Phonetic Variation Analysis Via Multi-Factor Sparse Plus Low Rank Language Model

Curtis Fielding, Joshua Weaver and **Brian Hutchinson**
Computer Science Department
Western Washington University

1 Introduction

Phonetic transcriptions contain rich information about language. First, the sequential patterns in phonetic transcripts reveal information about the language's phonotactics. When combined with lexical information, this can help to grow or correct pronunciation dictionaries and to improve grapheme-to-phoneme prediction. Second, the places where pronunciations deviate from the norm can be equally informative; for example, by providing cues for speaker traits such as accent, dialect or sociolect. Interesting in itself, detecting speaker characteristics can also be used to improve speech recognition system performance (Biadsy, 2011).

In this extended abstract we describe on-going work to automatically analyze both the regularities and the exceptions (deviations) in phonetic sequences. We use the Multi-Factor Sparse Plus Low Rank Language Model (Hutchinson et al., 2013), which was shown to effectively model regularities and exceptions in word sequences (e.g. by identifying lexical deviations characteristic of topic or speaker role). Preliminary results modeling commonalities and variation between dialects of American English are promising and suggest several extensions to this work.

2 Sparse + Low Rank Language Models

The Sparse Plus Low Rank Language Model (SLR-LM) is an exponential language model (Hutchinson et al., 2012) that defines the following conditional probability of a token x given a history h:

$$p(x|h) = \frac{\exp\left(\psi(x)^T(S+L)\phi(h)\right)}{\sum_{x'}\exp\left(\psi(x')^T(S+L)\phi(h)\right)}. \quad (1)$$

Here $\psi(x) \in \mathbb{R}^{d_\psi}$ is a feature function that maps token x to a feature representation, while $\phi(h) \in \mathbb{R}^{d_\phi}$ is the feature function for the history. Low rank matrix $L \in \mathbb{R}^{d_\psi \times d_\phi}$ and sparse matrix $S \in \mathbb{R}^{d_\psi \times d_\phi}$ are the parameters. L has a compact singular value decomposition, $L = U\Sigma V^T$, so

$$\begin{aligned} \psi(x)^T L\phi(h) &= \psi(x)^T U\Sigma V^T \phi(h) \\ &= \left(U^T\psi(x)\right)^T \Sigma \left(V^T\phi(h)\right) \\ &= \tilde{\psi}(x)^T \Sigma \tilde{\phi}(h). \end{aligned}$$

Therefore, the L matrix induces continuous low-dimensional representations $\tilde{\psi}(x)$ of tokens ($U^T\psi(x)$) and histories ($\tilde{\phi}(h) = V^T\phi(h)$). The probability of a token following a history is largely governed by $\tilde{\psi}(x)^T\Sigma\tilde{\phi}(h)$, the weighted inner product between these low-dimensional representations. By restricting the rank (i.e. dimension), tokens are soft-clustered through this embedding: the weight matrix L effectively models what *kind* of tokens follow what *kind* of histories. The patterns learned by these weights can be misleading, however, as in in the case of common multiword sequences (e.g. "united states"), where the probability of the bigram is *not* well described by the kind of token "states" is or the kind of history "united" is. The sparse matrix S plays the role of correcting weights (and thus probabilities) for such exceptional n-grams that are not well-modeled by the low rank matrix.

3 Preliminary Experiments and Results

Our preliminary experiments use the TIMIT Corpus (Garofolo et al., 1993), which provides dialect labels for each speaker as well as manual phonetic transcriptions of each utterance. Altogether, there are 462 speakers, with 10 utterances per speaker including two utterances spoken by each speaker explicitly designed to elicit dialectal variation. The training data totals 167,840 phone tokens, with a vocabulary of 60 phones.

We use a generalization of the SLR-LM known as the Multi-Factor SLR-LM (Hutchinson et al., 2013). This model incorporates additional sparse matrices, each designed to model distinct kinds of influences on the token sequence (e.g. due to a particular topic, or a particular speaker, etc.). In our case, in addition to the global low rank matrix L, and a global sparse matrix S_0, there are eight additional sparse matrices, one for each of eight dialect regions. The probability of token x following history h in the subset of the corpus corresponding to the ith dialect region is

