INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions

Rachael Tatman[1], *Conner Kasten*[2]

[1]The University of Washington, United States
[2]Zonar Systems, United States

rctatman@uw.edu, conner.kasten@emerssso.com

Abstract

This project compares the accuracy of two automatic speech recognition (ASR) systems–Bing Speech and YouTube's automatic captions–across gender, race and four dialects of American English. The dialects included were chosen for their acoustic dissimilarity. Bing Speech had differences in word error rate (WER) between dialects and ethnicities, but they were not statistically reliable. YouTube's automatic captions, however, did have statistically different WERs between dialects and races. The lowest average error rates were for General American and white talkers, respectively. Neither system had a reliably different WER between genders, which had been previously reported for YouTube's automatic captions [1]. However, the higher error rate non-white talkers is worrying, as it may reduce the utility of these systems for talkers of color.

Index Terms: automatic speech recognition, sociolinguistics, dialect, gender, race

1. Sociolinguistic bias in ASR

Variation in speech, including regional dialects, has long proven challenging for automatic speech recognition systems [2]. Despite significant advances in the overall accuracy of speech recognition, recent work suggests that dialect and gender biases still remain, at least for YouTube's automatic captioning [1]. The previous study did have several areas for improvement, however. First, it looked at accuracy for words spoken in isolation. Since it is standard for ASR systems to consider the linguistic context in which a word is spoken, only considering words spoken in isolation place systems at a disadvantage. Second, since the captions were generated over the span of several years, it does not provide a snapshot of a system at one point in time. Third, it only considered a single system. And, fourth, it does not include an acoustic analysis of the speech varieties being described in order to determine that that they do represent robust regional variation. This paper expands on that work by reevaluating YouTube's automatic captions using connected speech and includes an evaluation of a second system, Microsoft's Bing Speech API. Both systems are evaluated using the same recordings of connected speech transcribed over the period of a few minutes. It also includes an acoustic analysis of regional differences in the speech data used to evaluate the automatic speech recognition systems.

2. Four Varieties of American English: Data

Four distinct varieties of American English were chosen for this comparison: General American, Northern Cities, Southern and Californian English. These are all major, well-documented varieties of American English which are acoustically distinct.

Acoustic data for these varieties was taken from the Dialects of English Archive [3], which was created by Paul Meier as a resource for actors. As a result, many of the recordings are of lower quality than those typically used in phonetic research. While this is not ideal for a fine-grained acoustic analysis, it is reasonable for this study since it is better representation of the type of speech data a commercial ASR system encounters. All speech data is taken from talkers reading the passage "Comma Gets a Cure" [4], which was designed to include Wells Standard Lexical Set [5].

The discussion of acoustic dialectal differences observed in the data below shows that these are four distinct varieties of American English, and that they include phonetic features which have been established as being associated with these varieties in the sociophonetic literature. The strong differences between these varieties should provide a robust metric of the ability of automatic speech recognition (ASR) systems to handle dialectal variation.

A total of 39 talkers were included in this analysis: 11 from Alabama, 8 from California, 8 from Michigan and 12 General American talkers. There were slightly more male talkers than female talkers (22 men, 17 women). For 13 talkers, including all the General American talkers, their race was unreported. Among the remaining talkers, 13 were white, 8 African American, 4 of mixed race and 1 Native-American.

2.1. General American

General American English, also called Standard[ized] American English, Mainstream US English and Mainstream American English, is the prestige variety of English in the United States. Unlike in Britain, where the prestige variety of Received Pronunciation has well-defined target vowel pronunciations, General American is is characterized by a lack of stigmatized linguistic features [6]. The most stigmatized features in American English are lexical (e.g. "ain't", "skeeter") and morphological (e.g. double negation, double modals). However, some phonological features are also stigmatized. In particular, talkers who are participating in major, on-going vowel shifts, especially in the South, may be regarded as "less correct" by other talkers [7].

