CLASSIFICATION OF CHINESE DIALECT REGIONS FROM L2 ENGLISH SPEECH

Jiahong Yuan, Zhengqiang Rao, Hui Lin, Yang Liu
LAIX Inc.
{jiahong.yuan, zhengqiang.rao, hui.lin, yang.liu}@liulishuo.com

ABSTRACT

This paper presents an effort to classify Chinese speakers' L1 dialect regions from their L2 English speech. By applying LightGBM (a gradient boosting classifier based on decision trees) to softmax-based features from deep neural networks, our system achieved 68% accuracy on five dialect regions using one sentence and 82% accuracy using 23 words. The results represent a nearly 50% error reduction over a baseline system based on HMM/GMM and forced alignment. We demonstrated that modeling phone boundaries and vowel stress yielded a relative error reduction of 18%, with phone boundaries being more useful than vowels and consonants. Furthermore, in terms of classification models, LightGBM was extremely robust on this task, which we believe deserves further investigation.

Index Terms— L2, accent, classification, softmax, LightGBM

1. INTRODUCTION

Foreign accent has been a topic of interest to researchers in many fields. In linguistics, people are interested in the phonetic and phonological characteristics of second language (L2) speech and the factors that influence degree of foreign accent, aiming to better understand the mechanisms of speech production/perception and second language learning [1,2]; in speech technology, the focus is more on automatic assessment and identification of L2 speech and accent, from the perspective of improving ASR (Automatic Speech Recognition) and CALL (Computer Assisted Language Learning) systems [3-9]. Relatively few efforts have been made to integrate the results of these two research areas. In this paper, we present a study on classification of Chinese dialect regions from L2 English speech, combining insights from both linguistics and speech technology research.

Generally speaking, foreign accent results from the interference of the learner's first language (L1). Differences in phonemic inventory between L1 and L2 may cause negative L1 transfer in L2 production. For example, making a distinction between /l/ and /r/ is often difficult for Japanese learners of English because the two sounds are not distinguishable in Japanese [10]. Characteristics of phoneme pronunciation, therefore, play an essential role in foreign accent identification [11]. Besides phonemes, phone boundaries also contain useful information about speech characteristics. For example, the timing of voicing in stop consonants, measured by voice onset time (VOT), is a boundary-bound phonetic feature that has been extensively studied in linguistics [12]. The VOT of stops varies across languages. Individuals who learn an L2 later in life often fail to produce consonants with authentic VOT values in L2 [13]. [14] demonstrated that phone boundaries were helpful in automatic assessment of spoken language proficiency. [15] showed employing a special HMM for phone boundaries significantly improved forced alignment accuracy.

In L2 English, the placement and realization of lexical stress contribute to perceived proficiency and accent [16, 17]. Although lexical stress is suprasegmental and mainly realized on pitch, duration, and intensity, vowel quality and spectral characteristics also play an important role [18, 19]. [20] found that English unstressed vowels were produced more peripheral in the vowel space by late bilinguals of Spanish and English. Therefore, in automatic classification of L2 English accents, it may be helpful to treat stress as a property of vowels in the acoustic models.

We aim to automatically classify Chinese speakers' dialect regions from their L2 English speech, by incorporating phone boundaries and lexical stress into a state-of-the-art learning framework for the task. Compared to classification of native languages from L2 speech, classification of Chinese dialect regions is more challenging because of the impact of Mandarin [21]. Most Chinese speakers learned Mandarin in their childhood, and many speak Mandarin instead of the native dialect in their daily life. In this study, we obtain softmax-based features extracted from deep neural networks and employ LightGBM [22] classifier for dialect classification. Our experiments show significant error rate reduction compared to a baseline GMM/HMM approach. We also demonstrate the effectiveness of linguistically inspired features (stress and phone boundaries) and the LightGBM classification.

2. DATA

The dataset was compiled for a large-scale analysis of Chinese speakers' L2 English speech, which contains approximately 1.4M utterances (1600 hours) collected through a mobile app. With the app, a user reads a sentence after listening to it from a L1 speaker, and receives an automatic assessment of his/her speech. The dataset has 123 sentences and (on average) 12K speakers (mostly adults) per sentence, a total of 390K speakers from 120 cities in China. Based on our analysis of the data as well as Chinese dialectology, we grouped the 120 cities into five dialect regions (R1 to R5) as shown in Figure 1.

