Volume 61 http://acousticalsociety.org/

190th Meeting of the Acoustical Society of America

Philadelphia, Pennsylvania

11-15 May 2026

Speech Communication: Paper 2pSC10

The influence of transcriber and task factors on the quality of ASR datasets for minoritized language varieties

Rachel Hayes-Harb and Shannon Barrios
Department of Linguistics, University of Utah, Salt Lake City, UT, 84112, USA; r.hayes-harb@utah.edu; s.barrios@utah.edu

Neal Patwari and Maneesha Rani Saha
Kahlert School of Computing, University of Utah, Salt Lake City, UT, 84112, USA; neal.patwari@utah.edu; maneesha.saha@utah.edu

Automatic speech recognition (ASR) systems have become ubiquitous, and they provide an increasingly unavoidable means for human-computer interaction. Unfortunately, they typically perform better for speakers of historically favored languages and accents, with particularly inaccurate performance for so-called "non-native" accents. ASR performance depends on large and diverse speech corpora that include audio recordings with accurate labels, and solutions to ASR system bias typically call for larger datasets, assuming that more data from diverse speakers will reduce accent bias. However, more data alone does not solve the problem of disparate labeling errors for some accents. Our goal is to mitigate the biases that permeate ASR by exploring psycholinguistically-informed methods for dataset construction by examining the generalizability of listener and task factors that are known to modulate transcription accuracy in laboratory studies in contexts designed to resemble crowdsourced ASR dataset development. Over several experiments we have found mixed support for the hypotheses that transcription accuracy will be higher when: (1) talkers and listeners have similar first language backgrounds; (2) listeners self-identify as familiar with the talker's accent; and (3) audio recordings are presented in single-talker versus mixed-talker blocks.

Proceedings of Meetings on Acoustics, Vol. 61, 060002 (2026); https://doi.org/10.1121/2.0002361.

1. INTRODUCTION

Automatic speech recognition (ASR) systems have become ubiquitous, and provide an increasingly unavoidable means for human-computer interaction. Unfortunately, ASR services have disparate performance, typically performing better for speakers of historically favored languages (Sprugnoli et al., 2017) and accents (Markl, 2022), and for speakers with so-called "native" (as opposed to "non-native") accents (DiChristofano et al., 2023).

ASR performance depends on the availability of diverse and large high-quality speech corpora that include audio recordings of speech and their accurate transcriptions (also known as "labels" or "ground truth"). Sufficient transcription data remains critical in ASR, despite the success of semi-supervised approaches to reduce the required quantity of labeled data (Synnaeve et al., 2019; Radford et al., 2023). Generally, transcriptions for ASR datasets are obtained by crowdsourcing rather than by "expert" transcribers due to the expense of paying for transcriptions. Crowdsourced transcription can reduce the time and cost of transcription (Sprugnoli et al., 2017), and producing open datasets with low barriers to participation as a speaker or transcriber has the potential to increase the linguistic diversity represented in datasets. However, crowdsourced transcription tends to be of lower quality than expert transcription, and despite its promise, "the viability of methods for crowdsourcing speech transcription significantly depends on the target language" (Sprugnoli et al., 2017, p. 284).

A growing body of research on ASR systems documents bias in both the datasets and the performance of ASR systems. This work often leads to proposed solutions, including calls for greater diversity of speakers (Koenecke et al., 2020; Maison & Estève, 2023). For example, in response to criticisms of accent bias in Alexa, Amazon claimed that "as more people speak to Alexa, and with various accents, Alexa's understanding will improve" (Harwell, 2018). However, it is not clear, even if their products collect a more diverse set of speech samples, that the "accent gap" will close, since little work has addressed inaccuracies in the labeling of collected audio speech samples. In fact, research rarely acknowledges transcribers or familiarity with diverse accents, which can lead to disparate labeling errors for accents that transcribers are less familiar with (Prinos et al., 2024). In short, larger quantities of speech samples from minoritized language varieties are insufficient alone to reduce accent bias if transcription errors are higher for those samples. The goal of our ongoing project is to mitigate the biases that permeate ASR by exploring psycholinguistically-informed methods for high-quality dataset construction for marginalized accents.

A. THE ROLE OF THE LISTENER IN TRANSCRIPTION ACCURACY

Similarities between the linguistic backgrounds of speakers and listeners can support listening comprehension speed and accuracy. For example, Adank et al. (2009) found a listening comprehension advantage for speakers of so-called Standard British English listening to other speakers with their same accent compared to Glaswegian English-accented speech. Other work has demonstrated that language learners can be even more accurate than first language speakers when transcribing the speech of other language learners with whom they share a first language background. For example, Xie and Fowler (2013) found that listeners whose first language is Mandarin outperformed those whose first language is English when identifying words produced in Mandarin-accented English. However, evidence for a benefit of a shared language background has been mixed, with some finding the benefit for some but not all pairs of listeners and talkers (Munro et al., 2006; Hayes-Harb et al., 2008).

