Comparing the Use of Sociophonetic Variables in Speech and Twitter

Rachael Tatman (rctatman@uw.edu), Department of Linguistics

Work Supported by NSF Grant DGE-1256082. I would also like to extend my thanks to Nicole Chartier and the UW Sociolinguistics brownbag for their input.

Background

Research question: are orthographic representations of dialect features on Twitter accurate representations of the tweeter's speech?

Sociolinguistic variation on Twitter shows many of the same patterns as speech data:
- racial variation (Eisenstein 2015)
- accommodation (Johnson 2013)
- style-shifting (Schnoebelen 2012)
- dialectal variation (Tatman 2015)

But do tweets represent socio-phonetic variation in the same way? A better understanding of this will allow us to:
- Use Twitter data for dialectology in a principled way
- Possibly use Twitter data to automatically identify users' geographic location

Speakers

Since the aim of this project is to directly compare speech and Twitter data it was necessary to find parallel data sets that:

- Clearly showed use of well-studied sociophonetic variables in both tweets and speech
- Had a clear parallel between tweets and speech (in subject matter, intended audience, level of formality and authorship)

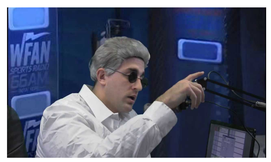
With that in mind the data presented here comes from two speakers:

Mike Francesca:
- popular New York sports radio pundit
- speaker of New York English

Age: 61 (born 1954)
Ethnicity: Caucasian
Location: Long Beach, NY

Bill Buchanan, "Mike Zaun"
- central member of Mongo Nation, Francesca's fan base
- produced videos in the persona of "Mike Zaun", where he impersonated Francesca's speech and mannerisms
- has an active Twitter account (@BigActionBill) where he often (≈50% of the time) tweets using variant spellings

Age: 26 (born 1989)
Ethnicity: Caucasian
Location: Massapequa, NY



New York sports radio pundit Mike Francesca (left) and fan Bill Buchanan (right) performing as "Mike Zaun" (a pun on "Mike's On", the title of Francesca's radio show).

Methods

Speech Data

- Speech data from Youtube
  - Original: Call-in radio show (Francesca 2013)
  - Imitation: Parody of call-in radio show (Buchanan 2013)
- Transcribed and annotated by hand
- Force aligned using FAVE-align (Rosenfelder et al. 2011) for acoustic measures

Twitter Data

- Collected with R script (available on the author's Github) using TwitteR (Gentry 2015) package
- The 100 most recent tweets (as of May 22, 2015) from Bill Buchanan's Twitter account
- Annotated by hand
- Dataset of all tweets available on author's github
- Notes on Twitter data:
  - Square brackets are author's annotations
  - All @ tags have been removed for ease of reading
  - Value in (parentheses) are indexes to dataset

Observed Variables

- Content analysis of Twitter data revealed a large number of phonological variables (stereotypes) associated with New York English (Labov 2006)
- All used with multiple lexical items

Variable	# in Tweets	Color	Example
[ɻ] deletion	38	Mustard	Beah [beer] (9)
ð stopping	9	Violet	Duh [the] (8)
Backed /aɪ/	7	Magenta	Woyld [wild] (25)
Distinct [ɔ] ("cawfee" vowel)	5	Blue	Nawt [not] (14)
G-dropping*	4	Green	Swarmin [swarming] (40)

[ɻ] deletion & ð stopping

Speech
- Zaun used both [ɻ] deletion & ð stopping at higher rates than Francesca
  - [ɻ]: $\chi^2(2, N = 82) = 8.11$, $p < 0.01$
  - ð: $\chi^2(2, N = 106) = 6.32$, $p = .0.01$

Tweets
- Zaun tweeted both at the same rate as his use in speech
  - [ɻ]: $\chi^2(2, N = 97) = 0.16$, $p = 0.68$
  - ð: $\chi^2(2, N = 61) = 2.26$, $p = 0.13$

Example Tweets
- "**befoah** we get to the graduates I'd like to touch on this **rainjizz** [Rangers] story" (3)
- listen folks um I hate to break **dis** to u but **dis** is a nice college not a great one. It's not st **jawns** [John's] ok (2)

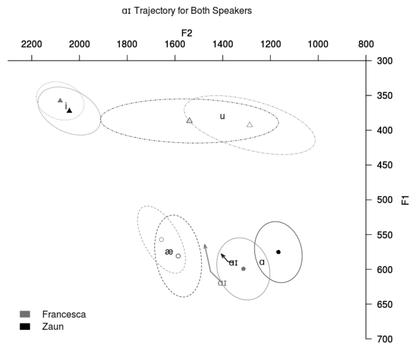
*Though not a dialect marker, this variable was represented extensively in both Twitter and speech data and was included for completeness. It patterns with the backed /aɪ/ and distinct [ɔ], as can be seen in the chart in the conclusions section.

