A data-driven approach to stylistic identification

Most quantitative sociolinguistic approaches to style can be described as top-down: external factors such as topic, context, and interlocutor are mapped onto sections of conversational speech, and then the task is to determine how the linguistic properties of those sections differ. We pursue a bottom-up approach to identifying stylistic shifts in speech: given a sequence of observations of a variable, we infer the sequence of underlying generative states by training a Hidden Markov Model (HMM). This methodology may be of particular interest in light of growing interest in the temporal-sequential dynamics of sociolinguistic variation (Podesva 2007, Sharma & Rampton 2011, Tamminga 2014).

The dataset we use contains 18,022 observations of the /dh/-stopping variable (*them ~ dem*) from sociolinguistic interviews with 42 white Philadelphia English speakers in the Philadelphia Neighborhood Corpus (Labov & Rosenfelder 2011). For each speaker, we consider the fit of HMMs with 2, 3, or 4 underlying states, where states may be thought of as styles that have a set probability of emitting a 1 ('them') or 0 ('dem'). The HMMs are initialized using a transition probability matrix that favors staying in the current state or moving to a nearby state over moving to a more distant state. We then train the HMMs on the observed data using 100 iterations of the Baum-Welch algorithm and infer the most probable state sequence from the algorithm's output using Viterbi decoding. Given the differences between the observed data, the Viterbi-decoded sequence of probabilities, and the number of model parameters, we are able to calculate the Bayesian Information Criterion (BIC) in order to assess whether the addition of more states improves the performance of the HMM enough to justify the increase in parameters.

We find that for about 70% of the speakers (29/42), the data is fit best by an HMM with just two states. This is strikingly convergent with the top-down analysis given by Labov (2001) for similar speakers from the same speech community. He finds that 8 categories of externally-determined styles (such as "narrative") form only two clusters in terms of their quantitative /dh/-stopping behavior. We therefore suggest that our new methodological approach offers the promise of extracting stylistic clusters automatically from observed data. The inferred state sequences, then, could be used as input to a qualitative investigation of the interview context, providing a more statistically robust basis for the linking of stochastic linguistic behavior to ethnographic insights. Alternatively, they might be used to factor out stylistic variation in order to facilitate the investigation of other sequential properties of the data, such as intraspeaker priming.

References:

Labov, W. 2001. The anatomy of style-shifting. In Eckert & Rickford (eds.), *Style and Sociolinguistic Variation*. Cambridge University Press.

Podesva, R. 2007. Three sources of stylistic meaning. In *Proceedings of SALSA 2007*.

Sharma, D. & B. Rampton. Lectal focusing in interaction: A new methodology for the study of superdiverse speech. *Queen Mary's OPAL #22.*

Tamminga, M. 2014. *Persistence in the Production of Linguistic Variation.* Ph.D. dissertation, University of Pennsylvania.