In work in the field of computer science, researchers have independently hypothesized the existence of an intelligibility benefit, recommending that researchers "ensure that transcribers are fluent in the accent of the speakers whose speech they transcribe" (Prinos et al., 2024, p. 1251). However, one challenge associated with crowdsourced speech labeling is the collection of useful and accurate self-reported language background information from both speakers and transcribers. For example, Mozilla Common Voice (Ardila

et al., 2020) asks account holders (who may act as speakers and/or transcribers) to identify one or more languages, and for each language identified, are asked "How would you describe your accent?". The latter is an open-ended question that provides a number of vague pre-set response options (e.g., United States English, England English). Inspection of Mozilla Common Voice datasets (available at mozilladatacollective.com) reveals that contributors use a wide range of descriptors for their accents, reinforcing the challenge of neatly and accurately capturing an individual's relevant linguistic background. In the experiment reported here, we ask whether the assessment of broad similarities between self-reported language backgrounds of speakers and listeners that are permitted by such platforms are associated with an intelligibility benefit (observed as increased transcription accuracy) in a crowdsourced transcription task.

We are further interested in whether there are alternative options for quickly and easily assessing language experience that will predict individuals' accuracy transcribing particular accents. Kennedy and Trofimovich (2008) found that listeners' experience with the speech of second language learners of English in general predicted their transcription accuracy for learner-accented English speech. Alexander and Nygaard (2019) found evidence that the transcription accuracy benefit associated with learner-accented speech experience is to some extent accent-specific: participants trained on Korean- or Spanish-accented speech exhibited higher transcription accuracy at test for the accent they were trained on relative to the other accent. Given such findings, here we ask participants to report their familiarity with several learner accents of English, with the hypothesis that these self-assessments may provide an additional and/or alternative way to predict their ability to accurately transcribe speech samples.

B. TASK FACTORS AFFECTING TRANSCRIPTION ACCURACY

Psycholinguistics research shows that listeners rapidly and robustly adapt to unfamiliar accents over the course of exposure, whether that exposure occurs in daily life (Kennedy & Trofimovich, 2008) or in the context of a listening experiment (Clarke & Garrett, 2004; Maye et al., 2008; Xie et al., 2017). A second well-documented finding in psycholinguistics is that speech processing is slower and/or less accurate when talker variability is high, than when talker variability is limited (Mullennix et al., 1989; McLaughlin et al., 2024). Specifically, transcriptions are more accurate in single-talker conditions (with speech samples blocked by speaker) than in mixed-talker conditions (where speech samples from multiple talkers are mixed; Mullennix et al., 1989), especially when the accent is unfamiliar to the transcribers and/or the signal is degraded by noise (Tamati & Pisoni, 2014). Here we investigate whether adjustments to the organization of speech samples presented to transcribers impacts transcription accuracy in crowdsourced data collection.

C. RESEARCH QUESTIONS

The experiment described below is designed to answer the following research questions: (1) Is transcription accuracy higher when talkers and listeners have similar first language backgrounds?; (2) Is transcription accuracy higher when listeners self-identify as familiar with the talker's accent?; and (3) Is transcription accuracy higher when audio recordings are presented in single- versus multi-talker blocks?

2. EXPERIMENT

Study materials (except for the materials downloaded from the ALLSTAR database), analysis code, and anonymized data are available on the Open Science Framework at https://osf.io/d8xy9/.

A. PARTICIPANTS

Three hundred and seven participants were recruited via Prolific.co and and paid at a rate of 12 USD per hour to complete the $\sim$ 20-minute experiment. All participants met the Prolific profile criterion of self-

reported fluency in English. Sixty-five of these participants were recruited with the additional specification that their Prolific profile indicated Mandarin as the participant's first language. Thirty-three participants were excluded for missing one or more of three attention check questions; fourteen additional participants who were specifically recruited based on their Prolific profiles indicating Mandarin as a first language were excluded because they indicated a different first language in the pre-task questionnaire; and five more participants were excluded because they indicated both English and Mandarin as first languages, and they did not make a large enough group to separately analyze as an early multilingual Mandarin and English group. Data from the remaining 255 participants were included in the analyses. Participants were randomly assigned to one of four conditions: three single-talker conditions and one multi-talker condition. They are summarized in Table 1, along with age and sex information collected during a pre-task questionnaire.

Table 1: Participant characteristics by talker condition and first language.

