Investigating perceptual biases, data reliability, and data discovery in a methodology for collecting speech errors from audio recordings

John Alderete, Monica Davies
Simon Fraser University

Collaborators: Stefan Frisch (South Florida) Alexei Kochetov (Toronto)

Thanks also to: Holly Wilbee, Jennifer Williams, Gloria Fan, Rebecca Cho
Abstract

This work describes a methodology of collecting speech errors from audio recordings and probes how some of its methodological decisions affect data quality and composition. Speech errors of all types (phonetic, phonological, syntactic, etc.) were collected by six data collectors from podcasts of natural and unscripted English speech. Analysis of these errors showed that (i) different listeners find different errors in the same audio recordings, but (ii) the frequencies of error types are similar across listeners; (iii) errors collected “online” in daily life without recourse to a recording are more likely to be affected by perceptual biases than “offline” errors collected from podcasts, and (iv) when properly trained, listeners are able to collect gradient phonetic errors that have not traditionally been included in speech error collections.
Speech error research

**Common stock of problems, English and other languages** (Shattuck-Huffnagel, Dell, Stemberger)

- Sound errors: source and error ‘close’, phonological similarity, syllable position constraint, word-onset effect; lexical bias, frequent sound bias, no category constraint
- Word errors: not necessarily ‘close’, category constraint, semantic bias, mixed errors

**Groundwork for organization of production processes** (Garrett, Fromkin, Dell, Stemberger)

- Lexical errors: lemma selection
- Word exchanges: function assignment
- Sound errors: phonological (and phonetic) encoding

**Linguistic insights** (Fromkin, Stemberger, Shattuck-Huffnagel)

- Psychological reality of phonemes, features, rimes, etc.; structural notions like ambisyllabicity and CV structure
Problems with observational techniques

Labor intensive
• Most large studies take several years to see results

Error prone
• Errors can be missed, incorrectly documented, and misclassified; about 1 in 4 need to be tossed out

Subject to perceptual biases, perhaps not representative
• Difficult to hear errors in certain contexts, certain structures
• Perceptual systems pull gradient patterns to poles
• Transcription methods not always best

Ambiguity of the data
• Many errors can be classified in more than one way, making enumeration of error types difficult

The turn from observational methods
➢ To mitigate some of these problems, many use experimental techniques: SLIP technique, tongue twister technique; also normal speaking, brain imaging
Goal: methodology that mitigates problems

Objective
Modify observational techniques to make them more efficient and less error prone
• Many data collectors: many talkers
• A rigorous training regime
• Separate data collection and classification
• Exclusive use of audio recordings: “offline” data collection

Progress in a year and a half
With team of five students, collected 8,500 errors of all types
Approx. 2,000 of 8.5k classified
Goals: Specific research questions

1. Given multiple data collectors, are there any important effects due to data collectors?
   - Expect to find differences, but not clear if they will affect the sample

2. Is data collected online different from data collected offline?
   - Expect to find differences, and that offline collection will result in a more accurate reflection of error levels and types.

3. Given offline collection method, what other uses might audio recordings have in language production studies?
   - Gradient phonetic structures, existence and relevance for principles of speech production.
Context: Problems with data reliability

Sources of incorrect errors or unrepresentative sample of actual patterns

- Failure to record errors
  - errors are not recognized/remembered
  - errors occur too fast, past time envelope
- False Positives
  - error was misheard (‘slip of ear’, Cutler 1981)
  - incorrect recall or transcription of error
  - error doesn’t meet definition “non-habitual deviation from speech plan”
- Misclassification
  - due to incorrect transcriptions
  - incorrect or insufficient contextual information
  - coding inconsistency

A reliability test: Ferber 1991

4 listeners, 45 minutes of German (online), also listened offline by Ferber

- Approx. 2/3 online errors heard by only one listener
- Important differences between two classes of listeners: one listener found no sound errors, two listeners found no word errors
- Many false positives: 24% (12/51 all online)
Context: Perceptual biases (Bock 1996, Pérez et al. 2007)

