analysis

The transcription returned by Google was scored as Correct, when the transcription contained the target word, Incorrect, when the transcription did not contain the target word, and Not Recognised, when no transcription was returned. Response accuracy was treated as ordinal data with Correct transcription being the best match for the target, and Incorrect transcription being a better match than Not Recognised. Although it is not known where in Google ASR the different errors of Not Recognised and Incorrect originated from (e.g., insufficient training data, incorrect pruning of search paths), a Not Recognised error indicates that the ASR could not map the stimulus onto anything, and an Incorrect indicates that the ASR could map the stimulus onto something, making Not Recognised worse than Incorrect.

We constructed an Ordinal Mixed Model (OMM) [16, 17] with the dependent variable Accuracy (zero for Not Recognised, one for Incorrect, and two for Correct). The independent variables were ASR Dialect, Speaker Dialect (both contrast coded, comparing SAusE to the baseline GenAm), Correspondence Score (CS, contrast coded, comparing One-to-Two to the baseline of One-to-One), R Dropping (contrast coded, comparing R Dropping to the baseline of No R Dropping), and Weighted Vowel Similarity (WVS, continuous from 0 for maximal dissimilarity to 1 for maximal similarity). Interactions between ASR Dialect and Speaker Dialect; ASR Dialect and Correspondence Score, R Dropping, and WVS; and Speaker Dialect and Correspondence Score, R Dropping, and WVS were included. Interactions between CS, WVS, and R Dropping were not included due to multicollinearity between these factors. Speaker and Target were added as random intercepts.

3. Results

In the OMM, the main effect of SAusE Speaker Dialect indicates that GenAm ASR is significantly less accurate at recog-

nising SAusE speakers compared to GenAm speakers ($\beta = -1.401, z_{0.575} = -2.439, p = 0.015$). The main effect of SAusE ASR Dialect indicates that SAusE ASR is significantly less accurate at recognising GenAm speakers than GenAm ASR ($\beta = -1.087, z_{0.547} = -1.988, p = 0.046$). The interaction between SAusE Speaker Dialect and SAusE ASR Dialect indicates that ASR Accuracy increased significantly when both ASR and Speaker Dialect were SAusE ($\beta = 2.907, z_{0.768} = 3.785, p < 0.001$). The main effects of Speaker Dialect, ASR Dialect, their interaction, are consistent with Hypothesis (1), stating that ASR accuracy is significantly better in the congruent conditions than in the incongruent conditions due to accent differences between training and test data (Figure 1).

However, the lower accuracy of the GenAm ASR on SAusE speakers may also be caused by the SAusE test data being overall more difficult. Therefore, we tested the effect of Speaker Dialect with respect to ASR Dialect using planned comparison with Bonferroni correction [18, 17]. The planned comparison shows that target words are recognised significantly more accurately in the congruent conditions ($p < 0.0001$ for GenAm ASR and $p < 0.005$ for SAusE ASR).

The lower accuracy of the SAusE ASR on GenAm speakers may also be caused by the overall lower quality of the SAusE ASR. Therefore, we tested the effect of ASR Dialect with respect to Speaker Dialect using planned comparison with Bonferroni correction [18, 17]. The planned comparison revealed that both GenAm and SAusE ASR perform significantly more accurately in the congruent conditions ($p < 0.0001$ both for GenAm and SAusE speakers). The results of the planned comparisons are consistent with Hypothesis (1), as they indicate that ASR accuracy is significantly better in the congruent conditions than in the incongruent conditions.

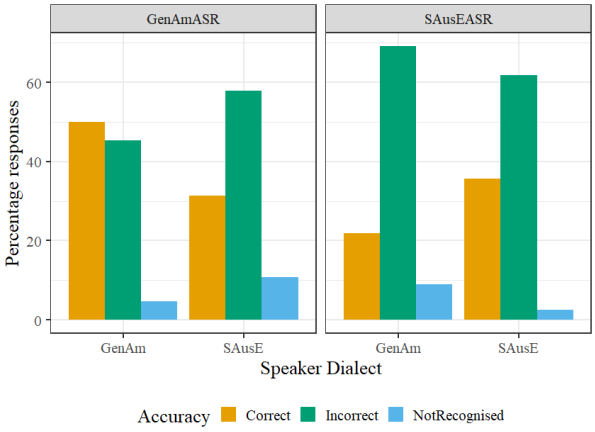

Figure 1: *Accuracy: The effect of Speaker- and ASR Dialect.*

Having tested the effect of dialect mismatch on ASR accuracy, the effect of CS, WVS, and R Dropping were analysed in the OMM. We found no significant main effect of CS, WVS, or R Dropping. That is, we found no evidence of CS, WVS, and R Dropping affecting the accuracy of GenAm ASR on GenAm speech. The lack of results are consistent with Hypotheses (2.1)-(2.3) as accent differences were not expected to affect the congruent conditions.