Talker Condition	First Language	N	Mean Age	Female / Male / Self-describe
Single: M35	English	34	35	24/10/0
	Mandarin	15	34	12/2/1
	Other	14	37	6/8/0
Single: M39	English	42	33	24/18/0
	Mandarin	8	33	5/3/0
	Other	12	31	6/6/0
Single: M43	English	29	37	19/10/0
	Mandarin	16	31	6/10/0
	Other	23	33	12/11/0
Multi	English	33	41	18/15/0
	Mandarin	12	34	8/4/0
	Other	17	34	8/9/0

B. SPEECH MATERIALS

The speech materials were 60 of the Hearing in Noise Test sentences produced by three L1 Chinese Mandarin talkers (M35, M39, M43) from the ALLSTAR corpus (Bradlow, n.d.). These materials were selected because they provide a matched set of read sentences for all three speakers, have accompanying transcriptions, and represent speech materials from a marginalized variety of English. The audio files were mixed with speech-shaped noise (Demonte, 2019) at a signal-to-noise ratio of zero using a script (McCloy, 2013) using Praat (Boersma & Weenink, 2020) in order to avoid ceiling effects on transcription that might obscure effects of the variables of interest. In the single talker conditions, all 60 sentences were produced by one of the three talkers. In the multi-talker condition, we created three counterbalance lists where the 60 sentences were randomly assigned to the three talkers such that each talker produced 20 of the sentences. Pilot transcription data involving talker M39 suggested that 20 sentences was sufficient to observe adaptation to the talker. Participants in the multi-talker condition were randomly assigned to the three counterbalance lists.

C. PROCEDURES

From Prolific, participants were redirected to the experiment, which was conducted in PsychToolkit (Stoet, 2010, 2017). Participants were first presented with information about the study and upon consenting to participate, they proceeded to the pre-task questionnaire. The pre-task questionnaire asked participants to indicate their age and sex, as well as use of and familiarity with several languages and language varieties, including English and Mandarin, and also included items assessing their beliefs about language and accents. Items relating to languages other than Mandarin and English (e.g., Russian, Turkish, Korean, etc.) were included as distractors and are not analyzed here. Two items were included in the pre-task questionnaire as attention checks. The relevant pre-task questionnaire items are listed in Table 2.

Table 2: Pre-task questionnaire items. The associated variable names are presented in italics.

Item	Response Options
Is English your first language, or one of your first languages? – *firstLang*	Yes, English is my first language No, I speak English as a second language. My first language(s) is(are): [text response]
Do you speak Mandarin? – *firstLang*	Yes, Mandarin is my first language No, I don't speak Mandarin Yes, I speak Mandarin as a second language
Use the slider to indicate how familiar you are with Mandarin-accented English – *mandAccentFam*	Not at all familiar – Very familiar (0–10 scale)
I am good at understanding accents that are different from mine. – *goodAtAccents*	Strongly disagree – Strongly agree (0–10 slider)
I interact frequently with people whose accents are different from mine. – *interactAccents*	Strongly disagree – Strongly agree (0–10 slider)
For each statement, indicate your level of agreement. (see Note for the beliefs statements) – *beliefsAccentGoals, beliefsAccentDifficult*	Strongly disagree – Strongly agree (0–10 slider)

The four beliefs statements: (a) The goal of a language learner should be to pronounce the language like a native speaker; (b) The goal of a language learner should be to speak with a standard accent in the language; (c) Non-native accents make it hard for me to understand speech; and (d) Non-standard accents make it hard for me to understand speech.

In the sentence transcription task, participants first heard an example sentence produced by a speaker of English as a first language for the purpose of adjusting their headphone volume. Next, they completed a practice trial produced by the same talker to get accustomed to the transcription task. On this and each of the sixty test trials, participants heard a sentence and saw the instructions: "Type the sentence you hear as completely and accurately as possible. Press RETURN/ENTER when you are done. You have 10 seconds to respond." A cursor appeared on the screen and participants typed their response with a timeout of 10 seconds, after which point the task continued automatically to the next trial. There was a 750 ms inter-trial interval, and a participant-controlled break halfway through the task. The suitability of the trial structure was confirmed via pilot testing. Finally, participants completed a post-task questionnaire. The post-task questionnaire included one attention check item, in addition to the two questions presented in Table 3.

Table 3: Post-task questionnaire items. The associated variable names are presented in italics.

Item	Response Options
To the best of your ability, please describe the dialects/accents of the speakers you just listened to, for example, by country, region, language background – *postTalkerLang*	[text response]
How familiar are you with the accent of the speaker you just listened to? – *postAccentFam*	Not at all familiar – Very familiar (0–10 scale)

D. ANALYSIS AND RESULTS

All analyses were conducted in R (R Core Team, 2025) version 4.5.2. The categorical variable *firstLang* was coded using participants' responses to the pre-task questions concerning English and Mandarin. Participants were assigned to the following first language groups: English, Mandarin, Other. As noted above, participants who indicated both Mandarin and English as first languages were excluded from the analysis because they numbered only six, which was too few to be included as a separate first language group. The scalar variables *mandAccentFam*, *goodAtAccents*, and *interactAccents* are derived from the raw Likert scale response to the associated questions.

