Gender and Dialect Bias in YouTube's Automatic Captions

Rachael Tatman
Department of Lingusitics
University of Washington
rctatman@uw.edu

Abstract

This project evaluates the accuracy of YouTube's automatically-generated captions across two genders and five dialects of English. Speakers' dialect and gender was controlled for by using videos uploaded as part of the "accent tag challenge", where speakers explicitly identify their language background. The results show robust differences in accuracy across both gender and dialect, with lower accuracy for 1) women and 2) speakers from Scotland. This finding builds on earlier research finding that speaker's sociolinguistic identity may negatively impact their ability to use automatic speech recognition, and demonstrates the need for sociolinguistically-stratified validation of systems.

1 Introduction

The overall accuracy of automatic speech recognition (ASR) has increased substantially over the past decade: a decade ago it was not uncommon to report a ASR error rates of 27% (Sha and Saul, 2007), while a recent Microsoft system achieved a word error rate (WER) of just 6.3% on the Switchboard corpus (Xiong et al., 2016). Have these strong gains benefited all speakers evenly? Previous work, briefly discussed below, has found systematic bias both by dialect and gender. This paper provides additional evidence that sociolinguistic variation continues to provide a source of avoidable error by showing that the WER is robustly different for male and female native English speakers from different dialect regions.

It is well established in the field of sociolinguistics that there is quantifiable variation in language use between social groups. Gender-based variation in language use, for example, has been extensively studied (Trudgill, 1972; Eckert, 1989, among many others). There is also robust variation in language use by native speakers across dialect regions. For instance, English varies dramatically between the United States (Cassidy and others, 1985), New Zealand (Hay et al., 2008) and Scotland (Milroy and Milroy, 2014).

Sociolinguistic variation has historically been a source of error for natural language processing. Differences across genders in automatic speech recognition accuracy have been previously reported, with better recognition rates reported for both men (Ali et al., 2007) and women (Goldwater et al., 2010; Sawalha and Abu Shariah, 2013). Previous work has also found evidence of dialectal bias in speech recognition in both English (Wheatley and Picone, 1991) and Arabic (Droua-Hamdani et al., 2012). In addition, there are many anecdotal accounts of bias against dialect in speech recognition. For example, in 2010 Microsoft's Kinect was released and, while it shipped with Spanish voice recognition, it did not recognize Castilian Spanish (Plunkett, 2010). This study investigates whether YouTube's automatic captions have different WER for native English speakers across two genders and five dialect regions.

2 Method

Data for this project was collected by hand checking YouTube's automatic captions (Harrenstien, 2009) on the word list portion of accent tag videos. Annotation was done by a phonetically-trained listener familiar with the dialects in the study. YouTube's automatic captions were chosen for three reasons. The first is that they're backed by Google's speech recognition software, which is both very popular and among the more accurate

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 53–59,
Valencia, Spain, April 4th, 2017.

proprietary ASR systems (Liao et al., 2013). The second is the fact that the accuracy of YouTube's automatic captions specifically are an area of immediate concern to the Deaf community and is a frequent topic of (frustrated) discussion: they are often referred to as "autocraptions" (Lockrey, 2015) due to their low accuracy and the fact that content creators will often rely on them instead of providing accurate captions. Finally, YouTube's large, diverse userbase allowed for the direct comparison of speakers from a range of demographic backgrounds.

2.1 Accent tag

The accent tag, developed by Bert Vaux and based on the Harvard dialect survey (Vaux and Golder, 2003), has become a popular and sustained internet phenomenon. Though it was designed to elicit differences between dialect regions in the United States, it has achieved wide popularity across the English-speaking world. Variously called the "accent tag", "dialect meme", "accent challenge" or "Tumblr/Twitter/YouTube accent challenge", videos in this genre follow the same basic outline. First, speakers introduce themselves and describe their linguistic background, with a focus on regional dialect. Then speakers read a list of words designed to elicit phonological dialect differences. Finally, speakers read and then answer a list of questions designed to elicit lexical variation. For example, one question asks "What do you call gym shoes?", which speakers variously answered "sneakers", "tennis shoes", "gym shoes" or whatever the preferred term is in thier dialect.

This study focuses on only the word list portion of the accent tag. Over time, the word list has been changed and appended, most notably with terms commonly used in on-line communities such as "GPOY" (gratuitous picture of yourself) or "gif" (graphics interchange format, a popular digital image format). Even with these variations, all videos discussed here used some subset of the word-list shown in Table 1.

