Automatic Detection of Speaker Attributes Based on Utterance Text

Wen Wang, Andreas Kathol, Harry Bratt

Speech Technology and Research Lab
SRI International, Menlo Park, CA, USA
{wwang,kathol,harry}@.speech.sri.com

Abstract

In this paper, we present models for detecting various attributes of a speaker based on uttered text alone. These attributes include whether the speaker is speaking his/her native language, the speaker's age and gender, and the regional information reported by the speakers. We explore various lexical features as well as features inspired by Linguistic Inquiry and Word Count and Dictionary of Affect in Language. Overall, results suggest that when audio data is not available, by exploring effective feature sets only from uttered text and system combinations of multiple classification algorithms, we can build high quality statistical models to detect these attributes of speakers, comparable to systems that can exploit the audio data.

Index Terms: speaker attributes, machine learning, nativeness, gender, age, region

1. Introduction

Identifying speaker attributes, including whether the speaker is speaking his/her native language, and gender, age, and region of the speaker, could be useful for improving the performance of speech and speaker recognition systems. For example, nonnative speakers typically degrade performance of speech recognition due to mismatch to (largely native) speakers in the training data. Hence, detecting speaker attributes would enable adaptation techniques to reduce mismatch between training and test data. Also, identifying these speaker attributes leads to important applications in their own rights, such as improved user modeling and automated customer service, and intelligence.

Sociolinguistic studies have observed that language varies across social groups and regional contexts. In natural human-human conversations in the format of conversational speech and text (e.g., chat, messaging), subject matter experts have observed difference between younger and older, male and female participants on a variety of lexical cues such as the average sentence length, grammaticality, structural complexity, and use of words in special categories. [1] investigated word ngram based features in Support Vector Machines (SVM) for detecting nonnative speakers and demonstrated that lexical variations are quite effective for predicting whether the speaker is speaking his/her native language. Theoretical interest derives from the question of how speakers' native (first) language (L1) influences a second language in which they are not native (L2). Following [1]'s work, we investigate the binary classification problem of native/nonnative identification, instead of the much more challenging task of detecting L1 of a speaker. Speaker gender detection is straightforward as a binary classification problem of female/male identification. For age detection, we follow the traditional sociolinguistic studies and segment age values into several distinctive, non-overlapping age groups and hence investigate the multi-class classification problem.

Exploring the relation between language and geography region has been of interest to the research community over the decades. The essence of earlier approaches is identifying variable pairs and measuring their frequencies. Different from the traditional approaches where the relevant linguistic variables have already been identified, [2] explored a multi-level generative model for identifying geographical locations of authors of web text. They explored models working directly from raw text, that jointly identifies words with high regional affinity, geographically-coherent linguistic regions, and the relationship between regional and topic variation.

Our work differs from [1] in that we aim at developing systems for predicting attributes of not only speakers in conversational speech, but also a broader range of conditions, including attributes of players in a virtual world, authors of web text (blogs, twitters etc). Hence, we don't assume the availability of audio data which can provide cepstral features (typically modeled by Gaussian mixture models or hidden Markov models), phone strings (typically modeled by language models), and a variety of prosodic features. We focus on uttered text alone and explore a much larger space of lexical and structural features, beyond n-grams explored in [1], to capture lexical variations between speakers with different attributes and measuring contributions of individual features. Our work also differs from [2] in that we exploit hypotheses from sociolinguistics and use them to guide generating the large feature space. Hence, our work is to investigate contributions of individual features in a large feature set already defined, instead of jointly identifying this set of features and predicting regions.

The rest of the paper is organized as follows. Section 2 describes the data we used for this study. Section 3 provides detailed illustrations of the features as well as the models. Experimental results and discussions appear in Section 4, and conclusions are summarized in Section 5.

2. Data

Many studies in identifying speaker attributes are based on small, locally collected data set. Similar to the work in [1], we used a subset drawn from the Linguistic Data Consortium (LDC) Fisher phase I corpus, identified as SRI-FSH. The LDC Fisher Phase 1 corpus contains a large number of speakers, a small, but significant portion of which are nonnative speakers. All conversations are conducted in English. All speakers who did not declare English as their native language according to LDC's corpus documentation on speaker information are included in the database as nonnatives. For natives, Shriberg et al. [1] chose a random subset of American English native speakers about equal in size to the nonnative set. The final selection comprised 749 nonnative speakers and 741 native speakers. For each speaker an average of 1.9 conversation sides were avail-

able, resulting in 1,451 conversation sides from native speakers and 1,431 conversation sides from nonnative speakers and 2,882 conversation sides in total. Note that Fisher conversations are 10 minutes in length, yielding about 5 minutes of speech per conversation side on average. Also note that all speaker information, including being native/nonnative in English, is based on self-reported L1s in the corpus documentation from LDC. Hence, it is likely that some speakers will be labeled as nonnatives based on their L1s yet having high English proficiency.

