Analytical choices for analyzing multidimensional behavior - many analyst test hypotheses about human speech.

First Author[#], Second Author[#], ...[#], & Last Author[#]

[1] #

[...] ...

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Second Author: Writing - Review & Editing; ...: Writing - Review & Editing; Last Author: Writing - Review & Editing.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: crowdsourcing science, data analysis, scientific transparency, speech, acoustic analysis

Word count: X

Analytical choices for analyzing multidimensional behavior - many analyst test hypotheses about human speech.

Introduction

In order to effectively accumulate knowledge, science needs to (i) produce data that can be recreated by using the same methods and (ii) arrive at conclusions about data that are robust. In recent coordinated efforts to replicate published findings, the scientific disciplines have uncovered surprisingly low succes rates for (i) (e.g., Open Science Collaboration 2015, Camerer et al. 2018, REF) leading to what is now referred to as the *replication crisis*. Beyond difficulties replicating scientific findings, more and more evidence suggests that the theoretical conclusions drawn from data are surprisingly variable even if researchers have access to reliable data (REFS). The latter situation has been referred to as the *inference crisis* (Rotello, Heit & Dubé 2015, Starns et al. 2019) and is, among other things, rooted in the inherent flexibility of data analysis, often referred to as researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011, Gelman & Loken 2013). Data analysis involves many different steps including inspecting, organizing, transforming, and modeling the data. Along the way, different methodological and analytical choices need to be made, all of which might change its final interpretation. These researcher degrees of freedom are a blessing and a curse at the same time.

There are a blessing, because they allow us to look at nature from many different angles allowing us to make important discoveries and generate new hypothesis (e.g. Box 1976, Tukey 1977, de Groot 2014). They are a curse because idiosyncratic choices can lead to categorically different interpretation which find their way into the publication record where they are taken for granted (Simmons et al. 2011). Recent projects have shown that the variability between different data analysts is immense and leads researchers to draw vastly different conclusions about one and the same dataset (e.g. Silberzahn et al. 2018, Starns et al. 2019, Botvinik-Nezer et al., 2020). These projects, however, might still underestimate the extend to which analysts vary because data analysis is not only restricted to statistical inference of data tables. Human behavior is complex and offers many ways to be translated into numbers. This is particularly true for fields that draw conclusions about human behavior and cognition from multidimensional data like speech or video data. In fields working on human speech production for example, researchers need to make many decisions about what to measure and how to measure it. This is not trivial given the temporal extension of the acoustic signal and its complex structural composition. Decisions about operationalizing the raw data might not only influence downstream decisions about statistical modelling but statistical results might lead researchers to go back and revise earlier decisions about the raw data.

In this article, we investigate the diversity in analytic choices when many analyst teams analyze the same speech production data. We explore the interaction between analytic choices at the stage of the operationalization of raw data and subsequent statistical modelling. Specifically we report the impact of analytic pipeline on research results obtained by ## teams that gained access to the same raw data set to answer the same research question.

Researcher degrees of freedom

Data analysis comes with many decisions like how to operationalize a given phenomenon or behavior, what data to submit to statistical modelling and which to exclude, what models to use or what inferential decision procedure to apply.However, if these decisions during data analysis are not specified in advance, we might stumble upon seemingly meaningful patterns in the data that are merely statistical flukes. This can be problematic because to err is human.

We have evolved to filter the world in irrational ways (e.g., Tversky and Kahneman

1974), seeing coherent patterns in randomness (Brugger 2001), convincing ourselves of the validity of prior expectations ("I knew it", Nickerson 1998), and perceiving events as being plausible in hindsight ("I knew it all along", Fischhoff 1975). In connection with an academic incentive system that rewards certain discovery process more than others (Sterling 1959, Koole & Lakens 2012), we often find ourselves exploring many possible analytical pipelines, but only reporting a select few. This issue is particularly amplified in fields in which the raw data lend themselves to flexible operationalizations (Roettger 2019). Combined with a wide variety of methodological and theoretical traditions as well as varying levels of statistical training across fields and subfields, the inherent flexibility of data analysis might lead to an vast plurality of analytic approaches.

