Cognition 149 (2016) 31–39

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Automatic analysis of slips of the tongue: Insights into the cognitive architecture of speech production

Matthew Goldrick[a,*], Joseph Keshet[b,*], Erin Gustafson[a], Jordana Heller[a], Jeremy Needle[a]

[a] Department of Linguistics, Northwestern University, USA
[b] Department of Computer Science, Bar-Ilan University, Israel

ARTICLE INFO

Article history:
Received 9 March 2015
Revised 9 September 2015
Accepted 4 January 2016
Available online 9 January 2016

Keywords:
Speech production
Automatic phonetic analysis
Speech errors
Machine learning
Structured prediction

ABSTRACT

Traces of the cognitive mechanisms underlying speaking can be found within subtle variations in how we pronounce sounds. While speech errors have traditionally been seen as categorical substitutions of one sound for another, acoustic/articulatory analyses show they partially reflect the intended sound. When “pig” is mispronounced as “big,” the resulting /b/ sound differs from correct productions of “big,” moving towards intended “pig”—revealing the role of graded sound representations in speech production. Investigating the origins of such phenomena requires detailed estimation of speech sound distributions; this has been hampered by reliance on subjective, labor-intensive manual annotation. Computational methods can address these issues by providing for objective, automatic measurements. We develop a novel high-precision computational approach, based on a set of machine learning algorithms, for measurement of elicited speech. The algorithms are trained on existing manually labeled data to detect and locate linguistically relevant acoustic properties with high accuracy. Our approach is robust, is designed to handle mis-productions, and overall matches the performance of expert coders. It allows us to analyze a very large dataset of speech errors (containing far more errors than the total in the existing literature), illuminating properties of speech sound distributions previously impossible to reliably observe. We argue that this provides novel evidence that two sources both contribute to deviations in speech errors: planning processes specifying the targets of articulation and articulatory processes specifying the motor movements that execute this plan. These findings illustrate how a much richer picture of speech provides an opportunity to gain novel insights into language processing.

1. Introduction

The acoustic and articulatory properties of speech vary from moment to moment; if you repeat a word several times, no two instances will be precisely the same. Hidden within this variation are traces of the cognitive processes underlying language production. For example, when repeatedly producing a word, you will tend to slightly reduce its duration—reflecting (in part) the ease of retrieving the word from long term memory (Kahn & Arnold, 2012; Lam & Watson, 2010). Such effects can also be found at the level of individual speech sounds within a word. One such effect can be observed in bilingual speakers’ pronunciations of second language speech sounds. Such sounds are more accented when speakers have recently produced a word in their native language, relative to cases where the same speaker has just produced sounds in the second language (Balukas & Koops, 2015; Goldrick, Runnqvist, & Costa, 2014; Olson, 2013). This suggests that the difficulty of retrieving words and sounds when switching languages can modulate how sounds are articulated.

Here, we focus on one source of evidence that has played a key role in theories of language production: speech errors (Fromkin, 1971, et seq.). Errors involving the mis-production of sounds (“pig” mispronounced as “big”) reveal the graded influence of intended productions on articulation. Errors simultaneously reflect acoustic/articulatory properties of both the target and error outcome (Frisch & Wright, 2002; Goldrick, Baker, Murphy, & Baese-Berk, 2011; Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; McMillan & Corley, 2010; McMillan, Corley, & Lickley, 2009; Pouplier, 2007, 2008). Such effects are consistent with theories of language production incorporating continuous, distributed mental representations in the cognitive process underlying the planning (Dell, 1986; Goldrick & Blumstein, 2006; Plaut & Shallice, 1993; Smolensky, Goldrick, & Mathis, 2014) and

[*] Corresponding authors at: Department of Linguistics, Northwestern University, 2016 Sheridan Rd., Evanston, IL 60208, USA (M. Goldrick). Department of Computer Science, Bar-Ilan University, Ramat Gan 52900, Israel (J. Keshet).

E-mail addresses: matt-goldrick@northwestern.edu (M. Goldrick), joseph.keshet@biu.ac.il (J. Keshet).

http://dx.doi.org/10.1016/j.cognition.2016.01.002
0010-0277/© 2016 Elsevier B.V. All rights reserved.

articulation of speech sounds (Goldstein et al., 2007; Saltzman & Munhall, 1989). According to these theoretical perspectives, articulation reflects subtle, gradient variation in the representational structures and cognitive processes underlying speech (e.g., variation in the degree to which the native language is activated can yield graded changes in the degree of accent in non-native speech; partial activation of target sounds can influence how errors are articulated).

While studies of phonetic variation have provided a rich source of information about language processing, most researchers have relied on manual annotation to obtain accurate data. This approach suffers from two critical flaws. It is highly resource intensive; a single experiment in our lab (Goldrick et al., 2011) required over 3000 person-hours for analysis. With respect to speech error studies (as discussed below), this has prevented researchers from obtaining the data required to reliably evaluate different hypotheses. Second, this approach is fundamentally subjective: manual labels reflect the judgments of annotators. This presents a barrier to replication.

Recent studies have aimed to address these issues through computational methods that automatically measure acoustic properties of speech (e.g., Gahl, Yao, & Johnson, 2012; Labov, Rosenfelder, & Fruehwald, 2013; Yuan & Liberman, 2014). These methods eliminate subjective judgments while enormously reducing the resources required for analysis. Although this has provided great advances in studies of phonetic variation, existing methods do not provide a comprehensive solution. They have not provided the fine granularity of measurement necessary to reliably measure differences at the level of individual speech sounds (specifically, consonant sounds). Furthermore, these existing methods require a complete transcription of the observed speech prior to phonetic analysis. This is a major burden, particularly for paradigms that are designed to produce tremendous variation in production (e.g., speech errors).