Backed /aɪ/

Speech:
- Used at the same rates in Zaun's and Francesca's speech
  - /aɪ/: $\chi^2(2, N = 62) = 0.27$, $p = 0.60$

Tweets:
- Zaun tweeted at a lower rate than his use in speech
  - /aɪ/: $\chi^2(2, N = 52) = 5.96$, $p = 0.05$

Example tweets:
- "...then we go to go to adventure land...and went on **awl** the rides with no **loynes** ok" (18)
- was very **koynd** ok (37)



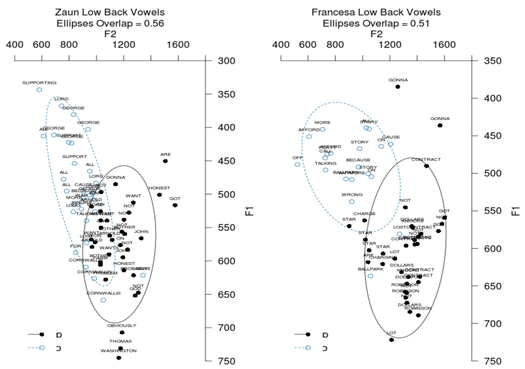
Distinct [ɔ]

- Only LOT and THOUGHT word classes counted, all tokens shown in chart
- Both very low, Parallel findings by Becker (2014)

Speech:
- Both used at same rate in speech
  - [ɔ]: $\chi^2(2, N = 64) = 0.51$, $p = 0.47$

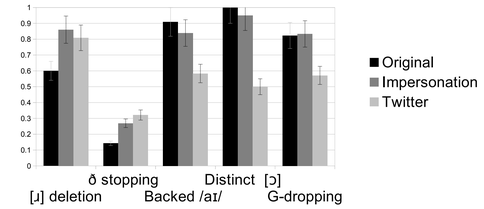
Tweets:
- Zaun used at lower rates in tweets than speech
  - [ɔ]: $\chi^2(2, N = 48) = 9.86$, $p < 0.01$

Example tweets:
- Truce Torre. I once provided ordio [audio] that was gonna be yoozed [used] as a funny bit **awn da** show. **Nevah** made it on thee **aih.** (32)
- **Weah** [we're] not **doin da** mets today!! Get **lawst**! (36)



Conclusion

Rate of Use of Variable by Source



Do speakers use variables the same way in tweets and speech?

Yes:
- Very salient variables are used at the same rate in speech and tweets
- Overshoot non-imitation production
- Patterns with performance registers (Schilling-Estes 1998)

No:
- Slightly less salient (but still stereotyped!) variables used at lower rates in tweets than speech
- Not used more highly in imitation production
- Patterns with earlier findings by Tagliamonte and Denis (2008) and Honeybone & Watson (2013)

References

Becker, Kara. (2014). The social motivations of reversal: Raised bought in New York City English. Language in Society, 43(04), 395-420.

Buchanan, Bill [Waituhseckin Productions] (2013, Sep. 30) *Mike Zaun 1776* [Video file]. Retrieved from https://www.youtube.com/watch?v=ptPBCSsPJR4

Eisenstein, Jacob (2015), Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19: 161–188. doi: 10.1111/josl.12119

Francesa, Mike [Ordio Mongo] (2013, Sep 17). *Mike Argues With A "Fair-Weathah Yankee Fan", tells him to get lawst* [Video file]. Retrieved from https://www.youtube.com/watch?v=JpH_31XBW_4

Gentry, Jeff (2015). twitteR: R Based Twitter Client. R package version 1.1.9.

Honeybone, Patrick & Kevin Watson (2013). Salience and the sociolinguistics of Scouse spelling: Exploring the phonology of the Contemporary Humorous Localised Dialect Literature of Liverpool. *English World-Wide*, *34*(3), 305-340.

Johnson, Ian (2013). Audience design and communication accommodation theory: Use of Twitter by Welsh-English Biliterates. *Social Media and Minority Languages. Convergence and the Creative Industries*, 99-118.

Labov, William (2006). *The social stratification of English in New York city*. Cambridge University Press.

Rosenfelder, Ingrid; Fruehwald, Joe; Evanini, Keelan and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. http://fave.ling.upenn.edu.

Schilling-Estes, Natalie (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. *Language in Society*, *27*(01), 53-83.

Schnoebelen, Tyler (2012). Do you smile with your nose? Stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*,*18*(2), 14.

Tagliamonte, Sali & Derek Denis (2008). "Linguistic ruin? LOL! Instant messaging and teen language". *American Speech* 83.1:3–34.

Tatman, Rachael (2015). # go awn: Sociophonetic Variation in Variant Spellings on Twitter. *Working Papers of the Linguistics Circle*, *25*(2), 97-108.

"I'm a **spawts guay**…
a better **spawts guay** than you" #KING (26)

Poster is available for download at http://tiny.cc/tatmanNWAV44