Speaker Identification and Geographical Region Prediction in Audio Reviews

Todd Sullivan, Ashutosh Kulkarni, Richa Bhayani

Department of Computer Science, Stanford University

todd.sullivan@cs.stanford.edu, ashutosh.kulkarni@cs.stanford.edu, rbhayani@stanford.edu

Abstract

Speaker identification and geographical region prediction is important for many tasks such as targeted advertising and personalization. In this paper, we propose using hundreds of small context-dependent Gaussian Mixture Models (GMMs) of MFCC features to predict the geographical region that a speaker currently lives in. We show a marked improvement over traditional large-Gaussian mixture model techniques that are often applied to speaker identification, dialect classification, and related tasks. In contrast to previous studies, we also use a new audio dataset of speakers giving 60 to 90 second product reviews where the products span hundreds of categories, the audio is generated from a multitude of noisy environments using various cheap webcams, and the vocabulary is unrestricted with no two speakers saying the same combination of words or sentences.

Index Terms: speaker identification, gaussian mixture models, geographical region prediction, product reviews

1. Introduction

With the advent of the read-write web, there has been a drastic increase in the amount of user-generated content present on the Internet. One emergent area of user-generated content is product reviews, which often consists of ratings, opinions, and textual descriptions. As home user bandwidth increases and video becomes more prevalent, more users are extending their product reviews to videos recorded with their webcams.

In this study we use the audio from video product reviews to identify users and predict the geographical region where each user lives. Audio from product reviews is an obvious choice for traditional speech research such as emotion recognition and emotion synthesis, but very little has been pursued outside of these fields.

Speaker identification for product reviews can help in a number of ways: it can reduce the amount of data entry for the user, help create a better speaker-independent rating model, help identify a user amongst a large number of users, and stop duplicate user profiles from being formed.

Geographical region prediction is the task of classifying the user as belonging to a particular geographical region. There have been some studies in the past related to dialect and accents, but none has been on product reviews which targets a different generation of users and includes speech from a wide range of topics.

We use GMMs for both tasks and introduce using hundreds of tiny context-dependent GMMs to improve accuracy in the geographical region prediction task. Our final aggregate classifier achieves an F-Measure of 65.6 in 10-fold cross validation averaged across the four U.S. regions of West, Midwest, Northeast, and South.

2. Related Work

Speaker identification using GMMs is a widely studied problem but no major study has been conducted on audio from product reviews. Amongst the various studies conducted, GMMs have been used as sole classifiers [1], or combined with other prosodic features and large vocabulary continuous speech recognition-based systems (LCVSR) [2]. Using prosodic features and LCVSR systems, allows one to tap the longer, more speaker specific characteristics as explained in [2].

Geographical region prediction is not a frequently studied problem and the closest analogy one can find is accent or dialect classification. There have been a number of studies on identifying and classifying foreign accents but dialects are a lesser studied problem [3].

Few studies have been conducted on how well humans perform on dialect classification tasks. Clopper and Pisoni's work on perceptual categorization [4] shows that listeners can only classify unknown speakers by dialect with 30% accuracy. [4]'s experiments involved English speakers in the U.S. Other experiments involving Dutch dialects place region of origin prediction at 60% and province prediction at 40% [5]. All experiments had each speaker speaking the same sentences.

Our work differs from previous work in that we are classifying the geographical region that a speaker currently lives in, which does not necessarily correspond to where the speaker grew up or a speaker's accent. Our audio is also from product reviews where the products span more than one hundred categories and no two speakers are speaking the same words or sentence combinations.

3. Dataset Collection

Our dataset comes from ExpoTV.com. ExpoTV.com is a video product review website. Users record themselves giving a 60 to 90 second review about a product and upload the video to the website.

We crawled approximately 200,000 pages of ExpoTV.com and extracted a subset of the users from the pages. Our dataset includes 9,073 reviews from 1,392 users. Table 1 details the information extracted for each review and user.

A review's rating is from 1 to 5. The category tree contains 237 categories and is two levels deep. There are 27 base categories including Arts, Books, Cars, Computers, Kitchen, and Sports. A review's title, description, and pro/con tags are free text. Most reviews do not contain any pros or cons.

1,762 reviews contain transcripts transcribed by humans. We used HTK [6] to train a speech recognizer and planned to use the extracted words from each review as features.

Unfortunately, the result was deemed unusable with only 12% of words correct on a held-out test set.

