



Can Conversational Word Usage be Used to Predict Speaker Demographics?

Dan Gillick

University of California, Berkeley
 Computer Science Division
 dgillick@cs.berkeley.edu

Abstract

This work surveys the potential for predicting demographic traits of individual speakers (gender, age, education level, ethnicity, and geographic region) using only word usage features derived from the output of a speech recognition system on conversational American English. Significant differences in word usage patterns among the different classes allow for reasonably high classification accuracy (60%-82%), even without extensive training data.

Index Terms: demographics, speech recognition, classification

1. Introduction

The Mixer corpora [1, 2] are primarily used in development and evaluation of speaker identification systems [3], where the goal is to model characteristics of individuals. But because these data contain demographic attributes of each participant, it is also possible to model characteristics of groups. Table 1 shows summary statistics for the demographic data collected as part of Mixer.

While acoustic cues may well be relevant for distinguishing between demographic groups, this work is concerned with classification based on simple word n-gram features derived from Automatic Speech Recognition (ASR) output. We find that word usage varies significantly between groups, providing surprisingly accurate classification for each demographic characteristic we studied: gender, age, education level, ethnicity, and geographic region. These differences are interesting from a theoretical perspective, and could have a variety of commercial applications. This work is intended as an introduction, a survey of the potential for demographic prediction, so the focus is on analysis rather than complex features or structured classification.

The paper is organized as follows: Section 2 outlines related work; Section 3 describes the Mixer data in more detail, addressing relationships between the demographic features; Section 4 describes feature selection and analyzes linguistic differences between classes; Section 5 shows classification results; Section 6 discusses future work.

2. Related Work

Inspired by Mosteller and Wallace’s classic work [4] on inferring authorship of the Federalist Papers, Doddington [5] showed that a relatively small set of word bigram features could help distinguish between speakers with surprising accuracy. This approach has proven especially useful in combination with acoustic models because the two information sources are fairly independent. Further, reasonable performance is maintained even given highly errorful ASR transcripts [6].

Trait	Summary Statistics
Gender	female (844); male (492)
Yr. of Birth	1922 - 1990; median=1974
Yrs. of Education	1 - 30; median=16
Native Language	U.S. English (765); 33 others
Occupation	Student (230); Homemaker (41) ...
Country Raised	U.S. (947); India (58); China (41) ...
State Raised	40 U.S. states represented
Ethnicity	White (457); Asian (388); Black (129) ...
Smoker	yes (142)
Height (cm)	134 - 198; median=168
Weight (kg)	36 - 160; median=68

Table 1: Demographic statistics for the segment of the Mixer data used in the NIST 2008 Speaker Recognition Evaluation (1336 total speakers). Bold features are analyzed in this work.

A second strand of related work involves predicting speaker age and gender using acoustic characteristics like pitch, jitter, shimmer, and spectral energy, along with behavioral features like speaking rate [7]. Müller’s Ph.D. research [8] involved the development of systems to improve mobile shopping and call-center dialogs using this sort of classification. Along similar lines, some linguistic research has focused on classifying dialects [9], a kind of regional inference problem.

In the text domain, Koppel et. al. [10] predict author gender in published fiction and non-fiction, and Schler et. al. [11] predict author gender and age in blog writing. For blogs, which tend to have the same sort of informal tone as spoken conversations, both gender and age classification (three classes: 13-17, 23-27, 33-42) give close to 80% accuracy¹.

Lastly, there is growing interest in automatic demographic profiling from other data sources. Cell-phone call logs, for example, may prove useful for classifying gender or socioeconomic status [12].

3. Data

For our experiments, we chose a subset of the 2008 NIST Speaker Recognition Evaluation data set (in turn, a subset of Mixer) containing 538 native speakers of American English. Each speaker participated in at least one phone conversation², and 5.6 conversations on average (standard deviation = 3.7): an average of 2500 words per speaker (standard deviation = 1800). ASR transcripts were generated using the SRI Decipher system [13], and give about 23% word error rate on our data.

¹Common word unigram features; average of 7000 words per author; 35000 training authors.

²Conversation partners were chosen randomly, given a suggested topic for informal discussion, and calls lasted at most 10 minutes.

10.21437/Interspeech.2010-421

Gender		Age		Education		Ethnicity		Region	
Male	39%	20-29	31%	High school or less	18%	White/Caucasian	63%	Northeast	40%
Female	61%	30-39	27%	College	52%	Hispanic/Latino	6%	South	21%
		40-49	20%	More than college	30%	Black/African-American	19%	Midwest	16%
		50+	22%			Asian	12%	West	23%

Table 2: Classes derived from demographic data.