**Content bias**: errors are easier to detect if they affect meaning

**Word onset**: errors are less noticeable if they occur at the end of words than the beginnings of words

**Predictability**: errors also easier to detect when they occur in highly predictable environments (e.g., ... *go smoke a cikarette*) or primed with words associated with the error word

**Bias for discrete symbolic categories**: especially for sound errors, biased toward assigning sounds to discrete phonological categories

**Compensation for coarticulation**: phonetic environments may enhance certain contrasts and lead to selection of some discrete sounds over others

**Attention bias**: lower level errors (phonetic or phonological) are often more difficult to detect and therefore require greater attention, substitution harder detect than exchange (e.g., *left lemisphere* vs. *heft lemisphere*)

**Feature biases**: sound errors with changes in some features are easier to detect that others, e.g., nasality and voicing easier than place; also more feature changes easier to detect than fewer
Context: Additional constraints and biases

Data collector bias: collectors differ in the rate of detection and types of errors (see perceptual biases)

Talker sampling bias: sample may be biased toward the errors of specific individuals, e.g., if just one data collector

Theoretical bias: purchase of a theory or specific hypothesis may affect the errors collected

Memory constraints: mentally very taxing to follow a conversation, track entities, and then detect and record errors

Native language bias: hard to detect errors that don’t fall into the categories of your own language
Context: Common ways of addressing problems

**Best practices:** only write down errors that have high degree of confidence in, 30 seconds after speech act, make conscious effort to collect (no multi-tasking)

**Intended words:** Immediately question speaker if intended word unclear (not always useful)

**Many data collectors:** Dell & Reich (1981) used 200 undergraduate students to collect a large sample; see also Pérez et al. 2007

**Collect lots of data:** 3 data collections have upwards of 6,000 errors

**Compare with experimental data** (see Stemberger): if fits with trends in experimental data, stronger statement; otherwise look for tasks effects or problems with data collection

**Objection**
Most large corpora are collected online by one or two individuals; therefore prone to (1) perceptual biases and (2) reliability problems.
**SFUSED: General methods**

**Multiple data collectors**
- reduces collector bias, allows it to be studied
- currently six data collectors, generating about 200 errors a week

**Offline collection from audio recordings**
- Errors collected from podcasts on different topics
- Podcasts selected for having natural unscripted speech, usually Western Canada and U.S. (Midlands dialect ‘Standard American’)
- Examples: *Astronomy Cast, Go Bayside, Accidental Tech, Rooster Teeth*
- Multiple podcasts (8 currently) with different talkers, approx. 50 hours of each podcast
- Record dialectal and idiolectual features associated with speakers (because habitual, so not an error); listeners develop expectations about individuals
- Older approach combined online and offline, but no longer collect online

**Training regime**
- Undergraduate students, introduction to formal linguistics, phonetics and phonology
- Given phonetic training in transcription and tested for transcription accuracy
- Introduction to speech errors, definition and illustration of all types
- Training through listening tests: assigned pre-screened recordings, asked to find errors; learn by reviewing correct list of errors. Trainees that reach a certain level of accuracy continue.
SFUSED: Addressing the problems

- **Follow best practices (online)**
  - direct observation (no secondary sources)
  - record error within 30 seconds
  - conscious effort

- **Collector bias**
  - multiple collectors
    - so biases can be studied
    - also, if bias exists, reduced to specific collectors

- **Talker bias**
  - reducing effect of specific individuals
    - multiple collectors
    - multiple recordings
    - larger number of talkers

- **Theory bias**
  - most data collectors not versed in language production theory
    - those that are not assigned a project

- **Symbolic categories**
  - we allow for gradient categories to be recorded

- **Perceptual biases**
  - inherent problem
    - can at least study given that researchers tagged, so can look for effects