The General American speech samples used in this project were produced by voice professionals (actors, voice coaches and speech language pathologists) who consciously avoided using stigmatized features. Though information on their dialect background was not available, none of the talkers included in this sample are participating in the three on-going vowel shifts in the US: the Northern Cities, Southern and Californian vowel shifts. This can be seen in Figure 1. Compared to other talkers, the male General American talkers showed a larger vowel space. This is probably due to their high degree of hyper-

http://dx.doi.org/10.21437/Interspeech.2017-1746

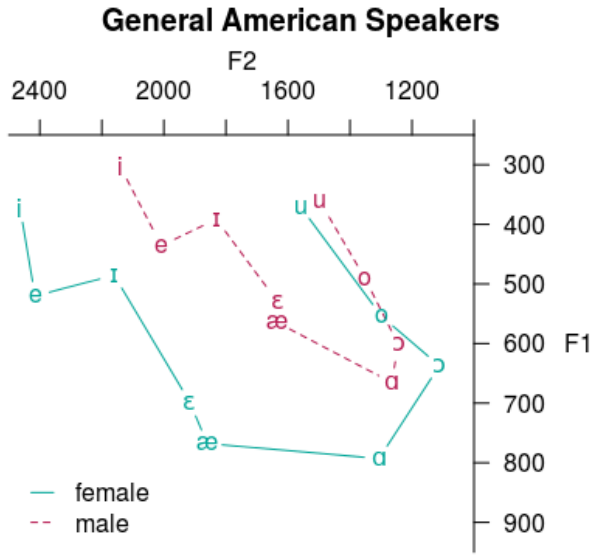

Figure 1: *Figure showing the vowel space of acoustic data from General American talkers in an F1 - F2 space, separated by gender. Note the larger vowel space relative to other varieties*

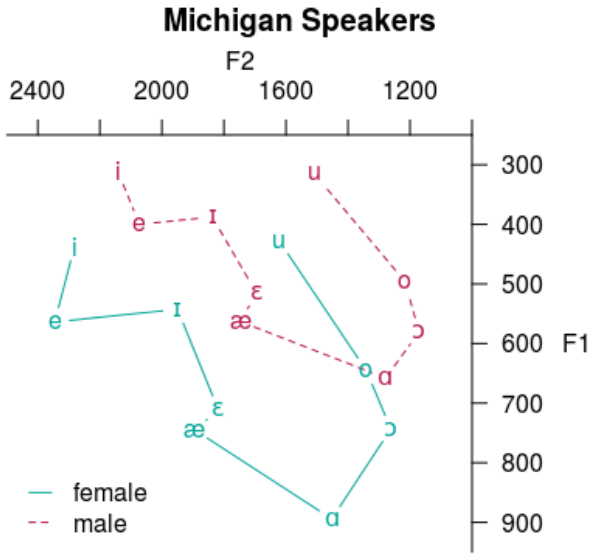

Figure 2: *Figure showing the vowel space of Michigan talkers in an F1 - F2 space, separated by gender. Note the raising of \æ\and backing of \ɛ\.*

articulation [8]. In addition, Figure 1 clearly shows that the male General American talkers, in particular, show a merger between \ɔ\and \ɑ\. This is unsurprising given that this merger is not a stigmatized feature.

Given that these talkers are both intentionally speaking clearly and also not users of stigmatized phonetic features, it seems likely that that ASR systems will transcribe their speech more accurately than that of talkers of other dialects. No race or ethnicity information was provided for these talkers.

2.2. Northern Cities

The Northern Cities vowel shift is an on-going vowel shift in the North East and Northern Midwest, especially in major cities in Michigan and Illinois, including Chicago [9]. It is characterized by lowering of \ɪ\to \ɛ\, backing of \ɛ\to \ʌ\, backing of \ʌ\to \ɔ\, lowering of \ɔ\to \ɑ\, fronting of \ɑ\to \æ\and raising of \æ\, sometimes as high as \i\[10]. The Michigan talkers included in this study are participating in parts of this on-going chain shift, as can be seen in Figure 2. In particular, \æ\is raised past the General American \ɛ\and \ɛ\is slightly backed. Of the Michigan speakers, 12 were Caucasian, 2 African-American and one was of mixed race.