Consequently - if methodologists are correct (e.g. Simmons et al. 2011, Gelman & Loken 2013) - there are many published papers that present overconfident interpretations of their data based on idiosyncratic analytic strategies. These interpretation are either associated with an unknown amount of uncertainty or lend themselves to alternative interpretation if analysed differently. However, instead of being critically evaluated, scientific results often remain unchallenged in the publication record. Critical reanalyses of published analytic strategies are uncommon because data sharing is still rare (Wicherts, Borsboom, Kats, & Molenaar, 2006, RECENT REF).

While this issue has been widely discussed both from a conceptual point of view (Simmons et al. 2011, Wagenmakers et al. 2012, Nosek and Lakens 2014) and its application in individual scientific fields (e.g. Wichert et al. 2015, Charles et al. 2019, Roettger 2019), there is still little known about the extend of analytical plurality in practice. Recent collaborative attempts have started to shed light on how different analyst tackle the same data set and revealed, not surprisingly, a large amount of variability.

Crowdsourcing alternative analyses

In a collaborative effort, Silberzahn et al. (2018) let twenty-nine independent analysis teams address the same research question. Analytic approaches and consequently the results varied widely between teams. 69% of the teams found support for the hypothesis, and 31% did not. Out of the 29 analysis strategies, there were 21 unique combinations of covariates. Importantly, the observed variability was neither predicted by the team's preconceptions about the phenomenon under investigation nor by peer ratings of the quality of their analyses. Their results suggest that analytic plurality is a fact of life and not driven by different levels of expertise or bias. Similar crowd-sourced studies recruiting independent analyst teams showed similar results.

Neuroscience Cognitive Modelling Clinical Predictive models

While these papers show a large degree of analytical flexibility with impactful consequences, these studies dealt with flexibility in inferential or computational modelling. In these studies the data tables were fixed and data collection or extraction could not be changed. However, in many fields the primary raw data is a complex signal that needs to be operationalized according to the research question. In social sciences, the raw observations correspond to recorded human behavior. In many cases, the behavior is recorded and stored as a complex visual and/or acoustic signal that is temporal extended exhibits a complex structure. Decisions about how to operationalize a theoretically relevant aspect of that behavior or the underlying cognitive processes might interact with downstream decisions about statistical modelling and vice verse. To understand how analytical flexibility manifests itself in a scenario where complex signals need to be operationalized, the present paper looks at an experimentally elicited speech data set

Operationalizing speech

(copy pasted from RDF paper and slightly shortened by myself. I think its a good point of departure. Maybe one of your can try to rephrase it in your own words?) RELEVANCE OF SPEECH PRODUCTION RESEARCH FOR COGSCI, AI, etc.

In order to understand speech, listeners have to map a continuous, transient signal onto discrete meanings. Speech offers a considerable number of perspectives and decisions along the data analysis pipeline.

IMAGE SHOWING A WAVE FORM WITH DIFFERENT DOMAINS (temporal) AND STRUCTURAL CUES (e.g. f0, dur, int) MAPPING ONTO FUNCTIONAL CONTRASTS

When conducting a study on speech production, the first important analytic decision to test a hypothesis relates to operationalization the relevant behavior, i.e. how to measure the phenomenon of interest. For example, how do we measure whether two sounds are acoustically identical (e.g. “bear” vs. “pear”), whether one word is more prominent than others (“He told YOU to be quite”), or whether two discourse functions are produced differently (“It’s raining.” vs. “It’s raining?”)? In other words, how do we quantitatively capture relevant features of speech?

This is not trivial. Speech categories are inherently multidimensional and vary through time. The acoustic parameters for one category are usually asynchronous, i.e. appear at different points of time in the unfolding signal and overlap with parameters for other categories (e.g. Jongman et al., 2000; Lisker, 1986; Summerfield, 1984; Winter, 2014). For example, the distinction between voiced and voiceless stops in English (i.e. /b/ and /p/ in “bear” vs. “pear”) can be manifested by many different acoustic (Lisker, 1977). Even temporally dislocated acoustic parameters correlate with this lexical contrast. For example, in the English words “led” versus “let”, voicing correlates can be found in the acoustic manifestation of the initial /l/ of the word (Hawkins & Nguyen, 2004).

The apparent multiplicity of phonetic cues grows exponentially if we look at larger temporal windows as is the case for suprasegmental aspects of speech. Studies investigating acoustic correlates of word stress (i.e. the difference between “insight” and “incite”), for example, have been using many different measurements including temporal characteristics (duration of certain segments or subphonemic intervals), spectral characteristics (intensity measures, formants, and spectral tilt), and measurements related to fundamental frequency (f0) (e.g. Gordon & Roettger, 2017).

