Sun, Hanyu, and Ting Yan. 2023. "Applying Machine Learning to the Evaluation of Interviewer Performance." *Survey Practice* 16 (1).
https://doi.org/10.29115/SP-2023-0007.

ARTICLES

Applying Machine Learning to the Evaluation of Interviewer Performance

Hanyu Sun[1], Ting Yan[1]

[1] Westat (United States)

Keywords: machine learning, interviewer monitoring, Computer Assisted Recorded Interviewing

https://doi.org/10.29115/SP-2023-0007

Survey Practice

Vol. 16, Issue 1, 2023

Survey organizations have long used Computer Assisted Recorded Interviewing (CARI) to monitor interviewer performance. Conventionally, a human coder needs to first listen to the audio recording of the interactions between the interviewer and the respondent and then evaluate and code features of the question-and-answer sequence using a pre-specified coding scheme. Although prior research found that providing feedback to interviewers based on CARI was effective at improving interviewer performance, such coding process tends to be labor intensive and time consuming. To improve the effectiveness and efficiency of using CARI to monitor interviewer performance, we developed a pipeline that heavily draws on the use of machine learning to process audio recorded interviews. In particular, machine learning is used to detect who spoke at which turn in a question-level audio recording and to transcribe conversations at the turn level. This paper describes how the pipeline was used to detect interviewer falsification and to identify problematic interviewer behavior in both recordings of mock interviews and actual field interviews. The performance of the pipeline was discussed.

Introduction

It is well documented in the survey literature that interviewers influence all aspects of survey data collection (e.g., West and Blom 2017). Survey organizations routinely monitor interviewer performance as a means to reduce error and improve data quality. Interviewer monitoring often aims to detect interviewer falsification and undesirable behaviors. Interviewer falsification is defined as the interviewer's intentional departure from the designed interviewing protocol (Groves 2004). Fabricating all or part of an interview is one type of interviewer falsification, which leads to biased estimates. In addition, interviewers may deviate from the standardized interviewing protocol such as *not* reading a survey question exactly as worded or using leading probes. Deviation from the standardized interviewing protocol has been shown to result in increased interviewer variance in the response data (West and Blom 2017).

Computer-Assisted Recorded Interviewing (CARI) has long been used by survey organizations to monitor interviewer performance (e.g., Hicks et al. 2010; Thissen 2013). CARI produces question-level audio recordings that capture the interactions between the interviewer and the respondent. Previous research has found that providing feedback to interviewers based on CARI was effective at improving interviewer performance (Edwards, Sun, and Hubbard 2020).

To use CARI, a human coder needs to first listen to the audio recording of the interactions between the interviewer and the respondent and then evaluate and code features of the question-and-answer sequence using a pre-specified coding scheme. The coding scheme can be designed to identify interviewer falsification by asking coders to indicate how likely the interviewer fabricated the whole or part of the interview. In addition, coders are asked to assess how closely the interviewer followed the standardized interviewing protocol, such as whether or not the interviewer read the question verbatim and whether or not the interviewer probed in a non-leading way. Coding of question-level recordings can be further aggregated to provide respondent-level or interviewer-level evaluations. Interviewers who are identified to have deviated substantially from the standardized interviewing protocol are called out for intervention. For instance, interviewers who are judged (by the coder) to have falsified interviews are pulled from data collection immediately. Interviewers who did not read questions verbatim are re-trained to correct their undesirable behaviors. Although it is an effective tool (Edwards, Sun, and Hubbard 2020), this conventional coding of CARI tends to be labor intensive and time consuming. As a result, in practice only a small proportion of completed interviews or a selected group of key questionnaire items are selected for coding and evaluation in a timely manner due to resource constraints.

To maximize the use of CARI in real-time, efficient processing of audio recordings is needed to process 100% of the recorded interviews as quickly as possible and as inexpensively as possible. Ideally, this process will also lead to automatic identification of recordings that are at a higher risk of being falsified or exhibiting undesirable interviewer behaviors, facilitating interviewer monitoring in field data collection. We developed a CARI machine learning (ML) pipeline to achieve exactly this goal.

As shown in Figure 1, the CARI ML pipeline includes two component processes built on machine learning: speaker diarization and speech-to-text. The speaker diarization process takes advantage of pre-trained models (Bredin et al. 2020) to determine who spoke at which turn in a question-level audio recording. At the end of the diarization process, the number of different speakers is output for each question-level audio recording. Since the paradigmatic question-answer sequence involves only two different speakers (a.k.a. an interviewer and a respondent), recordings that are detected to have 0 or only 1 speaker are automatically flagged by the pipeline to be at a risk of interviewer falsification.