2.3. Southern

The talkers from Alabama have phonetic features of Southern English, including both participating in the Southern Shift and maintaining a strong distinction between \ɔ\and \ɑ\. The components of the Southern Shift most apparent in these talkers is the raising and fronting of \ɪ\and \ɛ\and the lowering and backing of \e\(but not \i\, as has been observed in other talkers) [6]. In addition, both \u\and \o\are fronted. Though the formant measurements shown in Figure 3 were taken at the midpoint of vowels and thus do not show this, the Alabama talkers also had a high degree of \ai\monophthongization, a feature found in both white and Black Southern American English [11]. Of the Southern talkers, 7 were Caucasian and 6 African-American.

2.4. Californian

The Californian talkers in this sample are participating in parts of the California Vowel Shift [12], including fronting of \u\and \o\, and backing of of \ɑ\towards \ɔ\. Of the Californian talkers, five were Caucasian, 3 of mixed race, one Native-American and one did not have their race reported.

3. System Evaluations

3.1. Bing Speech (Project Oxford)

All of the speech files discussed above were transcribed using the Bing Speech API[1], a commercial speech recognition product offered through Microsoft Cognitive Services, previously Project Oxford. The API was accessed through a custom Android Application built for this analysis and developed using the Bing Speech Android SDK [2]. The app sent each file to be transcribed individually. In order to improve efficiency, the .wav files were down-sampled from 22050 Hz to 11025 Hz. The returned transcriptions were stored in a separate .txt file for each audio file. Not all files were transcribed, and of those that were, many had truncated transcriptions. Of the 39 files sent to the API, only 36 were returned with transcriptions, despite repeated efforts. It is unclear why this happened. However, given that there was a high correlation between the length of the returned transcription and the non-deletion errors in the text that was returned ($r = -0.45$, $n = 36$, $p < 0.05$), one plausible hypothesis is that the system keeps a running measure of confidence in the transcription and returns the transcription to the API only up to the point where that confidence measure decays below a predetermined point. Since this analysis was run soon after the

[1]https://www.microsoft.com/cognitive-services/en-us/speech-api
[2]https://github.com/Microsoft/Cognitive-Speech-STT-Android

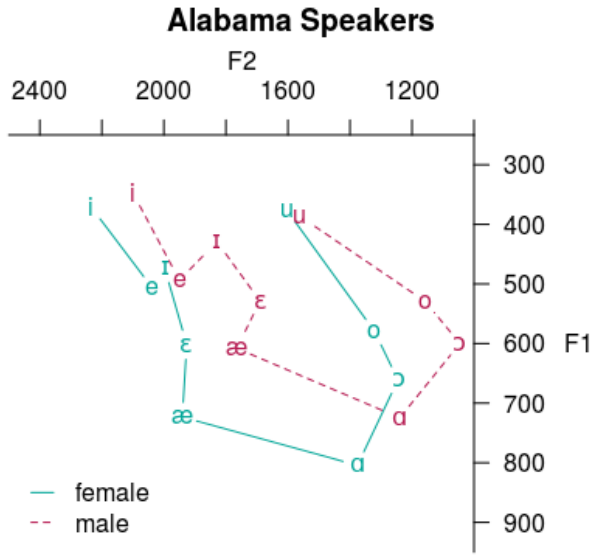

Figure 3: *Figure showing the vowel space of talkers from Alabama in an F1 - F2 space. Note the \ɪ\is raised while \e\is lowered and \ɛ\is raised.*

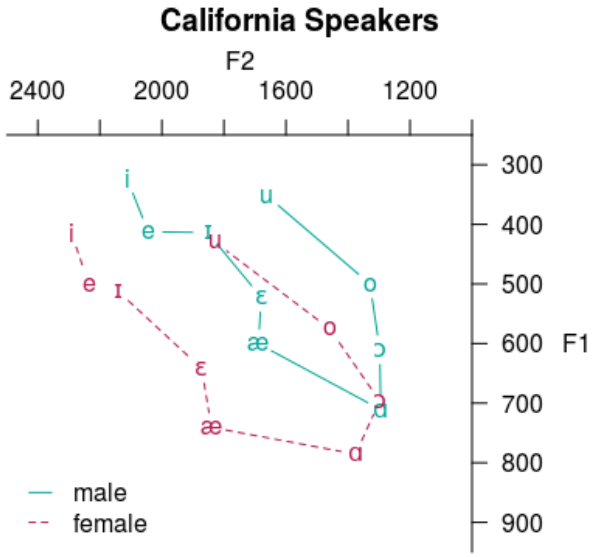

Figure 4: *Figure showing the vowel space of acoustic data from California in an F1 - F2 space. Note that bothu\and \o\are very fronted.*

launch of the API, it is also possible that the incomplete transcriptions were the result of a bug.

