Automatic Accent Quantification of Indian Speakers of English

Jian Cheng, Nikhil Bojja, Xin Chen

Knowledge Technologies, Pearson
4040 Campbell Ave., Menlo Park, California 94025, USA
jian.cheng@pearson.com

Abstract

In addition to measuring job candidates spoken English proficiency, quantifying the degree of accentedness may help companies assign employees to appropriate job categories, or identify employees who could benefit from additional speech training. In this paper, we discuss methods for automatic accent quantification of Indian English speakers. Similar to techniques used in speaker recognition, we used Gaussian mixture models (GMMs) for the modeling of accent spectral characteristics in different groups of subjects. Computationally, we verified that certain consonants in Indian English have more discriminative power than others in quantifying an Indian accent. As a result, we propose the idea of using GMMs to model only certain phonemes with high predictive power. By combining features from GMMs with others, we achieved a human-machine correlation coefficient of 0.84 at the participant level. The results validate the use of new proposed methods to quantify accents automatically.

Index Terms: accent quantification, GMM, Indian English, speech assessment

1. Introduction

Call centers and outsourcing firms are part of a fast-growing service industry. To be successful in this industry, there is an increased demand for employers to be able to detect the heaviness of an accent so that they can assign employees to appropriate job categories, or give them additional training to refine their accents as appropriate for their jobs.

Accent is the characteristic mode of pronunciation of a person or a particular language group. It comprises a distinctive quality of voice, pronunciation of vowels and consonants, stress, and prosody. Although different accents may have different grammar, semantics, vocabulary, and other language characteristics, our main focus here refers to the specific variation in pronunciation. In this paper, our goal is to detect the heaviness of an Indian accent in the speech of Indian speakers of English so that we can apply the technology to a real assessment environment.

Speakers of English from the Indian subcontinent have characteristic variations in their accent that are influenced by the nature of native languages spoken there. Most Indian languages are phonetic in nature and hence the pronunciation of English words tends to be slightly phonetic in nature. Indian English tends to have a reduced vowel system, and most pronunciation of Indian English is rhotic. Indian English may have more distinctive features in its consonants than in the vowels. These characteristics give rise to the prototypical Indian accent that differs slightly from other non-native accents of English. According to Trudgill & Hannah [1, p. 133] there are a number of generalizations regarding pronunciation patterns of Indian speakers of English: The alveolar consonants tend to be replaced by retroflex consonants; the voiceless plosives /p/, /t/, /k/ tend to be unaspirated; and Indian English tends to be syllable-timed rather than stress-timed. In our development process, we focused on these specific phonemes.

Although there has been a considerable amount of previous work on speaker recognition ([2] provides a detailed literature review) and some on accent classification [3, 4, 5, 6, 7, 8, 9, 10], to the best of our knowledge, no previous research has addressed the topic of accent quantification. Accent quantification is closely related to speaker recognition and accent classification, but it differs from these other topics in significant ways. While the motivation for speaker recognition is to identify a speaker, and while possible motivations for accent classification are to extract demographic information or to achieve better speech recognition results, the main motivation for accent quantification is assessment. Also, unlike accent classification, we deal with a single accent here and focus on measuring the heaviness of that accent. For this reason, we believe that the previously proposed parallel phone recognition [4] and/or phonotactic language models [10] for different accents may not be appropriate for the problem at hand. It is possible that the target subjects may have a similar accent class and tend to make the same phonotactic structural errors. Our work focuses on a regression problem instead of a classification problem.

In this paper, we showcase our work on automatic accent quantification by proposing a new method for modeling accent spectral characteristics. Borrowing from techniques used in the field of speaker recognition, we introduce GMMs to the accent quantification domain. Further, by applying the research from Trudgill & Hannah on pronunciation patterns of Indian English speakers, we propose a technique in which we model only certain phonemes that have better predictive power in quantifying an Indian accent. Given the promising results, we suggest that modeling spectral characteristics using our proposed method could offer a common solution for not only accent quantification but related fields.

The remainder of this paper is organized as follows: We first introduce GMMs and discuss how they can be used for accent quantification. We then propose the idea of using phonetic class-based GMMs. Next, we present some experiments in which we compared the performance from different numbers of mixtures in the GMMs and also explored the performance of different phonetic classes. The paper concludes with a discussion of the results.

