information

MDPI

Review

Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition

Philipp Gabler [1], Bernhard C. Geiger [1], Barbara Schuppler [2] and Roman Kern [1,3,*]

[1] Area of Knowledge Discovery, Know-Center GmbH, 8010 Graz, Austria
[2] Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria
[3] Institute of Interactive Systems and Data Science, Graz University of Technology, 8010 Graz, Austria
* Correspondence: rkern@know-center.at

Abstract: Superficially, read and spontaneous speech—the two main kinds of training data for automatic speech recognition—appear as complementary, but are equal: pairs of texts and acoustic signals. Yet, spontaneous speech is typically harder for recognition. This is usually explained by different kinds of variation and noise, but there is a more fundamental deviation at play: for read speech, the audio signal is produced by recitation of the given text, whereas in spontaneous speech, the text is transcribed from a given signal. In this review, we embrace this difference by presenting a first introduction of causal reasoning into automatic speech recognition, and describing causality as a tool to study speaking styles and training data. After breaking down the data generation processes of read and spontaneous speech and analysing the domain from a causal perspective, we highlight how data generation by annotation must affect the interpretation of inference and performance. Our work discusses how various results from the causality literature regarding the impact of the direction of data generation mechanisms on learning and prediction apply to speech data. Finally, we argue how a causal perspective can support the understanding of models in speech processing regarding their behaviour, capabilities, and limitations.

Keywords: automatic speech recognition; causality; speaking styles; data generation processes; annotation

check for updates

Citation: Gabler, P.; Geiger, B.C.; Schuppler, B.; Kern, R. Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. *Information* **2023**, *14*, 137. https://doi.org/10.3390/info14020137

Academic Editor: Tudor Groza

Received: 18 January 2023
Revised: 13 February 2023
Accepted: 15 February 2023
Published: 19 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Automatic speech recognition (ASR) has been a well-established and continuously advancing discipline for several decades [1–3]. Its core application is the transformation of speech signals into a textual representation, i.e., into an orthographic or other grapheme-based transcription. With time, this has progressed from finite vocabulary command recognizers to automatic transcription of spontaneous conversation, and from the fixed idiolect of a single speaker to various dialects of arbitrary speakers [4]. In parallel, mobile and ubiquitous computing devices have become more and more integrated into daily life, including natural language interfaces. Facilitated through the continuous increase in computation power and advances of machine learning technology, the utilization of ASR for speech-based conversational agents has come into close reach [5]. This will open new possibilities of interaction with machines, such as voice user interfaces, voice assistants, voice-based assistive technologies [6,7], and similar methods which can make information technology and computational power more accessible to diverse groups of people.

All these goals require speech recognition models which are strong, robust, and flexible. Such an objective may be partially in line with the competition for a better-than-state-of-the-art performance still mostly present in current-day machine learning, but requires methodology to go beyond it. While in some settings, ASR systems are able to reach human levels of recognition performance, there is still ground to be covered in many domains

and aspects. This could originate from the separate investigation of speech recognition as an engineering and modelling problem, where engineering is focused on recognition and de-emphasizing the environment in which speech data are being generated [8–10]. Especially in spontaneous or conversational settings, peculiarities of certain speaking styles, variation in and between speakers, audio quality and overlapping speech, and not least, imperfect transcription remain challenges at the frontier of speech technology. Notably, the properties complicating the recognition of spontaneous speech cannot just be dismissed as a different amount of noise; it is not that annotators are working sloppily, but the fact that transcription is necessary at all, and the perfect reconstruction of speech is fundamentally impossible, which distinguishes spontaneous from read speech data.

This review stands in line with the recent works of Jin et al. [11] and Feder et al. [12], which introduce causal reasoning into natural language processing, but is, to our knowledge, the first study to apply causal reasoning to ASR. By reviewing from a causal perspective the assumptions made by speech recognition models about data, we aim to contribute to ASR in a similar vein, covering three main areas:

1. Since the causal perspective requires a researcher to be explicit about the assumptions and relations of the data generation process, we rigorously analyse different scenarios of ASR, outline their influencing factors, and introduce a classification scheme, thus providing insight into how read speech and spontaneous speech do not conform to the same data generation process, and hence, differ in their sources of noise (Section 2). Notably, there is no real ground truth for annotations in the case of spontaneous speech data. Learning from it therefore constitutes a noisy-label problem, and must be qualitatively analysed and interpreted differently from read speech.
2. We present consequences of the causal mechanisms involved in the generation of speech data to allow for better interpretation and explain failures and successes in learning from it (with a focus on spontaneous speech and its transcription), and ultimately provide directions for future model architectures (Section 3). Concrete aspects of this involve theoretical considerations resulting from speech recognition in causal and anti-causal prediction settings, such as the effectiveness of semi-supervised learning, and a discussion of possible distribution shifts occurring in ASR.
3. The perspective of causal reasoning offers a foundation for the judgement of performance, robustness, and transferability of estimates—in short, the generalization behaviour of prediction models when the assumptions of inference diverge from the actual data generation. Through this, we hope to facilitate scientific interpretation and advance the analysis of learning behaviour, and reconcile the two viewpoints of engineering and modelling.

2. Some Characteristics of ASR Data

There are a few unsupervised approaches for speech recognition [13,14], but almost all ASR systems are supervised, and hence, trained from labelled data; a corpora of recorded speech chunks, together with the corresponding texts [15]. Yet, what exactly is the relation between speech and text in these corpora? Unlike in domains where existing pairs of data are often readily found (such as, e.g., medical images and diagnostic results), labelled ASR data almost always need to be produced specifically for the task. The acquired data typically fall into two broad categories, known as read speech and spontaneous speech. In read speech, pre-specified text prompts are used as stimuli and their read-aloud realizations are recorded; read speech corpora include, for instance, the widely used TIMIT [16] and LibriSpeech [17]. For spontaneous speech, on the other hand, annotators are asked to orthographically transcribe existing recordings of speakers talking freely (e.g., the Switchboard corpus [18], IMS GECO [19], or the spontaneous part of the Kiel corpus [20]). Other examples fall outside this dichotomy: speech data with different characteristics include rereading previously transcribed or artificially constructed free speech, often called “prepared speech” (used to create “spontaneous-like” data sets under very controlled conditions), or the transcription of read speech signals where the original text is unavailable (e.g., news broadcast

corpora such as those in Weninger et al. [21] and Radová et al. [22]). Furthermore, we can have a recitation of memorized text or content, constricted speech, such as when dictating or addressing a voice interface, and conversational speech—the case of spontaneous speech involving an interaction between two or more speakers [23–25].

To be able to talk more precisely about these nuances, we will characterize speech data through the two-dimensional schema illustrated in Figure 1. The first dimension, *style*, is used close to the typical meaning in the research community, and describes different degrees of (apparent) spontaneity, as they would be captured by the distribution of utterances modelled by a language model. We can consider spontaneity a coincidence of two components which strongly correlate in a specific context, and depend on modality, adaption of the speaker, and pragmatic and social constraints. From a speech production perspective, spontaneity expresses through directedness and locality—how much is an utterance planned ahead, or even revised (also referred to as "temporality"; cf. Auer [26]). In terms of reception, it typically aligns with variety or register, i.e., the distinction between (formal) written and (informal) spoken language. We will, henceforth, describe the axis as being delimited by the two poles of *unplanned* utterances, which are produced extemporaneously and informally, and *planned* utterances, which are formal and carefully deliberated (think of an unprepared monologue compared to a rhetorically crafted oral presentation). As the second dimension, which is not an established distinction, we use *mode* to refer to the directionality of the relationship between text and speech. This characterizes the process by which one part of the data is acquired from the other: is the text used as a stimulus to elicitate a *recited* speech signal, or is an existing speech signal *transcribed* to text? Under this schema, read speech is an example of stylistically planned speech in recited mode, whereas spontaneous speech can be categorized as transcribed speech in an unplanned style.

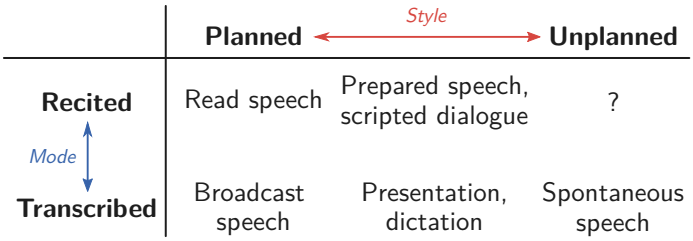

Figure 1. A finer categorization of ASR data by speech production styles and data acquisition modes, with exemplary instances. The conventional terms "read speech" and "spontaneous speech" constitute opposing poles. The style axis is a continuum: "semi-planned" situations, as in dictation or some news speech, are possible. It is unclear whether completely unplanned recited speech (marked with "?") is possible at all (perhaps the oral performance of certain stream-of-consciousness texts would come closest).

For both modes, one part of the data is given, independently chosen and directly manipulable, while the second part has to be derived from the first, specifically for the task, and with extra effort. In this way, style and mode complement each other: where the former is determined by the independently given part of the data, the latter characterizes the relation between it and derived part. The established terms of read and spontaneous speech constitute the prototypical cases on the opposing diagonal poles of this schema. In a sense, they follow the "natural" correlation that utterances in unplanned style are more likely to be transcribed; but in any case, in practise they constitute the data categories most frequently used for training ASR systems. Hence, from now on, our attention will be directed at these two kinds, especially insofar as they differ in mode, which will turn out to play an important role in our causal analysis.

