Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation

Nina Markl
Institute for Language,
Cognition and Computation
The University of Edinburgh
nina.markl@ed.ac.uk

Catherine Lai
Centre for Speech Technology Research
Linguistics and English Language
The University of Edinburgh
c.lai@ed.ac.uk

Abstract

Commercial Automatic Speech Recognition (ASR) systems tend to show systemic predictive bias for marginalised speaker/user groups. We highlight the need for an interdisciplinary and context-sensitive approach to documenting this bias incorporating perspectives and methods from sociolinguistics, speech & language technology and human-computer interaction in the context of a case study. We argue evaluation of ASR systems should be disaggregated by speaker group, include qualitative error analysis, and consider user experience in a broader sociolinguistic and social context.

1 Introduction

Automatic Speech Recognition (ASR) has become a common tool in human-computer interaction, enabling, for example, voice user interfaces and (imperfect) automatic captioning of multimedia content. As with other language technologies (e.g. Sap et al., 2019; Blodgett and O'Connor, 2017), rapid improvements in performance have not been equal for different user groups. As Blodgett et al. (2020) show, discussions of this "bias" are often poorly defined, not grounded in explicit normative judgments and divorced from socio-historical contexts, origins and harms of the system behaviours. In this paper, we argue that researchers at the intersection of speech and language technologies (SLT), human-computer interaction (HCI), and sociolinguistics are well-placed to consider the experiences and social context of different speaker/user groups in critical quantitative and qualitative evaluations of ASR systems. Knowledge about language variation and its relation to society coupled with expertise from HCI allows us to understand how predictive biases reflect larger social structures and ideologies about language, and how they affect users.

After presenting prior work on language variation and ASR, a case study of self-recorded audio diaries collected for the Lothian Diary Project[1] highlights the need for a context-sensitive approach to ASR evaluation which we outline.

2 Language variation, bias and ASR

Blodgett et al. (2020)'s critique notwithstanding, predictive bias, defined here as error and outcome disparities for different user groups (Shah et al., 2020), has become a research focus in SLT and other machine learning fields as applications are extended to high-stakes contexts such as hiring, policing and banking where they have been shown to (re)produce structural inequalities (see e.g. Benjamin, 2019). Predictive bias also appears to be prevalent in commercial ASR systems for English[2].

Recent work describes stark racial bias in commercial American English ASR systems (including Google's Cloud Speech) (Koenecke et al., 2020), with much higher word error rates (WER)[3] for speakers of African American English (AAE) than white speakers of (Californian) American English. Notably, these types of error disparities appear to be driven by under-representation of AAE training data both for the acoustic modelling (Koenecke et al., 2020) and the language model used to decode sequences of phones into utterances (Martin and Tang, 2020). "Regional" variation has also been reported as a source of unequal performance, with particularly high error rates reported on YouTube's captions for speakers from Scotland and (the US state) Georgia (Tatman, 2017). Similar to more recent work, YouTube captions have been found to perform worse for African American speakers (Tatman and Kasten, 2017). These problems are

[1] https://lothianlockdown.org/

[2] The focus here is on English, but predictive bias is likely to affect stigmatised and unstandardised varieties vis-a-vis standardised varieties of other languages too.

[3] WER is an edit-distance measure capturing the number of deletions, substitutions and insertions required per word to match a reference transcript.