The speech-to-text process uses pre-trained English models (Mozilla DeepSpeech 0.9.3) to transcribe conversations at turn-level. The pipeline subsequently compares turn-level transcript to the exact question wording from the survey instrument, yielding several distance measures on lexical and semantic similarities. The pipeline then uses distance measures to evaluate interviewer question reading behaviors; a larger distance measure suggests that

Figure 1. CARI machine learning (ML) pipeline

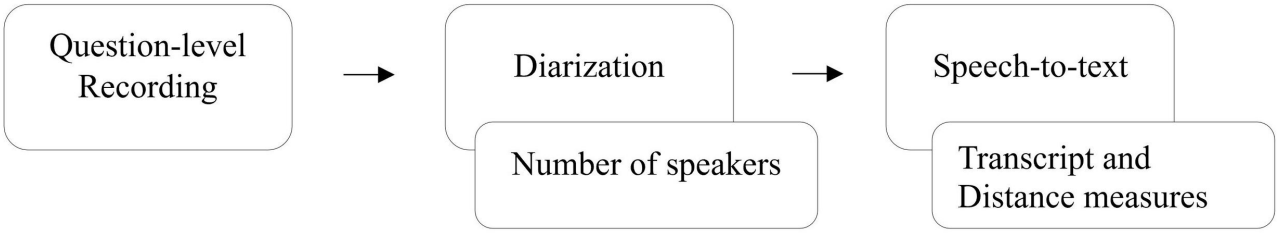

Figure 1. CARI machine learning (ML) pipeline

interviewers are more likely to *not* read questions exactly as worded. Interviewers with large distance scores will be automatically flagged as showing undesirable question-reading behavior.

We applied the CARI ML pipeline to recordings from both mock interviews conducted in a laboratory setting and actual in-person interviews from a national study. The study was conducted in 2020. The performance of the pipeline was evaluated and is presented next.

Study 1: Mock Interview Recordings

Data and Methods

We first applied the CARI ML pipeline to 34 mock interview recordings that were produced in a laboratory setting to explore its feasibility. In two scripts (scripts 5 and 7), interviewers were instructed to falsify by asking the survey question and then answering that question themselves. In addition, interviewers were instructed to read questions exactly as worded in three scripts, to make minor wording changes in one script, and to make major wording changes when reading the survey question in script 6. The deviations were scripted so that they were the same across interviewers. Table 1 presents the information about the mock interview recordings.

Table 1. Description of the mock interview recordings.

Script	Interviewers instructed to falsify	Actual number of speakers involved	Interviewers instructed to read questions	Number of question-level recordings
1	No	2	Exactly as worded	6
2	No	2	Exactly as worded	6
3	No	2	Exactly as worded	6
4	No	2	Minor wording change	6
5	Yes	1	Exactly as worded	2
6	No	2	Major wording change	6
7	Yes	1	Exactly as worded	2
Total	--	--	--	34

Two female interviewers were recruited to conduct the mock interviews. One of them is a native speaker of American English; she conducted interviews with three respondents (one native speaker of American English and two non-native speakers), yielding a total of 27 question-level recordings. The second interviewer is a non-native speaker of English and interviewed a respondent who is also a non-native speaker, producing 5 question-level recordings. Across these recordings, we manipulated features that are known to affect the quality of speech recognition, such as the presence of background noise (i.e., played a podcast in the background) and the respondent's distance to the digital voice recorder. In addition, this design allows us to explore how native vs. non-native speakers affects the performance of the speech-to-text component of the pipeline. All recordings were produced using the same digital voice recorder in a conference room on the same day. See Appendix 1 for more information.

To evaluate the performance of the CARI ML pipeline in detecting falsified interviews, we compared the number of speakers detected by the pipeline to the actual number of speakers involved in the recording. To evaluate the ability of the pipeline to assess interviewer question-reading behavior, we computed four distance measures. Three distance measures—the Jaro-Winkler distance, the Jaccard distance, and the Cosine distance—measure how different or similar the transcript from a recording is to the question wording in terms of words used. The fourth distance measure—word2vec—is a measure of semantic distance; that is, how different or similar in meaning the transcript is from the survey question. All four distance measures range from 0 to 1, with 0 indicating the transcript and the survey question wording are exactly the same and 1 indicating they are completely different.

Results

Results on interviewer falsification

Table 2 displays the number of speakers detected by the pipeline in comparison to the true number of speakers actually involved in each recording. For the purpose of interviewer falsification, recordings with zero or one speaker detected are flagged as falsified interviews, whereas recordings with two or more speakers detected are considered not falsified.