It should be noted that this is a particularly difficult ASR task. First, words are presented in isolation rather than within a frame sentence, which means that ASR systems cannot benefit from the use of language models. Second, the word-list portion of the accent tag challenge was intentionally constructed to only include words with multiple possible pronunciations and that serve as dialect markers. "Lawyer", for example, is generally pronounced [lɔɪ.jəˠ] in New England and California, but [lɔ.jəˠ] in Georgia (Vaux and Golder, 2003). These facts do place the ASR system used to generate the automatic captions at a disadvtange, and may help to explain the high error rates.

Again	Envelope	Potato
Alabama	Figure	Probably
Aluminum	Fire	Quarter
Arizona	Florida	Roof
Ask	Gif	Route
Atlantic	GPOY	Ruin
Attitude	Guarantee	Salmon
Aunt	Halloween	Sandwich
Auto	Image	Saw
Avocado	Iron	Spitting
Bandanna	Lawyer	Sure
Both	Marriage	Syrup
Car	Mayonnaise	Theater
Caramel	Muslim	Three
Catch	Naturally	Tomato
Caught	New Orleans	Twenty
Cool Whip	Officer	Waffle
Coupon	Oil	Wagon
Crayon	Oregon	Wash
Data	Pajamas	Water
Eleven	Pecan	

Table 1: Word list for accent tag videos.

2.2 Speakers

A total of eighty speakers were sampled for this project. Videos for eight men and eight women from each dialect region were included. The dialect regions were California, Georgia, New England (Maine and New Hampshire), New Zealand and Scotland. These regions were chosen based on their high degree of geographic separation from each other, distinct local regional dialects and (relatively) comparable populations. Of these regions, California has the largest population, with approximately 38.8 million residents, and New England the smallest, with Maine and New Hampshire having a combined population of approximately 2.6 million (although the United States census bureau estimates the population of New England as a region at over 14 million as of 2010 (Bogue et al., 2010)).

Sampling was done by searching YouTube using the exact term "accent challenge" or "accent

tag" and the name of the geographical region. Only videos which had automatic captions were included in this study. For each speaker, the word error rate (WER) was calculated separately. Data and code used for analysis is available online[1].

3 Results

The effect of dialect and gender on WER was evaluated using liner mixed-effects regression. Both speaker and year were included as random effects. Speaker was included to control for both individual variability in speech clarity and also recording quality, since only one recording per speaker was used. Year was included to control for improvements in ASR over time. Automatic captions are generated just after the video is uploaded to YouTube, and the recordings used were uploaded over a five year period, so it was important to account for overall improvements in speech recognition.

A model which included both gender and dialect as fixed effects more closely fit the data (i.e. had a lower Akaike information criterion) than nested models without gender (χ^2 (5, N= 80) = 31 p <0.01), without dialect (χ^2 (5, N= 80) = 14, p <0.01) or without either (χ^2 (5, N= 80) = 31, p <0.01). In terms of dialect, speakers from Scotland had reliably worse performance than speakers from the United States or New Zealand, as can be seen in Figure 1. The lower level of accuracy for Scottish English can not be explained by, for example, a small number of speakers of that variety. The population of New Zealand, the dialect which had the second-lowest WER, is roughly 80% that of Scotland. Nor is it factor of wealth. Scotland and New Zealand have a GDP *per capita* that falls within one hundred US dollars of each other.

There was also a significant effect of gender: the word error rate was higher for women than men (t(78) = -3.5, p <0.01). This is shown in Figure 2. This is somewhat surprising given earlier studies which found the opposite result (Goldwater et al., 2010; Sawalha and Abu Shariah, 2013).

In addition, there was an interaction between gender and dialect. Adding an interaction term between gender and dialect to the model above significantly improved model fit (χ^2 (5, N= 80) = 16, p <0.01). As can be seen in Figure 3, the effect of gender was not equal across dialects. Differences between genders were largest for speakers

[1]https://github.com/rctatman/youtubeDialectAccuracy

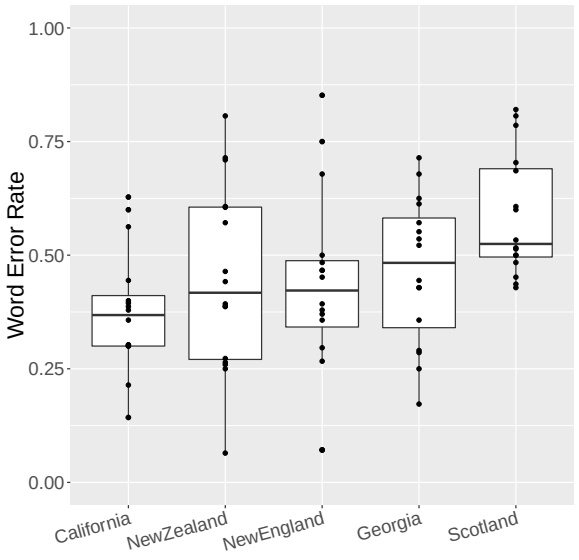

Figure 1: YouTube automatic caption word error rate by speaker's dialect region. Points indicate individual speakers.

