EXTENDING A NORTH AMERICAN ENGLISH CATEGORY LEARNER TO A NON-STANDARD VARIETY: CATEGORIZING VOWELS ACROSS SPEECH STYLES IN GLASWEGIAN ENGLISH

Shannon Mooney

Georgetown University
sm2842@georgetown.edu

ABSTRACT

Despite much research on the performance of distributional category learning models on Standard North American English (e.g., Feldman [2], deBoer and Kuhl [1], McMurray et al. [10], Vallabha et al. [16], and many others), statistical learning of vowel categories of other regional varieties remains vastly underaddressed in computational literature. This paper applies an unsupervised infinite mixture model (as developed in Feldman [2]) to vowels from a corpus of Glaswegian English sociolinguistic interviews. While originally developed for North American English vowels in carrier syllables devised by Hillenbrand et al. [4] to limit variation due to phonetic context, the distributional learner was also able to categorize vowels largely correctly across speech styles common to sociolinguistic interviews. This displays the ability of the distributional learner to operate relatively well on data with extensive overlap from running Glaswegian English speech, demonstrating that computational models of category acquisition can handle more complex inputs than minimal pair lists, and can be used with naturally-occurring speech from non-standard regional varieties.

Keywords: machine category learner, vowels, Scottish English, sociophonetics

1. INTRODUCTION

The distributional vowel category learner introduced in Feldman [2] that relied only on first and second formant information performed somewhat well on variable data coming from multiple speakers. Although later models included in Feldman [2] improved upon the initial model by also adding lexical information and computing the distributional configuration of phonemes based on their lexical constraints, in this paper the initial distributional model is used on even more variable data, showing that a distributional learner can also be applied to naturally occurring speech with some success. As Feldman [2] and others such as McMurray et al. [10] and Vallabha et al. [16] point out, distributional category learners typically only work well when there is a separation in formant space between vowel categories. The high overlap between vowel categories that occurs in realistic speech is not easily disambiguated by a category learner model, but for children acquiring language it has been shown to not be as difficult and to begin quite early in the acquisition process [6], despite that children have been shown to use a statistical learning process for auditory tones similar to the one used by most computational learners, including the distributional model under consideration in the present paper [13].

The training corpus used by Feldman [2] for the distributional learner takes its phonetic category parameters from the Hillenbrand et al. [4] study of American English vowels. One simulation is run with data from 45 male speakers, and another simulation is run with formant data from both the male and female speakers, totaling 139 individuals. The formant data that makes up the training corpus is not raw data, but is a random sampling of the Gaussian distribution of the means and covariances from the phonetic data for each vowel category, based on the frequency of that vowel category in parental speech in CHILDES [9]. The means and covariances of only male vowel formant values from Hillenbrand et al. [4] are sampled to create one 20,000 token corpus, and a second 20,000 token corpus is created from the means and covariances of male and female formant values combined. Feldman's [2] idea behind this is to create one corpus of less variable data (males only) and one corpus of more variable data (males and females), with the hypothesis that the distributional learner will perform better on the less variable corpus, with an encoded bias towards fewer phonetic categories and towards the center of the vowel space.

Although Feldman [2] considers the vowel data from Hillenbrand et al. [4] to be highly variable and representative of a realistic input that must be categorized in language acquisition, this data is only conservatively variable relative to naturally occurring vowel data such as is found in the sociolinguistic interview setting. The formant data in Hillenbrand et al. [4] come from vowels produced in h-V-d syllables in isolation, which eliminates variability due to stress and prosodic patterning, allophony and other influences of surrounding context, and sociolinguistic factors such as style and register. The vowel data in the current paper are thus considerably more variable than the data originally used, as the current data come from sociolinguistic interviews containing three separate styles (word list, reading passage, and conversation sections) with six speakers of Glaswegian English.