Responses to the four language and accent beliefs statements were analyzed using the `psych` package (Revelle, 2025). Responses were submitted to parallel analysis, indicating a two-factor solution. An exploratory factory analysis revealed that the statements relating to the accent goals of language learners (statements a and b) loaded strongly on one factor (λ=0.692 and 0.864, respectively), while the two statements relating to difficulty understanding accents (c and d) loaded strongly on a second factor (λ = 0.926 and 0.817, respectively). The two factors were weakly correlated (r = .17) and accounted for 69% of the total variance. Given this structure, we computed two separate composite scores (*beliefsAccentGoals* and *beliefsAccentDifficult*) by averaging responses to the two items within each factor.

Sentence transcription accuracy was computed using the `autoscore` package (Barrett et al., 2019) with the default settings. Proportion correct was computed for each sentence by dividing the number of correctly-transcribed words by the number of words in the target sentence. The categorical variable *postTalkerLang* was coded from open-ended text responses to the question asking participants to "describe the dialects/accents of the speakers you just listened to". Responses were assigned the value 1 if they contained the words "Chinese", "Mandarin", "China", "Hong Kong", and/or "Taiwan", and otherwise 0. The scalar variable *postAccentFam* was derived from the raw Likert scale response to the associated questionnaire item.

Figure 1 shows transcription accuracy by *talkerID* and *numTalkers*. In the single talker condition, the same talker is associated with all 60 trials. In the multi-talker condition, each talker is associated with 20 of the trials out of the full set of 60. Visual inspection of the figure indicates highly variable performance by individual participants, in addition to variation in accuracy among the three talkers.

To address our research questions, we fit several linear mixed-effects models (`lme4`; Bates et al., 2015), using maximum likelihood and the bobyqa optimizer. We first employed a model comparison approach to determine the growth model that best fit the shape of transcription accuracy change over time, following the recommendations in Nagle (2019). We created and scaled linear, quadratic, and cubic trial order variables, and then built separate unconditional linear, quadratic, and cubic models of log proportion correct over trial, including random intercepts for participant and target sentence. Model comparison using sequential likelihood ratio tests among these models as well as the null model revealed that the cubic model provided the best fit. We next created the full model from the cubic model, adding all variables as individual fixed

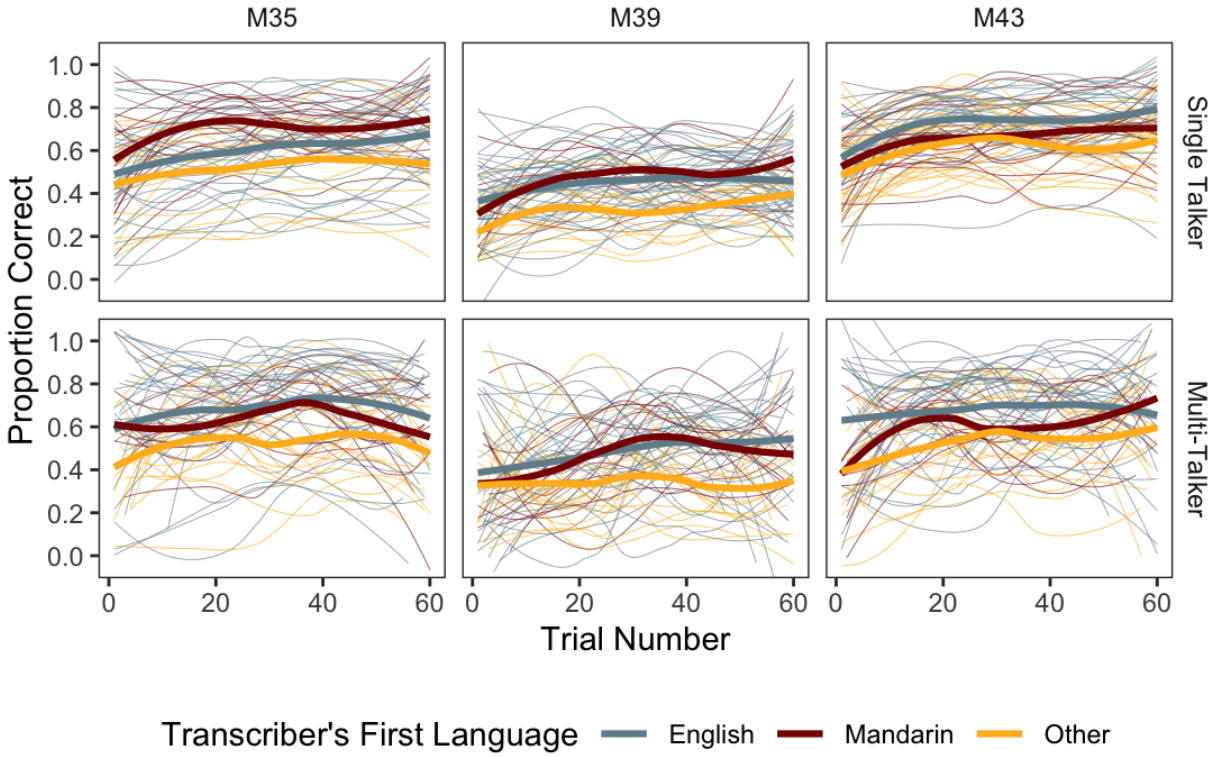

Figure 1: Smoothed (loess) trajectories of transcription accuracy (Proportion Correct) over time (Trial Number) for each participant (thin lines) and first language group (thick lines); colors indicate the first language of the transcriber (firstLang***). Facets indicate talker condition (columns:*** talkerID***; rows:*** numTalkers***).***

effects, in addition to the two-way interaction of *numTalkers* and *talkerID*. The model formula and Type III ANOVA analysis of the model results are provided in Table 4.