The demographic data is self-reported and some values are missing, but there is little reason to suspect its veracity. We chose five demographic traits to study and sub-divided them into classes as shown in Table 2. For age, education, and region, the classes represent intuitive, rather than data-driven, groupings; the regional classes represent the official U.S. regional mapping used by the Census Bureau.

Before diving into classification experiments, it is important to know how these demographic traits are related in the data. If, for example, age and education are highly correlated, it will be hard to tell whether the predictive education features are meaningful, or if they are just useful for predicting age. We used logistic regression models³ to quantify the relationship between these variables. Table 3 shows McFadden’s R^2 for each model, a statistic that attempts to generalize the standard notion of the fraction of total variance explained in a traditional regression model as:

$$R_{McFadden}^2 = 1 - \frac{\log \hat{L}(M_{full})}{\log \hat{L}(M_{intercept})} \quad (1)$$

This ratio of the model likelihood to the intercept-only model likelihood has the right quantitative properties: the value is between zero and one, and larger values imply better models. While all the values in Table 3 are relatively small, suggesting minimal correlations between the demographic traits, it is important to have some other point of reference. For comparison, we note that McFadden’s R^2 for a model predicting age from a random subset of 20 word bigram features is approximately 0.14. Further, the predictive power of the demographic-only models are unimpressive: Predicting age, for example, from the four other demographics yields only a nominal (2%) improvement over the majority class baseline.

	Gender	Age	Educ	Eth	Region	ALL
Gender	–	0.006	0.006	0.008	0.013	0.036
Age	0.003	–	0.021	0.035	0.022	0.069
Educ	0.004	0.023	–	0.034	0.013	0.069
Eth	0.005	0.046	0.033	–	0.049	0.116
Region	0.007	0.023	0.005	0.032	–	0.061

Table 3: McFadden’s R^2 values for each model. Each trait (row label) is modeled using each other trait (column label) as a single predictor, and as a weighted sum of all the other traits together (the ALL column).

The sign and scale of the regression coefficients allow us to draw the following conclusions about this particular demographic data set:

1. Age and education level are negatively correlated: Older subjects tend to be less well educated than younger subjects.

2. The Asian subjects tend to be younger and from the Western region, while the the Black/African-American subjects tend to be older and from the Southern region.
3. Black/African-American and Hispanic/Latino subjects tend to have less education; Asian subjects have more education (this latter effect is considerably stronger).

4. Feature Selection

For classification, we employed a simple bag-of-ngrams representation, and found that bigrams gave slightly better performance than unigrams. That is, each speaker is represented by a vector indicating the presence or absence of each bigram feature in their full conversation history. Experiments with feature representation suggested that (1) normalizing the feature vector to have unit length does not improve performance, and (2) using normalized or unnormalized bigram counts or log-counts (as opposed to presence/absence) degrades performance.

Since many bigrams are either very rare or non-informative, we selected the top 2000 bigrams according to Information Gain for each demographic trait. Equation (2) gives the Information Gain formula for a feature (bigram) f as the difference between the prior (class) Entropy and the Entropy of the posterior (class conditional on the feature):

$$IG_f = - \sum_c P(c) \log P(c) + P(f) \sum_c P(c|f) \log P(c|f) + P(\bar{f}) \sum_c P(c|\bar{f}) \log P(c|\bar{f}) \quad (2)$$

Table 4 shows some of the most discriminative features for each class. A few observations:

1. Most of the important gender features are bigrams used more frequently by women (and she, because I’m, my daughter); almost all the distinctly male features include “uh”.
2. Subjects in their twenties are characterized primarily by the use of the word “like”; older subjects tend to distinguish themselves by comments about lifestyle (children, grandchildren).
3. Subjects with less education tend to use more pronouns, especially plural (we, them, us); those with more education use “I” more, and speak abstractly or reflectively more often (it’s interesting, find that, that point).
4. Ethnic and regional differences are hard to summarize. Some regional features are geographic (in Philadelphia, Bay Area), but many are not.

³Using the mlogit package in R

Class	Bigrams	
Female	my husband	oh my
Male	uh that's	uh uh
20-29	and like	cool [laugh]
30-39	it definitely	was living
40-49	excuse me	for president
50+	many years	kids that
H.S. or less	not into	yes uhuh
College	I'm working	lot about
Grad. School	of other	yeah or
White/Caucasian	in touch	watch the
Hispanic/Latino	um more	my thing
Black/African-American	may have	them up
Asian	mm yeah	oh like
Northeast	gone to	was no
South	it's [laugh]	different in
Midwest	choose to	opposed to
West	cool and	about your

Table 4: Some sample high Information Gain bigrams are shown for each class.