Looking at even larger domains, the expression of pragmatic functions can be expressed by a variety of structurally different acoustic cues which can be distributed throughout the entire utterance. Discourse functions are systematically expressed by multiple local pitch modulations differing in their position, shape, and alignment (e.g. Niebuhr et al., 2011). They can also be expressed by global or local pitch modulations, as well as acoustic information within the temporal or spectral domain (e.g. van Heuven & van Zanten 2005). All of these phonetic parameters are potential manifestations of underlying functional contrasts like speaker’s intentions, levels of arousal or identity.

When testing hypotheses on speech production data, researchers need to make many decisions. The larger the functional domain (e.g. are we interested in lexical items or in whole utterances), the higher the number of possible operationalizations. These decisions are usually made prior to any statistical analysis, but can possible be revised in interaction with downstream analysis steps. To probe the variability in data analysis pipelines across researchers, we provided researchers with an experimentally elicited speech corpus that looked at a functional contrast that is potentially manifested across the whole utterance.

The data set - The prosody of redundant modifiers

Our data set was collected in order to answer the following research question: Do speakers acoustically modify a referring expression as a function of the typicality of the modifier of the noun (e.g. “a blue banana” vs. “a yellow banana”)?

Referring is one of the most basic and prevalent uses of language and one of the most widely researched areas in the language science. It is an open question how speakers choose a referential expression when they want to refer to a specific entity like a banana. The context within which an entity occurs (other non-fruits, other fruits, other bananas) plays a large part in determining the choice of referential expression. Generally, speakers aim to be as informative as possible to uniquely establish reference to the intended object (Grice 1975) and are therefore expected to only use for example a modifier if it is necessary for disambiguation (e.g. the adjective “yellow” when there is a yellow and a less ripe green banana).

Despite this coherent idea of rational and efficient speakers, there is much evidence that speakers are often overinformative: Speaker use referring expressions that are more specific than strictly necessary for the unambiguous identification of the intended referent (Sedivy 2003, Westerbeek et al. 2015, Rubio-Fernandez 2016), which has been argued to facilitate object identification and making communication between speakers and listeners more efficient (Arts et al. 2011, Paraboni et al. 2007, Rubio-Fernandez 2016). Recent findings suggest that the utility of a referring expression depends on how good it is for a listener (compared to other referring expressions) to identify a target object. For example, Degen et al. (2020) showed that modifiers that are less typical for a given referent (e.g. a blue banana) are more likely to be used in an overinformative scenario (e.g. there is just one banana).

This account, however, has mainly focused on content selection (Gatt et al. 2013), i.e. whether a certain referential expression is chosen or not. However, speech communication is so much richer. Even looking at morphosyntactically identical expressions, speakers can modulate this expression via suprasegmental acoustic properties like temporal and spectral modifications of the segments involved (e.g. Ladd 2008). Most prominently, languages use intonation to signal discourse relationships between referents (among other functions). Intonation marks discourse-relevant referents for being new or given information to guide listeners' interpretation of incoming messages. In many languages, speakers can use particular pitch movements to signal whether a referent has already been mentioned and is therefore referred back to, or a referent is newly introduced into the discourse. Many languages use intonation in order to signal if a referent is contrasting with one or more alternatives that are relevant to the current discourse. Content selection aside, in a scenario in which a speaker wants to refer to a banana when there is also a pear on the table, the speaker would most likely produce a high rising pitch accent on "banana" to indicate the contrastive nature of the *noun.* In a scenario in which the speaker wants to refer to a yellow banana when there is also a less ripe green banana on the table, the speaker would most likely produce a high rising pitch accent on "yellow" to indicate the contrastive nature of the *modifier.* In addition to a pitch accent, elements that are new and/or contrastive are often produced with additional suprasegmental prominence, i.e. segments are hyperarticulated, resulting in longer, louder and more clearly articulated acoustic targets.

INFORMATION ABOUT THE DATA SET

Research questions

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

Material

Procedure

Data analysis

We used R (Version 4.0.2; R Core Team, 2020) and the R-package *papaja* (Version 0.1.0.9997; Aust & Barth, 2020) for all our analyses.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/