- **Missed errors**
  - inherent problem, and difficult one
    - but available of audio recording reduces this considerably
    - Error detection rates enables us to determine if many errors were missed

- **Incorrect errors**
  - also difficult problem
    - reduce it with phonetic training
    - Must have a second check at classification stage
<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>Date</th>
<th>Example</th>
<th>Intended</th>
<th>Confidence</th>
<th>Gradient</th>
<th>Corrected</th>
<th>Podcast</th>
<th>File</th>
<th>Time</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFUSED</td>
<td>Raw collection phase</td>
<td>2015-06-13</td>
<td>If it feels like 'hey why not let's go out' I'll say I think you know what? People on okay cupid</td>
<td>think</td>
<td>y</td>
<td>whc02_2014-07-02</td>
<td>0.17:05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-06-13</td>
<td>yes and uh just send anything you've seen that you want us to talk about we have concerns</td>
<td>talk</td>
<td>who03_2014-04-04</td>
<td>0.25:35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-04-13</td>
<td>&quot;You're [certainly] shooting the girls this morning Dr. P.&quot;</td>
<td>certainly</td>
<td>who03_2014-04-04</td>
<td>0.02:41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-03-13</td>
<td>&quot;yes you kim's right. Kim's right. Uh she won he probably got [har] lost in the hash marth /will[...]/ lost, wildeness</td>
<td>with</td>
<td>who03_2014-04-04</td>
<td>0.07:02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-03-13</td>
<td>I feel like even if you're /fr[..]h [a] small group of people</td>
<td>with</td>
<td>who03_2014-04-04</td>
<td>0.08:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-02-13</td>
<td>Look, I'm making a house out of all of our mash potatoes</td>
<td>with</td>
<td>who03_2014-04-04</td>
<td>0.18:17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I feel like I'm already making a more efficient use of my time because I'm /washing the movie and an watching</td>
<td>with</td>
<td>who03_2014-04-04</td>
<td>0.06:18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I'd have sex with caveat [anna xxx Emma Stone]</td>
<td>Emma</td>
<td>y</td>
<td>who02_2014-04-04</td>
<td>0.07:41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>well how good is it now am I doing something</td>
<td>good</td>
<td>y</td>
<td>who02_2014-04-04</td>
<td>0.11:18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I think that you duh the short attention /fr[..] can be manipulated by shiny objects</td>
<td>span</td>
<td>who02_2014-04-04</td>
<td>0.13:50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>kids ask your mom what a radio station is xxx cause [fag] here comes another good one</td>
<td>(null)</td>
<td>who02_2014-04-04</td>
<td>0.16:51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>but that's what that's why you're gonna have to name your [kit xxx kid like [facebook]</td>
<td>kid</td>
<td>y</td>
<td>who02_2014-04-04</td>
<td>0.03:04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I had a guy in college say [i don't know] you don't wanna be Jeffrey</td>
<td>you</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.08:19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>isn't it annoying when somebody jux xxx when somebody does one and you didn't give them /per[..]</td>
<td>permission</td>
<td>who02_2014-04-04</td>
<td>0.10:57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>o.k. I'm here /fr[..]h [a] both an awesome clown entertainment personality and the brother of your [fr[..]l]</td>
<td>good</td>
<td>who02_2014-04-04</td>
<td>0.15:07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>not even you 'cause steve was /free bagging you was taken? [other speaker: not free bagging, be tea</td>
<td>you</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.17:40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>the [fr[..]l] xxx the original transformers surprised me by being moderately watchable</td>
<td>original</td>
<td>who03_2014-04-04</td>
<td>0.00:16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>you can see this in our every day lives where we xxx [fr[..]r] talk about celebrities in a certain way</td>
<td>trans</td>