2.1. The Case of Spontaneous Speech

Read and spontaneous speech could superficially appear to be just two complementary forms of producing the same kind of data—pairs of text and signals. Even so, they have

been observed to result in remarkable differences—see the historical trends in Figure 2. While for read speech, a human-like performance of less than 5% word error rate can be achieved, a worse limit seems to exist for spontaneous speech [27,28], where 20% to 50% have been typical at least until the arrival of deep learning, and values below 10% are still an exception [29–32]. Several factors are responsible for this gap in performance. Whereas, in terms of style, read speech mostly conforms to prescriptive norms (as the language of text prompts is usually in planned, written style), and speakers are more likely to apply clearer pronunciation, spontaneous speech tends to deviate from these characteristics. Being unplanned, it shows syntactic signs of its on-line production [26], higher degrees of segmental reduction [33], and more intricate forms of variation (speech rate, fillers, self-corrections, repetitions, etc.) than read, and thus, planned, speech. Consequently, already due to its grammatical and acoustic peculiarities, it is to be expected that spontaneous speech is harder to automatically process than read speech, and does not allow for simply reapplying models trained on read speech. This is a seemingly obvious but still widely ignored argument in the field of speech technology. For example, Schuppler [34] questions the transferability of improvements between speaking styles, and recommends that the nature of the data needs to be taken into account already when defining the concepts and basic assumptions of a speech recognition method.

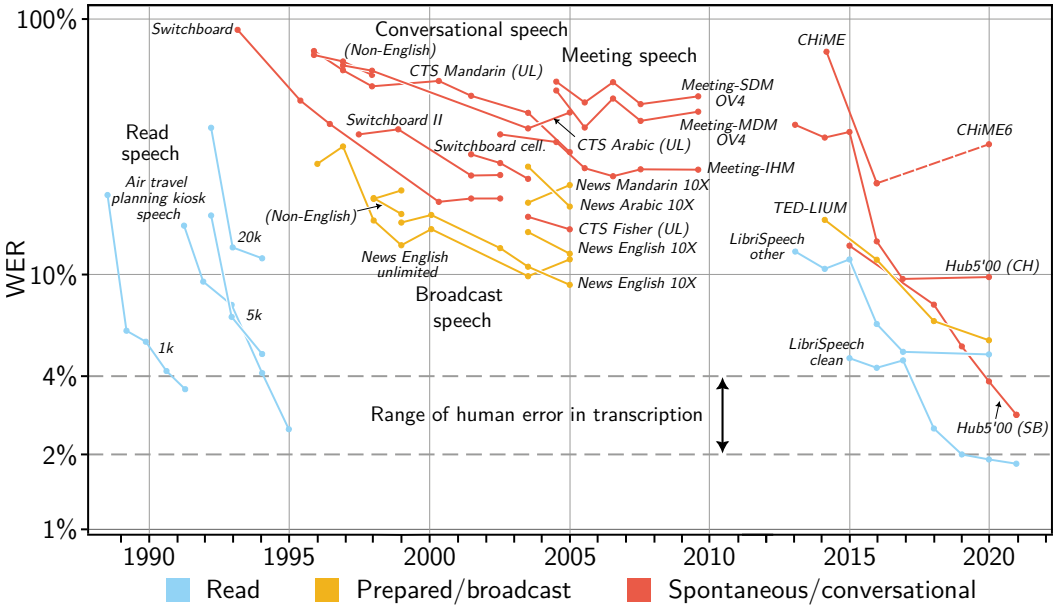

Figure 2. The historical progress of speech-to-text benchmarks, extended from Ajot and Fiscus [35] (as cited in Valenta and Šmídl [28]). We have added new points after 2009 from collected benchmark results of deep learning methods trained on newer data sets, using the best results for each year as listed in Synnaeve [36]. Of these, LibriSpeech is (mostly) read speech; TED-LIUM contains prepared and spontaneous speech; HUB5'00 is built from Switchboard and CallHome, which are types of conversational speech; CHiME and its successor CHiME6 consist of noisy low-resource conversational speech. We can see how performance gains become flatter the more spontaneous the speaking style becomes.

Besides these more directly noticeable differences in style, mode is also a factor complicating speech recognition for spontaneous speech: since transcriptions are added through a process of annotation, such speech lacks a real textual ground truth, or "gold standard". This lack manifests itself in the form of "noisy labels" (transcriptions, in our case), a known problem in machine learning [37] as well as in natural language processing, where learning under annotation noise has been theoretically shown to result in biases for certain classification problems [38]. Due to the more complicated grammatical and acoustic nature of unplanned utterances, ranging from simple ambiguity up to extreme reduction causing

"disappearance of words" [39], spontaneous speech transcription is also afflicted with errors of usually 5% to 10% [40]. Accordingly, negative impact on recognition can be observed at least on the lexical [28] and phonetic level [41]. In addition, the result of the transcription process, unlike with the arguable objectivity of the text prompt in data created through recitation, is not just noisy: it might not congrue with any concept present "in the head of the speaker", even if all annotators agree in their perception of what has been said. Any kind of annotation presupposes linguistic theory, as well as language competence and experience of the annotators, and is, therefore, an act of human interpretation. Hence, orthographic transcription can also never be perfect, even if apparently objective criteria are used [42] (Chapter 4.2); this is exactly what we refer to here as a "lack of a real ground truth".

To mitigate uncertainty induced by data collection under these conditions, annotation processes usually involve multiple iterations and corrections by trained annotators, using an annotation schema which should be as well-defined as possible [43]. The reliability of the annotated data can at least be judged by estimating the base level variability through the comparison of multiple annotated texts, and quantified with inter-annotator agreement measures [44]. Alternatively, explicit models of annotation can be used, which allow for more nuanced analyses of annotation noise and make the quantification of reliability more interpretable [45,46]. In speech science (phonetic studies, prosodic labelling, etc.), inter-annotator agreement is frequently reported, but to our knowledge, no studies regarding the relation between inter-annotator agreement and recognition performance for ASR systems exist. In application-oriented machine learning, such considerations seem to be taken less seriously than in (computational) linguistic research; for example, Geiger et al. [10] found that of all investigated publications on a certain text classification task, only about 70% report inter-annotator agreements at all, about 50% provide no information about annotation guidelines, 85% provide no details about the training of annotators, and 45% do not even specify the number of annotators.

The orthographic transcription of spontaneous speech is an extreme case of imperfect annotation, as the vocabulary and space of acoustic realization are both unbounded and the involved variation can be extremely diverse, leading even to unanswerable situations of stark disagreement between annotations, or labelling as "incomprehensible"—not to mention the case of a basic lack of competence, such as with unfamiliar dialects or topic knowledge. For illustration, consider two examples from the GRASS corpus [23], with inaccurate transcriptions of a German conversation involving a "Reflow-Ofen" (reflow oven; a certain technical apparatus):

- `musst einen <?>Ofen bauen` "[you] must build a <?> oven"
- `hast einen <*ENG>reflow offen` "[do you] have a reflow open"

In the first chunk, the full technical term was apparently unknown to the annotator and transcribed as partly incomprehensible with <?>. In the second one, the English part was understood and transcribed correctly, but "Ofen" was mistaken as a separate adjective "offen" (open), even though the initial vowels of both words are phonetically different. While in this case, the "correct" transcription is obvious to someone familiar with the term, there are clearly cases when this is impossible, and parts of utterances are completely unrecoverable.

3. Modelling beyond Prediction: The Relevance of Causal Reasoning

Were it only for different amounts of uncertainty arising from unplannedness and transcription, achieving human-level performance in recognition of spontaneous speech would pose merely an engineering problem, which could be hoped to be solved by using more data, more complex models, and more computational effort. However, there is a more fundamental factor of discrepancy at play: while the prediction problem solved by ASR is the same for read and spontaneous speech, going from speech to text, the two different modes of data correspond to oppositely directed data generation processes, which are not aligned with the direction of prediction, as depicted in Figure 3. Concretely, in ASR on data

from recited speech, the textual ground truth is sought to be recovered backwards from the output signal after the recitation by the speaker. (This distinction is similar to the notions of "forward" and "inverse" problems with respect to a physical theory; cf. Tarantola [47].) In contrast, in ASR on transcribed speech, a learner is fit to emulate the process of transcription, going forwards from the speech signal to the textual output. These two settings both contain a transformation mechanism and sources of stochasticity, and so could be considered as just two sides of the same coin, with the corresponding prediction problems convertible into each other by using the opposite directions of Bayes' rule—which is also what generally is carried out. However, this conversion cannot, in general, be held up when variables or relationships in the data generation process undergo change. For instance, in read speech, a change in the recording environment, skewing the signal, would not be mirrored in the corresponding text because it is the text prompt which causes the speech signal, while the same change could considerably impact the transcribed text in the spontaneous speech case. Conversely, exchanging annotators (with their idiosyncratic behaviour) for spontaneous speech cannot affect the underlying recording—as the transcription does not cause the original signal—as opposed to manipulating the text prompts in read speech, which *does* (in general) massively change the recordings. These examples illustrate how conditional probabilities—as in Bayes' rule, which may already be difficult to calculate in practise due to the potentially intractable posterior densities [48]—are not automatically compatible with changes in the causal mechanisms underlying the data. In addition, the different behaviour under manipulation also implies different effects on prediction. In order to provide a language to analyse these kinds of differences and their consequences through mathematical and graphical tools, we introduced causal reasoning.

Figure 3. Conceptual illustration of the different generative processes underlying read and spontaneous speech. Solid arrows indicate the direction of the process, dashed arrows the direction of the prediction.