A user has one of the five shopping styles shown in the table. A user may have zero or more hobbies, interests, and nutshells. There are 7 hobbies, 12 interests, and 8 nutshells, each listed in the table. All users have a U.S. state. In our geographical region predictions we map each U.S. state to the regions Northeast, Midwest, South, and West as specified by the U.S. Census Bureau.

Table 1. *Review and User Information.*

Review Data	
Rating	Title
Category	Description
Product	Pros/Cons

User Data	
Shopping Style	Big Spender, Shopoholic, Speedster, Sensible One, None
Hobbies	Arts and Crafts, Blogging, Cooking, Gaming, Home Improvement, Outdoor Activities, Photography
Interests	Animals, Beauty, Books, Cars, Electronics, Fashion, Fitness, Movies and TV, Music, Sports, Technology, Travel
Nutshells	Bargain Hunter, Gear Head, Greenie, Hipster, Parent, Researcher, Stay at Home Dad, Stay at Home Mom
U.S. State	CA, MO, OH, …

4. Gaussian Mixture Models

We implemented both diagonal covariance and full covariance Gaussian mixture models for our tasks. While the full covariance mixture models slightly outperformed the diagonal models, their training time was prohibitive given our time and computational constraints. All results and descriptions below use a diagonal covariance matrix.

Our feature set for all GMMs is the standard 39 MFCCs, taken from 25 millisecond windows of speech every 10 milliseconds using HTK. We train one GMM for each class in the dataset. Due to time and computational constraints, all of our results are from training on 100 randomly selected reviews from each class, which corresponds to 1 to 1.25 million windows for each GMM.

4.1. Training

We use EM to train each GMM. We restrict the maximum number of iterations to 100 and quit once the ratio between iterations' log likelihoods for the training set falls between 0.99 and 1.01.

For a given mixture size, we start by training a mixture of size 1. This first Gaussian is initialized by setting the mean for each feature to a random number between the feature's minimum and maximum values and the variance to the difference between the feature's maximum and minimum value divided by 2π. Once EM is complete on an n-Gaussian mixture, we split each Gaussian by creating two new Gaussians that have the same diagonal covariance matrix as their parent and means of $\pm 0.2\sigma$ for each feature. This train-split process is repeated until we reach the target mixture size.

4.2. GMM-based Classification

For a given classification task, we train one GMM for each class. When presented with a sequence of MFCCs as a test document, we use each class's GMM to compute the probability of the MFCC sequence being generated from the mixture and pick the class whose GMM outputs the highest probability.

4.3. Context-dependent Models

Aside from using one GMM for each class, we also employ context-dependent GMMs. For example, in our geographical region prediction task we use shopping style-dependent GMMs. In this case, for each class/shopping style pair we train a GMM. During the classification process, we only use the GMM from each class that corresponds to the test document's shopping style. In our geographical region predictions section we show that using context-dependent models where each class/context pair are trained on only 10 reviews produces impressive results over the no context GMM technique.

4.4. Aggregate Classifier

We use an aggregate classifier to combine the "plain" GMMs, context-dependent GMMs, and other features into a final classification decision. For the aggregate classifier, the features are the plain GMM-based classifier's decision, the context-dependent GMM-based classifiers' decisions, and any other explicit features such as the review's rating and the presence of each hobby, interest, or nutshell in the user's profile. We train our aggregate classifier on new data that was not used during any of the GMM training, which we have an abundance of since our dataset contains 9,073 reviews and we only use 100 reviews from each class for GMM training. We tried many classifiers but found random forests [7] via WEKA [8] produced the best results.

5. Speaker Identification

While an interesting topic in itself, the speaker identification task served as a confirmation that our GMM implementation works as intended. For our speaker identification dataset we randomly selected 101 users that each has 17 to 30 product reviews. For each user we evenly split their reviews into training and testing data with random assignments to the training and testing sets. We trained one 64-Gaussian mixture on each user's training data and then applied our GMM-based classification method of Section 4.2 to the test set to identify users.

The results of the speaker identification system were encouraging given the fact that the audio in our dataset is unprocessed, produced under different recording conditions/environments with various low-cost webcams, and the speech uses an unrestricted vocabulary and is about a wide range of products from kitchen appliances to movies to computers. Table 2 summarizes our speaker identification results, showing the mean and standard deviation of precision, recall, and F-Measure across all 101 users. All F-measure results presented here and in later sections use balanced F-Measure.