3.2. YouTube Automatic Captions

Audio files were also transcribed using YouTube's automatic captions [13]. These were generated by creating MP4 videos with the audio files as a soundtrack. The language was manually set to "English (United States)". Once generated, the automatic captions were then downloaded as .srt files and converted to plain text prior to analysis. The conversion to plain text did result in the removal of information about ASR recognition confidence, which is color-coded by word in automatic captions.

3.3. Word Error Rate

Because many of the transcriptions returned by Bing were partial, calculating Word Error Rate (WER) as the by-word edit distance from the correct transcription would have led to an artificially high WER. [3]. To correct for this, the WER was calculated as the number of non-deletion errors divided by the total number of words in the automatic transcription. So if the the correct transcription was "The lamb is cute" and the returned transcription was "The lamb shoots", the WER would be 0.33 (one substitution over three words) rather than 0.5 (one substitution and one deletion over four words).

Overall, the WER were quite high, especially given very high accuracy (under 0.07) recently reported by a team at Microsoft [15]. The mean WER for the Bing transcriptions was 0.45($\sigma = 0.18$). The YouTube WER was both lower and less variable ($\mu = 0.31, \sigma = 0.07$).

3.4. Difference in Accuracy by Dialect

WER did vary across dialects, as can be seen in Figure 5. For both systems, the lowest average WER was for General American talkers and the highest for talkers from California. While the former was expected, the latter is surprising given an earlier finding that YouTube's automatic captions had the highest accuracy for Californian talkers [1]. Differences in WER by dialect were not robust enough to be significant for Bing (under a one-way ANOVA) ($F[3, 32] = 1.6, p = 0.21$), but they were for YouTube's automatic captions ($F[3, 35] = 3.45, p < 0.05$). The differences between these two systems is not surprising given the far lower variance for the YouTube WER.

3.5. Difference in Accuracy by Talker Gender

Previous studies have found differences in ASR accuracy among genders [16, 17, 18], and earlier work found that men's voices were recognized with more accuracy by YouTube's automatic captions [1]. That effect was not replicated here for either system. Neither Bing ($F[1, 34] = 1.13, p = 0.29$), nor YouTube's automatic captions ($F[1, 37] = 1.56, p = 0.22$) had a significant difference in accuracy by gender.

3.6. Difference in Accuracy by Talker Race

The well-documented, systematic differences between General American and varieties such as African American English [19] and Chicano English [20] may be a source of preventable errors for automatic speech recognition systems. As can been seen in Figure 7, for both systems, error rates were lowest for white talkers as a group, and higher for African American and mixed race talkers. As with dialect, differences in WER between races were not significant for Bing ($F[4, 31] = 1.21, p = 0.36$), but were significant for YouTube's automatic captions ($F[4, 34] = 2.86, p < 0.05$). All talkers were native English speakers.

4. Discussion

This study evaluated two automatic speech recognition systems, Microsoft's Bing Speech and Google's YouTube automatic cap-