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 34–40

not limited to proprietary systems, as Mozilla's open source system DeepSpeech performs significantly worse for speakers of Indian English than "American English"[4] (Meyer et al., 2020), and also fails to transcribe AAE morpho-syntactic variation correctly (Martin and Tang, 2020). While some early research has suggested ASR performance differences based on (binary) speaker gender (Adda-Decker and Lamel, 2005; Benzeghiba et al., 2007; Tatman, 2017), it is unclear that gender by itself is a significant factor in recent systems (Tatman and Kasten, 2017; Meyer et al., 2020). Koenecke et al. (2020) suggest that the interaction of gender and race is significant, with differences between Black men and Black women being more significant than between white men and white women or men and women across race[5]. These results appear to be linked to speaker's speech styles (e.g. in Adda-Decker and Lamel, 2005) and use of dialect features (Koenecke et al., 2020), both of which have long been documented to pattern with gender (see Labov, 1990, for a classic paper) and could be correlated with gender in training and test sets. Other work in this space has focused on the potential of ASR to improve accessibility of audio media and digital technologies, looking at experiences of Deaf and hard of hearing users (Glasser, 2019) and dysarthric speakers (De Russis and Corno, 2019; Young and Mihailidis, 2010). For both groups commercial ASR systems perform quite poorly, though the severity and amount of errors varies by speaker. Research on predictive bias in commercial ASR for regional varieties of English beyond the United States and in the context of systems not exclusively trained on American English, as well as experiences of second language learners of English, and other groups who are potentially particularly reliant on ASR to access computing technologies such as elderly people, is sparse.

From a linguistic perspective, no language variety or speech style is inherently more difficult, incorrect, or inappropriate than any other. There are, however, powerful ideologies regarding the relative status of different varieties and styles which are rooted in broader socio-historical contexts and reflect the social status of the groups who speak them (Woolard and Schieffelin, 1994). In addition to being stigmatised in "traditional" contexts of power in society, varieties spoken by marginalised communities appear to be (not coincidentally) underrepresented in the data we use to build and evaluate speech technologies, leading to substantial predictive biases making speech technologies less accessible to already marginalised groups.

3 Lothian Diaries: A case study

The Lothian Diary project is an ongoing interdisciplinary research project inviting residents of the Lothians region of Scotland to contribute self-recorded audio and video diaries about their experiences of the COVID-19 pandemic. The more than 120 diaries collected so far are highly variable in recording quality, number of speakers and topics discussed, and participants are diverse[6] in terms of age, gender, linguistic background, ethnicity, socio-economic class and level of education. Edinburgh and the surrounding Lothians region are of particular interest for sociolinguistic research because of the capital region's status as a centre for higher education, finance and tourism. In addition to the variation within Scottish English[7] between different areas and different socio-economic groups within the city, there is also a wide range of other first and second language varieties of English, as well as other languages. The Lothian Diary project also includes many of these other varieties of English, rather than focusing on speakers with long residential histories in a particular area (as is often the case in sociolinguistic work) or first language speakers (as is usually the case in SLT evaluation). The recordings form a highly naturalistic and exceptionally varied data set. ASR is used here to facilitate social science research which requires accurate and complete transcriptions (achieved through manual correction).

So far, 13 diaries submitted by participants who agreed to have them made public, have been processed with the Google Cloud Speech-to-Text API[8] (GC STT). Diaries (16 kHz FLAC files) were processed in their entirety using the model used for long audio files which uses asynchronous speech recognition. WER was computed separately for

[4] Meyer et al. (2020)/Mozilla do not specify speaker race or region within the US.

[5] A finding which echoes work in other ML domains and other areas of SLT highlighting the way that multiple demographic axes linked to interacting structures of oppression (e.g. gender and race) cannot be considered separately (Buolamwini and Gebru, 2018; Jiang and Fellbaum, 2020)

[6] though not representative of the Scottish population

[7] "Scottish English" is used here as a broad term including the continuum between Scots and Scottish Standard English (see Stuart-Smith, 2004)

[8] https://cloud.google.com/speech-to-text

ID	G	Variety	WER
RF	F	Scottish English	46.4
La	F	Scottish English	35.7
CE	F	Scottish English	20.9
AA	M	Scottish English	29.8
MR	M	Scottish English	55.3
DL	M	Scottish English/Scots	88.9
Li	F	Southern British English	29.5
JW	M	Canadian English*	25.3
L	F	L2 English (L1: Lithuanian)	31.5
S	F	L2 English (L1: Cantonese)	27.7
MG	F	L2 English (L1: Italian)	35.3
JL	M	L2 English (L1: Filipino)	40.8
A	M	L2 English (L1: Chinese)	70.2

Table 1: Word Error Rates for different participants vary widely both across and within groups (lower is better). *Decoded using 'en-US' language option, for all others 'en-GB' was used

each speaker using `sclite`[9]. In the following section, we present a brief qualitative error analysis.