[1] focused on native/nonnative based on American English due to availability of data and models for American English from prior work in speech and speaker recognition. But their work and the work in this paper could be conducted on other languages as well. Here is a brief review of definitions of nonnative speakers for American English in [1]:

- A speaker whose first language is any dialect of American English, and who is speaking in that dialect in the conversations studied, is considered to be a native speaker.
- Talkers whose first language is not English, are considered nonnative speakers when speaking English.

Speakers whose first language is a non-American dialect of English (e.g., British, Australian, Indian) are removed from consideration, since they typically are not trying to modify their accent when speaking English. Also, bilingual and multilingual speakers who reported American English as one of their native languages were also removed from the study [1].

For gender detection, based on self-reported speaker information, among the 2,882 conversation sides, 1,408 conversation sides are from male speakers and 1,474 are from female speakers. For age detection, prior social science studies suggest breaking age values into three age groups, namely, Youth (under 18 years old), Young Adult (18 to 24 years old), and Adult (25 years and older). Among the 2,882 conversation sides, 2,304 are from Youth, 542 are from Young Adults, and 36 are from Adults. For regional information, LDC corpus documentation provided the self-reported country and state information of speakers. To separate from the native/nonnative detection task, we consider only speakers from the states of US. The U.S. Census Bureau divides these states into four regions: West, Midwest, Northeast, and South [3]. So we formulate the region detection problem into a four-class classification problem.

3. Features and Models

3.1. Features

Similar to [1], the utterance text from speakers in this work are automatic speech recognition 1-best hypotheses. We explored lexical and structural features for speaker attribute detection. We included a set of "lexical" features, including n-grams extracted from all of that speaker's utterances, denoted *ngram* features. We also included various lexical features extracted from language use constituents (LUC) annotations, denoted *LUC* features. LUCs show a variety of linguistic phenomena, including discourse markers, disfluencies, person addresses and person mentions, prefaces, extreme case formulations, and dialog act tags (DAT). We categorized dialog acts into statement, question, backchannel, and incomplete. We classified disfluencies (DF) into filled pauses (e.g., *uh*, *um*), repetitions, corrections, and false starts. Person address (PA) terms are terms that a speaker uses to address another person. Person mentions (PM) are references to non-participants in the conversation. Discourse markers (DM) are words or phrases that are related to the structure of the discourse and express a relation between two utterances, for example, *I mean*, *you know*. Prefaces (PR) are sentence-initial lexical tokens serving functions close to discourse markers (e.g., *Well, I think that...*). Extreme case formulations (ECF) are lexical patterns emphasizing extremeness (e.g., *This is the best book I have ever read*).

We built automatic annotation tools for these LUCs. The automatic annotation tools for discourse markers, extreme case formulations, and prefaces are rule-based systems integrating an HMM-based Part-of-Speech (POS) tagger, predefined tables, and heuristic rules. Person address terms and person mentions are automatically detected using a Conditional Random Fields based model using contextual N-grams and POS tags and annotation results from a named entity tagger. The named entity tagger is an HMM model following the Nymble model [4], classifying Person, Organization, GPE (geopolitical entities and locations which are also political units, such as countries, counties, and cities), and Location. The automatic disfluency detection model was a hybrid system combining hidden-event language models, Conditional Random Fields based models, and rule-based models, for predicting fillers, repetitions, revisions, and restarts, following the approaches described in [5].

After automatic annotation of the automatic transcripts with rich LUCs, we extracted features related to the presence of prefaces, the counts of types and tokens of discourse markers, extreme case formulations, disfluencies, person addressing events, and person mentions, and the normalized values of these counts by the number of sentences of a speaker. We also include a set of features related to the dialog act tags of sentences from a speaker.

Discourse marker related features could be reflective on complexities of sentences. To explicitly model grammaticality of sentences, we adapted state-of-the-art English syntactic parsers to conversational style English text by employing semi-supervised training approaches using LDC released newswire and switchboard treebanks and the unlabeled Fisher transcripts [6]. We then applied the parser to parse transcripts and compute the likelihood scores of 1-best parses.

We also developed a set of "structural" features, inspired by conversation analysis, to quantitatively represent the different participation patterns of speakers in a conversation. Structural features for a speaker include the percentage of sentences and turns spoken by that speaker out of all sentences and turns in the show respectively, and the average number of words per sentence and per turn of that speaker.