<td>who02_2014-04-04</td>
<td>0.04:44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>hey man I want /you to xox introduce you to my friend Carl</td>
<td>introduce</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.08:36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>yeah okay /kitxxx can't see it but I'm playing with one of those tech desk fingerboards right now</td>
<td>guys</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.11:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>if you've only [fr[..]l] we've argued three times and it hasn't gone good</td>
<td>3 times</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.16:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>oh you're wearing /fr[..]l xxx like tennis racket on your feet?</td>
<td>the like</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.06:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>if we can't [fr[..]l]boot for an ideal with regard to the gorilla arm ratio</td>
<td>shoot</td>
<td>who03_2014-04-04</td>
<td>0.10:21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>if you're about to say that you accidentally gave off a [fr[..]l] xox donkey, you're remembering the donkey</td>
<td>donkey</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.14:25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I saw bamboo for real xxx ah bamboo?</td>
<td>who03_2014-04-04</td>
<td>0.17:39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>/fr[..]xxx leave a review that says I believe in my heart that the host of this show could fight one head</td>
<td>koalas</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.05:18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>yeah where /where[es] joke does personality for morning?</td>
<td>necessity</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.11:51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I'm here xxx i'm here for the petting zoo</td>
<td>the</td>
<td>who03_2014-04-04</td>
<td>0.16:21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>there's been a lot of new scans and tests and studies on dogs that show they do have a very high</td>
<td>high</td>
<td>who03_2014-04-04</td>
<td>0.19:13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>God I hope I meet /[and] Allen</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.21:26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>[and] their conclusion that they came to is that it existed in this weird riff between species</td>
<td>the</td>
<td>who03_2014-04-04</td>
<td>0.00:56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>that's dolphin for leave/they love /itunes</td>
<td>economics</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.02:13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>we call it trickle up /fr[..] xxx economics</td>
<td>who03_2014-04-04</td>
<td>0.07:02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>nah you know what /fr[..]l [man this thing is weird]</td>
<td>man</td>
<td>who03_2014-04-04</td>
<td>0.09:14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>they're for fish commutes? That sounds like really inefficient</td>
<td>it's</td>
<td>who03_2014-04-04</td>
<td>0.06:19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>forty fish per minute uh shooting out across /through xxx the vastness of space into uh higher on the</td>
<td>the</td>
<td>who03_2014-04-04</td>
<td>0.06:58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>shake a xam in it xox salmon in to that</td>
<td>salmon</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.18:24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I don't think this is orally built for [fr[..]l] xox uh for food transportation</td>
<td>your mom</td>
<td>y</td>
<td>who03_2014-04-04</td>
<td>0.21:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>you sent your mom sends thousands of dollars through xxx that tube and then two lippolops con</td>
<td>your mom</td>
<td>who03_2014-04-04</td>
<td>0.21:18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015-01-13</td>
<td>I agree with you. I love the whiskey /in[b] xxx inside things</td>
<td>inside</td>
<td>who03_2014-04-04</td>
<td>0.00:44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SFUSED: Classification phase