2. An accent quantification system

To develop new methods for accent quantification, we used recordings of speech from two English tests: Pearson Test of English Academic [11] and Versant English [12]. For each re-

sponse recording, we asked human raters to judge the extent of the speakers accent. Based on these ratings, we assigned each speaker to one of six groups in which group 1 members were judged to have a native speaker accent (e.g, US, UK, Australian) and group 6 members were judged to have a heavy Indian accent. Our goal was to create models that would allow us to predict the likelihood that a new speaker belonged to a specific group. In this way, these likelihoods were used to quantify accent. We separated the data into three sets: a training set, a development set and a testing set. The training set was used to train our models, the development set was used to refine our models, and finally, the testing set was used to validate our approach by comparing the results of our model for a new set of speakers with human ratings.

2.1. Gaussian mixture models

Gaussian Mixture Models (GMMs) [13] are broadly used in speaker recognition and accent classification. Here we adopted the GMM framework to model accent explicitly. A GMM is composed of a finite mixture of multivariate Gaussian components:

$$P(\boldsymbol{x}|M) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (1)$$

where $\boldsymbol{x}$ is the feature vector, K is the number of Gaussian components, w_k is the weight of the k'th Gaussian component, $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$. A GMM model M is defined by these parameters: w_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$. After we extracted interested feature vectors from a recording: $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$, the averaged log-likelihood of $\boldsymbol{X}$ with respect to model M_i is defined as:

$$LL_{avg}(\boldsymbol{X}, M_i) = \frac{1}{T} \sum_{t=1}^{T} \log P(\boldsymbol{x}_t|M_i). \quad (2)$$

In our case, seven gender-independent GMM models were built: one Universal Background Model (UBM) and six other models for each of the six groups of speakers. For each utterance, we computed the LL_{avg} in each model and produced seven log-likelihoods. We were more interested in the posterior probability $P(M_i|\boldsymbol{x})$ instead of the likelihood $P(\boldsymbol{x}|M_i)$. By making the assumption that the prior probabilities of models are equally likely and by approximating the probability of the utterance using the UBM, Bayes theorem gives us the following:

$$logP(M_i|\boldsymbol{x}) \cong LL(\boldsymbol{X}, M_i) - LL(\boldsymbol{X}, M_{UBM}). \quad (3)$$

This is the general GMM-UBM framework for speaker recognition[14]. For each utterance, we produced the log-posterior probability $logP(M_i|\boldsymbol{x})$ in each speaker group model and treated these probabilities as input features for further machine learning. We tried using log-likelihood $logP(\boldsymbol{x}|M_i)$ in each speaker group model directly, but experimentally, we showed that the posterior probability was more robust to variations in speakers and acoustic channels.

2.2. GMM input features

The input features for our GMMs were based on 21 linear filterbank derived cepstra. The MFCC feature vector was composed of 65 coefficients, 21 static coefficients, 21 delta, 21 delta-delta, delta energy and the delta-delta energy. Cepstral mean and variance normalization were applied. Although we explored different settings we found that the regular setting for speaker recognition used here was more or less the optimal one.

Since our work is in the automatic assessment domain [12], we were interested in using products of automatic speech recognition (ASR). To this end, we used an augmented ASR system based on HTK [15] that was optimized for accuracy with non-native speech. The data sets used to develop the ASR system include native speakers of more than 60 different languages. They were different from the data sets in the current research. Because we were dealing with recordings in which speakers read aloud passages, we augmented the ASR system with rule-based language models [16, 17]. This augmented ASR system provided lexical units and spectral measures as well as a time-aligned phoneme transcription of the response. We used the phoneme alignments from this ASR system as a speech activity detector. We limited the feature vectors from the training files to be from specific phoneme frames.

2.3. GMM Model training and log-likelihood

Using all the training data, we built a UBM from the full set of feature vectors of interest. We then trained the accent heaviness dependent models by adapting the UBM using the training data from the specified groups via a MAP adaptation procedure [14]. Only mean vector adaptation was performed. For GMM model training and the log-likelihood computation, we used the open-source ALIZE toolkit [18].

2.4. Phoneme classes

To investigate whether or not certain phonemes have more discriminative power to quantify an Indian accent compared to others, we grouped phonemes together into different phoneme classes (Table 1).

Table 1: Different phoneme classes.