5. Classification Experiments

We trained a discriminative multiclass classifier for each demographic trait. Since the data set is fairly small, we used a form of bootstrapping to get a more reliable estimate of the classification error rate [14]: The data set is randomly permuted and split into training (400 speakers) and test (138 speakers); this process is repeated 50 times and the test error rates reported in Table 5 are averaged over all 50 iterations.

The classifier is our own implementation of the Margin In-fused Relaxed Algorithm (MIRA) [15], an online learner that, like the perceptron algorithm, updates a set of weights for the correct class and the predicted class when it makes mistakes. MIRA is “ultra-conservative” in the sense that it updates weights just enough that subsequent classification of the same example would yield the correct class. The resulting decision boundary is an approximation of the large-margin boundary learned exactly by a Support Vector Machine⁴. Our training procedure iterates through the training data 5 times, and takes the average of the weights after each iteration as the final weight set (the averaging tends to ameliorate over-training, a common problem for perceptron-like algorithms). Training with MIRA is fast—less than a second of computation per model—which makes this algorithm a good choice for our bootstrapped experiments. Other discriminative classifiers, like multiclass logistic regression, give comparable results but take longer to train; Naive Bayes yields poorer performance.

	Baseline	MIRA	% Improvement
Gender	39%	18%	54%
Age	66%	35%	47%
Education	49%	33%	33%
Ethnicity	38%	28%	26%
Region	62%	40%	35%

Table 5: Baseline error rates (always predict the majority class), classifier error rates, and relative improvements are shown for each demographic trait.

While the overall error rate is a single number, useful for

⁴MIRA also has no slack variables.

comparison, the distribution and types of common errors are more interesting. Table 6 shows normalized confusion matrices: the column labels represent the true classes and the row labels represent the predicted classes. Thus each column must sum to 100%, the diagonal values are correct predictions and off-diagonals are errors categorized by type.

	Actual Gender	
	Male	Female
Male	85%	9%
Female	15%	91%

	Actual Age			
	20-29	30-39	40-49	50+
20-29	81%	25%	13%	10%
30-39	11%	52%	17%	9%
40-49	4%	11%	49%	9%
50+	4%	12%	21%	72%

	Actual Education Level		
	H.S. or less	College	Grad. School
H.S. or less	56%	5%	6%
College	35%	79%	41%
Grad. School	9%	16%	53%

	Actual Ethnicity			
	White	Hispanic	Black	Asian
White	88%	47%	29%	42%
Hispanic	1%	14%	1%	1%
Black	5%	28%	61%	22%
Asian	6%	11%	9%	35%

	Actual Region			
	Northeast	South	Midwest	West
Northeast	70%	17%	26%	22%
South	10%	56%	20%	9%
Midwest	9%	14%	38%	7%
West	11%	14%	16%	62%

Table 6: Normalized confusion matrices for each demographic trait.

In general, the classifier tends to over-predict the majority classes (the 20-29 age group or the White/Caucasian ethnic group, for example) and under-predict minority classes (Hispanic/Latinos or Midwesterners). This is sensible for this data set, as it corresponds with the prior probabilities of the classes, but may not be desirable if new data does not match this underlying distribution. Still the results are fairly impressive, especially in comparison to a similar classification task using blog data [11] which achieved comparable results for age and gender prediction from considerably more training data.

Overall classification error rate disregards an underlying continuity in the classes. Mis-classifying a 29-year-old as belonging in the 30-39 class is less of a mistake than mis-classifying a 20-year-old in that class; similarly, confusing adjacent age groups is less troublesome than mistaking someone in the 20-29 range for someone in the 50+ range. Thus the confusion matrix shows fairly impressive performance for the age and education classifiers: most of the errors are in adjacent categories. Only 6% of subjects who have had at least some graduate school are classified as having 12 or fewer years of ed-