The significant quadratic and cubic order effects confirm that overall and as expected, participants adapted to the Mandarin-accented speech over the course of the transcription task. Concerning our first research question, we find a significant main effect of the first language of the transcriber. Follow-up pairwise comparisons using the `emmeans` package (Lenth & Piaskowski, 2025) revealed that the English group was significantly more accurate than both the Mandarin group (estimate=0.05, SE=.019, z=2.345, p=0.050) and the Other group (estimate=0.086 SE=0.014, z=6.350, p<0.0001), but that Mandarin and Other groups did not significantly differ in transcription accuracy (estimate=0.04, SE=0.021, z=1.897, p=0.140).

Our second research question asked whether transcribers' self-assessed pre-task familiarity with Mandarin-accented English would predict higher transcription accuracy. The effect of *mandAccentFam*, which was assessed in the pre-task questionnaire, was marginally significant. We therefore conducted a simulation-based power analysis for this fixed effect with the `simr` package (Green & MacLeod, 2016). We extended the dataset to simulate up to 350 participants, and used the powerCurve command to estimate power over sample size. Estimated power reached only 60.0% at 350 participants, indicating a very small effect size.

Our final research question concerned whether transcription accuracy was higher when sentences were presented in single- versus mixed-talker conditions (see Table 5). To follow up on the significant interaction of *numTalkers* and *talkerID*, we created three new mixed-effects models with the same structure as the final model (removing the fixed factor *talkerID* and the *numTalkers*talkerID* interaction). For talkers M35 and M39, the effect of *numTalkers* was not significant (F(1,126.1)=0.094, p=0.760 and F(1,126)=1.053, p=0.307,

Table 4: Type III ANOVA table for final model.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	P value
order	0.011	0.011	1	14987.936	0.332	0.565
orderQuadratic	2.424	2.424	1	14988.510	70.750	0.000***
orderCubic	1.118	1.118	1	14988.307	32.633	0.000***
mandAccentFam	0.119	0.119	1	254.409	3.482	0.063
numTalkers	0.097	0.097	1	254.409	2.821	0.094
talkerID	10.320	5.160	2	376.264	150.606	0.000***
firstLang	1.419	0.709	2	254.409	20.703	0.000***
goodAtAccents	0.016	0.016	1	254.409	0.460	0.498
interactAccents	0.007	0.007	1	254.409	0.203	0.652
beliefsAccentDifficult	0.092	0.092	1	254.409	2.689	0.102
beliefsAccentGoals	0.246	0.246	1	254.409	7.188	0.008**
postTalkerLang	0.999	0.999	1	254.409	29.150	0.000***
postAccentFam	0.036	0.036	1	254.409	1.047	0.307
numTalkers:talkerID	0.408	0.204	2	376.264	5.956	0.003**

*p< 0.05, **p< 0.01, ***p< 0.001. `final.model=lmer(logPropCorr~(1|participant)+(1|target)+order+orderQuadratic+orderCubic+mandAccentFam+numTalkers*talkerID+firstLang+goodAtAccents+interactAccents+beliefsDifficult+beliefsGoals+postTalkerLang+postAccentFam, data=scored.sum, REML=F, lmerControl(optimizer="bobyqa"))`

respectively). However, for talker M43, the effect of *numTalkers* was significant ($F(1,131.6)=13.058$, $p<0.001$), with higher accuracy in the single talker condition than in the multi-talker condition.

For the purpose of exploratory analyses, we also assessed variables that are not directly associated with the primary research questions, but which might be useful in predicting an individual's transcription accuracy. Neither participants' self-assessment of how good they are at understanding accents that differ from their own (*goodAtAccents*) nor their self-assessed frequency of interaction with people whose accents differ from their own (*interactAccents*) were significant predictors of transcription accuracy. As described above, we computed two beleifs factors. While participants' responses to statements regarding the difficulty they experience with so-called "non-native" and "non-standard" accents (*beliefsAccentDifficult*) was not a significant predictor, their beliefs about goals for language learners (*beliefsAccentGoals*) did significantly predict log transcription accuracy. Lower levels of agreement with the statements that learners should have pronouncing like a native or standard speaker as a goal were associated with more accurate log transcription accuracy ($r(253)=-0.257$, $p<0.001$).