Class	Phonemes
ap	all phonemes
vp	all vowel phonemes
cp	all consonant phonemes
ip	phonemes associated with an Indian accent

Following the generalizations in Trudgill & Hannah [1, p. 133], ip covers the labiodental fricative /v/, the bilabial approximant /w/, the plosives /p/, /t/, /k/, the alveolars /t/, /d/, /s/, /z/, /l/, /r/, the postalveolar fricatives /zh/, /sh/, and the postalveolar affricate /ch/. We expected the ip class to have more discriminative power than the other classes.

For each phoneme class and each group of speakers, we used Formula 3 to compute the log of the approximate posterior probability. We denoted these values as ap_i, vp_i, cp_i, ip_i, where i was one of the six scales.

2.5. Some other features

Prosodic features have been used in speaker recognition [2] and accent classification [3, 5] successfully. We explored the effect of prosodic features on quantifying an Indian accent based on our previous research results [19]. Some features that can be used directly are phoneme duration log-likelihoods: log_seg_prob and $iw\_log\_seg\_prob$ [19].

Our ASR system produced confidence scores at the word level [20]. We predicted that heavily accented speech from Indian speakers of English may produce lower confidence scores of recognized words, although any accent could cause this effect. We used two averaged confidence scores: lattice-based $lconf^{\mu}$ and acoustic-based $aconf^{\mu}$ [20] as extra features. We computed a few spectral likelihood features [12] as well.

2.6. Final models and performance measures

When developing different GMM models, overfitting to the training data is often unavoidable. To verify that some parameters could be better than others in our GMMs, we built both linear regression and backpropagation neural network models using six input features as shown in Formula 3. The models were built using the training data and then tested with the development set. For the final model, we used the optimal parameters and combined the training set and the development set for model training. The results were then reported on the test set. The test set was never used to train models.

Our target was human ratings. To measure performance, we compared Pearson correlation coefficients between machine scores and human ratings.

3. Experiments and Results

3.1. Experimental Data

For the experiments, we used recordings of speakers in real assessment environments as they read aloud passages from a high-stakes English test – Pearson Test of English Academic [11] and from Versant English [12]. The average number of words per passage was about 50. The sample rate for the recordings was 8 kHz with 8 bits (telephone band). We asked human raters to rate the responses according to the rating criteria listed in Table 2. The logic behind this scale was to consider native speaker pronunciation variants the least Indian accented. Two to three different human raters rated every response. The responses identified by human raters as silence, or irrelevant or completely unintelligible material were excluded from our study. On average, every subject provided about 2.3 valid responses.

Table 2: The Indian accent rubric used by human raters.

Scale	Criteria
6	*Heavy* Indian accent: The speaker has a very thick Indian accent.
5	*Medium heavy* Indian accent: The speaker sounds Indian and the accent is pronounced.
4	*Medium light* Indian accent: The speaker sounds Indian and the accent is clear.
3	*Light* Indian accent: The speaker sounds Indian but the accent is subtle.
2	Non-native accent: The speaker sounds like a learner of English (e.g. Korean, Japanese, Philippine), and is probably not Indian.
1	Native: The speaker sounds native (e.g, US, UK, Australian).
0	Unintelligible/Irrelevant: Silence or unintelligible material.

There was a total of 825 participants. We randomly assigned about one quarter of the data to a test set ($n = 208$), about one quarter to a development set ($n = 206$), and the rest to a training set ($n = 411$). There were $1,446$ valid responses in the training and development sets and 488 valid responses in the test set. All the final parameters used for testing (such as GMM parameters) were tuned on the development set.

Figure 1 shows the average Indian accent score distribution for individual participants in the test set. For the 8 pairs of raters whose number of ratings was more than 40, the average of the inter-rater correlations at the response level was 0.774. For each response, we randomly assigned one rating to one group and the other rating to another group and correlated the two. Using this approach, the average correlation was 0.778. This level of correlation indicates that the human raters made reasonable judgments about Indian accent.

Table 3 lists the total speech segment duration in different sets for different phoneme classes. It gives a rough idea about the real speech size we had to train different GMMs.

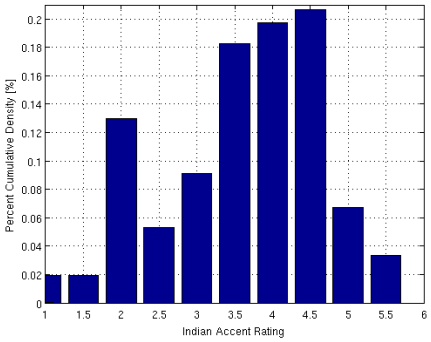

Figure 1: A histogram showing the average Indian accent score distribution for individuals in the test set as rated by humans.

