Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Causal Structure Discovery for Error Diagnostics of Children's ASR

Vishwanath Pratap Singh[1], *Md. Sahidullah*[2], *Tomi Kinnunen*[1]

[1]School of Computing, University of Eastern Finland, Finland
[2]Institute for Advancing Intelligence, TCG CREST, India
vsingh@uef.fi, sahidullahmd@gmail.com, tomi.kinnunen@uef.fi

Abstract

Children's automatic speech recognition (ASR) often underperforms compared to that of adults due to a confluence of interdependent factors: physiological (e.g., smaller vocal tracts), cognitive (e.g., underdeveloped pronunciation), and extrinsic (e.g., vocabulary limitations, background noise). Existing analysis methods examine the impact of these factors in isolation, neglecting interdependencies—such as age affecting ASR accuracy both directly and indirectly via pronunciation skills. In this paper, we introduce a causal structure discovery to unravel these interdependent relationships among physiology, cognition, extrinsic factors, and ASR errors. Then, we employ causal quantification to measure each factor's impact on children's ASR. We extend the analysis to fine-tuned models to identify which factors are mitigated by fine-tuning and which remain largely unaffected. Experiments on Whisper and Wav2Vec2.0 demonstrate the generalizability of our findings across different ASR systems.

Index Terms: children's ASR, speech foundational models, causal structure discovery, physiology, cognition, pronunciation

1. Introduction

Automatic speech recognition (ASR) is in the growing demand for child-centric technological solutions [1, 2, 3]. Unfortunately, the performance of ASR for children lags considerably behind that for adults [4, 5]. *Speech foundation models* (SFM) followed by fine-tuning is the current approach for improving performance for children's ASR [2, 6, 7]. Even after fine-tuning on children's speech, SFMs for children consistently produce higher *word error rates* (WERs)—the standard measure of transcription accuracy—compared to models optimized for adults [8, 6, 9, 2]. This disparity highlights the need for a **systematic framework** to identify the root causes of accuracy degradation in children's ASR.

Previous research has identified several factors that contribute to degraded ASR performance for children. Anatomical differences, such as shorter vocal tracts and lighter vocal cords, lead to higher fundamental frequencies and greater spectral variability, which vary with age and gender [10, 11]. Developing pronunciation skills further contribute to inconsistencies in articulation [12, 13]. Children may additionally struggle in pronouncing complex words despite mastering simpler vocabulary first [14, 7]. Additionally, in educational settings (where current children's speech data is collected), background babble from parental conversations or classroom sounds tends to interfere with recognition [15, 16]. The length of utterance also plays a role: shorter, fragmented speech tends to be more challenging [17, 7] for attention-based ASR systems [18].

The above factors are concurrently present in children's speech recordings and hardly impact ASR performance in isolation. Instead, they exhibit complex interdependencies. For instance, age influences both anatomical development and pronunciation ability. Still, most earlier studies [12, 17, 19] focus on these factors independently, underscoring a gap in current analysis approaches. Hence, a systematic analysis framework for understanding the causes of degradation should be inclusive of all these causes and should consider their interdependence.

Given that multiple co-existing factors contribute to children's ASR performance [20, 21], a **causal framework** [22, 23, 24] provides a natural approach for analyzing their interdependencies. Unlike purely *statistical*[1] methods that are limited in capturing correlations, a *causal* framework explicitly models cause-and-effect relationships of the underlying mechanisms affecting ASR performance. In practical terms, causal models are typically formalized using a directed acyclic graph (DAG) [24]. The variables under consideration are represented as *nodes* whereas *directed edges* indicate their causal relations. Specifically, a directed edge from node A to node B in a causal DAG means that A is a potential cause of B [25, 22].

For children's ASR analysis, addressed in our work, a causal graph may include nodes for variables such as age, pronunciation skills, background noise, and WER, with the directed edges illustrating how these factors influence one another. Using the example of age directly influencing a child's pronunciation skill, and the pronunciation skills in turn impacting WER, a causal DAG may reflect this assumption through a causal chain. The DAG helps distinguish **direct effects** (e.g., age $\rightarrow$ pronunciation) from **indirect effects** (e.g., age $\rightarrow$ pronunciation $\rightarrow$ WER), enabling a more precise analysis of the contributors to ASR performance. For the reader less familiar with causal methodology, we provide further detail in Section 2.