In addition to the pre-task self-assessment of familiarity with Mandarin-accented English speech, we collected two related post-task measures. While *postAccentFam*, where participants were asked to indicate their degree of familiarity with the accent they had just heard, was not significant, we did find a significant effect of *postTalkerLang*, where participants were asked to describe the accent they had just heard. Participants who were able to identify the accent as "Chinese" or "Mandarin", or who named a country in which Mandarin is widely spoken, exhibited significantly higher transcription accuracy (mean proportion correct=0.644) than those who did not (mean proportion correct=0.533).

Table 5: Mean proportion correct by number of talkers and talker ID, averaged across all other factors.

numTalkers	talkerID	mean propCorr
Single Talker	M35	0.609
Multi-Talker	M35	0.632
Single Talker	M39	0.428
Multi-Talker	M39	0.438
Single Talker	M43	0.671
Multi-Talker	M43	0.622

3. DISCUSSION

This experiment was designed to address three research questions concerning the roles of the listener and task factors in determining transcription accuracy. Our first question was: Is transcription accuracy higher when talkers and listeners have similar first language backgrounds? To address this question, we asked participants to identify their first language(s). Participants who responded English (but not also Mandarin) were assigned to the English *firstLang* group, those who responded Mandarin (but not also English) were assigned to the Mandarin group, those who responded with a language other than English or Mandarin were assigned to the Other group, and those who identified both English and Mandarin as first languages were excluded from the analysis (due to a small number). With linguistic background operationalized in this way, and with these speech materials and participants, we observed no advantage associated with shared first language background on transcription accuracy (indeed, the English first-language group was more accurate than both the Mandarin and Other groups). This may not be surprising given that evidence for such a benefit has been mixed. Importantly, individual talker and listener variation on various dimensions, including language proficiency and the acoustic-phonetic properties of the speech samples, have been found to modulate the degree — or even existence — of the benefit (Munro et al., 2006; Hayes-Harb et al., 2008; Xie & Fowler, 2013). Indeed, visual inspection of our data (see Table 5 and Fig. 1) suggests that transcription accuracy varies a great deal by talker and by listener. The results of our experiment thus reinforce the complexity of the relationship between listener and talker language backgrounds in determining transcription accuracy. Moreover, as noted above, our goal in this study was to use the type of assessment of the broad similarities between the self-reported language backgrounds of speakers and listeners that are available from crowdsourced platforms. Future research will need to examine whether alternative means of assessing the language background and experience of speakers and transcribers are associated with an intelligibility benefit in a crowdsourced transcription task.

Our second research question was: Is transcription accuracy higher when listeners self-identify as familiar with the talker's accent? We operationalized familiarity with the talker's accent via participants' pre-task self-assessment of their familiarity with Mandarin-accented English (0-10 Likert scale). We found no effect of this pre-task self-assessment, nor of a post-task self-assessment where participants were asked to rate their familiarity with the accent of the speaker they had just listened to. We thus found no evidence that participants were able to assess their own experience with Mandarin-accented English speech in ways that predicted their transcription accuracy. Given the potential value in having transcribers self-identify as skilled with respect to particular accents, future research should continue to investigate ways of assessing language experience that may be predictive of transcription skill.

Finally, we asked: Is transcription accuracy higher when audio recordings are presented in single- versus multi-talker blocks? Participants were assigned to one of three single-talker conditions where they heard the

same talker produce all 60 sentences, or to a multi-talker condition where they heard each of the three talkers produce 20 of the sentences. The number of talkers exhibited an interaction with talker ID. Specifically, we found that while talkers M35 and M39 did not show an effect of the number of talkers, transcription of talker M43's speech was more accurate in the single- than in the multi-talker condition. As such, our data are partially consistent with past reports that transcription accuracy is greater in single- vs. multi-talker conditions (Mullennix et al., 1989; Tamati & Pisoni, 2014). It is worth noting that the effect of single- vs. multi-talker conditions is observed only for the most intelligible of the three talkers (M43). Indeed, descriptively, the data pattern in the opposite direction for the other two talkers (M35, M39), with higher proportion correct scores for the multi-talker conditions than for the single-talker conditions. That the effect of the single- vs. multi-talker manipulation was dependent on the talker suggests that simply blocking speech materials by talker may not enhance crowdsourced transcription accuracy for all talkers. Future work will be needed to more closely examine this interaction of talker and number of talkers.

We asked participants several additional questions to allow for exploratory analyses of listener factors that may influence their transcription accuracy. It is worth noting that as with the lack of significant effects of participants' self-assessments of familiarity with Mandarin-accented English, there were no effects of participants' self-assessments of how good they are at understanding other accents or their frequency of interaction with people whose accents differ from their own. Indeed, the language and accent beliefs measures that ask participants to rate their level of difficulty with "non-native" and "non-standard" accents similarly did not significantly predict transcription accuracy. Our data provide no evidence that the type of pre-task assessments of listener factors of the sort we have incorporated here are predictive of transcription accuracy.