3.1. Formalization of Causal Models

Naive application of predictive statistical methods to observed data can lead to wrong or even paradoxical results when applied carelessly to scientific enquiry [49]. Causal reasoning attempts to conceptualize the causal assumptions of a domain, and provide the tools to remedy this situation by a mathematically sound framework [50,51]. In ASR literature, the term "causal" is often used to describe alignment with the flow of time, such as the prediction of future events given the past. For example, in Healy et al. [52], a "causal" algorithm may only use past time frames in the context of real-time speech processing, similarly to the notion of a causal filter in signal processing. Like Granger causality [53], these usages refer to causation in a weaker, observational sense than the counterfactual definition prevalent in causal reasoning. They generally only describe prediction, not causation, by capturing assumptions such as that the past may affect the future, but not vice versa, without considering the structure of underlying physical mechanisms. By providing new concepts and formalisms to align predictive modelling with the underlying data generating processes, causality extends the expressivity of models and the instruments available to analyse them [50]. As such, it is expected that causally correct modelling will yield solutions that are inherently more in line with the data generation process, and thus, more in line with changes in the environment [54,55], and can address the apparently divergent goals of prediction and explanation [56].

One of the main definitions of causality is based on the notion of *counterfactuals*: "We think of a cause as something that makes a difference, and the difference it makes must be

a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well." [57]. This perspective considers the hypothetical ("would have") outcomes of an event if the value of the cause had differed from what it actually was. In the case of causes resembling a binary treatment, counterfactual is the fictitious outcome if the treatment would have been absent. An alternative definition is given by Pearl and Mackenzie [58], who state that causality is simply the answer to the question: Why? What unifies the different schools of thought in regard to causality is that they all seek to rigorously formalize the connection between the observations—the data—and the representation of the data generation process in the form of causal models.

Such causal generative models support the interpretation and explanation of observed phenomena, especially when applied in scientific contexts [59,60]—being able to reason about the inner workings of a system is a necessary requirement for understanding its causal structure, which consists of the mechanisms and noise sources relating the individual components. This motivation is usually claimed for "plain" generative models, but results only from their causal interpretation (Pearl [51] (p. 22) notes that generative models are typically, even if unintentionally, constructed with causal mechanisms in mind). Causal models go one step further and decidedly require the applicant to make their assumptions about the real underlying mechanisms explicit. Defining a causal model means to specify the generative processes, which, in contrast to observed associations, are considered to correspond to real, physical mechanisms:

> [C]ausal relationships are more "stable" than probabilistic relationships. We expect such difference in stability because causal relationships are *ontological*, describing objective physical constraints in our world, whereas probabilistic relationships are *epistemic*, reflecting what we know or believe about the world. Therefore, causal relationships should remain unaltered as long as no change has taken place in the environment, even when our knowledge about the environment undergoes changes. [51] (p. 25)

Such information cannot simply be derived from data alone, but must generally be acquired through independent reasoning and domain knowledge, thereby serving as a potentially stronger inductive bias for learning. (Note that this is not an absolute preclusion; there is a growing body of work on causal discovery and structure learning (cf. Vowels et al. [61]), which we will not examine further. In the case of read and spontaneous speech, the directions of the causal mechanisms are clearly given.) Causal statements requiring such assumptions are often eschewed, be it on epistemological grounds or simply due to a lack of mathematical formalization, but can in fact be made statistically rigorous [62]. The important ingredients therefore are the concepts of mechanisms and manipulation: where non-causal predictive models only exploit associations ("association is not causation"), causal generative models aim to faithfully represent how data are created through a series of causal mechanisms [63]. These *independent causal mechanisms* (ICM) [64] are supposed to share no information and remain constant, even if the environment changes, also known as independent mechanism assumption [65]. Complex systems are often assumed to comprise several independent internal dynamics, but when observed, only a mixture of them can be studied. An example is given by Gresele et al. [66], who analysed the cocktail party problem, where the content of the speech signal is considered to be independent of the placement of the microphone and room acoustics, and furthermore, the effects of the positions of individual speakers on the recorded signal are independent. Their causal generative model can capture these dependence and independence relationships, and allows for an improved separation of speakers.

Different formalizations of causal structure can be used to mathematically express and analyse the consequences of noise and structure changes onto the behaviour of learned models; there are a variety of approaches to define causal generative models. So-called causal diagrams, as those in Figure 3, are a simple graphical way to capture the essential causal relationships [67,68], going back to Bayesian networks (i.e., directed graphical

models), with their nodes representing variables, and directed edges direct causal relations (i.e., mechanisms, not only statistical dependencies), together forming a directed acyclic graph. Causal diagrams offer an intuitive way to study the causal relationships, but the nature of the relations is left unspecified. Structural causal models (SCMs) represent another major causality framework, connecting causal graphs with structural equations, and are able to express the specific relations between variables in closed form [51]. SCMs also formalize do-notation [69], which aims to rigorously describe the effect of interventions modelled via random variables and their probabilities. A third formalization based on the counterfactual definition of causality is the so-called potential outcome framework [70].

Unlike a pure statistical model, which defines a specific probability distribution, a causal model includes a structural representation entailing multiple distributions [64]. The complete causal model then is considered a rendering of the data generation process, i.e., the associations in the data are considered the result of the causal effects. If the causal model is correct and no assumptions are violated, an improved interpretation of the observed data is possible, since it provides explanatory power ("why is … happening?") and affords counterfactual questions ("what will happen if …?"). This allows for us to reason about the real world even under hypothetical manipulation by explaining phenomena through entities connected through causal relationships. In contrast, a model not respecting causal relationships can only make predictions in the form of likely associations generalized from past observations. When wrong or incomplete mechanisms are assumed (by misspecifying or leaving out causal relationships), predictions based on these wrong assumptions might fail [68,71]. Not only does the insistence on working with a causally justified generative model facilitate the correct probabilistic interpretation of interventions, it can also provide a formal explanation for the occurrence of different biases in estimates derived from observational data (e.g., unaccounted confounders or selection bias), and means to their remediation.

3.2. A Causal Analysis of ASR Data Generation

It should by now be clear that read and spontaneous speech, insofar as they differ in mode, need to be described by separate models of data generation (cf. Figure 3). From a purely probabilistic standpoint, if we denote the complete text by W and the speech signal by S, the joint density $p_{W,S}$ can be validly decomposed as either $p_S(s)p_{W\mid S}(w \mid s)$ or $p_W(w)p_{S\mid W}(s \mid w)$ (ignoring more fine-grained factorizations, such as the internal structure of the language model or the intricacies of the transcription or speech production process). Choosing any one of these decompositions and fixing it as a generative model amounts to a causal interpretation: $W \to S$ for recited, and $S \to W$ for transcribed speech, with reversed roles of cause and effect. There are other causal models implying the same joint distribution, the simplest cases of which being $W \leftarrow U \to S$ with some additional unobserved variable U, and conditionally independent W and S, but the validity of the two named choices is quite intuitive from a counterfactual perspective. If, in a read speech setting, S is manipulated, say by cutting the cable of the microphone, the proper model of the situation will be $p_W(w)p'_S(s)$, with a new interventional distribution p'_S for S; because we have intervened on the effect S, the relationship to W is rendered independent, as p_W and $p_{S\mid W}$ are independent mechanisms. We have illustrated how such an intervention can be formally defined through structural equations in Figure 4. The joint distribution after an intervention is clearly different from $p_W(w)p_{S\mid W}(s' \mid w)$, which is only the probability of observing the signal of a cut microphone, s', under the old model. In reverse, if the cause—the text prompt—is modified, we arrive at an interventional distribution $p'_W(w)p_{S\mid W}(s \mid w)$, where the dependency of S on W is preserved. An opposite intervention can be imagined for transcribed speech $S \to W$, e.g., by editing a recorded signal S, which would directly affect the transcription mechanism. This effect is relevant especially in spontaneous speech, where there is no ground truth, and annotation errors arising from homophony, ambiguity, lack of competence or topic knowledge can skew data sets.

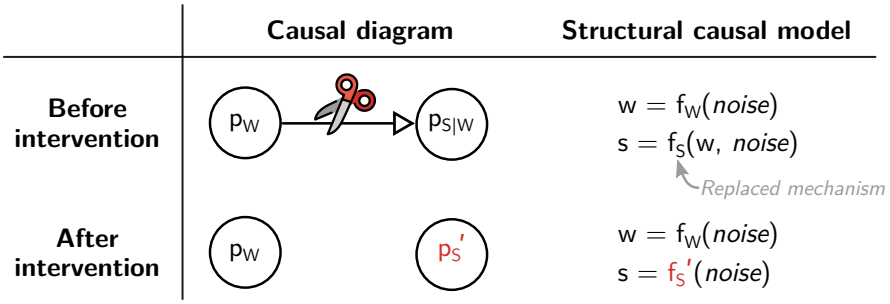

Figure 4. An intervention can be explained formally through its relation to structural causal models. Every conditional distribution with density $p_{Y \mid X}$ can be written in the form of a deterministic mechanism f_Y, of which is a function of the values of the parents X and an independent noise part. The joint distribution of a set of variables is thus given as a structural causal model by the set of mechanisms and noise distributions. An intervention in this formalism consists of replacing a mechanism. In the exemplified case—cutting a microphone cable—the dependency on the parent is removed, resulting in a mechanism independent of the W. The principle of an independent causal mechanism refers to the fact that W and S are statistically dependent, while their mechanisms f_W and f_S can be independently manipulated.

With this knowledge, we can discuss a number of implications of the direction of causality on downstream tasks. Jin et al. [11] in a study similar to ours highlight the importance of taking into account causal directions already in the data collection stage, and of subsequent causality-aware modelling. In their work, they analyse a number of common NLP tasks, motivated by the case of machine translation, where one sentence in one language is the cause of the corresponding sentence in another language. For the different directions in machine translation, they find empirical support for their assumed independent causal mechanisms by estimating minimum description lengths as approximations of Kolmogorov complexity [72] (Chapter 7). Interestingly, this setting corresponds directly to the case of recited speech and spontaneous speech (see Figure 3). Depending on the direction of translation in the ground truth and the direction of translation the learned system should make, this is either a case of causal or anti-causal learning, affecting semi-supervised learning and domain adaptation, of a similar kind to the arguments we provide next.