Add new variables (largely following Stemberger 1993)
  • Raw speech errors then given a set of 45 fields that classify the error

Example
  • “I read officially it's /[k]upposed to ^come out this afternoon”
    • intended: supposed

Classification:
  • level: sound, contextual (from come), nonlexical
  • type: anticipation
  • intruder: [k]
  • supplantedIntended: [s]
  • supplantedIntendedSyllableRole: onset
  • supplantedIntendedSyllabicPosition: word-initial, etc.

Variables:
  • Right lemma?, lexical, contextual, level, gradient, confidence, contentWord, intruder, supplantedIntended, syllableRole, syllable position, intendedPartOfSpeech, distance, clause-bound, word-bound
  • use a set of attributes enable us to test specific hypotheses about the structure of speech errors, e.g., phonological similarity

Verification
  • All errors are relisted and checked for validity and accuracy
SFUSED: Illustration of the interface
Experiment 1: Effect of data collector

Questions
- How reliable is our error collection methodology?
- Are data collectors collecting the same types of errors?
- What errors are heard by all?
- Which are heard by specific collectors?

Methods
- Listened to nine 30-40 min. podcasts (three different series), 380 minutes total
- Each podcast listened by three trained listeners (two split the nine)
- Made error submissions separately, then at least two listened to all errors and checked error validity

Basic findings

<table>
<thead>
<tr>
<th></th>
<th>Total ‘errors’ found</th>
<th>Total ‘errors’ incorrect</th>
<th>Total correct errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>380</td>
<td>94 (25%)</td>
<td>286 (75%)</td>
</tr>
</tbody>
</table>

*error every minute
*error every 1 minute 20 secs
Experiment 1: Accuracy and rate

<table>
<thead>
<tr>
<th></th>
<th>Total ‘errors’</th>
<th>False positives</th>
<th>% correct</th>
<th>Minutes per correct error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>50</td>
<td>16</td>
<td>68%</td>
<td>4.85</td>
</tr>
<tr>
<td>Listener 2</td>
<td>85</td>
<td>18</td>
<td>79%</td>
<td>3.21</td>
</tr>
<tr>
<td>Listener 3</td>
<td>177</td>
<td>33</td>
<td>81%</td>
<td>2.64</td>
</tr>
<tr>
<td>Listener 4</td>
<td>206</td>
<td>32</td>
<td>84%</td>
<td>2.18</td>
</tr>
</tbody>
</table>

Implications
- Can require 70% accuracy as criterion for continuation in data collection
- Individual collectors far about actual MPE, but impossible to expect collectors to find all errors
- Large number of false positives: need verification stage in classification
Experiment 1: Consistency in correct errors

<table>
<thead>
<tr>
<th>Out of 286 correct errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Heard by just one person</td>
<td>193 (67%)</td>
</tr>
<tr>
<td>Heard by specifically two people</td>
<td>53 (18%)</td>
</tr>
<tr>
<td>Heard by all three people</td>
<td>40 (14%)</td>
</tr>
<tr>
<td>Heard by more than one</td>
<td>93 (32%)</td>
</tr>
</tbody>
</table>

- About 2/3rds of the data were heard by just one person
- Raises the issue of whether there are differences in the types of errors heard by individual collectors?
Experiment 1: Error levels by collector

<table>
<thead>
<tr>
<th></th>
<th>Sound</th>
<th>Word</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>17 (48.6%)</td>
<td>14 (40.0%)</td>
<td>4 (11.4%)</td>
<td>35</td>
</tr>
<tr>
<td>Listener 2</td>
<td>38 (55.9%)</td>
<td>15 (22.1%)</td>
<td>15 (22.1%)</td>
<td>68</td>
</tr>
<tr>
<td>Listener 3</td>
<td>89 (61.4%)</td>
<td>40 (27.6%)</td>
<td>16 (11.0%)</td>
<td>145</td>
</tr>
<tr>
<td>Listener 4</td>
<td>100 (57.8%)</td>
<td>46 (26.6%)</td>
<td>27 (15.6%)</td>
<td>173</td>
</tr>
</tbody>
</table>

Definitions
- Sound errors: any sublexical error, usually most difficult to detect
- Word word: substitution, deletion, addition, exchange of a word
- Other: morphological, syntactic, morpho-syntactic, marginal categories

Observations
- Percentage of sound errors is broadly consistent, btw 50-61%
- Only outliers are word and other categories
Experiment 1: Salience measures for all errors

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Corrected errors</th>
<th>Gradient errors</th>
<th>Exchanges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>34</td>
<td>19</td>
<td>4</td>
<td>11.8%</td>
</tr>
<tr>
<td>Listener 2</td>
<td>67</td>
<td>34</td>
<td>8</td>
<td>12%</td>
</tr>
<tr>
<td>Listener 3</td>
<td>145</td>
<td>54</td>
<td>25</td>
<td>17.2%</td>
</tr>
<tr>
<td>Listener 4</td>
<td>173</td>
<td>73</td>
<td>25</td>
<td>14.5%</td>
</tr>
</tbody>
</table>