While causality in ASR has been previously studied for different tasks, such as to understand the impact of noise mitigation algorithms on ASR [26], the closest related work to ours is [7]. The authors in that study [7] used a predefined causal DAG to present prior knowledge motivated from psychological and social experiments [20, 21]. Their *hand-crafted* DAG hard-codes the assumed causal relations between the explanatory variables included. In a stark contrast, in this study we take a much less restrictive approach, facilitated by **causal structure discovery** [23]—methodology for identifying causal relations *automatically*. This considerably simplifies the task of

[1]Following Pearl [22], it is important to be clear on the distinction of correlational ('statistical') and 'causal' terminologies. The former focuses only on statistical associations (such as correlations) and *cannot differentiate between 'cause' and 'effect'*. Causal approaches 'go beyond' correlations and conditional probabilities by encoding explicitly the underlying causal mechanisms, possibly derived from the infamous 'domain knowledge'.

10.21437/Interspeech.2025-1967

Table 1: *Categorization of variables in analysis. Variables not available in metadata of CSLU Kids [16] are inferred, and their computation is discussed in Section 3.*

Type	Variable	Metadata/Inferred
Physiological	Age	Available
	Gender	Available
Cognitive	Pronunciation Ability	Inferred
Extrinsic	Signal-to-Noise Ratio	Inferred
	Vocab Difficulty	Inferred
	#Words in Audio	Available

the 'ASR performance analyst' who now only needs to decide which variables (measurements and/or metadata) to include to analysis—but without need to specify (potentially restrictive) assumptions on their cause-effect relations. We argue that reliable and automated identification of the important factors that influence ASR performance is helpful in designing better ASR systems: an ASR engineer benefits from knowing where to focus when designing new architectures, data augmentation or training recipes. Currently we primarily observe only the final outcome effect—high WERs in children ASRs—but lack accurate picture of the root causes.

We use the same set of variables and categorization (see Table 1) as [7] for easier comparability with that prior work. The ASR systems included consists of two well-known categories of open-source SFMs, representative of present state-of-the-art: (i) self-supervised – `Wav2Vec2.0` [27], and (ii) weakly-supervised – `Whisper` [28]. We consider both off-the-shelf (pretrained) and fine-tuned SFMs. We utilized CSLU kids corpus [16] for causal structure discovery, as it has gender and age metadata available from a diverse range. While we utilize MyST [15] for fine-tuning the SFMs.

2. A Primer in Causality

Causal analysis aims to establish cause-and-effect relationships that go beyond mere statistical (correlational) associations [25, 24]. Causal relations are formalized through a *directed acyclic graph* (DAG), whose nodes represent the variables and edges represent the cause-effect relations. Causal analysis is typically conducted in two stages: (1) Causal structure discovery, which involves identifying causal relationships among variables in a DAG—either hardcoded from prior knowledge [7, 31] or learned from data [29]; and (2) Causal quantification, which focuses on estimating the functional relationships between the connected variables. We review each step briefly.

2.1. Causal Graph

A causal graph, represented using DAG, with N random variables encodes how interventions on one affect others. It has generally four node types (Fig. 1). Formally, let G be a causal DAG with nodes $X_1, \ldots, X_n$. An edge $X_i \rightarrow X_j$ implies that X_i causally influences X_j. In the example shown in Fig. 1(a), X_1 has no parents, X_2 has X_1 as its parent (denoted by $\text{PA}_2 = \{X_1\}$), X_3 has X_2, X_1 as its parents ($\text{PA}_3 = \{X_2, X_1\}$), and X_2 and X_3 are descendants of X_1.

2.2. Causal Structure Discovery

Traditional methods for uncovering causal relationships rely on *interventions* or *randomized experiments* (studies where variables are deliberately manipulated to observe causal effects, such as randomized controlled trials, which randomly assign subjects to different groups). They can be impractical in machine learning due to cost and feasibility constraints—such as the need for extensive resources during data collection, and ethical considerations [23]. Causal structure discovery achieves this by inferring causal relationships from observational data—data collected *without* direct intervention, simply by observing variables as they occur naturally [35]. In this context (i.e., inferring from observational data), causal structure discovery serves as a *post-hoc* explainability approach, providing explanations for decisions *after* they have been made [36, 37].