As described earlier, four recordings were truly falsified recordings as interviewers were instructed to answer the question themselves. As shown in Table 2, the pipeline correctly detected the number of speakers involved in the four recordings as one and, thus, succeeded in flagging the four recordings as falsified interviews.

Thirty recordings truly had two speakers. The pipeline detected two or three speakers for 18 of them but erroneously detected only one speaker for 12 recordings. We further examined these 12 recordings to better understand why the pipeline only detected one speaker. In nine of them, the interviewer was a native speaker, but the respondent was a non-native speaker. In one recording,

Table 2. The number of speakers detected in mock interviews.

	True number of speakers involved in recording	
Number of speakers detected by pipeline	1	2
1	4	12
2 +	0	18
Total	*4*	*30*

both the interviewer and the respondent were non-native speakers. The pre-trained model used by the pipeline was developed and trained on a corpus of audio recordings in English. It is not surprising that the diarization process of the pipeline did not seem to work well for recordings involving non-native speakers. For the remaining two recordings where the pipeline failed to detect two speakers, both the interviewer and the respondent were native speakers, but there was either background noise in the recording or the respondent sat further away from the digital recorder microphone (i.e. far-field microphone). It is well documented that background noise and far-field microphones affect the performance of diarization (e.g. Haeb-Umbach et al. 2020; Park et al. 2021).

Even though the CARI ML pipeline did not detect the correct number of speakers for all 34 recordings, it did correctly identify four falsified recordings as falsified and 18 non-falsified recordings as non-falsified, yielding an accuracy rate of 73%.

Results on interviewer question reading behavior

Figure 2 plots the four distance measures for all 34 recordings by how interviewers actually read the survey questions in the recordings. The red line shows the means of distance measures for recordings in each group. As stated earlier, the closer the distance measures are to zero, the more likely the interviewer read the question verbatim.

As shown in Figure 2, the mean distance measures are closer to zero for recordings in which the interviewer was instructed to read the question verbatim and is smaller than the mean for recordings in which interviewers made minor wording changes, which is smaller than the mean for recordings in which interviewers made major wording changes. The pattern is consistent across all three lexical distance measures that only take words into account. The semantic distance measure that compares similarity in meanings did not seem to differentiate actual question reading behavior (panel d of Figure 2).

We then created a summary distance measure by taking an average of the four distance measures for each recording and plotted the summary distance measure in panel e of Figure 2. We see the same downward line, indicating that the summary distance measure is able to differentiate interviewer question-answer behaviors. Overall, the findings suggest that the CARI ML pipeline is a feasible approach to process question-level recordings.

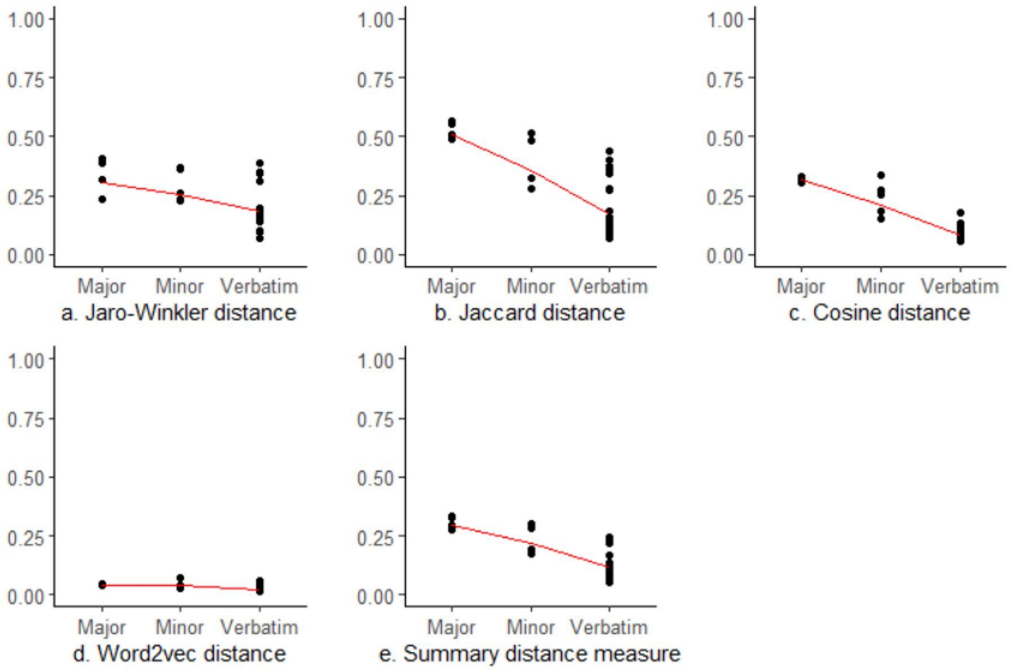

Figure 2. Distance measures by actual interviewer question reading behavior (major deviation, minor deviation, or verbatim)