[3] See [14] for a broader discussion of the shortcomings of WER

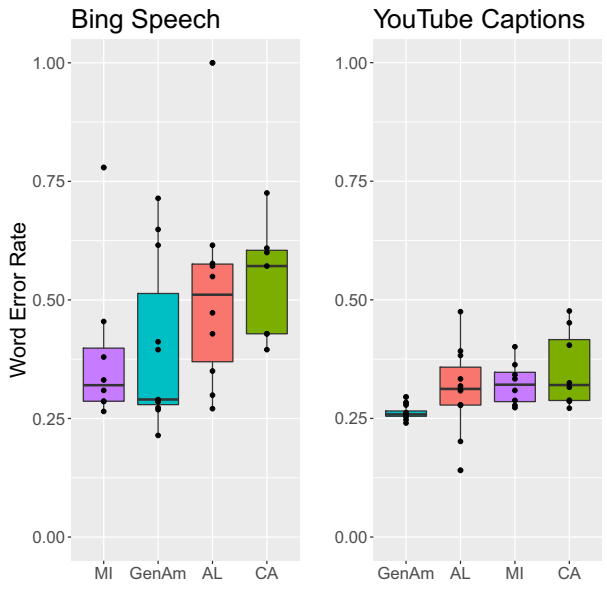

Figure 5: *Word error rate by region. Points represent individual talkers.*

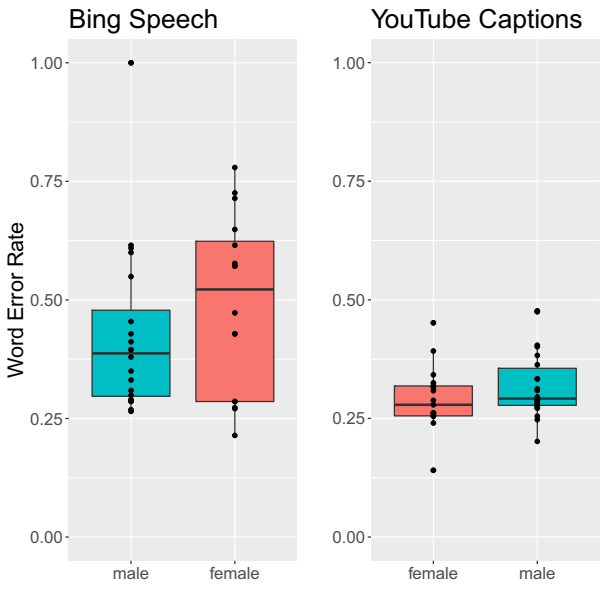

Figure 6: *Plot showing word error rate by gender. Points represent individual talkers.*

tions, on a sociolinguistically-stratified sample of talkers from different dialect backgrounds, genders and races. While both systems made errors, the rate at which errors were made varied based on talker's social identities, and in particular their dialect background and race. These results were only statistically reliable for YouTube's automatic captions, although given the high variability in Bing's WER, the sample size was too small to achieve high power. Given four dialect regions, an F of .35 (as observed for the YouTube captions) and a significance level of 0.05, at least 24 talkers per dialect should be sampled to obtain a power of 0.8 [21]. It was not possible to increase the sample size, however, given that all talkers in the archive who read "Comma Gets a Cure" [4] from each state were included.

However, even with the small sample size, some robust effects of talker ethnicity and dialect were observed, and the direction of the effects was the same across systems. Among the dialects, both systems had the lowest average WER for General American talkers, and among ethnicities, both systems had the lowest WER for white talkers. The former, of course, is possibly confounded by the fact that the General American talkers in this study were all voice professionals. As a result, they produced very clear, hyper-articulated speech. It is possible that the differences between dialects arose from this rather than a bias towards a particular dialect. This would not, however, account for biases towards white talkers, as the ethnicity of the General American talkers was unknown.

This study provides additional evidence that, despite dramatic improvements in the technology, automatic speech recognition systems continue to struggle to maintain high accuracy in the face of well-documented systematic sociolinguistic variation. This is particularly troubling given that the groups of talkers with the highest error rates, in particular African Americans and talkers who use non-standardized regional varieties, are those who have faced other kinds of discrimination as well. This highlights the need for further work on accent adaptation for these large-scale commercial systems, both for regional dialects [22] and ethnolects [23].

