Toward a holistic measure of reduction in spontaneous speech

Michael McAuliffe, michael.mcauliffe@alumni.ubc.ca
Molly Babel, molly.babel@ubc.ca

Department of Linguistics
University of British Columbia

What is reduction?

Frequent and predictable words are often reduced in speech. What does this mean? Reduction involves:

- Decrease in intelligibility, shortening of duration, articulatory undershoot, general lenition [Lieberman, 1963, Clopper and Pierrehumbert, 2008, and others]

The need for an improved measure

Measures of reduction have built-in assumptions:

Durational approaches have the fewest assumptions and are the most common, but ignore spectral content

Targeted acoustic approaches look at single acoustic measures, like vowel formants, but ignore sounds without those features

Segmental approaches use phonetic transcriptions, which have holistic judgments by transcribers, but no fine detail

There are some issues here:

- Given a framework that accommodates exemplars:
  - How do the current measures fit in?
  - If an exemplar is located in multidimensional acoustic space, does a single-dimension measure capture enough information?
- Targeted acoustic measurements are limited by class of sounds
  - Are stops reduced in the same way across places of articulation?
  - Do all vowels reduce to schwa? [Clopper and Pierrehumbert, 2008, Scarborough, 2010]
- Segmental transcriptions have difficulty capturing most reductions
  - Too detailed of a transcription system hinders intertranscriber reliability
  - Boundaries are close to nonexistent
  - Transcribers bring in a wealth of extra knowledge about the context and the intended word

Research Question

Can we create a measure of reduction (and hyperspeech) that uses holistic acoustic representations?

Data

- Buckeye Corpus of Spontaneous Speech [Pitt et al., 2007]
- Only use monosyllabic CVC words from Gahl et al. [2012]
- Waveforms excised from context and converted to Mel-frequency cepstrum coefficients (MFCC), based off of HTK implementation [Young et al., 1995]
- Surrounding speaking rates and contextual probabilities calculated as in previous work
- Frequency and neighbourhood density from IPHOD [Vaden et al., 2009]

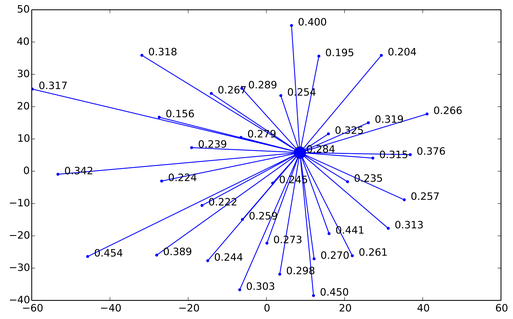

Figure 1: Example affinity propagation cluster for speaker s07's productions of the word *time*. Tokens are labelled with their durations in seconds.

Hypo-Hyperspeech measure

Separation of tokens based on Speaker and Word

Exclude when number of tokens per word per speaker is less than or equal to five

Distance matrix generated using dynamic time warping between MFCC representations

Prototype identification through affinity propagation algorithm, which identifies protoypes among data points and forms clusters of similar data points around them through belief propagation [Frey and Dueck, 2007]

Normalization of distance to prototype by subtracting the mean and assigning positive or negative sign based on difference in duration

Hypo-Hyperspeech measure is - for hypospeech and + for hyperspeech

Experiment 1: Statistical models

- Durational and spectral approaches have consistent predictors in spontaneous speech [Gahl et al., 2012]
- As the Hypo-Hyperspeech measure is calculated per word per speaker, we'll only look at contextual variables
- Preceding and following speaking rates are the syllables per second of running speech before and after the token
- Preceding and following conditional probabilities are the probability of the word given the preceding and following words

Two linear mixed-effects models were constructed with Speaker and Word as random effects and full random slopes specified.

Duration similar to Gahl et al. [2012]

Hypo-Hyperspeech as described in Procedure

Hypothesis

The Hypo-Hyperspeech measure will show contextual effects consistent with previous reduction measures.

Results

Factor	Duration	Hypo-Hyper
Prec. speaking rate	-7.38	-4.33
Foll. speaking rate	-8.30	-5.78
Prec. cond. prob.	-3.45	-2.38
Foll. cond. prob.	-7.74	-2.36

Table 1: Effect sizes (t-statistics) from linear mixed-effects models

Experiment 2: Listener judgments

- We want to make sure our measures of reduction correspond to listener judgments of reduction.