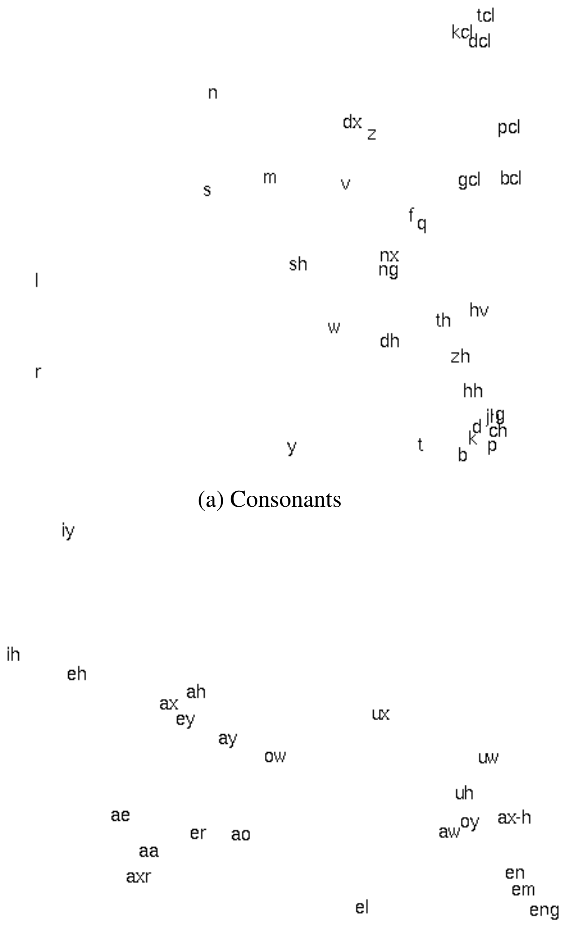
$$P(x|h) = \frac{\exp\big(\psi(x)^T(L + S_0 + S_i)\phi(h)\big)}{\sum_{x'}\exp\big(\psi(x')^T(L + S_0 + S_i)\phi(h)\big)}.$$

The non-zero elements of each sparse matrix will correspond to the n-grams whose probabilities deviate most from the global model; here, the exceptions most characteristic of a given dialect. We train the model as described in (Hutchinson et al., 2013), using one-hot (indicator) feature functions, ψ and ϕ. Upon training, we obtain a rank-4 L matrix, an S_0 matrix with 177 non-zero entries, and eight dialect-specific sparse matrices with between 2 and 37 non-zero entries.

The low-dimensional continuous phone representations are plotted in Figs. 1a (consonants) and 1b (vowels) after further reducing dimensionality for purposes of display using multidimensional scaling. The intuitive clustering of phones confirms that the low rank model is indeed learning phone sequence regularities.

The elements of S_0 capture sequences that often function as a single unit; in our case the largest non-zero entries global sparse matrix are the sequences of stop closure then stop release (e.g. "gcl g", "pcl p"). Other large values in S_0 are short words: "y ux" (you), "aa r" (are), "dh ax" (the).

Some of the non-zero elements in the dialect-specific sparse matrices capture true dialectal variation (Clopper and Pisoni, 2007); e.g. "iy z" in the Southern dialects, as in "g r iy z iy" versus "g r iy s iy". However, many exceptions in these matrices simply serve to boost the probability of phone n-grams that are frequent due to lexical content. While this may be desirable in spontaneous speech, where word choice itself can signal dialect, in a read speech corpus like TIMIT this is an artifact and would be better normalized out.



(a) Consonants

(b) Vowels

Figure 1: 2-d visualizations of 4-d continuous phone representations.

4 Conclusions and Future Work

The on-going work described in this abstract uses the Multi-Factor SLR-LM to analyze phonetic regularities and exceptions in American English. Qualitatively good continuous representations of phones are learned, and the global sparse weight matrix accurately identifies phone n-grams that often function as a single unit, but dialect-specific exceptions capture only a limited amount of true dialectal pronunciation variation. There are many ways that we plan to extend this work. More should be done to normalize the input data to reduce the effect of word choice on the learned weights. One could also use a Multi-Factor SLR-LM to jointly model a more diverse and overlapping set of influences on pronunciation: with appropriately annotated data, the model could capture exceptions due to age, gender, accent, education level, etc. One could even train multi-lingual models to jointly model phonotactic similarities and differences between languages. Finally, the robustness of the model to the high noise found in automatic phonetic transcripts should be explored.

References

Fadi Biadsy. 2011. *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*. Ph.D. thesis, Columbia.

Cynthia G. Clopper and David B. Pisoni. 2007. Free classification of regional dialects of American English. *Journal of Phonetics*, 35(3):421–438.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. 1993. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM.

B. Hutchinson, M. Ostendorf, and M. Fazel. 2012. A sparse plus low rank maximum entropy language model. In *Proc. Interspeech*.

B. Hutchinson, M. Ostendorf, and M. Fazel. 2013. Exceptions in language as learned by the multi-factor sparse plus low-rank language model. In *Proc. ICASSP*, pages 8580–8584.