Study 2: Field Interview Recordings

Data and Methods

To examine how the CARI ML pipeline works for recordings of interviews in actual data collection, we selected 1,182 recordings of question-answer sequences from a large-scale cross-sectional study of a nationally representative sample. These recordings were part of 53 in-person interviews conducted by four randomly selected field interviewers (three females and one male).

All 1,182 recordings went through conventional CARI coding by having coders listen to the actual recordings using a behavior coding scheme. Coders were asked to indicate how clearly they could hear the interviewer and the respondent in the question-level recording by selecting one of the four options: very clearly, somewhat clearly, not very clearly, and cannot hear the interviewer/respondent. We used the coder judgment to determine the actual number of speakers in the question-level recording. The number of speakers was zero if the coder selected that neither the interviewer nor the respondent could be heard. The number of speakers was one if either the interviewer or the respondent could be heard very clearly, somewhat clearly, or not very clearly. In addition, the number of speakers was two if both the interviewer and the respondent could be heard very clearly, somewhat clearly, or not very clearly.

Coders were also asked to evaluate whether or not the interviewer maintained the question meaning by selecting yes, no, or unclear. If the coder selected yes, the interviewer was determined to have maintained the meaning. By contrast,

Table 3. Description of the field interview recordings.

Interviewer included in the selected recordings	Number of interviews included selected recordings	Number of question-level recordings selected	Number of recordings the coder determined to have 2 speakers	Number of recordings the coder determined that question meaning was maintained
A	3	60	30	54
B	26	615	559	552
C	9	172	152	130
D	15	335	299	282
Total	*53*	*1182*	*1040*	*1018*

Table 4. Number of speakers detected in field interview recordings.

	Number of speakers by coder coding	
Number of speakers detected by pipeline	0 or 1	2
0 or 1	111	323
2 +	31	717
Total	*142*	*1040*

if the coder selected no, the interviewer was considered to have *not* maintained the question meaning. Table 3 presents the information about the 1,182 recordings.

Findings

Table 4 presents the true (as determined by human coders) and detected number of speakers in the question-level recordings. A total of 142 recordings had either zero or one speakers according to human coders, and the CARI ML pipeline successfully detected 111 of them as having either zero or one speaker. In other words, the pipeline correctly flagged 78% of falsified recordings as falsified. As to the 31 recordings that the pipeline failed to flag as falsified, most of them had background noise such as music played in the background, loud static noise, and loud keyboard typing sound.

In addition, the CARI ML pipeline falsely flagged as falsified 323 (out of the 1,040) recordings where human coders indicated that there were two speakers. We reviewed these recordings and noticed that most of the recordings include features that are known to affect diarization, such as overlapping speech, background noise, and the respondent having a softer voice than the interviewer. The overall agreement rate between human coders and the pipeline was 78%.

Distance scores are plotted in Figure 3. Recordings in which the interviewer was judged by human coders to have maintained the question meaning tended to have lower values on all four distance measures than recordings in which the interviewer was judged to not have maintained the question meaning.