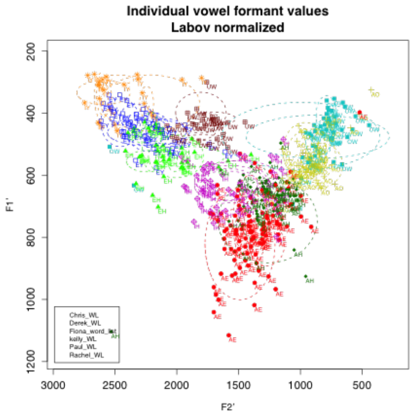
2. DATA

The input to the model is normalized vowel formant data from six sociolinguistic interviews recorded in a laboratory setting in Glasgow, Scotland in 2009. Hour-long interviews were conducted with three female and three male Glaswegians of working-class backgrounds aged 18 to 25. F1 and F2 measurements from the primarily and secondarily stressed vowels with duration over 50ms, excluding pre-rhotic vowels, diphthongs, and vowels in lexical items that tend to reduce (e.g., function words), were gathered through forced alignment and extraction with FAVE [12] set to use the *British English Example Pronunciation* dictionary [11], as recommended by Mackenzie and Turton [8]. In word list and reading passage styles, formant measurements were obtained manually. Two corpora of 20,000 F1–F2 pairs, following the methodology of Feldman [2], were designed: one with combined word list and reading passage data with phoneme frequencies adjusted to match token probabilities from the Unisyn lexicon [3] set to Edinburgh post-lexical rules, and one with unadjusted conversational data. The first simulation in this paper is run on word list data from all six of the interviews, and the second simulation is run on conversation style data from one interview. In addition, normalized formant values for each category are used in lieu of a random sampling of a Gaussian distribution, providing even more variability to the dataset through the existence of outliers. Despite these new challenges to the distributional learner, it is able to parse close to the correct number of vowel categories in each dataset, with an only somewhat diminished F-score.

3. SIMULATION WITH TRAINING SET 1: WORD LIST VOWELS

The formant values for this training set were taken from a word list of 80 tokens stratified by vowel and from six speakers, three males and three females. All speakers were native Glaswegian speakers of Scottish English of working-class background between the ages of 18 and 25 to control for sociolinguistic variation in Scottish English by age, class, and region. Diphthongs and schwa were excluded, although pre-rhotic vowels were included, as vowels in this context maintain contrasts and are largely unreduced in Scottish English. Figure 1 shows the formant values of the word list tokens across all speakers, normalized using the Labov (ANAE) method [7].

Figure 1: Normalized formant values in the current dataset.



Despite the careful style promoted by the wordlist task, there is still considerable category overlap among these tokens of vowels (though the dataset from Hillenbrand et al. [4] also shows overlap between vowel categories). Following the methodology of Feldman [2], a corpus was created from the tokens of vowels based on phonemic probability in CHILDES (modified for Scottish vowel categories). This resulted in a corpus of 735 tokens. Table 1 shows the token probability by vowel.

The IMM (infinite mixture model) distributional learner was run on two dimensions (first and second formants) of the 735 vowels from the word list corpus, adjusted for phoneme probabilities. There was a prior on the phonetic category parameters for category means towards the center of the vowel space—of 500 on the first dimension and 1500 on the second dimension, and a prior to bias the model towards fewer phonetic categories. A factor which

Table 1: Empirical probability of Scottish English vowels by word token (adapted from CHILDES Phonematized Parental Frequency Count.)

Vowel	Word token probability
a	.08
ɔ	.16
ɛ	.07
e	.04
ɪ	.18
i	.08
o	.06
ʉ	.12
ʌ	.18

differentiates the Scottish English data used in this study from the American English data used in the previous work is the number of vowel categories. While Feldman [2] based the success of her distributional learner on its ability to recover all 11 categories of American English vowels (once schwa and diphthongs /aɪ/ and /aʊ/ are removed), Scottish English contains 9 vowels not considering schwa and diphthongs. The distributional learner overcategorized this data after 10,000 iterations— into 10 separate vowel categories, rather than the 9 expected, while Feldman found that the learner undercategorized her data– finding 10 of 11 categories. The F-score of the distributional learner on this dataset was .32, less than the .45 F-score found by Feldman [2] for the distributional learner on the corpus of male and female tokens, but considering the increase in variation within categories in this dataset due to the vowel formant values being extracted from real words rather than from the artificial context of h–V–d used in Hillenbrand et al. [4], this reduction in the F-score, representing somewhat worse performance by the distributional learner, was expected.