In this work, we propose a novel computational framework for automatic analysis of speech appropriate for evaluating hypotheses relating to the phonetics of speech errors. This is based on a set of algorithms in machine learning (Keshet, Shalev-Shwartz, Singer, & Chazan, 2007; McAllester, Hazan, & Keshet, 2010; Sonderegger & Keshet, 2012). Our automatic approach matches the performance of expert manual coders and outperforms algorithms used in the existing psycholinguistic literature. The analyses reveal novel properties of the phonetics of speech errors. Furthermore, we show (via a power analysis) that reliable investigation of the properties of individual speech sounds requires datasets larger than those used in previous work. These findings show how automatic analysis creates an opportunity to gain a much richer, objective, and replicable picture of acoustic variation in speech.

1.1. Phonetic variation in sound substitution errors

One key source of evidence for the structure of the cognitive mechanisms underlying language production is speech errors (Fromkin, 1971). Sound substitution errors (e.g., intending to say *bet*, but producing *pet*; written as *bet*→*pet*) have been studied in the laboratory by asking participants to rapidly produce artificial tongue twisters composed of syllables with alternating contrasting sounds (*pet bet bet pet*; Wilshire, 1999). Based on transcriptions of speech, it was long assumed that such errors reflect the categorical substitution of one sound for another (Dell, 1986; Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979). However, more recent quantitative analyses of the phonetic (acoustic/articulatory) properties of errors have revealed that errors systematically differ from corresponding correct productions—a deviation that reflects properties of the intended sound (Frisch & Wright, 2002; Goldrick & Blumstein, 2006; Goldrick et al., 2011; Goldstein et al., 2007; McMillan & Corley, 2010; McMillan et al., 2009; Pouplier, 2007, 2008). For example, an important acoustic cue to the distinction between words like *pet* and *bet* is voice onset time (VOT), the time between the release of airflow (e.g., opening the lips) and the onset of periodic vibration of the vocal folds (Lisker & Abramson, 1964). In English, voiceless sounds like /p/ have relatively long VOTs whereas voiced sounds like /b/ have short VOTs (Lisker & Abramson, 1964). In a *bet*→*pet* error, the resulting /p/ sound is distinct from correct productions of the same sound (*pet*→*pet*). The error /p/ tends to have a shorter VOT—which makes it more similar to the intended sound /b/. The complementary pattern is found for errors like *pet*→*bet*; the error /b/ tends to have a longer VOT than the corresponding sound in *bet*→*bet*. Note that similar effects are found in non-errorful speech when a competitor word is explicitly primed (e.g., priming *top* while reading the word *cop* aloud yields a blend of /t/ and /k/ articulations; Yuen, Davis, Brysbaert, & Rastle, 2010).

These deviations have been attributed to one of two distinct types of cognitive processes that underlie the production of speech: (i) *planning processes* that construct a relatively abstract specification of the targets of articulation; or (ii) *articulatory processes* that specify the specific motor movements that execute this plan. To illustrate this division, when producing *pet*, planning processes might specify that the initial sound is /p/ but not the precise timing of the associated lip movements; these would be specified during articulatory processing. Below, we outline how different theories have proposed that deviations of errors from correct productions arise at each level of processing.

Within planning processes, many theories of speech production assume that representations are patterns of activation over simple processing units (Dell, 1986). For example, the contrast between *big* and *pig* is represented by graded patterns of activation over units representing speech segments /p/ and /b/. While this type of representation can express arbitrarily varying combinations of /p/ and /b/, theories typically incorporate mechanisms that constrain the patterns of activation. These mechanisms force planning processes to select relatively discrete representations for production (e.g., primarily activating /p/, with little activation of /b/). A variety of mechanisms have been proposed to account for this, including: boosting activation of one representation relative to alternatives (e.g., Dell, 1986); lateral inhibition that reduces the activation of alternative representations (see Dell & O'Seaghdha, 1994, for a review); and attractors over distributed representations (e.g., Goldrick & Chu, 2014; Plaut & Shallice, 1993; Smolensky et al., 2014). However, these constraints on activation are typically not categorical; while one unit may be highly active, others may remain partially active. This has been proposed as one possible mechanism for producing deviations in speech errors. If the specification of the intended target sound remains partially active, the phonetic properties of the error could be distorted towards the intended target (Goldrick & Blumstein, 2006; Goldrick & Chu, 2014; Smolensky et al., 2014). For example, in *bet*→*pet*, the speech plan could specify the target is 0.9 /p/ and 0.1 /b/—resulting in articulations that combine properties of both sounds.

Articulatory processes could provide an additional source of distortions in speech errors. Such processes specify the continuous, coordinated dynamics of articulator movements that execute the speech plan (Saltzman & Munhall, 1989). Tongue twisters require speakers to rhythmically alternate different configurations of speech gestures (e.g., altering the relative timing of lip opening and glottal movement for /p/ vs. /b/). Research across a variety of domains of action has suggested that alternating different movements is inherently less dynamically stable than repeating synchronous actions. When participants are asked to perform alternating movements under varying response speeds, they spontaneously shift from successful alternation to synchronized move-