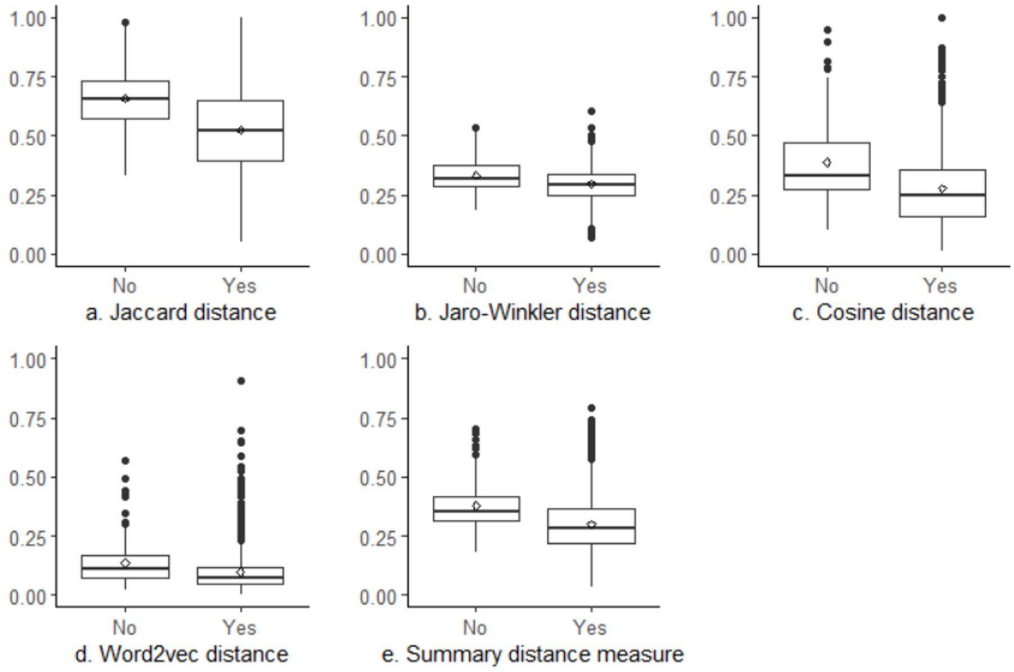

Figure 3. Distance measures for field interviews (Yes – Coder coded that the interviewer maintained the question meaning, No – Coder coded that the interviewer did not maintain the question meaning)

Summary and Discussion

We developed a CARI machine learning pipeline to identify falsifications and undesirable interviewer behaviors in a timely manner. We applied the pipeline to recordings from both mock interviews and actual interviews and found that the performance of the pipeline was robust on detecting the number of speakers to identify suspected falsifications and on yielding distance measures to assess how likely the interviewer read the question exactly as worded.

Compared to the conventional CARI coding, the advantage of using the CARI ML pipeline is real-time efficient and effective interviewer monitoring (see Appendix 2 for an illustration on how to aggregate data for interviewer monitoring; see Appendix 3 for efficiency comparison between the pipeline and the conventional CARI coding); 100% of the recorded interviews can be processed in real time, whereas the conventional CARI coding is typically performed on a small proportion of recordings. Our findings suggest that the CARI ML pipeline is a promising tool for interviewer monitoring and has the potential to substantially increase the effectiveness and efficiency of the conventional CARI coding.

Contact information

Hanyu Sun
Address: 1600 Research Blvd., Rockville, MD 20850
Email: hanyusun@westat.com
Phone: (301) 315-5916

Submitted: April 19, 2023 EST, Accepted: June 16, 2023 EST

REFERENCES

Bredin, Herve, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. "Pyannote.Audio: Neural Building Blocks for Speaker Diarization." *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May, 7124–28. https://doi.org/10.1109/icassp40776.2020.9052974.

Edwards, Brad, Hanyu Sun, and Ryan Hubbard. 2020. "Behavior Change Techniques for Reducing Interviewer Contributions to Total Survey Error." In *Interviewer Effects from a Total Survey Error Perspective*, edited by Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West, 77–90. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781003020219-9.

Groves, B. 2004. "Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection, and Repair of Its Effects." *Survey Research* 35 (1): 1–5.

Haeb-Umbach, Reinhold, Jahn Heymann, Lukas Drude, Shinji Watanabe, Marc Delcroix, and Tomohiro Nakatani. 2020. "Far-Field Automatic Speech Recognition." *Proceedings of the IEEE* 109 (2): 124–48. https://doi.org/10.1109/jproc.2020.3018668.

Hicks, W. D., B. Edwards, K. Tourangeau, B. McBride, L. D. Harris-Kojetin, and A. J. Moss. 2010. "Using CARI Tools to Understand Measurement Error." *Public Opinion Quarterly* 74 (5): 985–1003. https://doi.org/10.1093/poq/nfq063.

Park, T.J., N. Kanda, D. Dimitriadis, K.J. Han, S. Watanabe, and S. Narayanan. 2021. "A Review of Speaker Diarization: Recent Advances with Deep Learning." *Computer Speech & Language* 72: 101317. https://doi.org/10.48550/arXiv.2101.09624.

Thissen, M. Rita. 2013. "Computer Audio-Recorded Interviewing as a Tool for Survey Research." *Social Science Computer Review* 32 (1): 90–104. https://doi.org/10.1177/0894439313500128.

West, Brady T., and Annelies G. Blom. 2017. "Explaining Interviewer Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5 (2): 175–211. https://doi.org/10.1093/jssam/smw024.

SUPPLEMENTARY MATERIALS

Appendices

Download: https://www.surveypractice.org/article/84055-applying-machine-learning-to-the-evaluation-of-interviewer-performance/attachment/171815.docx