4. SIMULATION WITH TRAINING SET 2: CONVERSATION

The same IMM distributional learning model was then run on vowel data from the conversational section of a sociolinguistic interview. Future work will expand this to the conversational sections of all six interviews in this corpus. With a hypothesis that this further shift towards naturalistic speech will introduce even more variability among vowel categories, tokens from only one speaker were used to create the corpus in this dataset. Another dimension of variability in this dataset is that while the word list vowel tokens were extracted by hand with the first and second formants measured at the vowel midpoint, in the conversational data, due to time restrictions, FAVE, a forced alignment and extraction tool, [12] was employed. Because the length of the naturally-occuring speech from which the vowels were extracted was robust enough ($\sim$ 30 minutes) to approximately represent phonemic probability distributions, the training corpus was created simply from the non-normalized vowel tokens— they were not manipulated to replicate the phonemic probability for the word list training corpus as shown in Table 1 in the previous section. Reduced vowels, such as those that occur in function words, were removed, as well as vowels in unstressed position in words, diphthongs, and vowels less than 5ms in duration. The resulting vowel corpus consisted of the first and second formants of 2,182 vowels. Figures 4 and 5

Figure 2: Normalized vowel category means from Scottish English wordlist data.

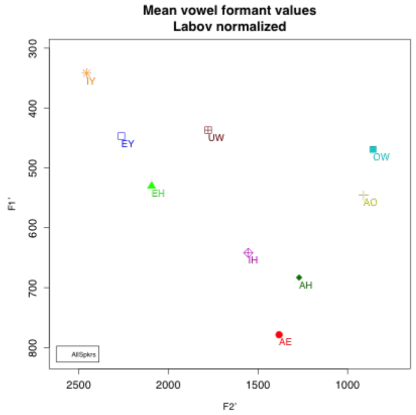
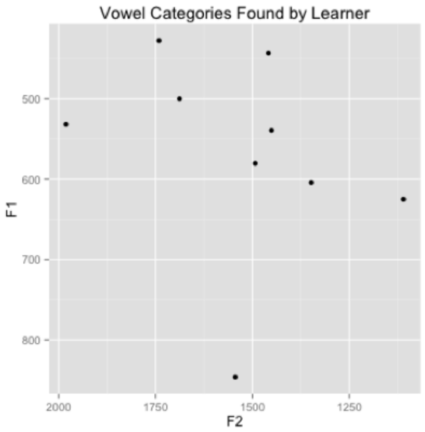


Figure 3: Categories found by distributional learner.



show the overall distribution of vowel tokens and the categories found by the distributional learner. The vowel learner recovered 6 vowel categories out of the 9 actual vowels present in the conversation, for an F-score of .47, higher than both the F-score from Feldman [2] for combined male and female phonetic parameters and the F-score found for the wordlist training data from males and females in the previous section of the paper.

Figure 4: All vowel tokens from conversation colored by actual vowel.

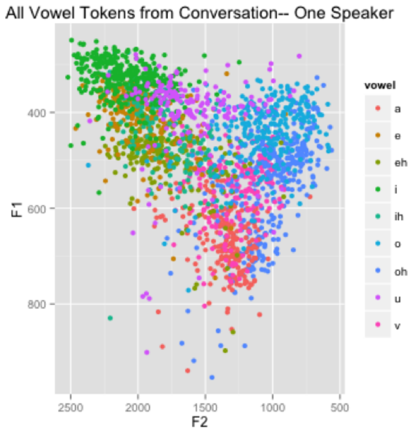
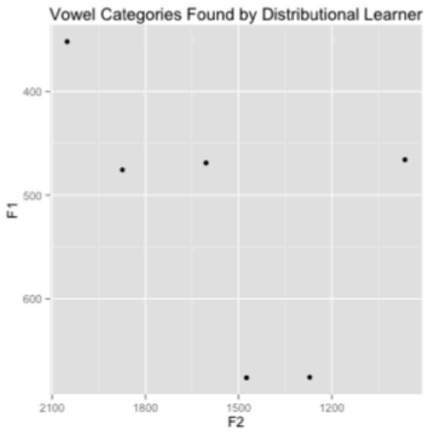


Figure 5: Vowel categories learned from conversation.