On the other hand, we did find a significant effect of whether participants believe that speaking with a "native-like" or "standard" accent should be a goal of the language learner, with those who believe that learners should try to achieve "native-like" or "standard" accents showing less accurate transcription accuracy. This finding is consistent with previous research showing that listeners who self-report more positive attitudes towards language learners and learner accents were more accurate when transcribing learner-accented speech (Ingvalson et al., 2017). Given this finding, a pre-task assessment of transcriber beliefs may provide a means of determining who is likely to transcribe speech samples accurately. However, additional research is required to clarify the contribution of listener's beliefs to transcription accuracy with a focus on which beliefs are most predictive of accurate transcription.

We found that participants whose metalinguistic knowledge led them to identify the accent of the talker(s) as "Mandarin" or "Chinese", or as originating from a country where Mandarin is a dominant language, exhibited significantly higher transcription accuracy than those who did not. Vaughn (2019) reports a related finding: participants who were told that a speaker's first language is Spanish led to higher transcription accuracy for Spanish-accented English speech relative to participants who were not. These results suggest that knowing the language background of a speaker (whether by inference or through information provided by others) may activate accent-specific expectations that improve transcription accuracy. In crowdsourced transcription, it is possible that posing such post-task questions to a transcriber could help direct their assignment to future transcription tasks to increase overall accuracy (Hettiachchi, 2022). Alternatively, identifying the language background of speakers to transcribers in advance may improve transcription accuracy. Such possibilities require research to determine their efficacy in enhancing the accuracy of crowdsourced transcription.

Importantly, we found that the transcribers did improve in their transcription performance over the course of the experiment; however, this improvement was not linear, and a cubic growth model best captured participants' performance. Thus, these untrained, paid participants showed the adaptation that is well-documented in the literature (Clarke & Garrett, 2004; Bradlow & Bent, 2008). While the experiment reported here focused on selected listener and task factors, there are additional motivation-related factors affecting adaptation that may also be worth pursuing. For example, Afghani et al. (2023) demonstrated that accuracy-based incentive pay increased transcription accuracy for an unfamiliar accent. Other studies have explored the im-

pact of telling listeners that their transcription accuracy will improve over time, with mixed results (Vaughn, 2019; Hayes-Harb et al., submitted). Research is needed to better understand the role of motivation-related effects on crowdsourced transcription accuracy.

We will close with a discussion of some essential considerations in this work. First, we used the terms "native" and "standard" in relation to accents in the questionnaires for the purpose of probing participants' language and accent ideologies. However, these terms are both scientifically inadequate and harmful to language users and communities (Cheng et al., 2021; Namboodiripad et al., 2026). Here we opted to group our participants according to their self-identified "first language" backgrounds; however, the complexity of human linguistic experience cannot be readily reduced to such simple descriptors. This is an especially important issue in the context of attempting to identify and elicit information that may enhance transcription accuracy in crowdsourced platforms. A second issue is that of consent and compensation in the development of datasets. Should an individual speaker be able to consent (whether knowingly or not) on behalf of a minoritized language community, given that their speech may be used to build, for example, accent recognition technology that can harm the community? Given variation among communities in their language values and practices, is it correct to assume that all communities want this technology? How do we ensure that communities benefit from technology built on their language and speech? A final consideration is the term "ground truth" itself. For ASR service providers, developers use the "ground truth" to describe a transcription, but the truth of what was said is neither uncontested nor known by the listener alone. As AI auditing pioneer Dr. Joy Buolamwini says, "those with the power to build AI systems do not have a monopoly on truth" (2023). It is critical that we interrogate what we mean by "ground truth" and who is able to participate in its construction. The answers to the questions posed above (and many others) must shape careful ASR research and system development. Ultimately, the overarching objective of our work is to explore ways of mitigating the biases that permeate ASR while continuing to problematize ASR systems for the risks they pose to talkers, transcribers, users of the technology, and their communities.

ACKNOWLEDGMENTS

We are grateful to the members of the Speech Acquisition Lab as well as the attendees of the the 190th Meeting of Acoustical Society of America for their feedback on this project. This work was made possible by seed funding from the University of Utah Office of the Vice President for Research.

REFERENCES

Adank, P., Evans, B. G., Stuart-Smith, J. & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, 35(2), 520–529.

Afghani, C., Baese-Berk, M. M. & Waddell, G. R. (2023). Performance pay and non-native language comprehension: Can we learn to understand better when we're paid to listen? *Applied Psycholinguistics*, 44(4), 593–609.

Alexander, J. E. D. & Nygaard, L. C. (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, 145(6), 3382.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4218-4222.

Barrett, T. S., Borrie, S. A. & Yoho, S. E. (2019). Automating with Autoscore: Introducing an R package for automating the scoring of orthographic transcripts. *PsyArXiv*. https://doi.org/10.31234/osf.io/d8au4.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1–48.

Boersma, P. & Weenink, D. (2020). Praat: Doing phonetics by computer (Version 6.1.37).

Bradlow, A. R. (n.d.). ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings [Data set]. https://speechbox.linguistics.northwestern.edu/allsstar.