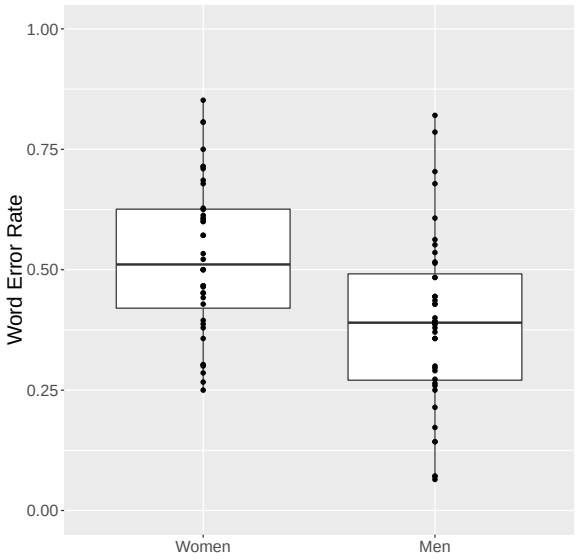

Figure 2: YouTube automatic caption word error rate by speaker's gender. Points indicate individual speakers.

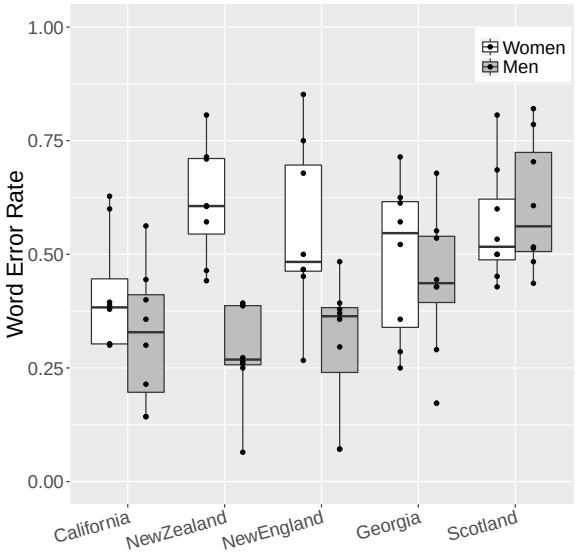

Figure 3: Interaction of gender and dialect. The difference in Word Error Rates between genders was largest for speakers from New Zealand and New England. In no dialect was accuracy reliably better for women than men.

from New Zealand and New England.

Given the nature of this project, there is limited access to other demographic information about speakers which might be important, such as age, level of education, socioeconomic status, race or ethnicity[2]. The last is of particular concern given recent findings that automatic natural language processing tools, including language identifiers and parsers struggle with African American English (Blodgett et al., 2016).

4 Effects of pitch on YouTube automatic captions

One potential explanation for the different error rates found for male and female speakers is differences in pitch. Pitch differences are one of the most reliable and well-studied perceptual markers of gender in speech (Wu and Childers, 1991; Gelfer and Mikos, 2005) and speech with a high fundamental frequency (typical of women's speech) has also been found to be more difficult for automatic speech recognizers (Hirschberg et al., 2004; Goldwater et al., 2010). A small experiment was carried out to determine whether pitch differences were indeed underlying the differing word error rates for male and female speakers.

First, a female speaker of standardized American English was recorded clearly reading the word list shown in Table 1. In order to better approximate the environment of the recordings in the accent tag videos, the recording was made using a consumer-grade headset microphone in a quiet environment, rather than using a professional-grade microphone in a sound-attenuated booth. The original recording had a mean pitch of 192 Hz and a median of 183 Hz, which is slightly lower than average for a female speaker of American English (Pépiot, 2014). The pitch of the original recording was artificially scaled both up and down 60 Hz in 20 Hz intervals using Praat (Boersma and others, 2002). This resulted in a total of seven recordings: the original, three progressively lower pitched and three progressively higher pitched. These resulting sound-files were then uploaded to YouTube and automatic captions were generated. The video, and captions, can be viewed on YouTube[3].