5. CONCLUSIONS

This work points towards some interesting future directions for the possibility of using statistical vowel learners on more realistic speech data. The learner recovered 6 out of 9 vowel categories in conversational running speech with a relatively high F-score of .47 for the conversational speech training dataset, which was comprised of raw, non-normalized formant frequencies that had not been randomly sampled by mean and covariance of a Gaussian distribution, but occurred just as they did in the conversational section of the sociolinguistic interview, including outliers. It outperformed both the distributional model applied to the word list training corpus and Feldman's [2] original reported F-score of .45 when the distributional learner was applied to a training corpus based on a combination of male and female phonetic parameters (Feldman's [2] training corpus made of just male vowel tokens, though, had a higher F-score of .699 using just the distributional model). The lower F-score for the wordlist training set, and the higher number of categories found, exceeding even the actual number of vowel categories in Scottish English, may have been due to the effect of surrounding phonetic context on the vowels of the relatively short word list.

In future work, it will be important to create a training corpus from a random sampling of the Gaussian distribution of formant values for each vowel category of the word list, rather than using all of the normalized frequencies, which may include allomorphic effects. The finding that the distributional learner was able to recognize six distinct vowel categories in the conversational style corpus of one speaker is important, as it indicates that just based on thirty minutes of non-normalized, naturally occurring speech, with no lexical effects, using statistical learning, language acquirers can begin to segment the distribution of vowel categories in their language somewhat successfully. This F-score will certainly improve with the addition of lexical information using one of Feldman's [2] lexical-distributional models, but these are unexpectedly successful results for the distributional learner alone, applied to variable, naturally-occurring formant data. Future directions for this work will also be to include normalized conversational-style vowel tokens from all six sociolinguistic interviews in a training corpus for the distributional and lexical-distributional learners, and perhaps to include an F3 dimension to attempt to disambiguate the Scottish English fronted /ʉ/ from the other, non-round front vowels, as it most likely does not differ from ɪ very much in F1 or F2. Although Feldman [2] ultimately improves her original distributional learner to account for lexical constraints, this paper shows how the IMM distributional learner, run on only two formant frequency dimensions, can discover a number of vowel categories in spite of the extensive overlap involved in conversational-style speech.

6. REFERENCES

[1] de Boer, B., Kuhl, P. K. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online* 4(4), 129–134.

[2] Feldman, N. H. 2013. A role for the developing lexicon in phonetic category acquisition. *Psychological Review* 120(4), 751–778.

[3] Fitt, S., Isard, S. 1999. Synthesis of regional english using a keyword lexicon. *Proceedings: Eurospeech 99* 2, 823–826.

[4] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. 1995. Acoustic characteristics of american english vowels. *Journal of the Acoustical Society of America* 97, 3099–3111.

[5] Kendall, T., Thomas, E. R. 2010. Vowels: Vowel manipulation, normalization, and plotting in r. R package, version 1.1http://ncslaap.lib.ncsu.edu/tools/norm/.

[6] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255(5044), 606–608.

[7] Labov, W., Ash, S., Boberg, C. 2005. *The atlas of North American English: Phonetics, phonology and sound change.* Walter de Gruyter.

[8] MacKenzie, L., Turton, D. 2013. Crossing the pond: Extending automatic alignment techniques to british english dialect data. *New Ways of Analyzing Variation 42*.

[9] MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates 3 edition.

[10] McMurray, B., Aslin, R. N., Toscano, J. C. 2009. Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science* 12, 369–378.

[11] Robinson, A. J. 1997. British english example pronunciation (beep). Available from ftp://svrftp.eng.cam.ac.uk/pub/comp.speech/dictionaries.

[12] Rosenfelder, I., Fruehwald, J., Evanini, K., Yuan, J. 2011. Forced alignment and vowel extraction (fave) program suite. Available from http://fave.ling.upenn.edu.

[13] Saffran, J. R., Johnson, E. K., Aslin, R. N., Newport, E. L. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition* 70(1), 27–52.

[14] Scobbie, J. M., Gordeeva, O. B., Matthews, B. 2006. Acquisition of scottish english phonology: an overview. *QMU Speech Science Research Centre Working Papers*.

[15] Team, R. C. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[16] Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., Amano, S. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America* 104, 13273–13278.

[17] Wells, J. C. 1982. *Accents of English* volume 1. Cambridge: Cambridge University Press.

[18] Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Spring New York.

[19] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. Elan: a professional framework for multimodality research. *Proceedings of LREC 2006 Fifth International Conference on Language Resources and Evaluation*.