3.2.1. Inference in Causal and Anti-Causal Settings

In semi-supervised speech recognition, the goal is to improve the performance of a supervised approach by utilizing additional information from unlabelled inputs [73–75]. This is feasible since, typically, labelled instances are few and costly—they need to be recorded or transcribed—while suitable unlabelled audio data can be obtained with much less effort. The setting of semi-supervised learning is of particular interest in the context of causal and anti-causal learning, extensively studied by Schölkopf et al. [65], to which we refer to for more technical background information. We can transfer their theoretical results to ASR as follows: For the causal model of speech and text in spontaneous speech, $S \rightarrow W$ with cause S (signal) and effect W (text), the joint probability $p_{S,W}$ can be causally factorized into $p_{W \mid S}$ and p_S. Assuming the independence that causal mechanisms hold, the conditional mechanism $p_{W \mid S}$ will then be (structurally) independent from the mechanism of the cause, p_S. Thus, the cause does not contain information about the mechanism producing the effect, and when estimating $p_{W \mid S}$ via a learning algorithm, knowledge about p_S through unlabelled instances is not expected to improve this estimate. For the anti-causal case of read speech, $W \rightarrow S$, the situation is different: when predicting the cause from the effect, $p_{W \mid S}$, an independence of p_S is no longer expected, and hence, can be exploited in a learning context. Hence, knowledge about p_S may improve the prediction in the anti-causal direction via unsupervised methods to capture the marginal distribution of the effect, so that in practice, unlabelled training examples (i.e., signals only) may help to improve the

predictions in the anti-causal setting. Observations from recent semi-supervised deep-learning-based ASR systems, such as Zhang et al. [32], are in line with this prediction: even with extremely large unlabelled data sets, better results are still achievable on the test data of recited modality. Furthermore, in a certain sense, the "self supervision" approach that is taken when using a model such as Wav2Vec [76] as the basis for a fine-tuned predictor is a variant of the same technique, and follows the same pattern.

The causal direction also interacts with the distinction of generative and discriminative methods, i.e., learning via the joint or the conditional distribution. A purely discriminative approach aims to directly estimate the mechanism that maps the input variables to the output variable. Depending on the direction of the underlying data generation process, this may either align with the causal or the anti-causal direction. Blöbaum et al. [77] study the relationship of the causal direction and the two approaches in machine learning. Their work is based on the hypothesis that for anti-causal problems, generative approaches are expected to perform better than for a problem setting where the direction of the causal mechanism and prediction align. This hypothesis is rooted on the ICM assumption, where for causal inference, the mechanism of the cause (p_S in spontaneous speech) is not expected to be informative: knowledge about the structure of the signal alone does not immediately improve the ability to infer correct annotations. In contrast, in anti-causal inference, knowing the marginal distribution of the effect (p_S in read speech) will help in the estimate of the joint probability of cause and effect ($p_{S,W}$). Kilbertus et al. [78], on the other hand, investigate the ability of machine learning models to generalize to unseen data. For anti-causal problems, they find that what they call strong generalization (model identification, not only imitation) may not be achieved, even with state-of-the-art deep learning approaches and large amounts of data. The needed extrapolation outside of the support of the input is found to be especially problematic for generative approaches. This prediction is in line with the empirical observation that, while models trained for read or at least stylistically planned speech perform well on inputs close to their training data, they badly generalize to other settings such as spontaneous speech. The authors argue that this should be mitigated by integrating causal knowledge in order to guide search and validate predictions.

3.2.2. Causal and Anti-Causal Inference under Shifts

Recognizing the potential results of distributional changes and appropriately mitigating their effects is key to stability, generalizability, and transferability of learned models—we need to understand all external influencing factors and their relations. It should be emphasized that such changes are not a purely theoretical concern; on the contrary, the production of speech corpora involves many sources of stochasticity, and thus, many variables which, on purpose or by accident, shift their distribution. Potential sources of change are speakers, recording equipment or acoustic environment, or the annotators. Indirectly, even the passing of time itself continuously manipulates the data generation processes, as the involved annotators or speakers change behaviour (gaining experience or getting more used to the processes).

Causally, all such changes, be it from manipulation or environment change, can be treated as interventions. As there are only four possibilities, resulting from the combinations of causal direction (i.e., mode) and the location of the intervention, we propose an exhaustive terminology of shifts for the bivariate ASR setting following Castro et al. [79], visualized in Figure 5. Intuitive examples are straightforward to find: prevalence shift could result from changing text prompts, acquisition shift from different reciting speakers or recording equipment, population shift from different spontaneous speakers or recording equipment, and annotation shift from change in annotators or annotation practises. The appropriate causal model allows for us to identify these shifts and to take steps to improve robustness under distribution change. The abstract bivariate setting under which both modes of ASR can be subsumed is studied in detail by Schölkopf et al. [65]. By assuming only rather weak assumptions about noise structure, such as additive noise models [80],

they provide a set of mathematical modules and recipes for inference under shifts, covering well-known concepts such as concept drift (change of mechanism, i.e., acquisition or annotation shift in our terms), covariate shift (i.e., population shift), or semi-supervised learning and transfer learning from a causal perspective.

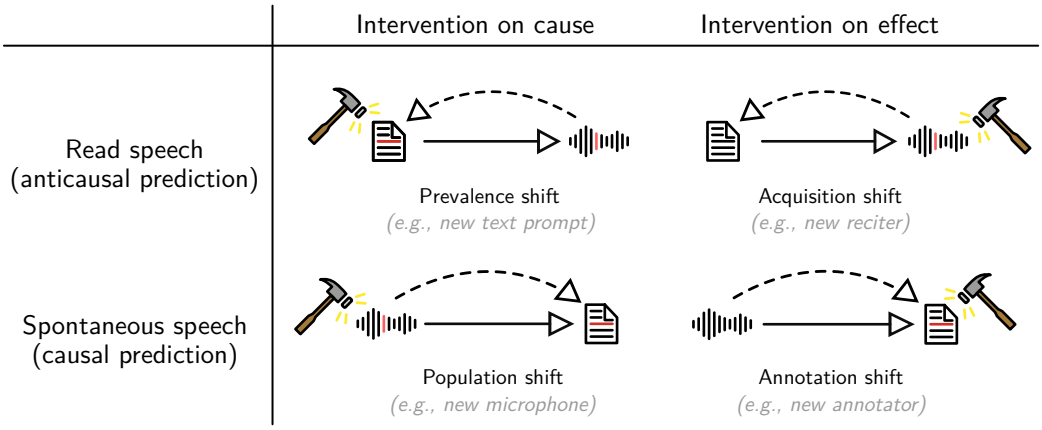

Figure 5. Classification of shifts in causal and anticausal prediction settings (solid arrows are causal mechanisms, dashed arrows are prediction directions). Upon intervention on speech signal or text, a change can propagate via the mechanism from cause to effect. Depending on the direction of prediction, shifts can have different effects on prediction and call for different measures of mitigation.

On a similar note, Cinelli et al. [71], while also concerned with more general causal graphical models, investigate several variants of bivariate effect estimation under the influence of "controls"—third variables directly causally related to cause or effect in a bivariate model, which can be seen as shifting either variable. Using only non-parametric assumptions, they classify several possible cases as "good", "neutral", or "bad", analysing the effect of including the third variable in regression. Depending on the case, such inclusion (or lack thereof) can lead to improved or degraded precision, or even bias. For example, including a child of the effect (Z in $S \to W \to Z$) results in what is elsewhere known as "case-control bias". We can relate this with annotation shift in our terminology, although it is not caused by an actual intervention; instead, by skewing the observed distribution through the inclusion of a superfluous variable, the regression acts as if under a changed environment.

3.3. A Causal Perspective on Learning in ASR Models

The methodology of ASR has always kept up with and innovated upon the developments in neighbouring disciplines (cf. Furui [81]). Early ASR systems employed a combination of linguistic theory (phonological and grammatical rules), signal processing techniques, and search algorithms (e.g., by dynamic programming), which can arguably be seen in relation to contemporary artificial intelligence and the rise of generative grammar, trying to mould speech recognition after theories of human language understanding. From the 1970s on, a trend to algorithmization can be observed, involving more sophisticated, less explanatory template matching, pattern recognition, and the rise of statistical techniques. The current state-of-the-art in ASR can roughly be split into two factions.

On one side, a range of more traditional approaches centred around Hidden Markov models (HMMs), which are a prominent example of generative models. HMMs have been introduced to speech recognition in the 1980s, and are accompanied by well-developed probabilistic and algorithmic theory (such as the Baum–Welch and Viterbi algorithms [82]). They enjoy consistent popularity and are available as toolkits such as HTK [83] and Kaldi [84]. These models are parametric generative sequence models, with a factorization implying specific conditional independence properties' "hidden" states (so called because they are usually the target of prediction), and observed values dependent on them. In the context of speech processing, the Markov model of hidden states is often called the *language model*,

and the observation model the *acoustic model*, because we can imagine a speech production process first generating an underlying sequence of segments, and then turning it into a wave form (such as when reading aloud). In practise, these parts are modelled as stochastic finite state automata selecting conditional mixtures of Gaussians for emission of observations. Given sufficiently strong transition and observation models (i.e., enough states and mixture components), HMMs are able to accurately approximate any observation sequence; however, this does neither mean that they are good models for speech synthesis, nor that they are well-calibrated representations of the generative process of speech [85].