Table 2. *Mean and Variance of the precision, recall and F-Measure for speaker identification.*

	Precision	Recall	F-Measure
Mean	83.6	81.1	80.6
Standard Deviation	13.7	15.5	12.6

6. Geographical Region Predictions

As mentioned in Section 3, each user in our dataset contains the state that he or she currently lives in. We mapped each state to one of four geographical regions (West, Midwest, South, and Northeast) in accordance with the U.S. Census Bureau. We segmented our overall dataset by region and randomly placed each user into the region's training or test set with a 70%, 30% split. This ensures that a user's reviews do not span both the training and test set. Table 3 shows the resulting user and review counts for each region in the training and test sets. For all results in this section we tested our classifiers on 200 randomly selected reviews from each region (for a total of 800 reviews). While the reviews were randomly selected, they were the same set of reviews for all classifiers.

Table 3. *Dataset Sizes*

Region	Training Set		Testing Set	
	Users	Reviews	Users	Reviews
West	211	1,171	53	217
Midwest	304	2,081	76	500
Northeast	193	1,177	49	294
South	404	2997	102	636

6.1. Four 256-Gaussian Mixtures

Using the same approach as in our speaker identification task, we randomly selected 100 reviews (roughly one million 25ms window frames) from each region's training data and trained a 256-Gaussian GMM. We tested the resulting GMM-based classifier on the test dataset and found discouraging results, slightly better than 'by chance' (randomly assigning a region to an audio). This dismal performance is summarized in Table 4.

Under the assumption that random frames might better represent the class than random reviews, we tried randomly selecting one million 25ms windows from all of the training data for each class. The performance with this second approach was not any better than selecting 100 random reviews.

Table 4. *Results for geographical region prediction with 256-Gaussian mixtures*

Precision	Recall	F-Measure	Class
35.9	14.0	20.1	West
18.8	24.0	21.1	Midwest
29.0	27.0	28.0	Northeast
23.6	33.0	27.5	South
26.8	24.5	24.2	Macro

6.2. Context-Dependent GMMs

To improve performance we used context-dependent GMMs as described in Section 4.3. The contexts in our case were the available metadata associated with each review. This metadata includes the presence/absence of each hobby, interest, and nutshell in the user's profile, the user's shopping style, the review's rating, and the base category that the review is in (i.e. Electronics, Kitchen, etc.)

We assumed independence of the contexts as a simplifying assumption, which results in 30 context groups: 7 hobbies, 12 interests, 8 nutshells, 1 shopping style, 1 review rating, and 1 base category. Each group has at least two contexts. For example, presence and absence for the hobby, interest, and nutshell groups and 27 categories for the base category context group. This results in 87 contexts in total. We trained a GMM for each context/region pair, with 348 GMMs in total. We restricted our mixtures to 8 Gaussians because of the time and computational constraints. To further decrease processing time we trained each GMM on only 10 randomly selected reviews that matched the context/region pair.

In order to combine the decisions from the 30 GMM-based classifiers, we used an aggregate classifier as described in Section 4.4. The aggregate classifier was trained/tested on the 800-review test set using 10-fold cross validation. Table 5 shows that using context-dependent GMMs greatly improves performance over the four 256-Gaussian mixture approach.

Table 5. *Results for context-dependent GMMs*

Precision	Recall	F-Measure	Class
55.7	64.0	59.5	West
51.2	52.0	51.6	Midwest
58.0	54.5	56.2	Northeast
49.7	44.5	47.0	South
53.6	53.8	53.6	Macro

6.3. Explicitly Including Context Features

To further improve performance, we built an aggregate classifier that included as features the four 256-Gaussian classifier's decision, the 30 context-dependent classifiers' decisions, and the user's hobbies, interests, nutshells, and shopping style. The review's rating and base category were left out as they did not improve performance. Like our context-dependent aggregate classifier, this final classifier was trained/tested on the 800-review test set using 10-fold cross validation. Table 6 summarizes the results, which shows significant improvement over the four 256-Gaussian classifier and the context-dependent aggregate classifier.

Table 6. *Results for aggregate classifier performance using explicit context features*

Precision	Recall	F-Measure	Class
63.3	72.5	67.6	West
66.5	69.5	68.0	Midwest
70.0	66.5	68.2	Northeast
63.4	54.5	58.6	South
65.8	65.8	65.6	Macro

6.4. Ablation Study

To analyze the contribution of each of the features we used, we performed an ablation study on the aggregate classifier. The performance of the system for each such ablation is listed in Table 7.