We used a "Dictionary of Affect in Language" (DAL) [7] to derive DAL based features. The Dictionary of Affect in Language is an instrument designed to measure the emotional meaning of words and texts. It does this by comparing individual words to a list of 8,742 words that have been rated by people for their activation, evaluation, and imagery. Each word in the lexicon receives a score according to "pleasantness", "activity", and "imagery". Then we computed the average of these scores for the sentences contributed by a speaker, and used them as DAL features. We also included average counts of words belonging to each of these three categories in sentences uttered by that speaker.

In order to provide an efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples, a text analysis application called Linguistic In-

quiry and Word Count, or LIWC, was developed [8]. The selection of words defining the most recent LIWC categories involved multiple steps over several years. The initial idea was to identify a group of words that tapped basic emotional and cognitive dimensions often studied in social, health, and personality psychology. With time, the domain of word categories expanded considerably. We designed LIWC inspired features by computing the average number of occurrences of words of each LIWC category in utterances contributed by a speaker.

Besides features based on DAL and LIWC, we also extracted features based on sociolinguistic studies on some speaker attributes. For example, compared to female speakers, male speakers could use more self-affirmation lexical cues and could use more homophobia and taunting words. We defined a list of words and phrases in these categories and extracted a set of features denoting the counts of types and tokens of these words and phrases. After feature extraction, the (very sparse) feature vectors are rank-normalized [9] along each dimension.

3.2. Support Vector Machines

To compare to the work in [1], we also employed linear kernel SVMs for classification. Models of how frequently speakers use certain words or phrases (idiolect) have been proposed by [10] and, although poor models by themselves, were found to improve speaker recognition systems in combination with other knowledge sources. [1] employed a version of n-gram based features in SVMs, which they reported gave better results than language-model-based approaches.

4. Experiments

To make efficient use of all available data, we employed 10-fold cross-validation. We randomly assigned speakers to ten roughly equal size partitions. Each partition in turn was used as the test set, while the other nine partitions were used to train models. Overall results were computed by averaging the outcomes over the ten test partitions. Performance is measured as classification accuracy as the percentage of correctly labeled samples for each class of the attributes.

Table 1 shows the classification accuracies of SVMs using three sets of features, that is, using word N-gram features alone, using all the other features described in Section 3 excluding word N-gram features, and the combination of word N-gram and all the other features, for detecting nativeness, gender, age group, and region classes for the speakers. As can be seen from the table, the SVM performances from word N-gram features outperform those from all the other features exclluding word N-gram features, for all attributes. Yet the other features provided complementary discriminative information to word N-gram features and the combination of all features significantly improved SVM performance for all attributes over just using word N-gram features. Note that the single best system in [1] is the MLLR transform SVM which uses the speaker maximum likelihood linear regression (MLLR) adaptation transforms employed by a large-vocabulary automatic speech recognizer (ASR) as features [11]. This system used full word recognition hypotheses to compute MLLR transforms, and clustered Gaussians into eight phone clusters, denoted *mllr,8class*. Note that this best single system exploited the audio data. The performance from the mllr,8class system within the same 10-fold setup produced accuracy as 88.24% for native/nonnative detection.

In addition to the feature level combination, which is represented as the *All* condition of features in Table 1, we also experimented with system level combination by combining the SVM using Word N-gram features and the SVM using all the other features. The results are shown in Table 2. The system combiner is a single layer neural network (perceptron) that produces a new score from the scores output by each of the two individual SVM systems. For each test set $\mathcal{T}$ of the 10-fold setup, we further randomly split it into two roughly equal sized folds and conducted 2-fold cross-validation on it, by training the combiner on one half and testing the combiner on the other half. The final score is an average over the two folds of the certain test set $\mathcal{T}$, and then further averaged over the 10 folds. We observed that the system-level combination produced significant improvement over the feature-level combination and now the SVM performance on native/nonnative detection is 88.25%, comparable to the 88.24% accuracy from the best single system mllr,8class.

Table 1: **Classification Accuracy** (%) from **SVM** on various attributes of speakers using various feature sets. Results are from 10-fold cross-validation.

Attribute	**Features**	Accuracy (%)
nativeness	Word N-gram	82.69
	Other Features	78.35
	All	86.87
	mllr,8class [1]	88.24
gender	Word N-gram	80.49
	Other Features	77.65
	All	83.24
age	Word N-gram	84.55
	Other Features	83.42
	All	87.13
region	Word N-gram	62.74
	Other Features	56.75
	All	64.57

Table 2: **Classification Accuracy** (%) from **SVM** on various attributes of speakers using feature-level and system-level combinations. Results are from 10-fold cross-validation.