The generative assumption underlying the vanilla HMM is the model $W \rightarrow S$, corresponding to the causal generative model of recited (read) speech, together with more specific assumptions about the factorization, namely W being Markov, and the elements of S being conditionally independent given W. This implies the interpretation that a well-defined sequence of segments is a "concept" within the speaker's mind, which we should be able to reconstruct. Young et al. [83] (p. 3), as a rare occurrence, make this explicit in the documentation of HTK: "[s]peech recognition systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols", and say their goal is to "effect the reverse operation of recognizing the underlying symbol sequence given a spoken utterance [...]". They also illustrate this traditional view of the speech recognition process, as seen in Figure 6. Such an interpretation may be causally valid in some cases, and may be a viable model for inference in even more, but ignores the fact that the transcribed case really calls for models of transcription, not of production. In NLP, several authors suggest that explicit annotator models may be used to improve the estimation of annotation noise [45,46,86]. From the perspective of causal reasoning, we see that such annotator models may have yet more advantages: they faithfully represent causal mechanisms in the data generating process, and may partake in causal analysis.

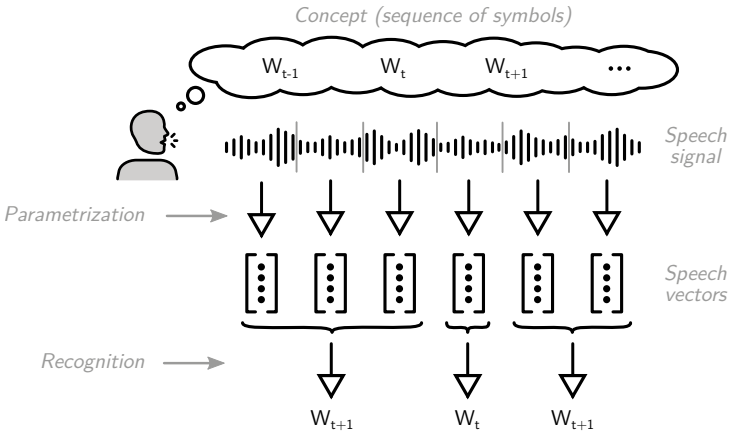

Figure 6. The causal assumptions built into HMM-style models made explicit, as illustrated in the HTK book (reproduced from Young et al. [83]).

The success which we see from HMMs is in a large part due to their capability of discrimination: they allow to rather easily learn likelihoods which can be used for the efficient and accurate prediction of the hidden state sequence, without the need to model every nuance of the distribution of possible utterances. Yet, in practise, ideal, accurate HMMs for realistic speech distributions can be too complex to use when acceptable predictive performance is required. Hence, several modifications to improve predictive capabilities have been proposed [87–90]. Although such models deviate from the generative assumptions implied by the original HMM, they improve the predictive performance, presumably by increasing the mutual information between observations and states [91]. These modifications instead lead to generative models which are not anymore as readily interpretable as a "read and pronounce" model of ASR data generation.

On the other end of the current state-of-the-art, we have ASR based on neural networks, which regained popularity during the 2000s and have profited from a fortunate combination

of increased computational power and technical advances in learning algorithms and network architectures (convolutional networks, regularization, pre-training, etc. [92,93]), as well as the availability of larger amounts of speech data. Neural networks have now improved their ability to deal with the temporal structure of sequential data, due to advances in recurrent networks [94] and other sequence models, such as the transformer architecture [95]. This has also led to some hybrid approaches combining neural networks with genuine speech processing techniques, such as connectionist temporal classification (CTC) [96,97]. The latest branch of neural ASR is based on deep learning, applying end-to-end deep neural networks such as Deep Speech 2 [98], or recent architectures based on self-supervised learning, such as Wav2Vec [76,99] or Whisper [100], which base their success on fine-tuning general pre-trained feature extractors [75]. A large part of the success of deep learning comes from the ability to directly train universal approximators as discriminative problem solvers, utilizing abundant amounts of data. Deep learning has always crossbred with other machine learning paradigms, and even monolithic end-to-end systems may contain components corresponding to the language and acoustic models in the HMM-style paradigm and use established decoding techniques such as Viterbi decoding or beam search. Yet, these architectures mostly forego any pretence of generative interpretability, of which is intrinsic in causal models, and are designed to directly maximize the predictive performance of a discriminative classifier.

Built on the advances in these technologies, recognition rates in ASR are ever-growing (although read speech still consistently beats spontaneous speech). However, this growth can also be interpreted as a sign of a preferential focus on prediction instead of explanatory modelling, which seems to be ingrained into machine learning and its applications [101,102]. A large part of machine learning progress is measured in terms of performance, and less attention is being paid to analysing and interpreting the generative processes assumed to underlie the data (cf. Chen and Asch [103]). This indifference to the origins of data has since long been called out as a "garbage in, garbage out" mentality, such as by Geiger et al. [10]. A striking example of this includes current deep learning models in natural language processing, which are trained as black boxes and then retroactively examined with "probes", using correlation to come up with possible interpretations of their inner workings, instead of theory guiding their construction [104].

To a certain extent, the same tendency also holds for ASR [8]. While some approaches, especially the ubiquitous hidden Markov-type models, can be interpreted as assumptions about data generation [85], this does not hold for many predictive models such as support vector machines and most neural networks [105]. Ostendorf [106] and Scharenborg [107] already call for more theory-guided modelling: to go beyond the "beads on a string" view of sequential speech data, and towards more cognitively motivated—and hence naturally causal—approaches, guided by insights about human perception (HMMs, in contrast, are decidedly models of production). Deng and Jaitly [108], who compare deep discriminative and generative models in speech recognition, argue from a learning theoretical perspective. They say that generative models limit the class of distributions that can be expressed by a model, serving as inductive priors. When applied under correct distributional assumptions, they are generally more interpretable, facilitate alternative learning modes (e.g., unsupervised or semi-supervised learning), and can achieve faster convergence with fewer data, given good priors. Discriminative models, on the other hand, who directly express complex posterior distributions, generally have a superior performance at the test time, can have better noise robustness, and have more scalable architectures and training algorithms at their disposal. The authors suggest that for combining their advantages and utilizing recently available computational methods, an integration of more generative approaches into the primarily discriminative deep learning models is necessary. Manning [109], at least for part-of-speech tagging (a task conceptually similar to transcribed speech recognition, since it involves labelled data), suggests to apply more descriptive linguistics in classification tasks based on annotated data, and reflect on the appropriateness of tag categories in order to close the gap between predictive classifiers and human annotators:

> Notwithstanding the significant progress that can be made by removing errors and improving the consistency of the treebank, there are interesting foundational linguistic issues as to which decisions are linguistically well-justified, and which turn into arbitrary conventions of treebank annotation.

Switching "treebank" to "speech corpus" is a critique which might equally be applied to transcribed speech data, and resonates with our arguments in Section 2: the absence of noise and consistency of transcription are not sufficient criteria for quality, and are not enough to close the performance gap to humans. Furthermore, orthography and even specialized annotation schemata are not adapted (well enough) to phenomena of spontaneous speech such as disfluencies, fillers, speaker noise of paralinguistic function, etc. As Reidsma and Carletta [110] note, annotation noise in linguistic data is typically not at random, and "[machine learning] makes systematic disagreement dangerous, because it provides an unwanted pattern for the learner to detect", making pure disagreement measures insufficient and inflating performance measures. Compare this to the transcription example at the end of Section 2.1: what would a "word error" even mean in a case like this, where the transcribed tokens, which are used as a reference, are inconsistent and questionable themselves?

4. Conclusions

The aim of this work is to provide a new perspective on speech recognition: the introduction of causal reasoning about data and models, and a first overview of its consequences. Causal reasoning urges us to be very accurate in matters relating to the data generation process in the modelling and development of systems—in the case of ASR specifically, the dependency between speech and text. As we have demonstrated, even in such a simple case as the bivariate setting with one cause and one effect, there is much to say about the relations between causal structure and modelling. In Section 2.1, we pointed out how annotation is a fundamental element determining the nature of spontaneous speech data. Not only is it advisable to take more seriously the lack of a ground truth and noisy nature of its labels, but further research into its interaction with prediction is necessary: how and where do annotators and ASR models make errors? Are they the same? What would be a meaningful measure of correctness, or performance, in the absence of a gold standard?

Section 3.2 indicates several possible directions of future research rooted in the literature on causal inference and machine learning. The structure of the independent causal mechanisms determines conditions under which unlabelled data may be beneficial, and hence, when semi-supervised learning can be applied successfully. Our analysis suggests that the direction of these relations should influence the choice and expected performance of machine learning modelling, and impacts the suitability of generative or discriminative approaches, which in turn result in different degrees of interpretability. The actual extent of these effects remains to be studied more systematically in the field of speech recognition, as well as the details of the means to deal with distributional changes or "shifts" occurring in situations such as transfer learning, or just naturally as artefacts during data collection or generation. By contrasting the generative processes of different kinds of speech data, we can classify shifts (matching different kinds of error sources in ASR data), formalize them, make theoretical statements about their consequences in predictive modelling, and give directions for possible mitigation strategies. More research is required to make the necessary assumptions explicit, and amend the modelling of all relevant sources (causally) influencing observed signals. Such results could then be used to arrive at more robust systems, performing well even in the presence of shifts in the data reflecting changes in the environment.

From the more abstract observations in Section 3.3, we can highlight that the interaction between the characteristics of data generation and modelling in ASR needs to be investigated further. We need to make explicit the underlying assumptions and mechanisms of speech data acquisition, with its two distinct forms of generation processes, and consequently compare them to the models underlying typical inference algorithms,

then lay out the specific sources of stochasticity occurring in each, and finally try to give an interpretation for some of the observable differences between ASR on read and spontaneous speech. A concrete step in this direction is the same as suggested by Jin et al. [11]: to always annotate or indicate the causal direction when collecting new speech data. More theoretical analyses are needed regarding the interplay between learning theory, causality, and ASR (for an exemplary argument in this direction, consider Appendix A). The question of generative vs. discriminative and black-box vs. interpretable modelling finally also leads back again to the questions we started with: what to make of data in which the labels themselves are the product of an intricate, non-uniform, and ultimately "subjective" stochastic process, such as the orthographic transcription of spontaneous speech?

