Foreign Accent Classification

CS 229, Fall 2011

Paul Chen
pochuan@stanford.edu

Julia Lee
juleea@stanford.edu

Julia Neidert
jneid@stanford.edu

ABSTRACT

We worked to create an effective classifier for foreign accented English speech in order to determine the origins of the speaker. Using pitch features, we first classify between two accents, German and Mandarin, and then expanded to a set of twelve accents. We achieved a notable improvement over random performance and gained insights into the strengths of and relationships between the accents we classified.

1. INTRODUCTION

Accented speech poses a major obstacle for speech recognition algorithms [4]. Being able to accurately classify speech accents would enable automatic recognition of the origin and heritage of a speaker. This would allow for robust accent-specific speech recognition systems and is especially desirable for languages with multiple distinct dialects. Accent identification also has various other applications such as automated customer assistance routing.

In addition, analyzing speech data of multiple accents can potentially hint at common linguistic origins. When an individual learns to speak a second language, there is a tendency to replace some syllables in the second language with more prominent syllables from his native language. Thus, accented speech can be seen as the result of a language being filtered by a second language, and the analysis of accented speech may uncover hidden semblances among different languages.

Spoken accent recognition attempts to distinguish speech in a given language that contains residual attributes of another language. These attributes may include pitch, tonal, rhythmic, and phonetic features [3]. Given the scale constraints of this project and the difficulty of extracting phonemes as features, we start by extracting features that correspond to pitch differences in the accents. This is a common approach when it comes to speaker and language identification and calls for feature extraction techniques such as spectrograms, MFCCs, and LPC.

2. PREVIOUS WORK

A previous CS229 class project [6] experimented with Hierarchical Temporal Memory in attempting to classify different spoken languages in transcribed data. They preprocessed their data using a log-linear Mel spectrogram and classified it using support vector machines to achieve above 90% accuracy. Although their project focuses on classifying completely different languages and we would like to classify different accents, their results can serve as a good frame of reference.

Research presented in a paper by Hansen and Arslan [3] used Hidden Markov Models and a framework that they termed "Source Generator" which attempts to minimize the deviation of accented speech from neutral speech. They used a large number of prosody based features. In comparing accented speech to neutral speech, they found that pitch based features are most relevant. Their work suggests that it is possible to classify accented speech with good accuracy using just pitched-based features.

A paper by Gouws and Wolvaardt [2] presented research that also used Hidden Markov Models to construct a speech recognition system. Their results elucidated some of the relations between training set size and different feature sets. They showed that the performance of using LPC and FBANK actually decrease with increasing number of parameters, while LPCEPSTRA increased and MFCC stayed the same. These results give us a better guidance for our choice of feature sets and amount of data.

Research by Chen, Huang, Chang, and Wang [1] used a Gaussian mixture model in order to classify accented speech and speaker gender. Using MFCC's as their feature set, they investigated the relationship between the number of utterances in the test data and accent identification error. The study displays very impressive results, which encourages us to think that non-prosodic feature sets can be promising for accent classification.

3. DATA AND PREPROCESSING

All training and testing were done with the CSLU: Foreign Accented English v 1.2 dataset (Linguistic Data Consortium catalog number LDC2007S08) [5]. This corpus consists of American English utterances by non-native speakers. There are 4925 telephone quality utterances from native speakers of 23 languages.

Three independent native American English speakers ranked and labeled the accent strength of each utterance. We used the Hidden Markov Model Toolkit (HTK) for feature extraction, MATLAB for preprocessing, and LibSVM and the Waikato Environment for Knowledge Analysis (Weka) for classification.

Data points were taken from 25 ms clips of utterances and were averaged over a window of multiple seconds to form features. Various preprocessing techniques were attempted, including sliding windows, various window lengths, standardization, and the removal of zeros from data points. The four second, non-sliding windows with standardization was chosen for use in further work as it gave the best results on our baseline classifier.

4. CLASSIFYING TWO ACCENTS

We began by assessing feature set quality and classifier performance based on classification accuracy between two accents. Aiming to select accents that are more easily differentiable, we initially selected the Mandarin and the German accent. Our initial feature sets were Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Filterbank Energies (FBANK) features, as they were the most frequently used features in other previous works, especially MFCC and LPC. FBANK features represent the prominence of different frequencies in a sound sample, while MFCCs normalize these to take human perception of sound into account. LPC features also represent sound as frequencies, but separate the sound into a base buzz and additional formants.

4.1 Establishing a Baseline

For our baseline classification, we ran Naive Bayes, logistic regression, and SMO classifiers[1] each on FBANK, MFCC, and LPC feature sets for German and Mandarin accented speech files. For each pair of classifier and feature set we obtained the results shown in Table 1.