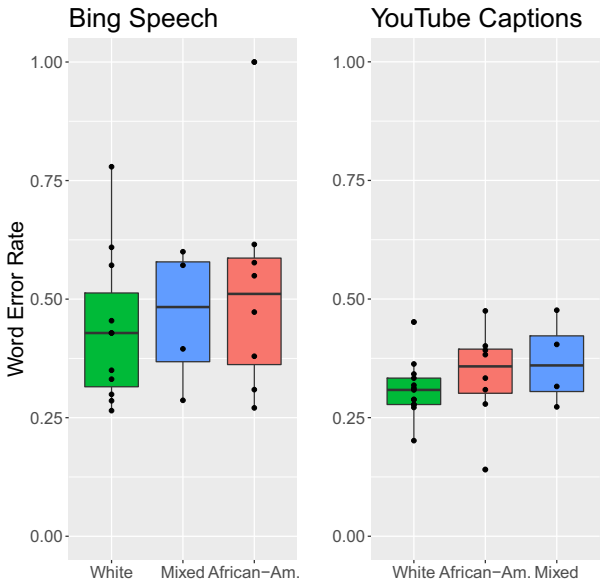

Figure 7: *Plot showing word error rate by region and talker race (excluding talkers of unknown race and the one Native American talker). Points are individuals.*

5. Acknowledgements

The authors would like to acknowledge the support of Emily Bender, Richard Wright, Gina-Anne Levow and Alicia Beckford Wassink for their support and guidence. Remaining errors and omissions are the authors'. This work was supported by NSF grant DGE-1256082.

6. References

[1] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *First Workshop on Ethics in Natural Language Processing*. ACL, 2017.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10, pp. 763–786, 2007.

[3] P. Meier and S. M. Muller, "IDEA: International dialects of English archive," *Accessed May*, vol. 17, p. 2005, 1998.

[4] D. Honorof, J. McCullough, and B. Somerville, "Comma gets a cure: A diagnostic passage for accent study," *Retrieved March*, vol. 20, p. 2017, 2000.

[5] J. C. Wells, *Accents of English*. Cambridge University Press, 1982, vol. 1.

[6] W. Wolfram and N. Schilling, *American English: dialects and variation*. John Wiley & Sons, 2015, vol. 25.

[7] D. R. Preston, "Perceptual dialectology: Aims, methods, findings," *Trends in Linguistics Studies and Monographs*, vol. 137, pp. 57–104, 2002.

[8] K. Johnson, E. Flemming, and R. Wright, "The hyperspace effect: Phonetic targets are hyperarticulated," *Language*, pp. 505–528, 1993.

[9] W. Labov, M. Yaeger, and R. Steiner, *A quantitative study of sound change in progress*. US Regional Survey, 1972, vol. 1.

[10] M. J. Gordon, *Small-town values and big-city vowels: A study of the Northern Cities Shift in Michigan*. Duke Univ Press, 2001.

[11] V. Fridland, "Tie, tied and tight: The expansion of /ai/ monophthongization in African-American and European-American speech in Memphis, Tennessee," *Journal of Sociolinguistics*, vol. 7, no. 3, pp. 279–298, 2003.

[12] P. Eckert, "Where do ethnolects stop?" *International journal of bilingualism*, vol. 12, no. 1-2, pp. 25–42, 2008.

[13] K. Harrenstien, "Automatic captions in youtube," *The Official Google Blog*, vol. 11, 2009.

[14] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 577–582.

[15] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 Conversational Speech Recognition System," *arXiv preprint arXiv:1609.03528*, 2016.

[16] S. Ali, K. Siddiqui, N. Safdar, K. Juluru, W. Kim, and E. Siegel, "Affect of gender on speech recognition accuracy," in *American Journal of Roentgenology*, vol. 188, no. 5. American Roentgen Ray Society, 2007.

[17] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[18] M. Sawalha and M. Abu Shariah, "The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus," in *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*. Leeds, 2013.

[19] J. R. Rickford, *African American vernacular English: Features, evolution, educational implications*. Wiley-Blackwell, 1999.

[20] F. Peñalosa, "Chicano sociolinguistics: A brief introduction." 1980.

[21] S. Champely, *pwr: Basic Functions for Power Analysis*, 2016, r package version 1.2-0. [Online]. Available: https://CRAN.R-project.org/package=pwr

[22] M. Najafian, A. DeMarco, S. J. Cox, and M. J. Russell, "Unsupervised model selection for recognition of regional accented speech." in *INTERSPEECH*, 2014, pp. 2967–2971.

[23] M. Lehr, K. Gorman, and I. Shafran, "Discriminative pronunciation modeling for dialectal speech recognition." in *INTERSPEECH*. Singapore, 2014, pp. 1458–1462.