The results clearly show that the speech data and explicit context features complement each other in achieving high accuracy. Of the explicit context features, the nutshell presence features and the shopping style feature are most important. Removing all GMMs, which translates to removing all speech-related features, results in a macro F-Measure drop of 14.8. This shows that speech-related features are indeed important in geographical region prediction. As expected from our previous classifier results, the context-dependent GMMs are much more important than the four 256-Gaussian classifier (listed here as "Plain GMMs").

Table 7. *Ablation study of the feature set*

Feature Set	Macro Precision	Macro Recall	Macro F-Measure
All Features	65.8	65.8	65.6
No GMMs	53.6	52.0	50.8
No Plain GMMs	64.2	64.0	63.9
No Context-Dependent GMMs	59.2	58.1	57.5
No Shopping Style	62.1	62.0	61.9
No Hobbies	64.5	64.4	64.2
No Interests	65.1	65.0	64.9
No Nutshells	60.0	60.0	59.9

7. Conclusions

The results for the speaker identification task correspond to what one would expect for the size of the training and test set as seen in literature. This confirms that our GMM implementation is performing properly.

Our work on geographical region prediction shows that acoustic features are indeed helpful for classifying regions. We showed that context-dependent GMMs, each trained on only 10 reviews and only including 8 Gaussians, can vastly outperform a more general large Gaussian mixture approach. We also showed that using an aggregate classifier to combine decisions from multiple GMM-based classifiers and to include additional explicit features greatly improves performance.

Our dataset does not include any gender-specific information, so we were not able to separate the male and female users. Traditionally, speaker identification using Gaussian Mixture Models has shown to bring better results with separate models for male and female data. This causes us to suspect that our model would also do better on the region prediction task if we were to train separate models for males and females.

As compared to results of humans classifying dialect [4], our model gives an overall accuracy of around 67%, which is high compared to human accuracy of 30%. While classifying geographical region does not directly transfer to classifying dialect, we believe they are comparable and, at the very least, that geographical region prediction from audio is more difficult. This result is also surprisingly good, classifying 2 out of every 3 people correctly for a user base where each user's home state may not correspond to the place they were born and the "dialect" that they speak. Additionally, most of the users in our dataset are most likely city-based and would therefore not be expected to have traditional dialects.

8. Future Work

Due to time and computational constraints, our large-Gaussian GMMs were only 256 Gaussians and our acoustic features were only the 39 MFCC features. We believe that even better performance can be achieved by using additional acoustic features to characterize the audio and by using larger mixtures. With 9,073 reviews we certainly have enough data for training.

A human study using a subset of our dataset would also be useful to compare with our results. While previous human studies show human performance at or below our system's performance, all previous studies have included only a few speakers that each spoke the same standardized sentences. The human classifiers in these previous studies also did not have access to each speaker's shopping style or list of hobbies, interests, and nutshells. A study where human classifiers have access to this metadata and listen to audio in a wide range of categories from a large open vocabulary would provide an ideal comparison for our system's results. An additional study on the effect of mixture size and training set size for the context-dependent GMMs would also be useful.

9. Acknowledgements

We would like to thank ExpoTV.com for not choking our crawler and Professor Dan Jurafsky for his great discussions and guidance. Todd Sullivan is supported by an NDSEG Fellowship sponsored by the DOD and AFOSR.

10. References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech, and Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.

[2] Frederick Weber,Linda Manganaro, Barbara Peskin, Elizabeth Shriberg, "Using Prosodic and Lexical Information for Speaker Indentification," ICASSP-2002

[3] Tao Chen, Chao Huang, Eric Chang, Jingchun Wang, "Automatic accent identification using Gaussian mixture models," in IEEE Workshop on ASRU , 2001.

[4] Cynthia G. Clopper, David B. Pisoni, "Some acoustic cues for the perceptual categorization of American English regional dialects," Journal of Phonetics, Volume 32, Issue 1, January 2004, Pages 111-140

[5] Van Bezooijen, R., & Ytsma, J., "Accents of Dutch: Personality impression, divergence, and identifiability," Belgian Journal of Linguistics, 13, 105\u2013129, 1999.

[6] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., "The HTK Book (for HTK version 3.2)," Cambridge University Engineering Department, Cambridge (2002)

[7] Leo Breiman, "Random Forests," Machine Learning. 45(1):5-32, 2001.

[8] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.