Table 1. Testing accuracy for baseline classifiers and features

	FBANK	LPC	MFCC
ZeroR	50.25	51.46	51.46
Naïve Bayes	57.21	58.85	60.64
Logistic Regression	69.3	59.70	60.28
SMO	66.53	59.61	60.45

4.2 Assessing Data Quality

To determine whether insufficient data was causing poor accuracy, we divided our feature data into a testing set (30%) and a training set (70%). We measured classification accuracy for the testing set when each classifier was trained on increasing fractions of the training data. We observed that accuracy increased when the classifier was trained with more data, but decreasing accuracy gains suggested that insufficient data was not the primary cause of poor accuracy (see Figure 1).

We also tested whether the accent data was too subtle, as some speech samples barely sound accented even to a human listener. Each speech sample was previously rated by 3 judges on a scale from 1 (negligible or no accent) to 4 (very strong accent with hindered intelligibility) [5], so we extracted FBANK features (which produced higher baseline accuracies than MFCC and LPC) from 3 different subsets of the more heavily-accented data with stronger accents and measured classification with our baseline classifiers. Specifically, we selected speech samples with average ratings greater than 2.5 and greater than 2.7. However, classification accuracy saw little improvement, perhaps due to the effect of a reduced data set size (see Table 2). Consequently, we continued to use all data available for Mandarin and German accented speech.

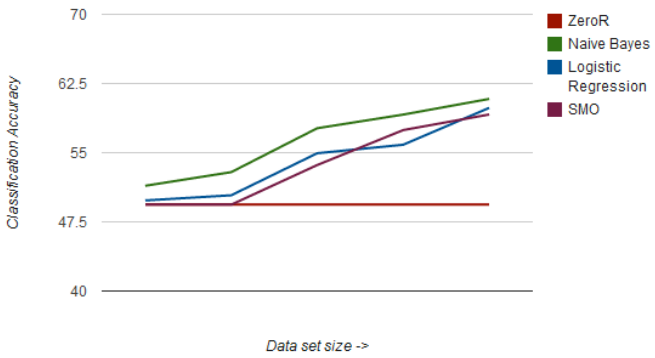

Figure 1. Significance of data set size.

Table 2. Classifier accuracies using most heavily accented data and FBANK features

Classifier	Accent Strength > 2.5		Accent Strength > 2.7	
	Training Accuracy (%)	Testing Accuracy (%)	Training Accuracy (%)	Testing Accuracy (%)
ZeroR	56.6	56.3	59.6	59.0
Naïve Bayes	55.7	56.2	61.2	50.7
Logistic Regression	63.0	57.9	69.1	52.5
SMO	61.9	59.3	66.9	58.4

4.3 Improving Feature Set Selection

Next we considered the quality of our features and expanded our MFCC feature set to include deltas, accelerations, and energies ("TARGETKIND = MFCC_E_A_D" in HTK configuration files). This again achieved little improvement over MFCC. By plotting training accuracy vs. testing accuracy (see Figure 2), we observed that training accuracy was also low, showing us that we were under-fitting the data. Thus, we attempted to boost accuracy by first over-fitting our training data before trying any optimization. We merged the individual feature sets (expanded MFCC, LPC, and FBANK) into a single set, but found that training error still did not improve substantially (see Table 3). We subsequently ran feature selection algorithms (including Correlated Features Subset Evaluation and Subset Evaluation using logistic regression and SMO) to try to remove all but the strongest features. This improved the accuracy on the training data, but not the testing data, which suggests that classifying on stronger accents using a larger data set could help.

[1] Unless otherwise specified, default Weka values were used for classifier parameters.

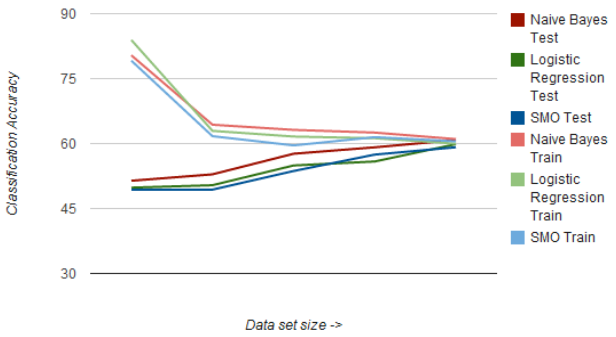

Figure 2. Classifier training and testing accuracies vs. training set size.

Table 3. Accuracy of baseline classifiers on merged feature set containing MFCC, LPC, and FBANK features.