Overall, the automatic captions for the word list were very accurate; there were a total of 9 errors across all 434 tokens, for a WER of .002. Though it may be due to ceiling effects, there was no significant effect of pitch on accuracy. The much higher accuracy of this set of captions may be due to improvement in the algorithms underlying the automatic captions or the nature of the speech in the recording, which was clear, careful and slow. More investigation with a larger sample of voices is necessary to determine if pitch differences, or perhaps another factor such as intensity, are what is underlying the differences in WER for male and female speakers. That said, even if gender-based differences in accuracy between genders can be attributed to acoustic differences associated with gender, that would not account for the strong effect of dialect region.

5 Discussion

The results presented above show that there are differences in WER between dialect areas and genders, and that manipulating one speaker's pitch was not sufficient to affect WER for that speaker. While the latter needs additional data to form a robust generalization, the size of the effect for the former is deeply disturbing. Why do these

[2]Speakers in this sample did not self-report their race or ethnicity and, given the complex nature of race and ethnicity in both New Zealand and the US, the researcher opted not to guess at speaker' race and ethnicity.

[3]https://www.YouTube.com/watch?v=eUgrizlV-R4

differences exist? From a linguistics standpoint, no dialect is inherently more or less intelligible. The main factor which determines how well a listener understands a dialect is the amount of exposure they have had to it (Clarke and Garrett, 2004; Sumner and Samuel, 2009); with sufficient exposure, any human listener can learn any language variety. In addition, earlier research that found lower WER for female speakers shows that creating such ASR systems is possible (Goldwater et al., 2010; Sawalha and Abu Shariah, 2013). Given that there is also a difference between dialects, these differences are most likely due to something besides the inherent qualities of the signal.

One candidate for the cause of these differences is imbalances in the training dataset. Any bias in the training data will be embedded in a system trained on it (Torralba and Efros, 2011; Bock and Shamir, 2015). While the system behind YouTube's automatic captions is propriety and it is thus impossible to validate this supposition, there is room for improvement in the social stratification of many speech corpora. Librevox, for example, is a popular open-source speech data set that "suffers from major gender and per speaker duration imbalances" (Panayotov et al., 2015). TIMIT, the most-distributed corpora available through the linguistic data consortium, is balanced for speaker dialect but approximately 69% of the speech in it comes from male speakers (Garofolo et al., 1993). Switchboard (Godfrey et al., 1992) undersamples women, Southerns and non-college-educated speakers. Many other popular speech corpora such as the Numbers corpus (Cole et al., 1995) or the AMI meeting corpus (McCowan et al., 2005) don't include information on speaker gender or dialect background. Taken together, these observations suggest that socially stratified sampling of spekaers has historically not been the priority during corpus construction for computational applications.

One solution to imbalanced training sets to focus on collecting unbiased socially stratified samples, or at the very least documenting the ways in which samples are unbalanced, for future speech corpora. This is already being addressed in the data collection of some new corpora such as the Automatic Tagging and Recognition of Stance (ATAROS) corpus (Freeman et al., 2014).

This does not help to address existing imbalances in training data, however. One way of doing this is to include information about speaker's social identity, such as the geographic location of the speaker (Ye et al., 2016) or using gender-dependent speech recognition models (Konig and Morgan, 1992; Abdulla and Kasabov, 2001).

Regardless of the method used to correct biases, it is imperative that the NLP community work to do so. Robust differences in accuracy of automatic speech recognition based on a speaker's social identity is an ethical issue (Hovy and Spruit, 2016). In particular, if NLP and ASR systems consistently preform worse for users from disadvantaged groups than they do for users from privileged groups, this exacerbates existing inequalities. The ideal would be for systems to preform equally well for users regardless of their sociolinguistic backgrounds.

Differences in performance derived from speakers' social identity is particularly concerning given the increasing use of speech-analysis algorithms during the hiring process (Shahani, 2015; Morrison, 2017). Given the evidence that speech analysis tools preform more poorly on some speakers who are members of protected classes, this could legally be discrimination (Ajunwa et al., 2016). Error analyses that compare performance across sociolinguistic-active social groups, like the one presented in this paper, can help ensure that this is not the case and highlight any imbalances that might exist.

Acknowledgments

The author would like to thank Richard Wright, Gina-Anne Levow, Alicia Beckford Wassink, Emily M. Bender, the University of Washington Computational Linguistics Laboratory and the reviewers for their helpful comments and support. Any remaining errors are mine. This work has been supported by supported by National Science Foundation grant DGE-1256082.