Our examples were reduced to the bivariate setting, and thus limited to a very coarse granularity; however, we can also take a more fine-grained view and model the internal structure of the causal relations between variables at each time step. As suggested by the style/mode distinction, the processes of speech production are structured on two axes: the "vertical" direction of mode (signal to text or vice versa), and the horizontal time structure, roughly described by style. For example, the acoustic model for read speech respects temporal ordering: every frame can depend only temporally on previous tokens. The Markov assumption used in HMM-style models is a causally valid model for this, but more complex alternatives are possible. The transcribed tokens in spontaneous speech, on the other hand, may also depend on the "future" of the speech signal: annotators are able to move forwards and backwards in recordings, giving the process a window-like (or attention-like) structure. Even more complicated causal models are needed when moving beyond single-speaker read and spontaneous speech: in conversations, crosswise interactions between multiple participants must be integrated. The transcription of conversational speech, or multi-stage annotation processes such as typically applied to increase confidence, can lead to even more complex causal annotation models. Having gone this far, the integration of speech synthesis as a module in a deep causal model of participants and their observable and unobservable states offers itself. We, furthermore, see the explanatory nature of the causal model structure as a starting point for an extended view on the relations between ASR and speech perception and production, and support better linking of linguistic insight and technical aspects.

As the next step in this direction, we suggest to develop practically realizable causal models and feasible inference methods for such fine-grained modular systems. Different approaches, from approximate Bayesian computation to discriminative machine learning, could achieve this goal; the concrete methodology of applying them in causal inference is still a topic of current research (see, for example, Toth et al. [111]). With such methods available, the discussed topics relating to generalization can be studied: the feasibility of transfer learning, stability under environment changes or out-of-distribution observations, or other questions of learning behaviour. Considering a causal perspective may yield more robust results in conditions such as mobile applications, varying background noise, or the expected increased interaction with autonomous agents in real-life conditions. On the other hand, the presented causal framework can serve as the basis for experimental questions such as the investigation of the relations between utterance style, language complexity, and predictive performance, the behaviour of annotators and annotation models, or the intrinsic differences between planned and unplanned speech.

As far as limitations go, we should note that a causally correct model cannot be expected to automatically outperform an associative model in tasks where shifts can be ruled out; the goal is restricted to prediction only, and the interpretability of the results is not a concern [78]. Yet, the benefits of the causal approach, as we have tried to argue, are numerous, such that we dare say that causal thinking puts the modeller in a significantly better position in any case, even if resorting to off-the-shelf inference methods in the end. Despite causality's well-developed theory having been available for some decades, causal machine learning can still be considered a newcomer. Reasons for this include that causal computational methods are still behind traditional ones, and education and statistical

practise have not yet picked up or fully appreciated causal concepts. This paper can only be a starting point, indicating possible directions of continuation. However, it makes an important step towards the introduction of causal reasoning and causal methods to speech recognition, and towards our understanding of applied ASR models and their behaviour.

Author Contributions: Conceptualization, all authors; formal analysis, B.C.G.; writing—original draft preparation, P.G.; writing—review and editing, all authors; visualization, P.G.; supervision, R.K.; project administration, B.S.; funding acquisition, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work by P.G. and B.C.G. was funded by the Austrian Science Fund (FWF Grant P32700).

Data Availability Statement: Data sharing not applicable.

Acknowledgments: All emojis are designed by OpenMoji, license: CC BY-SA 4.0. The authors would like to thank Gernot Kubin for his insights and fruitful discussion.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. A Note on Model Misspecification

Here, we present a draft for an argument on how ignoring knowledge about the data generation process may also lead to a potential model misspecification, or at least to a sub-optimal probabilistic model for inference. As discussed in Section 3.3, HMMs, while causally correctly capturing the data generation process of read speech, are not automatically suited for speech synthesis. This could be due to the language and emission models being separately too weak for generation, but still result in a jointly feasible discriminator [85]. In recognition, HMMs or their extensions are appropriate models for the process that produces read speech: the text prompt W is identified with a ground truth for hidden states, from which a speech signal S is observed.

In contrast, for spontaneous speech, we argue that an HMM is a misspecified model. Indeed, while the common preconception still assumes that there is a sequence of tokens W "in the head of the speaker" (cf. Figure 6), these tokens are inaccessible to the ASR system. Instead, training an ASR system on spontaneous speech requires a sequence of annotations W_A, and the ASR system is trained to infer the annotation W_A from the observed speech utterances S. For read speech, though, learning is anti-causal: we want to infer the variable W causing the speech utterances. Conversely, for spontaneous speech, learning is causal: we want to infer the variable W_A that are caused by the speech utterances. We cannot aim to directly infer the tokens W in the head of the speaker, because we have no training data for them. This fact suggests that HMMs may be inadequate models for recognizing spontaneous speech. Indeed, if at all, then a causally correctly specified HMM for spontaneous speech should have the speech utterances S as states and the annotations W_A as observations. Such an HMM eludes us, however, since the emission model would need to characterize the annotation process, whereas all our current understanding concerns the speech production process in the "standard, read speech-type" HMM.

It has become a common understanding that ASR systems relying on the standard formulation of HMMs perform poorly on spontaneous speech, and that satisfactory recognition rates can only be achieved using extensions deviating from the original generative assumptions, or end-to-end learning of neural ASR [93]. While some of this can, as we have laid out above, be traced back to the stylistic peculiarities of spontaneous speech (speech rate, fillers, non-canonical pronunciation, etc.), we believe that at least some part of the shortcomings of HMMs and the superiority of neural ASR can be explained by model misspecification. Indeed, inference in an HMM essentially relies on "inverting" the probabilistic model using Bayes' rule. The structure of the HMM, however, greatly limits the class of distributions which can be inferred by the designed ASR systems. Specifically, parameters ϕ in the "classic" HMM model, consisting of the transition and emission probabilities, are trainable using a data set $\mathcal{D}$. Setting (incorrectly) the state process to the

sequence of annotations W_A, the optimal parameters of the HMM-type model are thus obtained by

$$\phi^* = \arg\max_{\phi} \prod_{(w_a, s)\in\mathcal{D}} p_\phi(w_a \mid s) \tag{A1}$$

in the maximum likelihood sense, via the model parameters which best explain the observed data. The fact that the conditional probability $p_{\phi^*}(w_a \mid s)$ is obtained from the HMM model for joint probability $p_{\phi^*}(w_a, s)$ makes explicit the limitations of the HMM assumption. This assumption, as mentioned above, is not problematic for read speech where the process of generating training data conforms, at least approximately, to the model inherently assumed within the HMM.

In contrast, an end-to-end neural ASR directly models the conditional probability as a function $\hat{p}_\theta(w_a \mid s)$ parametrized over θ. The structure of this conditional probability is not limited to conform to the joint probability via Bayes' theorem, but is only constrained by the architecture of the neural network implementing it. One can thus reasonably assume that for spontaneous (or, in general, transcribed) speech we have

$$\prod_{(w_a, s)\in\mathcal{D}} p_{\phi^*}(w_a \mid s) < \max_{\theta} \prod_{(w_a, s)\in\mathcal{D}} \hat{p}_\theta(w_a \mid s), \tag{A2}$$

while for read (recited) speech we may still have

$$\prod_{(w_a, s)\in\mathcal{D}} p_{\phi^*}(w_a \mid s) \approx \max_{\theta} \prod_{(w_a, s)\in\mathcal{D}} \hat{p}_\theta(w_a \mid s). \tag{A3}$$

This may give an additional explanation for why neural ASR systems or non-generative HMM extensions outperform HMM-based ASR on spontaneous speech, but not necessarily on read speech. (We do not speculate about the size of this effect; it may be small, since the stylistic peculiarities of spontaneous speech may still outweigh effects due to a misspecified probabilistic model.) As an empirical example, Linke et al. [29] find a deep-learning-based ASR approach indeed performs better than an HMM on conversational speech, but still worse than on read speech.

References

1. Pierce, J.R. Whither Speech Recognition? *J. Acoust. Soc. Am.* **1969**, *46*, 1049–1051. https://doi.org/10.1121/1.1911801.
2. Roe, D.; Wilpon, J. Whither Speech Recognition: The next 25 Years. *IEEE Commun. Mag.* **1993**, *31*, 54–62. https://doi.org/10.1109/35.256880.
3. Hannun, A. The History of Speech Recognition to the Year 2030. *arXiv* **2021**, arXiv:2108.00084.
4. Furui, S. History and Development of Speech Recognition. In *Speech Technology: Theory and Applications*; Chen, F., Jokinen, K., Eds.; Springer: New York, NY, USA, 2010; pp. 1–18. https://doi.org/10.1007/978-0-387-73819-2_1.
5. Galitsky, B. *Developing Enterprise Chatbots: Learning Linguistic Structures*; Springer International Publishing: Cham, Switzerland, 2019. https://doi.org/10.1007/978-3-030-04299-8.
6. Corbett, E.; Weber, A. What Can I Say? Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. In Proceedings of the 18th International Conference on Human–Computer Interaction with Mobile Devices and Services, Florence, Italy, 6–9 September 2016 ; Association for Computing Machinery: New York, NY, USA, 2016; pp. 72–82. https://doi.org/10.1145/2935334.2935386.
7. Messaoudi, M.D.; Menelas, B.A.J.; Mcheick, H. Review of Navigation Assistive Tools and Technologies for the Visually Impaired. *Sensors* **2022**, *22*, 7888. https://doi.org/10.3390/s22207888.
8. Furui, S. Future Directions in Speech Information Processing. *J. Acoust. Soc. Am.* **1998**, *103*, 2747–2747. https://doi.org/10.1121/1.422797.
9. King, S.; Frankel, J.; Livescu, K.; McDermott, E.; Richmond, K.; Wester, M. Speech Production Knowledge in Automatic Speech Recognition. *J. Acoust. Soc. Am.* **2007**, *121*, 723–742. https://doi.org/10.1121/1.2404622.
10. Geiger, R.S.; Yu, K.; Yang, Y.; Dai, M.; Qiu, J.; Tang, R.; Huang, J. Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 325–336. https://doi.org/10.1145/3351095.3372862.