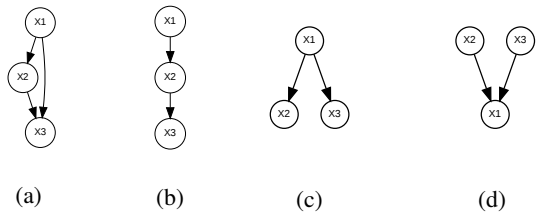


Figure 1: *Examples of directed acyclic graphs (DAGs) illustrating the causal relationships. (a.) Confounder: where X_2 is a parent of X_3, X_1 is a parent of both X_2 and X_3, making X_1 a confounder. X_2 and X_3 are descendants of X_1. (b.) Chain: a node that lies sequentially between two other nodes. (c.) Fork: indicates the nodes that have multiple children. (d.) Collider: a node is influenced by two or more nodes.*

A naive approach to causal structure discovery would involve evaluating all possible edge configurations among N variables. Since each directed edge can either be present or absent, the total number of possible configurations scales as $\mathcal{O}(2^N)$, which is exponential in N. This makes brute-force exploration infeasible, even for relatively small N. Consequently, more efficient algorithms are necessary to infer causal structures in a computationally tractable manner [33].

Two commonly used causal structure discovery algorithms include *Peter-Clark* (PC) [33] and *Fast Causal Inference* (FCI) [34] algorithms. Both leverage conditional independence [38] tests to extract information about the underlying causal structure. The main difference between the two is that PC assumes so-called *causal sufficiency* (i.e., no hidden confounders), whereas FCI accounts for latent variables and selection bias, making it more suitable for scenarios with unmeasured confounders.

2.3. Causal Quantification

Once the causal relationship between variables is established (i.e., DAG is formed), the next step is to estimate the functional relationship between the nodes, a process known as *causal quantification* [24]. Whereas a directed edge (whether handcrafted or automatically discovered) from age to WER in a DAG signifies that age has a causal effect on WER, causal quantification further determines the *magnitude* of this effect–e.g., how much an increase in age by one year increases (or decreases) WER.

A common approach to quantifying causal strength in DAGs is the *average causal effect* (ACE) [24], which measures the expected outcome difference between two scenarios when a node in DAG takes different values (e.g., 0 vs. 1). ACE is defined as:

$$\text{ACE}(X_i \rightarrow X_j) = \mathbb{E}[X_j \mid X_i = x_1] - \mathbb{E}[X_j \mid X_i = x_0] \quad (1)$$

Table 2: *Summary of Causal methods being used for explainable machine learning, and comparison with our proposed approach.*

Work	Task	Causal Structure Discovery Method	Causal Quantification Method
[29]	Animal Behavioral Modeling	PC-MI [30]	Graph Neural Network
[31]	Explainability in Recommendation System	Domain Knowledge	Logistic Regression
[32]	Sentiment Classification in NLP	Domain Knowledge	Bayesian Network
[7]	Causes of degradation in children's ASR	Prior Knowledge	Bayesian Network
Our	Causes of degradation in children's ASR	PC [33] and FCI [34]	Bayesian Network

This allows us to quantify the causal impact of specific variables on ASR performance, helping identify which factors considerably influence outcomes such as word error rates.

2.4. Causality in Machine Learning

Causality-aided explainable machine learning has gained popularity recently. A prior study [7] provides an overview of causality in ML; we briefly compare these studies with our approach in Table 2.

For direct comparison with prior work [7], we use *Bayesian inference* for causal quantification. First, we perform causal structure discovery and quantification to assess how open-source SFMs respond to variations in children's physiology (age, gender [20]), cognition (e.g., pronunciation [21]), and extrinsic factors. Then, we repeat the evaluation with fine-tuned SFMs.

3. Experimental Setup

3.1. Dataset

Following [7], we use CSLU Kids [16] for causal analysis due to the availability of diverse age groups and gender metadata. Since no standard protocol (training/development/test split) is available for CLSU Kids, we use the publicly available protocol from [3, 7]. We also include MyST [15] dataset for fine-tuning the SFMs to enhance generality of findings reported. For MyST, a standard protocol is available. Different from CSLU Kids, unfortunately, age and gender metadata labels are not available. This limits the use of the otherwise more recent and larger MyST in causal analysis.

3.2. Automatic Speech Recognition Systems

We consider two speech foundation model (SFM) based ASR systems in our experiments: `Wav2Vec 2.0` and `Whisper`. These two models complete the spectrum of SFMs, representing the two broad categories of self-supervised and weakly-supervised approaches. We utilize the open-source pre-trained `Whisper-Small` [2] and `Wav2Vec2.0-large` [3] for evaluation and fine-tuning. First, we use open-source SFMs "as is" in analysis then we fine-tune on MyST dataset [15].