Table 3: The total speech segment duration in different sets for different phoneme classes, in hours. n is the number of subjects in the corresponding sets.

Phoneme class	ap	vp	cp	ip
Training ($n = 411$)	4.44	1.67	2.77	1.80
Development ($n = 206$)	2.21	0.83	1.38	0.90
Test ($n = 208$)	2.28	0.85	1.43	0.93

3.2. GMM parameters

In this subsection we explore the relationship between the number of GMM components and performance using different phoneme class GMMs. Usually the larger the number of GMM components, the better the performance when we have enough training data. The limitation in training data often dictates that we use a smaller number of GMM components. Otherwise, we might overfit our model. Also, if the number of GMM components becomes too large, computation time may become too costly.

From Figures 2 and 3 we can see that overfitting becomes severe with the increase in the number of GMM components. Results from both linear regression (Figure 2) and backpropagation neural networks (Figure 3) show that the ip class has the best discriminating power even when the amount of training data is significantly less compared to the ap class. Although the ap class is a superset of the ip class, adding irrelevant phonemes seems to introduce noise to the models. The consonant phoneme class (cp) is better than the vowel phoneme class (vp). The vowel phoneme class (vp) performs the worst. These numbers validate the observations from phoneticians: Indian English has fewer peculiarities in its vowel sounds than its consonant sounds. Certain consonants have more discriminative power than others [1, p. 133].

Comparing Figure 3 to Figure 2, the results from neural network models are significantly better than those from linear regression models. It is clear that we are dealing with a non-linear problem. In Figure 3 we compared the performance difference between using the log of the approximate posterior probability and the log likelihood. From the results, we see that the best performance was achieved by using the log of the approximate posterior probability. For our data size, the best number of GMM components could be around 128 and 256. Here we used 128 for tuning parameters on the development set, but we used 256 for the final models since we combined both training data and development data to train the final models.

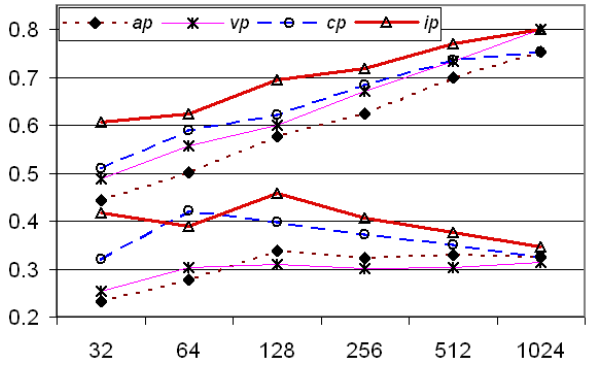

Figure 2: Performance using linear regression. The models are trained using the training set with six GMM features. The upper four lines are the results on the training set and the other four lines are the results on the development set. The x axis is the number of GMM components. The y axis is the correlation between human ratings and machine scores.

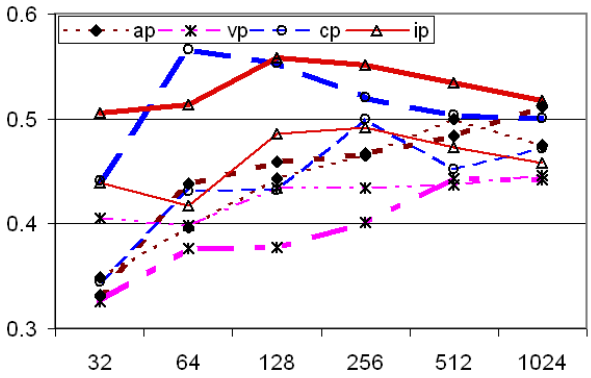

Figure 3: Performance of the development set using neural networks. The models are trained using the training set with six GMM features. The four thick lines are the results using the log of the approximate posterior probability. The four thin lines are the results using the log likelihood. The x axis is the number of GMM components. The y axis is the correlation between human ratings and machine scores.