A significant negative interaction between SAusE Speaker Dialect and CS ($\beta = -0.434, z_{0.187} = -2.327, p = 0.02$) indicates that the accuracy of GenAm ASR on SAusE Speakers significantly decreases when the target word contains a vowel phoneme with a one-to-two vowel correspondence (Figure 3). This finding is consistent with Hypothesis (2.1), stating that ASR accuracy would be negatively affected by phonemic differences in the incongruent conditions.

The lack of significant interaction between Speaker Dialect and WVS indicates that the accuracy of GenAm ASR on SAusE Speakers is not significantly affected by phonetic vowel differences (Figure 2). This result does not support Hypothesis (2.2), stating that acoustic-phonetic differences would negatively impact ASR in the incongruent conditions.

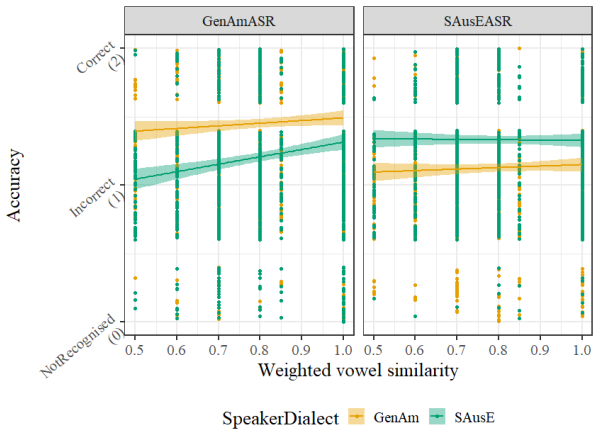

Figure 2: *Accuracy by ASR- and Speaker Dialect: No effect of phonetic similarity.*

The significant negative interaction between Speaker Dialect and R Dropping ($\beta = -0.536, z_{0.207} = -2.586, p = 0.01$) indicates that the accuracy of GenAm ASR on SAusE Speakers significantly decreases when the target word contains a postvocalic /ɹ/ in GenAm, but not in SAusE (Figure 3). This finding is consistent with Hypothesis (2.3), stating that differences in rhoticity would negatively affect ASR accuracy in the incongruent conditions.

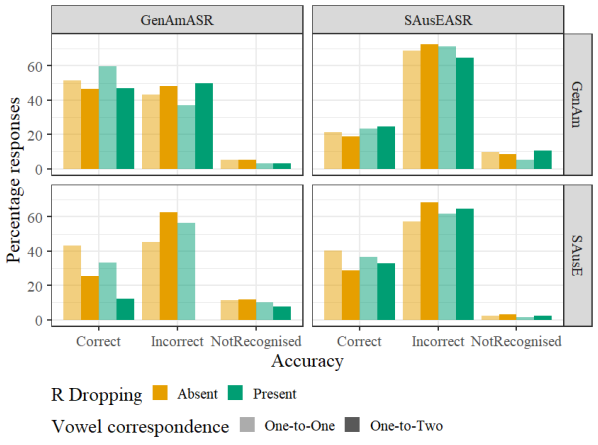

Figure 3: *Accuracy by Speaker- and ASR-Dialect: The effect of vowel correspondence and R Dropping.*

We found no significant interactions between SAusE ASR Dialect and any of the factors capturing accent differences, indicating that the accuracy of SAusE ASR on GenAm Speakers is not affected significantly by accent differences. These

results are not consistent with Hypotheses (2.1)-(2.3), stating that accent differences would negatively impact ASR accuracy in the incongruent conditions. The three-way interactions between SAusE ASR Dialect, SAusE Speaker Dialect, and either CS, WVS, or R Dropping were not significant.