Classifier	Training Accuracy (%)	Testing Accuracy (%)
ZeroR	50.37	53.56
Naïve Bayes	59.60	58.65
Logistic Regression	61.33	59.90
SMO	61.39	59.62

4.4 Selecting a Better Classifier

To improve training error, we tried using K-Nearest Neighbors (KNN) as well as LibSVM. KNN performed poorly, but we observed dramatic improvements in training set classification accuracy using a LibSVM classifier with a Gaussian kernel (see Table 4).

Table 4. Accuracy of initial LibSVM classifiers using Gaussian kernels.

Feature Set	Training Accuracy (%)	Testing Accuracy (%)
FBANK	63.43	59.81
LPC	96.19	57.12
MFCC (expanded)	82.48	57.88
All	89.63	57.45

Although training accuracy increased significantly, we did not see similar gains in testing accuracy. In order to boost testing accuracy, we optimized parameters of our LibSVM classifier (see Figure 3). Optimizing gamma versus C (the coefficient for the penalty of misclassification), we finally saw an improvement. We achieved a testing accuracy of 63.3% with C=128 and gamma = 0.000488 as parameters of the Gaussian kernel. We experimented with sigmoid and polynomial kernels and various parameter sets, but computing resources limited the range of parameters tried, so we did not achieve better accuracy in our preliminary optimizations.

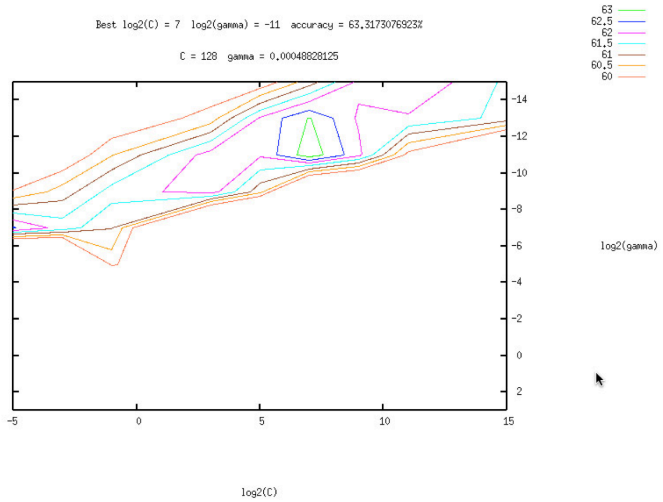

Figure 3. Optimizing gamma and C parameters of the LibSVM Gaussian kernel.

5. CLASSIFICATION ACROSS MULTIPLE LANGUAGES

We proceeded to process a dozen accents from our dataset, choosing only ones that have at least 200 utterances. We obtained a classification accuracy of 13.26% by reselecting parameters for LibSVM, which is a significant improvement over the baseline accuracy of random guessing (8%). Further, the confusion matrix across these twelve accents displayed interesting results. Figure 4 plots the percentage of cases in which each language on the y-axis was classified as a language on the x-axis. While we do not see a particularly distinct diagonal indicating correct classifications, this plot does illuminate some interesting relationships in our accent database.

The resulting figure shows that the Cantonese accent is very distinctive in our dataset and is easiest to classify with our features. It suggests that our Hindi accent samples share many similar aspects with other languages such that many instances of the other accents were classified as Hindi, while the opposite is true for German. This suggests that our initial choice of German and Mandarin for the two-class problem may have resulted in better results if we had chosen other accents.

This figure also hints at the similarity of accents from countries of geographic proximity. For example, the German accent is most frequently confused as the French and the Swedish accents, and the Japanese accent was often confused with the Cantonese and Mandarin accents. However, it also reveals that geographic proximity does not absolutely determine accent semblance. For example, the French accent is actually least likely to be confused with the German accent despite the fact that France and Germany are bordering countries.