Hypothesis

Measures of reduction should correlate well with expert listener judgements of reduction.

Procedure

- Two speakers (one male, one female) from the corpus data
- Three words (*back*, *said*, *time*)
- Three expert listeners
- Blocked by speaker and word
- Listeners heard all tokens of a word from a speaker, and then were presented with each token individually for rating.
- Orthographic presentation of the word co-occurred with auditory presentation.
- Rated tokens from 1 = "extremely reduced" to 9 = "extremely hyperarticulated"

Results

- Interrater reliability was high $[\chi^2(202) = 417, p < 0.05, W = 0.688]$
- Duration was well correlated overall with listener judgments of reduction $[t(147) = 5.97, p < 0.05, r = 0.44]$
- Hypo-Hyperspeech measure was also correlated overall with listener judgments of reduction $[t(147) = 2.94, p < 0.05, r = 0.24]$

References

Cynthia G Clopper and Janet B Pierrehumbert. Effects of semantic predictability and regional dialect on vowel space reduction. *The Journal of the Acoustical Society of America*, 124(3):1682–1688, 2008.

Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

Susanne Gahl, Yao Yao, and Keith Johnson. Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806, 2012.

Philip Lieberman. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6(3):172–187, 1963.

Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. Buckeye corpus of conversational speech (2nd release)[www. buckeyecorpus. osu. edu] columbus, oh: Department of psychology. *Ohio State University (Distributor)*, 2007.

Rebecca Scarborough. Lexical and contextual predictability: Confluent effects on the production of vowels. *Laboratory phonology*, 10:557–586, 2010.

KI Vaden, HR Halpin, and GS Hickok. Irvine phonotactic online dictionary, version 2.0 [data file], 2009.

S Young et al. The htmm toolkit (htk)(software and manual), 1995.

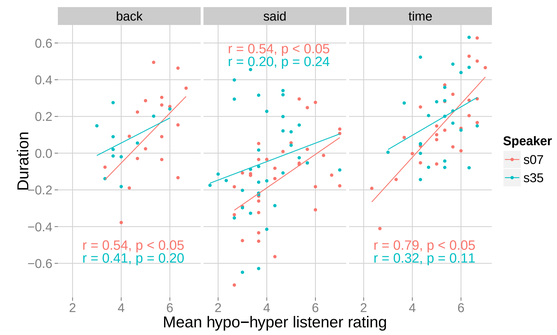

Figure 2: Correlations between duration and listener judgments of hypospeech and hyperspeech, separated by speaker and word

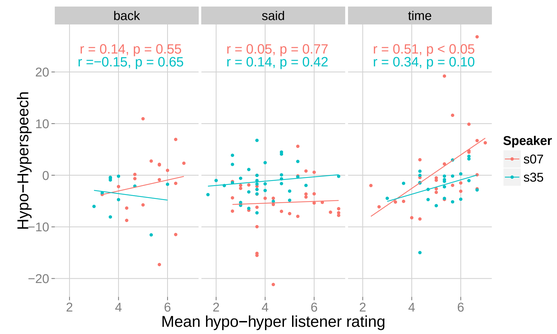

Figure 3: Correlations between the Hypo-Hyperspeech measure and listener judgments of hypospeech and hyperspeech, separated by speaker and word

In a future experiment we will use intelligibility as measured by a word identification task to better assess how the Hypo-Hyperspeech measure fits with listeners' assessment of reduction.

Discussion & Conclusions

- This Hypo-Hyperspeech measure works as well as duration, in terms of directionality and effect sizes.
- The Hypo-Hyperspeech measure leads to non-trivial relationships
  - Speaking rate is measured in syllables per second, so duration in seconds being affected speaking rate is trivial in some sense
  - Speaking rate affecting acoustic realization is a much more compelling relationship
- In addition to working as well as duration, the Hypo-Hyperspeech measure better matches our assumptions about the multidimensional nature of reduction.

Acknowledgements

Thanks to everyone in the UBC Speech in Context Lab, especially Jamie Russell and Jen Abel, and at Malaspina Labs.