4. Discussion and Conclusion

Automatic recognition of GenAm and SAusE speech by GenAm and SAusE ASR was tested, yielding two congruent (GenAm ASR - GenAm Speaker and SAusE ASR - SAusE speaker), and two incongruent conditions (GenAm ASR - SAusE speaker, SAusE ASR - GenAm speaker). Target words systematically represented accent differences.

Hypothesis (1) predicted that ASR accuracy would be lower when ASR Dialect and Speaker Dialect are incongruent. Hypothesis (1) is born out, as the collective results of our OMM and planned comparisons show that GenAm speakers are recognised by GenAm ASR more accurately than by SAusE ASR and vice versa. In addition, GenAm ASR is better at recognising GenAm speakers than SAusE ASR and vice versa. Therefore, reduced accuracy is attributed to accent differences, rather than to the overall better quality of one ASR or one test dataset. Reduced ASR accuracy caused by ASR- and speaker accent differences is consistent with the existing body of literature, and it can be caused by using incongruent training data, acoustic model, and/or pronunciation dictionary [3, 4, 5, 6].

Having confirmed that ASR accuracy is adversely impacted by accent differences, we analysed the effect of phonemic, phonetic, and rhoticity differences because, to the best of our knowledge, accent differences have not been used to predict speech recognition errors. In the congruent conditions, accent differences were not expected to affect ASR accuracy (Hypotheses (2.1)-(2.3)), and there was no signficant effect of accent differences on ASR accuracy in the congruent condition.

In the incongruent conditions, an imbalanced picture emerges: accent differences negatively impact target accuracy when GenAm ASR is used on SAusE speech, but not in the other direction. Hypothesis (2.1) predicted that ASR accuracy would be reduced in the incongruent conditions due phonemic differences caused by the larger number of vowel phonemes in SAusE. Hypothesis (2.1) is partially supported, as words containing vowel phonemes with one-to-two mappings between GenAm and SAusE are recognised less accurately when a GenAm ASR is applied to SAusE speech. However, no effect of vowel correspondence was found when a SAusE ASR was applied to GenAm speech. This can be attributed to GenAm ASR not having suitable acoustic models for the vowels that are only present in SAusE. In contrast, the acoustic models of the smaller vowel inventory of GenAm might be included in the models of the larger SAusE vowel inventory.

Hypothesis (2.2) predicted that ASR accuracy would be reduced in the incongruent conditions, due to acoustic vowel differences. Hypothesis (2.2) is not supported by our results, as vowel similarity had no significant effect on the recognition of GenAm target words by SAusE ASR or on the recognition of SAusE target words by GenAm ASR.

Hypothesis (2.3) predicted that ASR accuracy would be reduced in the incongruent conditions, due to differences in rhoticity. Our results partially support hypothesis (2.3), as words containing a postvocalic /ɹ/ in GenAm but not in SAusE were recognised less accurately when GenAm ASR was applied to SAusE speech. However, differences in rhoticity did not affect recognition accuracy of GenAm speech by SAusE ASR. The adverse impact of R Dropping on the recognition of SAusE words by GenAm ASR can be attributed to the GenAm ASR not being trained on non-rhotic accents, as postvocalic /ɹ/ is always present in GenAm. In contrast with the GenAm training data, postvocalic /ɹ/ must have been part of the SAusE training data, as word-final /ɹ/ is pronounced in connected speech in SAusE when /ɹ/ is was followed by a vowel at the beginning of the next word. For example, /ɹ/ is pronounced in the phrase *far away* in SAusE. Rhoticity differences may also contribute to the low recognition accuracy of non-rhotic African American Vernacular English by YouTube's automatic transcription [5, 19].

The results collectively show that phonemic and rhoticity differences impact word recognition accuracy, whereas fine-grained vowel differences do not. Therefore, segment-level accent differences, namely the presence of a vowel phoneme unknown to GenAm ASR, and the absence of postvocalic /ɹ/ seem to negatively impact word recognition, whereas subsegmental differences, such as a change in vowel quality, do not.

However, the effects of phonemic, phonetic, and rhoticity differences are challenging to separate due to the correlations between these factors caused by historical language changes affecting pre-/ɹ/ vowels in SAusE but not in GenAm [8, 20]. As a result of these changes, lexical sets containing a postvocalic /ɹ/ in GenAm, but not in SAusE, always show phonemic differences with a one-to-two vowel correspondence, as well as larger acoustic-phonetic differences. Therefore, only a few lexical sets contain phonetically different vowels without phonemic or rhoticity differences, and these might not have been sufficient to show how phonetic differences affect ASR quality.