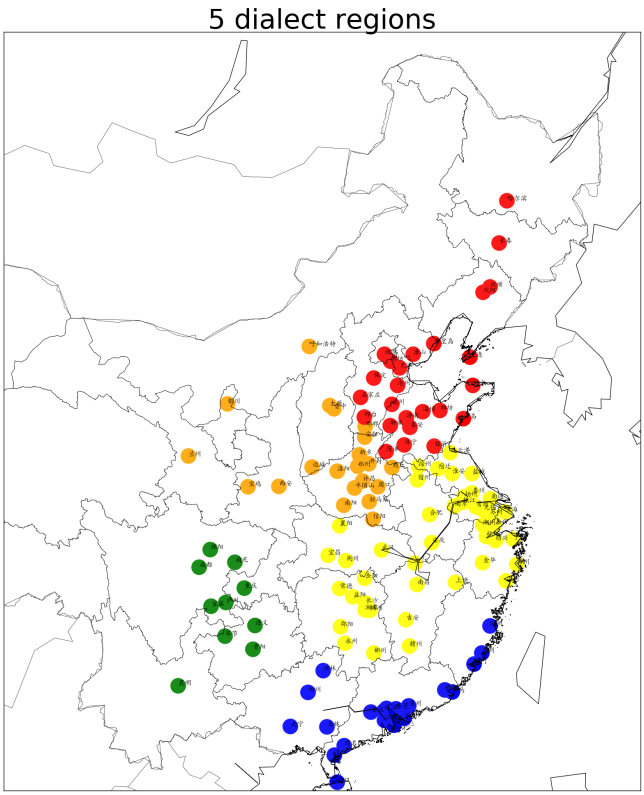

Figure 1. Chinese dialect regions: R1 (red), R2 (orange), and R4 (green) are different Mandarin dialects, R3 (yellow) includes Wu, Gan, Xiang, and some Mandarin dialects, and R5 (blue) mainly includes Min, Yue, and Hakka.

3. EXPERIMENTS ON UTTERANCES

3.1. Training and test division

We divided the data into training and test sets based on speakers to ensure no speaker overlap between the two sets. 50K speakers were randomly selected to form the test set, and the remaining were used for training. The distribution of the data on the five dialect regions is listed in Table 1.

Table 1. The number of utterances in training and test sets.

Dialect region	Training set	Test set
R1	262,812	38,214
R2	217,386	31,836
R3	443,660	64,609
R4	94,933	14,351
R5	206,372	30,364
Total	1,225,163	179,374

3.2. GMM/HMM baseline

A baseline system was built based on GMM/HMM and forced alignment. We trained five GMM/HMM acoustic models (of monophones plus phone boundaries), one for each dialect region using the training data of that region. Training was done with HTK [23], following the procedure in [15]. For every utterance in the test set, we ran forced alignment five times with the five acoustic models respectively, and selected the dialect region whose model had the highest alignment score. The overall classification accuracy on the test set was 33.3%.

There were 123 unique sentences in the data. The number of utterances (i.e., speakers) per sentence in the test set ranged from 1,235 to 1,587, with an average of 1,458. The classification accuracies on the 123 sentences varied greatly, from 26.3% to 40.5%. The sentence with the best accuracy (40.5%) was *I'm happy to have the opportunity to introduce myself here today*, which had 1,490 utterances (speakers).

3.3. Classification on softmax-based features

We propose to use a classifier for this task. We trained a Kaldi TDNN (nnet3) model to extract softmax-based features for classification [24]. The training data for the TDNN model consisted of about 1K hours of LibriSpeech L1 English speech. One softmax vector was computed at the center of every phone, and one computed at every phone boundary, then all the softmax vectors in an utterance were concatenated to form a feature vector to feed into a classifier. Phone centers and boundaries were determined by the baseline GMM/HMM system.

The LibriSpeech lexicon contains 69 phonemes, each of which has three HMM states, plus a "silence" phoneme with five HMM states. Therefore, the total number of HMM states (pdfs) in the model is 212, i.e., each softmax vector has 212 dimensions. We then summed the softmax values for all the states of the same phoneme, which reduced the dimensions of each softmax vector to 70. The procedure is illustrated in Figure 2.

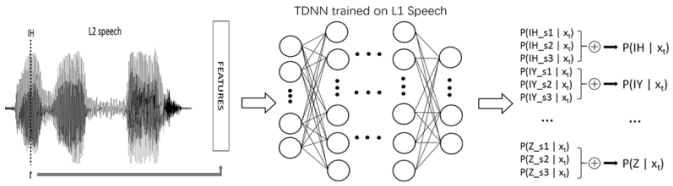

Figure 2. Softmax-based feature extraction.