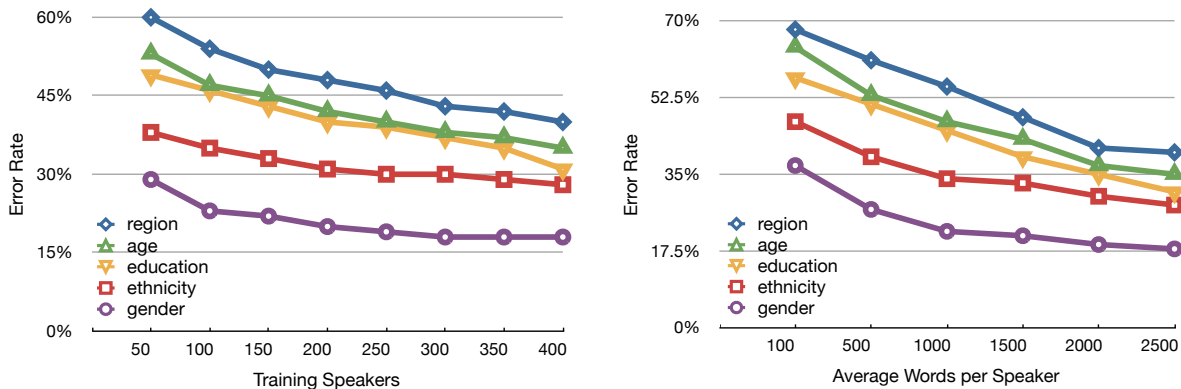


Figure 1: The demographic models are data-starved: Classification error rate continues to decline as the number of training speakers increases (left) and the average number of words per speaker, both for training and test, increases (right).

ucation; 4% of subjects in the 20-29 age range are classified as belonging to the 50+ range.

Lastly, adjusting the amount of data the classifier can use for training, both in terms of the number of speakers and the number of words per speaker (see Figure 1), makes clear that the models are data-starved. Error rates in every demographic category, with the possible exception of Gender, would almost certainly decline given more data. Increasing the number of training speakers to 800, for example, could well cut error rates by 3%-10% (absolute). We might also expect to gain by increasing the size of our feature set as the training set grows: given more data, the classifier should be able to learn meaningful weights for less common features.

6. Conclusion

We have provided an introduction to classifying demographic traits using word-based features in the Mixer corpus. Experimental results imply that intuitive demographic groups have quantifiably different word usage patterns that can be used to predict group membership with surprising accuracy. Of course, the groups we chose by hand, like age groupings and regional boundaries, may not yield the most accurate classification. Future work could involve identifying age groupings and contiguous geographic regions that produce the best classification results—a kind of linguistic model-based redistricting.

Our features are derived from ASR performed on informal conversations. It would be interesting to compare results from manual transcriptions, and from a different conversational style. While manual transcriptions do not exist for all the Mixer data, there is a set of interviews that might reveal quite different patterns of word usage. Lastly, combining word-based predictions with acoustic-based predictions is likely to give considerable improvement.

7. References

- [1] C. Cieri, W. Andrews, J. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki *et al.*, “The mixer and transcript reading corpora: Resources for multilingual, crosschannel speaker recognition research,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 117–120.
- [2] C. Cieri, L. Corson, D. Graff, and K. Walker, “Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora,” in *Proceedings of Interspeech*, 2007.
- [3] M. Przybocki, A. Martin, and A. Le, “NIST speaker recognition evaluations utilizing the mixer corpora, 2004, 2005, 2006,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.
- [4] F. Mosteller and D. Wallace, “Inference in an authorship problem,” *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, 1963.
- [5] G. Doddington *et al.*, “Speaker recognition based on idiolectal differences between speakers,” in *Seventh European Conference on Speech Communication and Technology*, 2001.
- [6] D. Reynolds, J. Campbell, W. Campbell, R. Dunn, T. Gleason, D. Jones, T. Quatieri, C. Quillen, D. Sturim, and P. Torres-Carrasquillo, “Beyond cepstra: exploiting high-level information in speaker recognition,” in *Proceedings of the Workshop on Multimodal User Authentication*, 2003, pp. 223–229.
- [7] C. Müller, “Automatic Recognition of Speakers’ Age and Gender on the Basis of Empirical Studies,” in *Ninth International Conference on Spoken Language Processing*. ISCA, 2006.
- [8] C. Müller, “Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender],” Ph.D. dissertation, Computer Science Institute, University of the Saarland, Germany, 2005.
- [9] D. Miller and J. Trischitta, “Statistical dialect classification based on mean phonetic features,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4, 1996.
- [10] M. Koppel, S. Argamon, and A. Shimoni, “Automatically categorizing written texts by author gender,” *Literary and Linguistic Computing*, vol. 17, no. 4, p. 401, 2002.
- [11] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, “Effects of age and gender on blogging,” in *2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.
- [12] J. Blumenstock, D. Gillick, and N. Eagle, “Who’s Calling? Demographics of Mobile Phone Use in Rwanda,” in *Proceedings of the AAAI Artificial Intelligence for Development Symposium*, 2010.
- [13] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sonmez, F. Weng *et al.*, “The SRI March 2000 Hub-5 conversational speech transcription system,” in *Proc. NIST Speech Transcription Workshop*, 2000.
- [14] B. Efron, “Bootstrap methods: another look at the jackknife,” *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
- [15] K. Crammer and Y. Singer, “Ultraconservative online algorithms for multiclass problems,” in *Computational Learning Theory*, 2003, pp. 99–115.