Observations

- Corrected errors (by speaker) are a red flag to listener; some variation, but listeners with more practice (3, 4) heard more uncorrected errors
- Gradient errors (cf. phonetic errors): also difficult to hear because involve subphonemic structure; good consistency
- Exchanges: only 1 out of 286 errors; very salient
Experiment 1: Salience measures for sound errors

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Errors in stressed syllables</th>
<th>Errors in initial segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>17</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Listener 2</td>
<td>38</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>Listener 3</td>
<td>89</td>
<td>73</td>
<td>31</td>
</tr>
<tr>
<td>Listener 4</td>
<td>100</td>
<td>77</td>
<td>44</td>
</tr>
</tbody>
</table>

Observations
- Errors in stressed syllables are presumably more salient, and also very common; relatively similar percentages (76-82%); still large number of errors heard in unstressed syllables (20%)
- Errors in initial segments are also more common and apparently more salient; 34-44%, with relative consistency
Experiment 1: What errors heard by all?

<table>
<thead>
<tr>
<th></th>
<th>Errors heard all three $N = 40$</th>
<th>All errors $N=286$</th>
</tr>
</thead>
<tbody>
<tr>
<td>% corrected by talker</td>
<td>77.5% (31/40)</td>
<td>34.6%</td>
</tr>
<tr>
<td>% sound errors</td>
<td>57.5%</td>
<td>58.0%</td>
</tr>
<tr>
<td>% gradient errors</td>
<td>2.5% (only one)</td>
<td>17.1%</td>
</tr>
</tbody>
</table>

Examples of uncorrected errors

rock $\rightarrow$ [b]rock
Right, the /[b]rock just gets xxx consumed when it strikes the ground

try $\rightarrow$ [k]try
... you write grant proposals, you /[k]try to get the time ...

technically $\rightarrow$ [d]echnically
... did come out before Left4Dead /[d]echnically, so xxx yeah

photographic $\rightarrow$ photographic
... to look through /photogaphic plates and go oh you know
Experiment 1: Take home messages

1. Many false positives: 25%
   Require a second check on error validity and accuracy

2. Data collectors find different errors
   Approx. 2/3rds of all errors heard by just one data collector

3. Specific errors differ, but error types and contexts broadly consistent
   - Sound errors 50-61%, closer to 60% for experienced listeners
   - Corrected errors: 37-56%, closer to 40% for experienced listeners
   - Gradient errors: 12-17%
   - Errors in stressed syllables: 76-82%
   - Errors in initial segments: 34-44%
Experiment 2: Online vs offline

To Do: I’m going to redo this section with more data; please suggest new measures

Expectation
➢ Offline method of collecting speech errors will be more reliable and more representative of the natural occurrence of errors

Method (“online” on-the-spot collection)
• Collectors: SFU linguistics undergraduates, after introduction to linguistics and phonetics and phonology (volunteers proceed into paid positions)
• Training: given one hour introduction to speech errors, including exposure to all major types of errors. Asked to find 1 hour period during their normal daily routine to collect these errors.
• Instructed in best practices (be sure of error, within 30 seconds, make conscious effort)
• Some errors collected at times other than ‘set aside time’, though typical practice was do this in a set time
• Students submitted raw errors with specific set of variables over email, monitored and vetted by first author
Experiment 2: Error levels by method

<table>
<thead>
<tr>
<th></th>
<th>Sound</th>
<th>Word</th>
<th>Other</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>97</td>
<td>85</td>
<td>17</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td>48.7%</td>
<td>42.7%</td>
<td>8.6%</td>
<td></td>
</tr>
<tr>
<td>Offline</td>
<td>166</td>
<td>75</td>
<td>45</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>58.1%</td>
<td>26.2%</td>
<td>15.7%</td>
<td></td>
</tr>
</tbody>
</table>