3.3. Inferred Variables

In this section, we describe how inferred variables in Table 1 are computed. For *pronunciation ability*, we use the Goodness of Pronunciation (GoP) score [39], a posterior probability of the tartget phone normalized by the maximum posterior of all phones. Similar to [7], per-phoneme scores are averaged to produce an utterance-level score, which is further discretized into *Low*, *Average*, and *High* values. For *Vocabulary Difficulty*, we employ a rarity-based metric [40] using word frequencies extracted from multiple text corpora. A sentence-level difficulty score is obtained by averaging individual word scores and discretizing the result into *Low*, *Average*, and *High* levels. Finally, for signal-to-noise (SNR) estimation, we use NIST's toolkit [4]. Similar to GoP and Vocabulary Difficulty, we discretize SNR to three categories: *Clean* (SNR $\geq$ 20 dB), *Average* ($5 \leq$ SNR $\leq$ 20 dB), and *Noisy* (SNR $\leq$ 5 dB). Discretization allows us to stratify the GoP, Vocabulary Difficulty, and SNR into meaningful groups, which helps in identifying cause-effect relationships more precisely [41].

4. Results and Discussion

Unlike [7], which uses a hardcoded DAG and focuses solely on causal quantification, we perform both causal structure discovery and causal quantification.

4.1. Causal Structure Discovery

First, we examine the differences between the hardcoded DAG used in a prior study [7] and the data-driven DAG obtained using PC and FCI, as shown in Figure 2. Below, we discuss the similarities and differences for each node:

Similar to the hardcoded DAG, the automatically inferred DAG suggests that physiological factors, such as age, influences both cognitive factors (e.g., pronunciation variability) and ASR errors. However, unlike the hard-coded DAG, which assumes that age affects all three types of ASR errors, the data-driven DAG indicates that age variability primarily contributes to substitution errors.

Regarding gender, the hardcoded DAG assumes it to influences ASR errors, whereas the data-driven DAG indicates no causal relationship between gender and ASR errors. This aligns with previous empirical studies [7, 5], which found no considerable differences in ASR errors between boys and girls speakers.

For pronunciation variation (GoP), both DAGs identify Vocabulary Difficulty and Age as influencing pronunciation variability. However, the data-driven DAG also incorporates an additional factor—the number of spoken words (i.e., sentence length). Moreover, the data-driven DAG suggests that mispronunciations predominantly lead to substitution and insertion errors.

Regarding SNR, the hardcoded DAG assumes that SNR affects all three types of ASR errors, whereas the data-driven DAG suggests that SNR (representing babble noise in a classroom setting) primarily contributes to insertion errors. This aligns with intuition, as overlapping background speech can introduce unintended words into ASR transcriptions, leading to insertion errors.

Lastly, concerning utterance length, the hardcoded DAG assumes that all three types of ASR errors are influenced by the

[2] https://huggingface.co/openai/whisper-small

[3] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

[4] https://www.nist.gov/information-technologylaboratory/iad/mig/nist-speech-signal-noise-ratio-measurements



Figure 2: *Causal structure framework for explaining the ASR errors for children. (a.) Hard-coded: directed acyclic graph (DAG) used in prior study [7]. (b.) Data-driven: Causal structure discovery-based DAG using PC and FCI algorithms. Both PC and FCI have produced the same DAG.*

Table 3: *Comparison between the average causal effect (ACE) for hardcoded DAG vs data-driven DAG for open-source and fine-tuned* `Whisper-small` *and* `Wav2Vec2.0-large` *models. Empty cells corresponds to absent edges in data-driven DAG.*

Model	**Cause**	**Open-source**						**Fine-tuned**					
		Hardcoded DAG			**Data driven DAG**			**Hardcoded DAG**			**Data driven DAG**		
		Subs	**Del**	**Ins**	**Subs**	**Del**	**Ins**	**Subs**	**Del**	**Ins**	**Subs**	**Del**	**Ins**
`Wav2Vec2.0`	Age	-4.36	-0.10	-1.60	-5.21	–	–	-2.12	-0.04	-2.81	-3.58	–	–
	Gender	0.80	0.17	0.87	–	–	–	0.25	0.06	0.22	–	–	–
	GoP	-1.07	-0.25	-0.88	-1.27	–	-1.08	-0.63	-0.14	-0.54	-0.84	–	-0.45
	SNR	-1.08	0.09	0.28	–	–	-2.38	-0.58	0.11	0.09	–	–	-1.75
	# Words	-9.20	0.30	-4.23	-12.1	0.25	–	-7.14	0.36	-3.12	-9.74	0.08	–
`Whisper`	Age	-3.25	-0.20	-2.16	-3.97	–	–	-3.0	0.32	-2.22	-2.31	–	–
	Gender	0.62	0.37	1.15	–	–	–	1.90	0.10	1.33	–	–	–
	GoP	-1.25	-0.18	0.35	-0.9	–	-0.22	-1.10	-0.49	-0.05	-0.65	–	-0.18
	SNR	-1.21	0.14	-0.21	–	–	-1.84	-1.02	0.15	-0.36	–	–	-1.54
	# Words	-5.10	1.02	-3.45	-4.76	0.85	–	-5.53	1.20	-3.59	-4.80	0.95	–