References

Waleed H. Abdulla and Nikola K. Kasabov. 2001. Improving speech recognition performance through gender separation. *changes*, 9:10.

Ifeoma Ajunwa, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*.

S. Ali, K. Siddiqui, N. Safdar, K. Juluru, W. Kim, and E. Siegel. 2007. Affect of gender on speech

recognition accuracy. In *American Journal of Roentgenology*, volume 188. American Roentgen Ray Society.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *arXiv preprint arXiv:1608.08868*.

Benjamin Bock and Lior Shamir. 2015. Assessing the efficacy of benchmarks for automatic speech accent recognition. In *Proceedings of the 8th International Conference on Mobile Multimedia Communications*, pages 133–136. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Donald J. Bogue, Douglas L. Anderton, and Richard E. Barrett. 2010. *The population of the United States*. Simon and Schuster.

Frederic Gomes Cassidy et al. 1985. *Dictionary of American Regional English*. Belknap Press of Harvard University Press.

Constance M. Clarke and Merrill F. Garrett. 2004. Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6):3647–3658.

Ronald A. Cole, Mike Noel, Terri Lander, and Terry Durham. 1995. New telephone speech corpora at CSLU. In *Eurospeech*. Citeseer.

Ghania Droua-Hamdani, Sid-Ahmed Selouani, and Malika Boudraa. 2012. Speaker-independent asr for modern standard arabic: effect of regional accents. *International Journal of Speech Technology*, 15(4):487–493.

Penelope Eckert. 1989. The whole woman: Sex and gender differences in variation. *Language variation and change*, 1(03):245–267.

Valerie Freeman, Julian Chan, Gina-Anne Levow, Richard Wright, Mari Ostendorf, Victoria Zayats, Yi Luan, Heather Morrison, Lauren Fox, Maria Antoniak, et al. 2014. ATAROS Technical Report 1: Corpus collection and initial task validation. *U. Washington Linguistic Phonetics Lab*.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. TIMIT acoustic-phonetic continuous speech corpus. *Linguistic data consortium, Philadelphia*, 33.

Marylou Pausewang Gelfer and Victoria A. Mikos. 2005. The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice*, 19(4):544–554.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Ken Harrenstien. 2009. Automatic captions in YouTube. *The Official Google Blog*, 11.

Jennifer Hay, Margaret Maclagan, and Elizabeth Gordon. 2008. *New Zealand English*. Edinburgh University Press.

Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1):155–175.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 591–598.

Yochai Konig and Nelson Morgan. 1992. Gdnn: A gender-dependent neural network for continuous speech recognition. In *Neural Networks, 1992. IJCNN., International Joint Conference on*, volume 2, pages 332–337. IEEE.

Hank Liao, Erik McDermott, and Andrew Senior. 2013. Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 368–373. IEEE.

Michael Lockrey. 2015. YouTube automatic craptions score an incredible 95% accuracy rate! *medium.com*, July. [Online; posted 25-July-2015].

Iain McCowan, Jean Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

James Milroy and Lesley Milroy. 2014. *Real English: the grammar of English dialects in the British Isles*. Routledge.

Lennox Morrison. 2017. Speech analysis could now land you a promotion. *BBC News*, Jan.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Erwan Pépiot. 2014. Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. In *Speech Prosody 7*, pages 305–309.

Luke Plunkett. 2010. Report: Kinect Doesn't Speak Spanish (It Speaks Mexican). September.

M. Sawalha and M. Abu Shariah. 2013. 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.

Fei Sha and Lawrence K. Saul. 2007. Large margin hidden Markov models for automatic speech recognition. *Advances in neural information processing systems*, 19:1249.

Aarti Shahani. 2015. Now algorithms are deciding whom to hire, based on voice. *All Tech Considered: Tech, culture and connection*, March.

Meghan Sumner and Arthur G. Samuel. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4):487–501.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE.

Peter Trudgill. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in society*, 1(02):179–195.

Bert Vaux and Scott Golder. 2003. The Harvard dialect survey. *Cambridge, MA: Harvard University Linguistics Department*.

Barbara Wheatley and Joseph Picone. 1991. Voice Across America: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, 1(2):45–63.

Ke Wu and Donald G. Childers. 1991. Gender recognition from speech. part i: Coarse analysis. *The journal of the Acoustical society of America*, 90(4):1828–1840.

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

Guoli Ye, Chaojun Liu, and Yifan Gong. 2016. Geolocation dependent deep neural network acoustic model for speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5870–5874. IEEE.