WER for individual speakers varies dramatically (see Table 1). Some of these errors appear to be related to accent differences. For example, Scottish speakers' pronunciations of *I* or *I've* are frequently mistranscribed as *ah* or *of* and other accent-based errors include: *cat* [ka?] > *car*, *living* > *leaving*, *hating our* > *heating are*. However, there is also significant variation within each accent group. GC STT fails to transcribe filled pauses (*uh, um*) and word fragments and occasionally deletes false starts and repetitions. Furthermore, errors appear to be more prevalent in the vicinity of hesitations and repetitions. As a result speakers who produce more hesitations and repetitions tend to have higher error rates, while people who appear to read from prepared notes tend to be more fluent and have lower error rates. The highest WER in this sample derives from a recording by a Scottish English speaker who produces many false starts, word fragments and a number of Scots words (which the system likely would not recognise under any circumstances). Words are also often substituted by a wrong (but often grammatically appropriate) inflectional form (e.g. past tense > present tense).

All of these errors are particularly challenging for the accurate and complete transcription of spontaneous and conversational speech, especially for social science research where researchers (users) might consider hesitations, false starts and filled pauses important as they convey pragmatic information. Considering impacts of this predictive bias, transcripts of speakers who produce more "fluent" speech are much more easily interpretable. Retrieving speech content and speech style of less fluent speakers as well as some second language speakers, on the other hand, requires more labour and time, potentially negating any benefits of ASR.

[9] https://github.com/usnistgov/SCTK

4 Proposed methods

To document predictive bias in ASR in a way that is mindful of 1) user experience, 2) socio-historical and (socio)linguistic context, 3) (potential) harms (re)produced by the system, and 4) technical aspects of ASR, we need to draw on methodologies and knowledge from HCI, sociolinguistics, research on fairness in AI, and SLT.

4.1 Intersectional benchmarks

ASR systems are usually evaluated in terms of their WER, for one or more unseen test sets (often including well-established benchmark sets). As seen in the case study above, word error rates vary strongly across individual recordings and speakers, and (benchmark) test sets (e.g. Barker et al., 2017) are becoming increasingly naturalistic and (potentially) diverse; a recent state-of-the-art system by Google (Chiu et al., 2018) was trained and tested on "representative" data drawn from Google's voice-search traffic. However, even assuming that the test sets are representative of the developer's users, it is 1) not clear that the intended or current user base is reflective of all use cases or potential users (especially if the system is sold to third parties as with GC STT), and 2) possible or even likely that significant variation in performance between user groups is hidden by reporting an average across all tested recordings. Importantly, as Black feminist scholarship has pointed out, multiple demographic axes linked to interlocking structures of oppression (e.g. race and gender) cannot be considered separately (Crenshaw, 1991). It is thus important that in addition to disaggregating by language variety to also consider, for example, gender to create an "intersectional" benchmark (see also Costanza-Chock, 2020). This approach has been successful in highlighting disproportionate predictive bias for particular subgroups in other ML domains (e.g. darker-skinned women in facial analysis: Buolamwini and

Gebru, 2018; Raji and Buolamwini, 2019), and SLT (Jiang and Fellbaum, 2020).

To apply an intersectional benchmark to a larger sample of the Lothian Diaries, we intend to match short audio snippets with the same reference transcript produced by different speaker groups to isolate pronunciation effects, and look systematically at potential differences in content and speech style (following Koenecke et al., 2020).