Attribute	**Combinations**	Accuracy (%)
nativeness	Feature-level	86.87
	System-level	88.25
	mllr,8class [1]	88.24
gender	Feature-level	83.24
	System-level	84.58
age	Feature level	87.13
	System-level	89.44
region	Feature-level	64.57
	System-level	67.35

Since this research also includes conducting feature analysis on contributions of individual lexical features to speaker attribute detection, we employed two feature selection approaches and investigated the joint of feature selection output. The first feature selection approach is straightforward by computing the cosine of the angle between the normal to the hyperplane and each of the feature dimensions. The idea is that the closer this

value is to zero, the more parallel the corresponding dimension is to the hyperplane and hence less discriminative that feature is. The second approach is the standard Support Vector Machine Recursive Feature Elimination (SVM-RFE) approach [12]. The frame-work of RFE consists of three steps:

1. Train the classifier;

2. Compute the ranking of all features with a certain criterion in term of their contribution to classification;

3. Remove the feature with lowest ranking. Goto step 1 until no more features.

Note that in step 3, only one feature is eliminated each time. It may be more efficient to remove several features at a time, but at expense of possible performance degradation. There are many feature selection methods besides RFE. However, in the context of SVM, RFE has been proved to be one of the most suitable feature selection methods by extensive experiments.

The most significant features, selected from all the other features excluding word N-gram features, are list below, for native/nonnative, gender, age group, and region detections.

nativeness: average number of disfluencies per sentence from the speaker, average number of discourse makers per sentence from the speaker, syntactic parse 1-best likelihood scores, percentage of sentences that are incomplete sentences from the speaker

gender: average number of disfluencies per sentence from the speaker, average number of person name addressing per sentence from the speaker, syntactic parse 1-best likelihood scores

age: average number of disfluencies per sentence from the speaker, average number of discourse makers per sentence from the speaker, syntactic parse 1-best likelihood scores

region: average number of person name addressing per sentence from the speaker, syntactic parse 1-best likelihood scores, percentage of self-affirmation sentences

It is interesting to notice that some features, including average number of disfluencies, discourse markers, person name addressing per sentence from a speaker and the syntactic parser 1-best likelihood scores, were selected as significant features for multiple attributes.

5. Conclusions

We have investigated models for detecting various attributes of a speaker based on uttered text alone. These attributes include whether the speaker is speaking his/her native language, the speaker's age and gender, and the regional information reported by the speakers. We explored various lexical features as well as features inspired by Linguistic Inquiry and Word Count and Dictionary of Affect in Language. Overall, results suggest that when audio data is not available, by exploring effective feature sets only from uttered text and system combinations of multiple classification algorithms, we can build high quality statistical models to detect these attributes of speakers, comparable to systems that can exploit the audio data.

6. Acknowledgements

We thank Luciana Ferrer for sharing the setup and results of the work in [1]. We thank Nick Taylor, Suzanne de Castell, Jennifer Jenson, Geoffrey Raymond, Kristin Precoda for their suggestions on sociolinguistic studies. This research is sponsored by the Air Force Research Laboratory. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

7. References

[1] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches", *in Proceedings IEEE Odyssey-08 Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, Jan. 2008.

[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation", *in Proceedings of EMNLP*, 2010.

[3] "Census regions and divisions of the united states", http://www.census.gov/geo/www/us_regdiv.pdf.

[4] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learning name-finder", *in In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 194–201, 1997.

[5] W. Wang, G. Tur, J. Zheng, and N. F. Ayan, "Automatic disfluency removal for improving spoken language translation", *in Proc. ICASSP*, 2010.

[6] W. Wang, "Weakly supervised training for parsing Mandarin broadcast transcripts", *in Proceedings of Interspeech*, pp. 2446–2449, Brisbane, Australia, September 2008.

[7] C. Whissell, "Whissell's dictionary of affect in language: Technical manual and user's guide", Laurentian University, http://www.hdcus.com/manuals/wdalman.pdf.

[8] "Linguistic inquiry and word count", http://www.liwc.net/liwcdescription.php.

[9] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification", *in Proc. ICASSP*, pp. 1577–1580, Las Vegas, Apr. 2008.

[10] G. Doddington, "Speaker recognition based on idiolectal differences between speakers", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, pp. 2521–2524, Aalborg, Denmark, Sep. 2001.

[11] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987–1998, Sep. 2007.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines", *Machine Learning*, vol. 46, pp. 389–422, 2002.