Bradlow, A. R. & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707–729.

Buolamwini, J. (2024). *Unmasking AI: My mission to protect what is human in a world of machines*. Random House.

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A. & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, 12(1), 715843.

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

Demonte, P. (2019). *Speech shaped noise master audio - HARVARD speech corpus*. https://doi.org/10.17866/rd.salford.9988655.v1

DiChristofano, A., Shuster, H., Chandra, S. & Patwari, N. (2023). Performance disparities between accents in Automatic Speech Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13), 16200–16201.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498.

Harwell, D. (2018, July 19). How Amazon's and Google's smart speakers leave certain voices behind. *The Washington Post*. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/

Hayes-Harb, R., Smith, B. L., Bent, T. & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics*, 36(4), 664–679.

Hayes-Harb, R., Townsend, A. & Newbold, C. Submitted. Examining an infographic-based intervention to improve listeners' adaptation to an unfamiliar accent.

Hettiachchi, D., Kostakos, V. & Goncalves, J. (2022). A survey on task assignment in crowdsourcing. *ACM Computing Surveys (CSUR)*, 55(3), 1-35.

Ingvalson, E. M., Lansford, K. L., Federova, V. & Fernandez, G. (2017). Listeners' attitudes toward accented talkers uniquely predicts accented speech perception. *The Journal of the Acoustical Society of America*, 141(3), EL234.

Kennedy, S. & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, 64(3), 459–489.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J., Jurafsky, D. & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14), 7684-7689.

Lenth, R. V. & Piaskowski, J. (2025). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://rvlenth.github.io/emmeans/

Maison, L., & Estève, Y. (2023). Some voices are too common: Building fair speech recognition systems using the Common Voice dataset. *arXiv [eeSS.AS]*. http://arxiv.org/abs/2306.03773

Markl, N. (2022, June 21). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. *2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea*. https://doi.org/10.1145/3531146.3533117

Maye, J., Aslin, R. N. & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cognitive Science*, 32(3), 543–562.

McCloy, D. (2013). Mix speech with noise [Praat Script]. https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat

McLaughlin, D. J., Colvett, J. S., Bugg, J. M., & Van Engen, K. J. (2024). Sequence effects and speech processing: Cognitive load for speaker-switching within and across accents. *Psychonomic Bulletin & Review*, 31(1), 176-186.

Mullennix, J. W., Pisoni, D. B. & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, 85(1), 365–378.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, 28(1), 111-131.

Nagle, C. (2019). An introduction to fitting and evaluating mixed-effects models in R. In J. Levis, C. Nagle & E. Todey (Ed.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 82–105).

Namboodiripad, S., Kutlu, E., Babel, A., Babel, M., Baese-Berk, M., Bassuk, P. B., Block, A., Pérez, R. C., Carlson, M. T., Carraturo, S., Cheng, A., Cheng, L. S. P., Combiths, P., Foushee, R., Frederiksen, A. T., Grammon, D., Hayes-Harb, R., Higby, E., Kendro, K. & Wright, K. E. (2026). Finding our ROLE: How and why to reframe essentialist approaches to language. *Cognition*, 271, 106444. https://doi.org/10.1016/j.cognition.2026.106444

Prinos, K., Patwari, N. & Power, C. A. (2024, June 3). Speaking of accent: A content analysis of accent misconceptions in ASR research. The 2024 ACM Conference on Fairness, Accountability, and Transparency. *FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro Brazil. https://doi.org/10.1145/3630106.3658969

Radford, A., Kim, J. W., Xu, T., Brockman, G. McLeavey, C. & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervisionLinks to an external site. *International Conference on Machine Learning (ICML)*, 28492-28518.

R Core Team (2025). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org

Revelle, W. (2025). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University. https://CRAN.R-project.org/package=psych

Sprugnoli, R., Moretti, G., Bentivogli, L. & Giuliani, D. (2017). Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation*, 51(2), 283–317.

Stibbard, R. M. & Lee, J.-I. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *The Journal of the Acoustical Society of America*, 120(1), 433–442.

Stoet, G. (2010). PsyToolkit: a software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096–1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* (Columbia, Mo.), 44(1), 24–31.

Synnaeve, G., Xu, Q., Kahn, J., Likhomanenko, T., Grave, E., Pratap, V., Sriram, A., Liptchinsky, V., & Collobert, R. (2019). End-to-end ASR: From supervised to semi-supervised learning with modern architectures. *arXiv [cs.CL]*. http://arxiv.org/abs/1911.08460

Tamati, T. N. & Pisoni, D. B. (2014). Non-native listeners' recognition of high-variability speech using PRESTO. *Journal of the American Academy of Audiology*, 25(9), 869–892.

Vaughn, C. R. (2019). Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America*, 145(5), 3218.

Xie, X. & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics*, 41(5), 369–378.

Xie, X., Theodore, R. M. & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology. Human Perception and Performance*, 43(1), 206–217.