number of words in an audio sample. In contrast, the data-driven DAG suggests that content length primarily affects substitution and deletion errors.

4.2. Causal Quantification

In this section, we compare the analysis results obtained using these two different frameworks. Quantification of ASR errors using ACE for children is presented in Table 3. First, for open-source SFMs, we observe that similar to the hardcoded DAG, the ACE for Age is negative for both `Wav2Vec2.0` and `Whisper` models in the data-driven DAG, indicating that an increase in Age reduces ASR errors. However, the hardcoded DAG shows a larger absolute ACE for substitution errors and a smaller ACE for deletion and insertion errors. This suggests that Age plays a lesser role in these two types of errors—interestingly, these are the very errors for which the data-driven DAG does not have a direct edge from Age.

Similarly, in the hardcoded DAG, the ACE for deletion and insertion errors is smaller for the *number of words in audio* (No. of Words), while the data-driven DAG lacks a direct edge between No. of Words and these two error types. Similar observations can be made regarding the lower ACE for Gender in hardcoded DAG while absent edges for Gender in data-driven DAG. Hence, forced causal associations in hardcoded DAG [7] result into weaker causal relationships (in terms of ACE) while these weaker edges are not present in the data-driven DAG.

Finally, we present the ACE for both hardcoded and data-driven DAGs for fine-tuned `Whisper` and `Wav2Vec2.0` models in Table 3. Our findings indicate that similar to hard-coded DAG, data-driven DAG also shows a reduction in ACE for for fine-tuned model. Specifically, after fine-tuning both DAGs have shown considerably lower ACE for Age (than that of open-source model in Table 3), indicating that fine-tuning reduces the impact of Age on ASR errors. However, ACE for *No. Words* node in both DAGs remain very high, even after fine-tuning indicating the limitation of fine-tuning.

5. Conclusion

We presented an approach for the construction of causal graphs for analyzing ASR errors in children. Unlike prior studies with hardcoded causal link assumptions, our data-driven method learns the causality relations automatically and removes unnecessary edges from the causal graph, thereby simplifying the analysis.

ACE analysis identifies key factors impacting ASR performance, guiding future research on data selection, model adaptation, and preprocessing for improved accuracy. For instance, fine-tuning addresses the age factor in ASR errors, while ACE for shorter utterances (which are typically in interactions of children's digital devices) remains very high. Hence, future research can be oriented towards addressing this concern by including short utterances in training or architecture suitable for short utterances.

One limitation of our study is its reliance on the CSLU Kids dataset due to the scarcity of child speech datasets with necessary metadata. Future work will explore additional datasets and extend causal inference to broader speech-processing tasks.

6. References

[1] J. H. Graafland, "New technologies and 21st century children," *Organization for Economic Co-operation and Development (OECD)*, no. 179, 2018. [Online]. Available: www.oecd-ilibrary.org/content/paper/e071a505-en

[2] R. Fan, N. Shankar, and A. Alwan, "Benchmarking children's asr with supervised and self-supervised speech foundation models," in *Proc. Interspeech*, 2024.

[3] V. P. Singh, M. Sahidullah, and T. Kinnunen, "Childaugment: Data augmentation methods for zero-resource children's speaker verification," *J. Acoust. Soc. Am.*, vol. 155, pp. 2221–2232, 2024.

[4] T. Patel and O. Scharenborg, "Improving end-to-end models for children's speech recognition," *Applied Sciences*, vol. 14, 2024.

[5] V. P. Singh *et al.*, "Spectral modification based data augmentation for improving end-to-end asr for children's speech," in *Proc. Interspeech*, 2022.

[6] R. Jain *et al.*, "Adaptation of whisper models to child speech recognition," in *In Proc. Interspeech*, 2023.