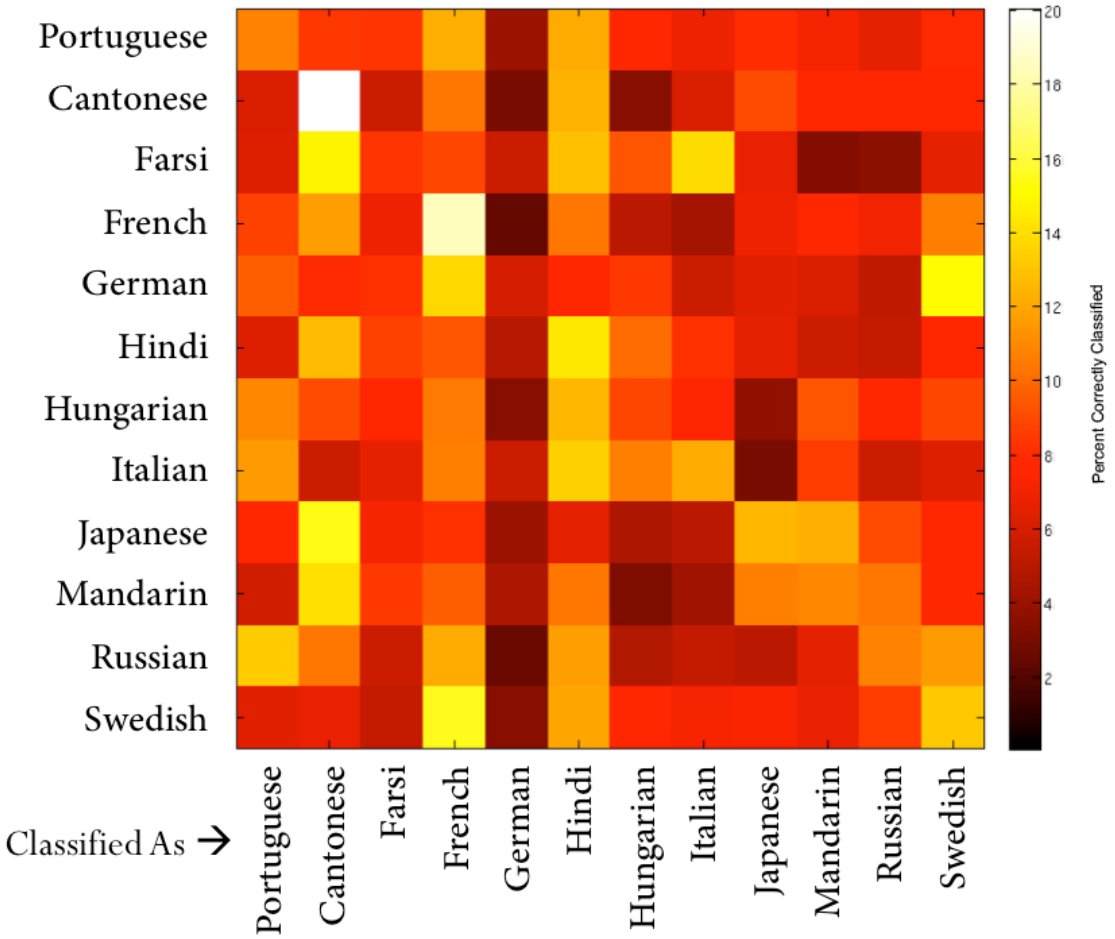

Figure 4. Confusion matrix for 12-way accent classification.

6. FUTURE WORK

We tried many different approaches in order to arrive at the best possible accent classifier using a set of features based solely on pitch. In the end, our training error was still significantly higher than our testing error, so these results might still be improved. To do this, we would want to use a larger data set with stronger accents. Performing more intensive feature selection using Subset Evaluation on LibSVM, which was infeasible with our limited computing and time resources, would likely prove helpful, as would performing more intensive parameter selection for different kernels.

In addition, the accent classification problem could be significantly different from other speech classification problems, and thus, other feature sets might be more informative. At this point, we would need to work with linguists and sociologists to generate these relevant features from scratch.

Altering the problem slightly, we could cluster accents from a common geographic region and work to identify between those groups. Inversely, further analysis of our current classification results and how those are correlated with geographic and historical data could uncover or reinforce these insights into the structures and origins of different languages and the histories of different peoples.

7. CONCLUSION

There is much need for improvement before an accent classifier could be used definitively in a speech recognition system. In our work, however, we have made progress in this area and have also uncovered insights into the relationships between accents and their origins. This suggests that in the future, there is hope for further improvement and an increased understanding of how we speak and where we come from.

8. ACKNOWLEDGMENTS

Thanks to Andrew Maas for his support and advice in this project throughout the process!

9. REFERENCES

[1] T. Chen, C. Huang, C. Chang, and J. Wang, “On the use of Gaussian mixture model for speaker variability analysis,” presented at the Int. Conf. SLP, Denver, CO, 2002

[2] E. Gouws, K. Wolvaardt, N. Kleynhans, and E. Barnard, “Appropriate baseline values for HMM-based speech recognition,” in Proceedings of PRASA, November 2004, pp. 169–172

[3] J. H. L. Hansen and L. M. Arslan, “Foreign accent classification using source generator based prosodic features,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 1, 1995, pp. 836–839.

[4] C. Huang, T. Chen, S. Li, E. Chang and J.L. Zhou, “Analysis of Speaker Variability,” in Proc. Eurospeech, 2001, vol.2, pp.1377-1380, 2001

[5] T. Lander, 2007, CSLU: Foreign Accented English Release 1.2. Linguistic Data Consortium, Philadelphia

[6] D. Robinson, K. Leung, and X. Falco, “Spoken Language Identifcation with Hierarchical Temporal Memory.” http://cs229.stanford.edu/proj2009/FalcoLeungRobinson.pdf