4.2 Qualitative error analysis

Intersectional benchmarks alone are not enough however, as WER does not account for the context or effect of an error. Understanding the context of errors is useful since errors are both more likely to occur and to be severe in particular phonetic, prosodic and lexical contexts. Like us (though working with a very different system and data), Goldwater et al. (2010) find that words before or after hesitations, repetitions and word fragments, turn-initial words and infrequent words are more likely to be misrecognised and that erroneous substitutions are often different forms of the same lexeme (e.g. *ask/asked*). While some of these errors can be easily disambiguated through context, others (e.g. *can/can't*) could be quite disruptive to communication. Word errors can also lead to domino effects, where one wrongly decoded word feeds into further erroneous predictions (Martin and Tang, 2020). While metrics which are more sensitive to the type and context of the error or directly model human evaluations have been proposed (Nanjo and Kawahara, 2005; Morris et al., 2004; Mishra et al., 2011; Kafle and Huenerfauth, 2020) they are not widely adopted and extensive qualitative error analysis is rare. A context-sensitive approach would be particularly interested in the type of error and its effect given the linguistic context.

4.3 User experience

Evaluations of SLT systems rarely reflect explicitly on how users interact with them[10]. However, because both (perceived) severity and impact as well as prevalence of errors depends on recording and task, understanding how people use ASR-based technologies in their daily life is important. Future work concerning predictive bias in ASR would benefit from incorporating HCI methodologies like interviews, ethnography and qualitative surveys to gain a deeper understanding of users' experiences. So far, researchers in HCI have been particularly interested in how people interact with voice user interfaces (e.g. Porcheron et al., 2018; Luger and Sellen, 2016), though little attention has been paid to the role of accent and dialect. Furthermore, especially given the context of the recent shift to increased remote work and education, applications of cloud-based speech recognition for personal or business use extend beyond voice user interfaces to automatic captioning of audio and video lectures and meetings. Domain-general and naturalistic recordings of continuous spontaneous speech pose a particular challenge to ASR systems, and insights into what types of errors users perceive to be particularly disruptive and common depending on their linguistic and demographic background should inform development and evaluation of ASR systems. For example, in the context of the Lothian Diary Project the goal of ASR is to produce transcriptions which can be used by linguists and other social science researchers to analyse both what participants are saying and how they are saying it. Every aspect of their speech, including disfluencies and repetitions as well as specific lexical choices (e.g. past tense vs present tense) are relevant to this analysis and should as such be preserved in a transcript. Furthermore, because most speech in this context is largely unplanned, higher error rates around disfluent or informal speech are particularly disruptive. When applying the proposed methodology to other use cases (e.g. automatic captioning of video lectures or business meetings) interviews with stakeholders can clarify what types of errors are particularly disruptive.

4.4 Considering context and impacts

Considering the broader societal context in which an ASR system is developed and implemented allows us to identify the specific harms it could inflict on users and (sometimes at least) see the underlying societal structures giving rise to predictive bias. Identifying risk and causes in turn allows us to mitigate harms (and, in future systems, bias). In the case of commercial ASR (in English), research suggests that predictive bias is a result of under-representation of varieties of marginalised speaker groups in proprietary training and test sets. For many open source and licensed corpora used to train and benchmark ASR systems, incomplete documentation makes it difficult to es-

[10] Though intended use is sometimes implicit in the choice of training and test data: e.g. Google's use of voice search data (Chiu et al., 2018)

timate representation; the commonly used Switchboard (Godfrey and Holliman, 1993) and TIMIT corpora (Garofolo et al., 1993) (both US English) and Mozilla's recent open-source Common Voice corpus[11], for example, do not record speaker race. The speaker characteristics of training sets depends on the broader societal context. For example, use of commercial speech recognition (e.g. in the case of Google's system) and participation in scientific studies (e.g. the licensed corpora) or crowd-source tasks (e.g. Mozilla Common Voice) differs across demographic groups (for example based on income and education). Imbalanced corpora are also tied to ideologies around whose ways of speaking are considered "legitimate", "correct" or "native".