[7] V. P. Singh, M. Sahidullah, and T. Kinnunen, "Causal analysis of asr errors for children: Quantifying the impact of physiological, cognitive, and extrinsic factors," Available at SSRN, 2024. [Online]. Available: https://ssrn.com/abstract=5125557

[8] R. Jain *et al.*, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46 938–46 948, 2023.

[9] R. Fan *et al.*, "Towards better domain adaptation for self-supervised models: a case study of child asr," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, 2022.

[10] S. Lee, A. Potamianos, and S. S. Narayanan, "Analysis of children's speech, pitch and formant frequency," *Journal of the Acoustical Society of America*, vol. 101, 1997.

[11] S. Lee *et al.*, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, 1999.

[12] Q. Li and M. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc. ICSLP*, 2002.

[13] T. Bent *et al.*, "How pronunciation distance impacts word recognition in children and adults," *J Acoust Soc Am*, vol. 150, no. 6, p. 4103, 2021.

[14] R. Shi *et al.*, "Function words in early speech perception," in *Proc. International Congress of Phonetic Sciences*, 2003.

[15] S. Pradhan *et al.*, "My Science Tutor (MyST) – a large corpus of children's conversational speech," in *Proc. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024.

[16] K. Shobaki *et al.*, "The OGI Kids' speech corpus and recognizers," in *Proc. Interspeech*, 2000.

[17] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[18] P. Karmakar *et al.*, "Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition," *Intelligent Systems with Applications*, vol. 23, 2024.

[19] J. Cao *et al.*, "A comparative analysis of automatic speech recognition errors in small group classroom discourse," in *Proc. of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 2023.

[20] Institute of Medicine (IOM) and National Research Council (NRC), *Transforming the Workforce for Children Birth Through Age 8: A Unifying Foundation*, L. Allen and B. B. Kelly, Eds. The National Academies Press, 2015.

[21] M. Zhang and J. Hudson, "The development of temporal concepts: linguistic factors and cognitive processes," *Frontiers in Psychology*, vol. 9, p. 2451, 12 2018.

[22] J. Pearl, "The foundations of causal inference," *Sociological Methodology*, vol. 40, pp. 75–149, 2010.

[23] Y. Zheng *et al.*, "Causal-learn: Causal discovery in python," *Journal of Machine Learning Research*, vol. 25, no. 60, 2024.

[24] D. Janzing, *et al.*, "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[25] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 5, 2021.

[26] G. Zhou *et al.*, "Causal analysis of speech recognition failure in adverse environments," in *Proc. ICASSP*, 2002.

[27] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *In Proc. Advances in Neural Information Processing Systems*, 2020.

[28] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *In Proc. ICML*, 2023.

[29] G. Gendron *et al.*, "Behaviour modelling of social animals via causal structure discovery and graph neural networks," in *In. Proc. International Conference on Autonomous Agents and Multiagent Systems*, 2024.

[30] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[31] S. Xu, Y. Li, S. Liu, Z. Fu, Y. Ge, X. Chen, and Y. Zhang, "Learning post-hoc causal explanations for recommendation," *arXiv preprint*, vol. arXiv:2006.16977, 2020.

[32] J. Zhou, Y. Lin, Q. Chen, Q. Zhang, X. Huang, and L. He, "Causalabsc: Causal inference for aspect debiasing in aspect-based sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 830–840, 2024.

[33] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, Prediction, and Search*. MIT Press, 2000.

[34] P. Spirtes *et al.*, "Causal inference in the presence of latent variables and selection bias," in *Proc. UAI*, 1995, pp. 499–506.

[35] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, p. 524, 2019.

[36] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proc. Advances in Neural Information Processing Systems*, 2016.

[37] A. Renkl, "Toward an instructionally oriented theory of example-based learning," *Cognitive Science*, vol. 38, no. 1, pp. 1–37, 2014.

[38] K. Zhang *et al.*, "Kernel-based conditional independence test and application in causal discovery," in *Pro. UAI*, 2011, pp. 804–813.

[39] X. Wei *et al.*, "Automatic speech recognition and pronunciation error detection of Dutch non-native speech: Cumulating speech resources in a pluricentric language," *Speech Communication*, vol. 144, pp. 1–9, 2022.

[40] O. Duskin, "Estimating word difficulty in english using python: A practical guide," 2023, available online: https://medium.com/@omerduskin/estimating-word-difficulty-in-english-using-python-\protect\@normalcr\relaxa-practical-guide-8f6812de5122.

[41] C. M. Bishop, *Deep Learning: Foundations and Concepts*. Springer, 2024.