The limitation of our study is using commercially available ASR. As Google ASR is a proprietary software, it is not known how the models were built and what data they were trained on. For instance, the LibriSpeech audio corpus, used for testing the recognition of GenAm speech in this experiment, is often used to train ASR systems, such as [21], and it might have been used to train Google's GenAm ASR as well. When the same data is used for training and testing ASR, spuriously high word recognition accuracy is expected, leading to better recognition of GenAm words by the GenAm ASR. The SAusE ASR might have been trained on a combination of GenAm and SAusE datasets to compensate for the comparative lack of SAusE data, as combining multiple dialects has improved ASR for low-resource varieties in [22, 4, 23, 24]. Training SAusE ASR on both dialects could explain why accent differences did not affect recognition of GenAm speech by SAusE ASR.

To conclude, errors of GenAm ASR on SAusE speech can be predicted from the accent differences in the number of vowel phonemes and in rhoticity. Therefore, when adapting an existing GenAm ASR for SAusE, we recommend ASR to be specifically trained on a dataset that emphasises phonemic vowel differences and differences in rhoticity. Creating targeted training dataset focusing on these linguistic features might allow the use of a smaller training set and thus, it might reduce the data need of ASR. In the future, we aim to improve the performance of our UNSW ASR, currently trained on GenAm data, on SAusE by training it on a SAusE dataset systematically representing accent differences.

5. Acknowledgements

This research was funded by the grant DP200103006. We thank James Grama, Amy Ramage, and the students at the Dept. of Communication Sciences and Disorders, University of New Hampshire for their help in selecting the GenAm speakers.

6. References

[1] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[2] J. Keshet, "Automatic speech recognition: A primer for speech-language pathology researchers," *International journal of speech-language pathology*, vol. 20, no. 6, pp. 599–609, 2018.

[3] L. t. Bosch, "ASR, dialects, and acoustic/phonological distances," in *Sixth International Conference on Spoken Language Processing*, 2000.

[4] A. Nogueiras, M. Caballero, and A. Moreno, "Multi-dialectal Spanish speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–841.

[5] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions." in *Interspeech*, 2017, pp. 934–938.

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

[7] I. Nigmatulina, T. Kew, and T. Samardzic, "Asr for non-standardised languages with dialectal variation: the case of swiss german," in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 2020, pp. 15–24.

[8] J. C. Wells, *Accents of English*. Cambridge: Cambridge University Press, 1982.

[9] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English, Phonetics, Phonology and Sound Change*. Berlin, Boston: Gruyter Mouton, 2008.

[10] F. Cox and J. Fletcher, *Australian English pronunciation and transcription*. Cambridge: Cambridge University Press, 2017.

[11] J. M. Hillenbrand, "American English: Southern Michigan," *Journal of the International Phonetic Association*, vol. 33, no. 1, pp. 121–126, 2003.

[12] F. Cox and S. Palethorpe, "Australian English," *Journal of the International Phonetic Association*, vol. 37, no. 3, p. 341–350, 2007.

[13] 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.

[14] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner *et al.*, "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box." ISCA, 2011.

[15] J. L. Preston, H. L. Ramsdell, D. K. Oller, M. L. Edwards, and S. J. Tobin, "Developing a weighted measure of speech sound accuracy," 2011.

[16] R. H. B. Christensen, "ordinal—regression models for ordinal data," 2019, R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[18] R. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2019, R package version 1.3.4. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[19] L. N. Hinton and K. E. Pollock, "Regional variations in the phonological characteristics of African American Vernacular English," *World Englishes*, vol. 19, no. 1, pp. 59–71, 2000.

[20] B. Gick, "A gesture-based accounts of intrusive consonants in English," *Phonology*, vol. 16, no. 1, p. 29–54, 1999.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[22] A. Messaoudi, H. Haddad, C. Fourati, M. B. Hmida, A. B. E. Mabrouk, and M. Graiet, "Tunisian dialectal end-to-end speech recognition based on deep speech," *Procedia Computer Science*, vol. 189, pp. 183–190, 2021.

[23] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8619–8623.

[24] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.