Some of the more obvious specific harms of predictive bias include difficulties using voice user interfaces, which for some users are crucial assistive technology. As ASR spreads into high-stakes contexts such as hiring, substantial harms could be incurred if systems perform worse for already marginalised groups, effectively encoding "accentism" and linguistic prejudice in automatic systems. Even assuming no prediction bias across different speaker groups, the use of ASR in automatic analysis of video interviews to recommend or rank applicants (e.g. HireVue[12]) risks real harm in the case of even small recognition errors and potentially entrenches existing language ideologies around "professional", "fluent" or "competent" speech patterns. For example, *HireNet* (Hemamou et al., 2019) extracts information about prosody and speech fluency to predict "hireability" (as annotated by recruiters). Other harms include less usable automatic captions and potential downstream effects as described in our case study.

5 Conclusion

We have proposed an approach to ASR evaluation which considers the experiences of different user/speaker groups, sociolinguistic context and potential impacts of predictive bias. We argue that this interdisciplinary approach is necessary to significantly advance our understanding of ASR usability. We particularly invite perspectives from the fields of human-computer interaction in order evaluate speech and language technologies as systems situated in specific sociolinguistic and socio-technical contexts which perform specific tasks for specific (language) users.

[11] available here: https://commonvoice.mozilla.org/en/datasets

[12] https://www.hirevue.com/

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? *9th European Conference on Speech Communication and Technology*, (January 2005):2205–2208.

Jon P. Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2017. *The CHiME Challenges: Robust Speech Recognition in Everyday Environments*, pages 327–344. Springer International Publishing, Cham.

Ruha Benjamin. 2019. *Race after technology : abolitionist tools for the New Jim Code*. Polity Press, Newark.

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. 2007. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778.

Sasha Costanza-Chock. 2020. Design values: Hard-coding liberation? In *Design Justice*. MIT Press. Https://design-justice.pubpub.org/pub/3h2zq86d.

Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299.

Luigi De Russis and Fulvio Corno. 2019. On the impact of dysarthric speech on contemporary ASR cloud platforms. *Journal of Reliable Intelligent Environments*, 5(3):163–172.

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 LDC93S1.

Abraham Glasser. 2019. Automatic speech recognition services: Deaf and hard-of-hearing usability. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–6, New York, NY, USA. Association for Computing Machinery.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62.

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.

Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):573–581.

May Jiang and Christiane Fellbaum. 2020. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online). Association for Computational Linguistics.

Sushant Kafle and Matt Huenerfauth. 2020. Usability evaluation of captions for people who are deaf or hard of hearing. *ACM SIGACCESS Accessibility and Computing*, (122):1–1.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2(2):205–254.

Ewa Luger and Abigail Sellen. 2016. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, page 5286–5297, New York, NY, USA. Association for Computing Machinery.

Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proc. Interspeech 2020*, pages 626–630.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Taniya Mishra, Andrej Ljolje, Mazin Gilbert, Park Avenue, and Florham Park. 2011. Predicting Human Perceived Accuracy of ASR Systems. In *INTERSPEECH-2011*, August, pages 1945–1948.

Andrew C Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition. In *INTERSPEECH-2004*, pages 2765–2768.

H. Nanjo and T. Kawahara. 2005. A new asr evaluation measure and minimum bayes-risk decoding for open-domain speech understanding. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/1053–I/1056 Vol. 1.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. *Voice Interfaces in Everyday Life*, page 1–12. Association for Computing Machinery, New York, NY, USA.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing. In *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Jane Stuart-Smith. 2004. Scottish English. In Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar W. Schneider, and Clive Upton, editors, *A*

Handbook of Varieties of English, pages 47–67. Mouton de Gruyter, Berlin; Boston.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender race on accuracy of bing speech and youtube automatic captions. In *Proc. Interspeech 2017*, pages 934–938.

Kathryn A. Woolard and Bambi B. Schieffelin. 1994. Language ideology. *Annual Review of Anthropology*, 23:55–82.

Victoria Young and Alex Mihailidis. 2010. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112. PMID: 20698428.