Table 4 lists some correlations in the development set. Neural networks were used to combine different features. For comparison, the two kinds of correlations between human raters in the development set were 0.739 and 0.717. From the table we can see that the best features for quantifying an Indian accent were GMM features, whereas the worst were phoneme duration log-likelihoods (log_seg_prob, $iw\_log\_seg\_prob$). We tried methods mentioned in [19] to explore the features from energy and fundamental frequency, but they did not improve performance either. Our hypothesis is that different types of accents may have different prosodic features that help in accent classification [3, 5], but prosodic features in our study did not help distinguish subjects from within the same accent group.

Not all the features mentioned in this paper were used in the final model. We selected the optimal features based on results from the development set. The final features were ip_{1-6}, cp_1, cp_3, cp_6, ap_{1-4}, vp_1, vp_{4-5}, $aconf^{\mu}$, $lconf^{\mu}$ and a few spectral likelihoods [12].

Table 4: Correlations at the response level using different features in the development set.

Features	Correlation
ip_{1-6}	0.558
$aconf^{\mu}$	0.234
$lconf^{\mu}$	0.184
log_seg_prob	0.035
$iw\_log\_seg\_prob$	0.057
ip_{1-6}, cp_1, cp_3, cp_6, ap_{1-4}, vp_1, vp_{4-5}	0.610
the final model	0.674

3.3. Results

Table 5 lists several important correlations in the test set using neural network models trained on the training and development sets with the parameters mentioned in Subsection 3.2.

Table 5: Correlations at the response level using different features in the test set.

Features	Correlation
ip_{1-6}	0.733
ip_{1-6}, cp_1, cp_3, cp_6, ap_{1-4}, vp_1, vp_{4-5}	0.741
the final model	0.779

If we use the average of all human ratings as the participant's final human score and the average of all machine scores as the participant's final machine score, at the participant level, the final correlation was 0.84. This result was achieved by using only about 2.3 read-aloud items.

3.4. Discussion

The GMM models used here are gender-independent. We expect that gender-dependent models may perform better as gender-dependent models were trained frequently in accent classification tasks [4, 6, 8]. Compared to the performance of GMM models that were trained using the training set, the significant performance improvement observed when using both the training and development sets reveals that collecting more data may be able to help improve performance. When we have enough data, we may increase the number of GMM components to further improve the performance.

4. Conclusions

We used GMMs successfully for modeling accent spectral characteristics in different groups of subjects. We proposed the idea of using GMMs to model only certain phonemes that may have better predictive power in quantifying an Indian accent. We verified computationally that Indian English has more distinctive features in consonants than in vowels, and that certain consonants have more discriminative power than others. We concluded that prosodic features may not help to quantify an accent. By combining features from GMMs with other features such as spectral likelihoods at the phoneme level, we achieved a human-machine correlation coefficient of 0.78 at the response level and 0.84 at the participant level. As a comparison, the correlation coefficient between human raters was 0.77 at the response level. The results support our hypothesis that our new proposed methods can successfully quantify an accent automatically.

5. References

[1] P. Trudgill and J. Hannah, *International English: A guide to the varieties of standard English*, 5th ed. London: Hodder Education, 2008.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12–40, 2010.

[3] J. H. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *ICASSP 1995*, pp. 836–839.

[4] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, pp. 1784–1787.

[5] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *ICASSP 1999*, pp. 221–224.

[6] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, pp. 343–346.

[7] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2004)*, pp. 41–44.

[8] J. H. L. Hansen and P. Angkititrakul, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, pp. 634–646, 2006.

[9] G. Choueiter, G. Zweig, and P. Nguyen, "An empirical study of automatic accent classification," in *ICASSP 2008*, pp. 4265–4268.

[10] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages (CASL 2009)*, pp. 53–61.

[11] Pearson. (2013, Feb.) Skills and scoring in PTE Academic. [Online]. Available: http://pearsonpte.com/SiteCollectionDocuments/US_Skills_Scoring_PTEA_V3.pdf

[12] J. Bernstein, A. Van Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[13] 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, pp. 72–83, 1995.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[15] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge, England: Cambridge University, 2000.

[16] J. Cheng and B. Townshend, "A rule-based language model for reading recognition," in *SLaTE 2009*.

[17] J. Cheng and J. Shen, "Towards accurate recognition for children's oral reading fluency," in *IEEE-SLT 2010*, pp. 91–96.

[18] J. F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *ICASSP 2005*, pp. 737–740.

[19] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Interspeech 2011*, pp. 1589–1592.

[20] J. Cheng and J. Shen, "Off-topic detection in automated speech assessment applications," in *Interspeech 2011*, pp. 1597–1600.