**Observations**

- Method had significant effect on levels ($x^2 = 16.29, P=0.0003$): ratio of sound/word errors changes quite a bit
- Interpretation: offline method enabled listeners to detect more difficult to hear sound and ‘other’ errors; percentage of word errors much higher with online method, perhaps because more salient
Experiment 2: Salience measures for all errors

<table>
<thead>
<tr>
<th></th>
<th>Corrected</th>
<th>Not corrected</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>98</td>
<td>101</td>
<td>199</td>
</tr>
<tr>
<td>Online</td>
<td>49.2%</td>
<td>50.8%</td>
<td></td>
</tr>
<tr>
<td>Offline</td>
<td>99</td>
<td>187</td>
<td>286</td>
</tr>
<tr>
<td>Offline</td>
<td>34.6%</td>
<td>65.4%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Exchange errors</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>11</td>
<td>5.5%</td>
</tr>
<tr>
<td>Online</td>
<td>11</td>
<td>49.2%</td>
</tr>
<tr>
<td>Offline</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>Offline</td>
<td>1</td>
<td>34.6%</td>
</tr>
</tbody>
</table>

Observations

- Method also affected the percentage of errors corrected errors ($x^2= 9.82, P=0.0017$): 1/2 for online, cf. 1/3 for offline
- Interpretation: listeners relied more on talker correction for detected errors; likely to actually undershoot this difference because context under-reported in online errors
- Easy to hear exchange errors also have much higher rate with online collection
More on exchanges

<table>
<thead>
<tr>
<th></th>
<th>Offline</th>
<th>online</th>
</tr>
</thead>
<tbody>
<tr>
<td>sound</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>morpheme</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>word</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>phrase</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>total</td>
<td>719</td>
<td>1092</td>
</tr>
<tr>
<td>Total/N</td>
<td>.28%</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

Much higher percentage of exchanges in online, across all levels.

> Easy to hear errors have higher rate of occurrence in online method; probably skewed, because careful listening of offline method much lower.
# Experiment 2: Salience measures for sound errors

<table>
<thead>
<tr>
<th></th>
<th>Stressed syllables</th>
<th>Unstressed syllables</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>79</td>
<td>81.4%</td>
<td>97</td>
</tr>
<tr>
<td>Offline</td>
<td>131</td>
<td>78.9%</td>
<td>166</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Error in initial segment</th>
<th>Error not initial segment</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>41</td>
<td>56</td>
<td>97</td>
</tr>
<tr>
<td>Offline</td>
<td>71</td>
<td>95</td>
<td>166</td>
</tr>
</tbody>
</table>

**Observations**
- Method didn’t affect counts in stressed vs. unstressed syllables
- Method did affect counts in initial-non-initial contexts
- Perhaps not the best measures of salience in sound errors; need deeper analysis of error types, features, complexity
Experiment 2: Take home messages

• Online method seems to be skewed towards easier to hear errors (word vs. sound errors, exchanges, corrected errors)
• But some measures of sound errors showed no effect or the opposite effect
• Perhaps need more sensitive measures: features, examine specific sound errors (anticipations vs. perseverations, etc.)
Possible additional questions

• Do online and offline differ in N-gram probability, e.g., *smoke a cigarette*? Expect difference between online and offline, since errors in highly predictable contexts are easier to hear.

• What about differences in the featural differences in sound errors, e.g., voicing (easier) vs. place (harder)?

*Your comments here ...*
Overall advantages of offline method

- Many studies: “I need more context”!
- **Greater accuracy** through recourse to a sound recording: can remove false positives and better understand the context of an error
- Can assess **error detection rates** by data collector: minutes per error; can posit threshold on MPE to ensure greater coverage of the errors in a recording
- Errors collected offline likely to **better reflect natural occurrence of errors** in human communication; seems to be easier to detect hard to hear errors
- Supports **phonetic analysis** (prosody, pause durations, segmental acoustics, etc.)
- Easier to find **gradient errors** that fall between discrete phonological categories
- Only real drawback: **can’t question speakers** on context, e.g., intended word.
Gradient errors: Motivation for research

Phonological vs. phonetic errors

Growing interest in fine-grained phonetic detail of speech errors (Buchwald, Goldrick, Frisch, Pouplier), but natural speech error corpora don’t transcribe this level because it’s impossible to do online.