11. Jin, Z.; von Kügelgen, J.; Ni, J.; Vaidhya, T.; Kaushal, A.; Sachan, M.; Schölkopf, B. Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 9499–9513. https://doi.org/10.18653/v1/2021.emnlp-main.748.
12. Feder, A.; Keith, K.A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M.E.; et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1138–1158. https://doi.org/10.1162/tacl_a_00511.
13. Glass, J. Towards Unsupervised Speech Processing. In Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012, pp. 1–4. https://doi.org/10.1109/ISSPA.2012.6310546.
14. Baevski, A.; Hsu, W.N.; Conneau, A.; Auli, M. Unsupervised Speech Recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 27826–27839.
15. Malik, M.; Malik, M.K.; Mehmood, K.; Makhdoom, I. Automatic Speech Recognition: A Survey. *Multimed. Tools Appl.* **2021**, *80*, 9411–9457. https://doi.org/10.1007/s11042-020-10073-7.
16. Zue, V.W.; Seneff, S. Transcription and Alignment of the TIMIT Database. In *Recent Research Towards Advanced Man–Machine Interface Through Spoken Language*; Fujisaki, H., Ed.; Elsevier: Amsterdam, The Netherlands, 1996; pp. 515–525. https://doi.org/10.1016/B978-044481607-8/50088-8.
17. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964.
18. Godfrey, J.J.; Holliman, E.C.; McDaniel, J. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In Proceedings of the Acoustics, Speech, and Signal Processing, IEEE International Conference On IEEE Computer Society, San Francisco, CA, USA, 23–26 March 1992; pp. 517–520. https://doi.org/10.1109/ICASSP.1992.225858.
19. Schweitzer, A.; Lewandowski, N. Convergence of Articulation Rate in Spontaneous Speech. In Proceedings of the INTERSPEECH 2013, Lyon, France, 25-29 August 2013; pp. 525–529.
20. Simpson, A.P.; Kohler, K.J.; Rettstadt, T. *The Kiel Corpus of Read/Spontaneous Speech: Acoustic Data Base, Processing Tools and Analysis Results*; Technical Report 1997; Universtität Kiel: Kiel, Germany, 1997.
21. Weninger, F.; Schuller, B.; Eyben, F.; Wöllmer, M.; Rigoll, G. A Broadcast News Corpus for Evaluation and Tuning of German LVCSR Systems. *arXiv* **2014**, arXiv:1412.4616.
22. Radová, V.; Psutka, J.; Müller, L.; Byrne, W.; Psutka, J.V.; Ircing, P.; Matoušek, J. *Czech Broadcast News Speech LDC2004S01*; Linguistic Data Consortium: Philadelphia, PA, USA, 2004. https://doi.org/10.35111/9r4k-j562.
23. Schuppler, B.; Hagmüller, M.; Zahrer, A. A Corpus of Read and Conversational Austrian German. *Speech Commun.* **2017**, *94*, 62–74. https://doi.org/10.1016/j.specom.2017.09.003.
24. Ernestus, M.; Kočková-Amortová, L.; Pollak, P. The Nijmegen Corpus of Casual Czech. In Proceedings of the LREC 2014: 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014, pp. 365–370.
25. Torreira, F.; Ernestus, M. The Nijmegen Corpus of Casual Spanish. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010; European Language Resources Association (ELRA): Paris, France, 2010; pp. 2981–2985.
26. Auer, P. On-Line Syntax: Thoughts on the Temporality of Spoken Language. *Lang. Sci.* **2009**, *31*, 1–13. https://doi.org/10.1016/j.langsci.2007.10.004.
27. Furui, S.; Nakamura, M.; Ichiba, T.; Iwano, K. Why Is the Recognition of Spontaneous Speech so Hard? In *Proceedings of the Text, Speech and Dialogue*; Matoušek, V., Mautner, P., Pavelka, T., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; pp. 9–22. https://doi.org/10.1007/11551874_3.
28. Valenta, T.; Šmídl, L. Word Confusions in the Transcription and Recognition of Spontaneous Czech. In *Tackling the Complexity in Speech*; Niebuhr, O., Skarnitzl, R., Eds.; Opera Facultatis Philosophicae Universitatis Carolinae Pragensis; Faculty of Arts, Charles University: Prague, Czech Republic, 2015; Volume XIV.
29. Linke, J.; Garner, P.N.; Kubin, G.; Schuppler, B. Conversational Speech Recognition Needs Data? Experiments with Austrian German. In Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 4684–4691.
30. Szymański, P.; Żelasko, P.; Morzy, M.; Szymczak, A.; Żyła-Hoppe, M.; Banaszczak, J.; Augustyniak, L.; Mizgajski, J.; Carmiel, Y. WER We Are and WER We Think We Are. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3290–3295. https://doi.org/10.18653/v1/2020.findings-emnlp.295.
31. Likhomanenko, T.; Xu, Q.; Pratap, V.; Tomasello, P.; Kahn, J.; Avidov, G.; Collobert, R.; Synnaeve, G. Rethinking Evaluation in ASR: Are Our Models Robust Enough? *arXiv* **2021**, arXiv:2010.11745.
32. Zhang, Y.; Park, D.S.; Han, W.; Qin, J.; Gulati, A.; Shor, J.; Jansen, A.; Xu, Y.; Huang, Y.; Wang, S.; et al. BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition. *IEEE J. Sel. Top. Signal Process.* **2022**, 1–14. https://doi.org/10.1109/JSTSP.2022.3182537.

33. Nakamura, M.; Iwano, K.; Furui, S. Differences between Acoustic Characteristics of Spontaneous and Read Speech and Their Effects on Speech Recognition Performance. *Comput. Speech Lang.* **2008**, *22*, 171–184. https://doi.org/10.1016/j.csl.2007.07.003.
34. Schuppler, B. Rethinking Classification Results Based on Read Speech, or: Why Improvements Do Not Always Transfer to Other Speaking Styles. *Int. J. Speech Technol.* **2017**, *20*, 699–713. https://doi.org/10.1007/s10772-017-9436-y.
35. Ajot, J.; Fiscus, J. Speech-To-Text (STT) and Speaker Attributed STT (SASTT) Results. In Proceedings of the NIST Rich Transcription Evaluation Workshop, Gaithersburg, MD, USA, 2009. Available online: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation) (accessed on 12 October 2022).
36. Synnaeve, G. wer_are_we. Available online: https://github.com/syhw/wer_are_we/tree/a5d4a30100340c6c8773f329b438017403d606ad#readme (accessed 6 February 2023).
37. Natarajan, N.; Dhillon, I.S.; Ravikumar, P.K.; Tewari, A. Learning with Noisy Labels. *Adv. Neural Inf. Process. Syst.* **2013**, *26*.
38. Beigman, E.; Beigman Klebanov, B. Learning with Annotation Noise. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; Association for Computational Linguistics: Singapore, 2009; pp. 280–287.
39. Kohler, K.J. The Disappearance of Words in Connected Speech. *ZAS Pap. Linguist.* **1998**, *11*, 21–33.
40. Zayats, V.; Tran, T.; Wright, R.; Mansfield, C.; Ostendorf, M. Disfluencies and Human Speech Transcription Errors. In Proceedings of the Interspeech 2019, ISCA, Graz, Austria, 15–19 September 2019; pp. 3088–3092. https://doi.org/10.21437/Interspeech.2019-3134.
41. Raymond, W.D. An Analysis of Coding Consistency in the Transcription of Spontaneous Speech from the Buckeye Corpus. In Proceedings of the Workshop on Spontaneous Speech: Data and Analysis, 2003; The National Institute for Japanese Language: Tokyo, Japan, 2003; pp. 55–71.
42. Stefanowitsch, A. *Corpus Linguistics: A Guide to the Methodology*; Number 7 in Textbooks in Language Sciences; Language Science Press: Berlin, Gemany, 2020. https://doi.org/10.5281/zenodo.3735822.
43. Hovy, E.; Lavid, J. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *Int. J. Transl.* **2010**, *22*, 13–36.
44. Artstein, R.; Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.* **2008**, *34*, 555–596.
45. Passonneau, R.J.; Carpenter, B. The Benefits of a Model of Annotation. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 311–326. https://doi.org/10.1162/tacl_a_00185.
46. Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; Poesio, M. Comparing Bayesian Models of Annotation. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 571–585. https://doi.org/10.1162/tacl_a_00040.
47. Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2005. https://doi.org/10.1137/1.9780898717921.
48. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773.
49. Pearl, J. Understanding Simpson's Paradox. *SSRN J.* **2013**. https://doi.org/10.2139/ssrn.2343788.
50. Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; Zhang, A. A Survey on Causal Inference. *ACM Trans. Knowl. Discov. Data* **2021**, *15*, 1–46. https://doi.org/10.1145/3444944.
51. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
52. Healy, E.W.; Tan, K.; Johnson, E.M.; Wang, D. An Effectively Causal Deep Learning Algorithm to Increase Intelligibility in Untrained Noises for Hearing-Impaired Listeners. *J. Acoust. Soc. Am.* **2021**, *149*, 3943–3953. https://doi.org/10.1121/10.0005089.
53. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438. https://doi.org/10.2307/1912791.
54. Peters, J.; Bühlmann, P.; Meinshausen, N. Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2016**, *78*, 947–1012. https://doi.org/10.1111/rssb.12167.
55. Bühlmann, P. Invariance, Causality and Robustness. *Stat. Sci.* **2020**, *35*, 1–36. https://doi.org/10.1214/19-STS721.
56. Schölkopf, B. Causality for Machine Learning. *arXiv* **2019**, arXiv:1911.10500.
57. Lewis, D. Causation. *J. Philos.* **1974**, *70*, 556–567.
58. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*; Basic Books: New York, NY, USA, 2018.
59. Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In *Proceedings of the ECML PKDD 2020 Workshops*; Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R.P., Gavaldà, R., et al., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, Switzerland, 2020; pp. 417–431. https://doi.org/10.1007/978-3-030-65965-3_28.
60. Burkart, N.; Huber, M.F. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* **2021**, *70*, 245–317. https://doi.org/10.1613/jair.1.12228.
61. Vowels, M.J.; Camgoz, N.C.; Bowden, R. D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *arXiv* **2021**, arXiv:2103.02582.
62. Pearl, J. Causal Inference in Statistics: An Overview. *Statist. Surv.* **2009**, *3*, 96–146. https://doi.org/10.1214/09-SS057.
63. Bareinboim, E.; Correa, J.D.; Ibeling, D.; Icard, T. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 1st ed.; Association for Computing Machinery: New York, NY, USA, 2022, pp. 507–556.

64. Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press: Cambridge, MA, USA, 2017.
65. Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; Mooij, J. On Causal and Anticausal Learning. *arXiv* **2012**, arXiv:1206.6471.
66. Gresele, L.; von Kügelgen, J.; Stimper, V.; Schölkopf, B.; Besserve, M. Independent Mechanism Analysis, a New Concept? *arXiv* **2022**, arXiv:2106.05200.
67. Greenland, S.; Brumback, B. An Overview of Relations among Causal Modelling Methods. *Int. J. Epidemiol.* **2002**, *31*, 1030–1037. https://doi.org/10.1093/ije/31.5.1030.
68. Suzuki, E.; Shinozaki, T.; Yamamoto, E. Causal Diagrams: Pitfalls and Tips. *J. Epidemiol.* **2020**, *30*, 153–162. https://doi.org/10.2188/jea.JE20190192.
69. Pearl, J. The Do-Calculus Revisited. In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, 14–18 August 2012; pp. 4–11.
70. Rubin, D.B. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *J. Am. Stat. Assoc.* **2005**, *100*, 322–331.
71. Cinelli, C.; Forney, A.; Pearl, J. A Crash Course in Good and Bad Controls. *SSRN J.* **2020**. https://doi.org/10.2139/ssrn.3689437.
72. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley Series in Telecommunications; Wiley: New York, NY, USA, 1991.
73. Chen, Y.; Wang, W.; Wang, C. Semi-Supervised ASR by End-to-End Self-Training. In Proceedings of the Interspeech 2020, ISCA, Shanghai, China, 25–29 October 2020; pp. 2787–2791. https://doi.org/10.21437/Interspeech.2020-1280.
74. Karita, S.; Watanabe, S.; Iwata, T.; Ogawa, A.; Delcroix, M. Semi-Supervised End-to-End Speech Recognition. In Proceedings of the Interspeech 2018, ISCA, Hyderabad, India, 2–6 September 2018; pp. 2–6.
75. Synnaeve, G.; Xu, Q.; Kahn, J.; Likhomanenko, T.; Grave, E.; Pratap, V.; Sriram, A.; Liptchinsky, V.; Collobert, R. End-to-End ASR: From Supervised to Semi-Supervised Learning with Modern Architectures. *arXiv* **2020**, arXiv:1911.08460.
76. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**, arXiv:2006.11477.
77. Blöbaum, P.; Shimizu, S.; Washio, T. Discriminative and Generative Models in Causal and Anticausal Settings. In *Proceedings of the Advanced Methodologies for Bayesian Networks*; Suzuki, J., Ueno, M., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9505, pp. 209–221. https://doi.org/10.1007/978-3-319-28379-1_15.
78. Kilbertus, N.; Parascandolo, G.; Schölkopf, B. Generalization in Anti-Causal Learning. In Proceedings of the NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning, Montreal, QC, Canada, 7 December 2018.
79. Castro, D.C.; Walker, I.; Glocker, B. Causality Matters in Medical Imaging. *Nat. Commun.* **2020**, *11*, 3673. https://doi.org/10.1038/s41467-020-17478-w.
80. Hoyer, P.; Janzing, D.; Mooij, J.M.; Peters, J.; Schölkopf, B. Nonlinear Causal Discovery with Additive Noise Models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–11 December 2008; Curran Associates, Inc.: Red Hook, NY, USA, 2008; Volume 21.
81. Furui, S. 50 Years of Progress in Speech and Speaker Recognition Research. *ECTI Trans. Comput. Inf. Technol. (ECTI-CIT)* **2005**, *1*, 64–74. https://doi.org/10.37936/ecti-cit.200512.51834.
82. Levinson, S.E.; Rabiner, L.R.; Sondhi, M.M. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell Syst. Tech. J.* **1983**, *62*, 1035–1074. https://doi.org/10.1002/j.1538-7305.1983.tb03114.x.
83. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; et al. *The HTK Book*; Technical Report; Cambridge University Engineering Department: Cambridge, UK, 2009.
84. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011; IEEE Signal Processing Society: Piscataway, NJ, USA, 2011; Number CONF.
85. Bilmes, J.A. What HMMs Can Do. *IEICE Trans. Inf. Syst.* **2006**, *E89-D*, 869–891.
86. Beigman Klebanov, B.; Beigman, E. From Annotator Agreement to Noise Models. *Comput. Linguist.* **2009**, *35*, 495–503. https://doi.org/10.1162/coli.2009.35.4.35402.
87. Varga, A.; Moore, R. Hidden Markov Model Decomposition of Speech and Noise. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990; Volume 2, pp. 845–848. https://doi.org/10.1109/ICASSP.1990.115970.
88. Ghahramani, Z.; Jordan, M. Factorial Hidden Markov Models. In *Proceedings of the Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1995; Volume 8.
89. Wellekens, C. Explicit Time Correlation in Hidden Markov Models for Speech Recognition. In Proceedings of the ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, USA, 6–9 April 1987; Volume 12, pp. 384–386. https://doi.org/10.1109/ICASSP.1987.1169614.
90. Bridle, J.S. Towards Better Understanding of the Model Implied by the Use of Dynamic Features in HMMs. In Proceedings of the Interspeech 2004, ISCA, Jeju Island, Republic of Korea, 4–8 October 2004; pp. 725–728. https://doi.org/10.21437/Interspeech.2004-281.
91. Bilmes, J.A. Graphical Models and Automatic Speech Recognition. *J. Acoust. Soc. Am.* **2002**, *112*, 2278–2278.

92. Deng, L. Deep Learning: From Speech Recognition to Language and Multimodal Processing. *APSIPA Trans. Signal Inf. Process.* **2016**, *5*, E1. https://doi.org/10.1017/ATSIP.2015.22.
93. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* **2019**, *7*, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880.
94. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
95. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
96. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 369–376. https://doi.org/10.1145/1143844.1143891.
97. Vyas, A.; Madikeri, S.; Bourlard, H. Comparing CTC and LFMMI for Out-of-Domain Adaptation of Wav2vec 2.0 Acoustic Model. *arXiv* **2021**, arXiv:2104.02558.
98. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*; Balcan, M.F., Weinberger, K.Q., Eds.; PMLR: New York, NY, USA, 2016; Volume 48, pp. 173–182.
99. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. Wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv* **2019**, arXiv:1904.05862.
100. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. *Robust Speech Recognition via Large-Scale Weak Supervision*; Technical Report; OpenAI: San Francisco, CA, USA, 2022.
101. Mackenzie, A. The Production of Prediction: What Does Machine Learning Want? *Eur. J. Cult. Stud.* **2015**, *18*, 429–445. https://doi.org/10.1177/1367549415577384.
102. Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus Machine Learning. *Nat. Methods* **2018**, *15*, 233.
103. Chen, J.H.; Asch, S.M. Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *N. Engl. J. Med.* **2017**, *376*, 2507–2509. https://doi.org/10.1056/NEJMp1702071.
104. Ma, D.; Ryant, N.; Liberman, M. Probing Acoustic Representations for Phonetic Properties. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 311–315. https://doi.org/10.1109/ICASSP39728.2021.9414776.
105. Padmanabhan, J.; Johnson Premkumar, M.J. Machine Learning in Automatic Speech Recognition: A Survey. *IETE Tech. Rev.* **2015**, *32*, 240–251. https://doi.org/10.1080/02564602.2015.1010611.
106. Ostendorf, M. Moving beyond the "Beads-on-a-String" Model of Speech. In Proceedings of the IEEE ASRU Workshop, Merano, Italy, 13–17 December 1999; pp. 79–84.
107. Scharenborg, O. Reaching over the Gap: A Review of Efforts to Link Human and Automatic Speech Recognition Research. *Speech Commun.* **2007**, *49*, 336–347. https://doi.org/10.1016/j.specom.2007.01.009.
108. Deng, L.; Jaitly, N. Deep Discriminative and Generative Models for Speech Pattern Recognition. In *Handbook of Pattern Recognition and Computer Vision*; World Scientific: Singapore, 2015; pp. 27–52. https://doi.org/10.1142/9789814656535_0002.
109. Manning, C.D. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *Proceedings of the Computational Linguistics and Intelligent Text Processing*; Gelbukh, A.F., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; pp. 171–189. https://doi.org/10.1007/978-3-642-19400-9_14.
110. Reidsma, D.; Carletta, J. Reliability Measurement without Limits. *Comput. Linguist.* **2008**, *34*, 319–326. https://doi.org/10.1162/coli.2008.34.3.319.
111. Toth, C.; Lorch, L.; Knoll, C.; Krause, A.; Pernkopf, F.; Peharz, R.; von Kügelgen, J. Active Bayesian Causal Inference. *arXiv* **2022**, arXiv:2206.02063.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.