Phonetic analysis as a check on perceptual bias: Frisch and Wright 2002 found z > s errors greater than z > s errors, cf. Stemberger 1991, found the opposite ‘anti-frequency bias’; likely that high incidence of s > z due to perceptual bias towards classifying any voicing as a voiced category.

Phonetic errors: deviations in some continuous parameter greater than two standard deviations from normal range, but not within normal range of another category; may not be pre-articulatory, speakers may not even be conscious of error (not investigated).

Phonological errors: outside normal range of intended category, and inside range of a different category; prearticulatory, speakers are conscious of the error, but can’t help it.

Questions:

• Are phonetic and phonological errors discrete classes (different mechanisms) or do they fall on an continuum? How do phonetic errors differ from phonological errors?
• What factors predict frequency and structure of gradient errors? (Frisch: stability).
• Is it even possible to record gradient errors in spontaneous speech? How frequent are they, and which structures tend to show gradience?
Gradient errors: Coding and pilot results

Transcription conventions
Follow protocol used for child language acquisition (see Stoel-Gammon 2001)

- [A|B] for ambiguous sounds between two poles A and B, but closer to A.
- [A-B] for transitional segments (other than affricates and diphthongs), that move from one structure A to another B
- Intrusion: weak sounds, short in duration, that are clearly audible but don’t have the full status of full consonants or vowels
- Use IPA for all transcriptions, not limited to sounds of English

Of 286 errors found in Experiment: 49 (17.1 %) gradient errors
- 34 ambiguous sounds (mostly vowels, nasal sounds)
- 12 transitional segments
- 3 intrusive segments
Projection: 15% of 4,000 = 600 phonetic errors, 10,000 = 1,500 phonetic errors

Out of 1811: 66 ambig, 29 transitional, 3 intrus, 4 other
Gradient errors: Illustrations

Gradient intrusion/epenthesis
GB012: “To everyone, including the /cam\[^b\]ra lens” (camera)
RT010: “... to make sure that no one else has ever /stu\[^d\]id\[^i\] fu ...” (studied)

Ambiguous segments [X|Y], X Y poles, X is ‘closest’ category
AC70: “Science is a ^collaborative /[^ae]\[^e\]f xxx effort xxx very few people work in any sort of isolation”
GB012: “I'm hoping you were /watch[w\[^r\]jing this when you were a teenager”
GB012: “I will say you are definitely the first guest to /[k\[^g\]ompare anything saved by the Bell to either Robert Altman or Paul Thomas Anderson”

Transitional segments [X-Y], starts out as X, moves to Y
AC70: “a night assistant just automatically gets you your data for you and ships it to you either /o[r-v]er the internet or perhaps on a DVD”
GB12: “Cuz now they're observing their /f[er-\[^ae\] xxx fake family, and that's when he's like . . .”
Gradient errors: Future prospects

Likely to get a lot of data
- If pilot results are right, should have several hundred examples of specific structures
- Better training: difficult to instruct data collectors about casual speech and coarticulation (which are habitual, false positives)

Potential differences between phonetic and phonological errors
- Focus on specific structures, e.g., front vowels, s-z, etc.
- Salience measures, e.g., corrected, stress, etc., fully conscious?
- Perceptibility: are phonetic errors recognized as errors like phonological errors?
- Context effects: blends, anticipations and perseverations?
- How fall in continuous parameter space?
- Are the poles in phonetic errors phonologically similar?

Correlate error patterns with non-error patterns
- Compare normal speech of same speaker with errorful speech
- Compare normal articulations of a population (e.g., Buckeye corpus)