For this experiment we only used the sentence with the best accuracy from the baseline system, "*I'm happy to have the opportunity to introduce myself here today*". This sentence has totally 90 phones and boundaries (i.e., 90 softmax-based vectors), therefore the feature vector has 6,300 dimensions (90*70). There were 11,998 utterances of this sentence in the dataset (from 11,998 speakers), 10,508 in the training set and 1,490 in the test set.

We compared three classifiers using the softmax-based features: SVM, MLP, and LightGBM. All classifiers were trained on the training utterances (10,508 utterances) and tested on the test utterances (1,490 utterances). The results showed that LightGBM was significantly better than the other two classifiers, with 8-10% absolute error reductions.

With LightGBM, the classification accuracy on the test utterances was 68.4%, demonstrating an absolute error reduction of 27.9% and a relative error reduction of 46.9% over the baseline system.

4. EXPERIMENTS ON WORDS

4.1. Data

The classification experiment above was based on only one sentence. We expect the accuracy to be better when more sentences are used for a speaker. Unfortunately, there were not many speakers in the dataset who spoke the same two or more sentences. To overcome this problem, we constructed new training and test sets by selecting words from different utterances to form a sequence of words. In the new data, every word sequence consists of 23 words: *morning, pleasure, sure, really, singer, right, night, tears, introduce, opportunity, agree, borrow, pencil, choose, about, song, never, favorite, throwing, butter, glad, afternoon, were*. We randomly selected them from different utterances to form the word sequences for five dialect regions respectively. The new training set contained 9,480 word sequences and the test set contained 1,254 word sequences. These were formed respectively following the original training and test division.

4.2. Effect of stress and phone boundaries

To evaluate the effect of stress, we trained two TDNN models using the Librispeech data. The difference between the two models is whether vowel phonemes are stress dependent or independent. If vowels are stress-dependent, then /AA0/, /AA1/, and /AA2/, for example, are treated as three different phonemes, otherwise there is only one /AA/ phoneme. We note that schwa (AH0/AX) is an independent phoneme in both cases. Including silence, the stress-dependent phoneme set consists of 70 phonemes, and the stress-independent set consists of 41 phonemes.

Softmax-based features were extracted in the same way as described in Section 3.2 on all utterances. Then the features of the selected words (from different utterances) were concatenated. In the 23 words we used, there were totally 105 phones (43 vowels and 62 consonants) and 82 phone boundaries. Two sets of softmax features were extracted, one from phone centers only, and the other from both phone centers and phone boundaries. Between-word boundaries, i.e., phone boundaries between adjacent words, were excluded because words were randomly selected so the same word might precede or follow different phonemes.

Overall, we created four sets of features on two aspects: stress vs. no stress, and phones vs. phones plus boundaries. The dimensions of these features are 4,305 (no stress, phones), 7,350 (stress, phones), 7,667 (no stress, phones and boundaries), and 13,090 (stress, phones and boundaries).

LightGBM models were trained on the four feature sets respectively with the same parameter settings as follows: *learning_rate*=0.05, *num_leaves=31* (max number of leaves in one tree), *max_bin*=255 (max number of feature bins), *max_depth=-1* (no limit on tree depth), *num_trees=5000* (the number of trees/iterations). The models were trained on the training data and tested on the test data. Because of the robustness of LightGBM on this task (discussed in detail in Section 4.5), no validation data were used for early stopping, and all models finished 5,000 iterations of training.

Classification results on the four feature sets are listed in Table 2. We calculated two types of accuracies: *1-best accuracy*: the best hypothesis is correct; *2-best accuracy*: one of the two best hypotheses is correct.

Table 2. Classification accuracies (%): 1-best / 2-best.

	phones	phones & boundaries
no stress	78.2 / 94.4	79.8 / 95.8
stress	78.6 / 94.7	82.1 / 95.7

From Table 2, we can see that employing vowel stress and phone boundary features improved 1-best accuracy to 82.1% from 78.2% when no stress or boundaries is used, i.e., a relative error reduction of 18% (from 0.218 down to 0.179), and improved 2-best accuracy from 94.4% to 95.7%, that is, an error reduction of 23%.

With no stress, phone boundaries led to an improvement of accuracy from 78.2% to 79.8%, whereas with stress the improvement was much greater, from 78.6% to 82.1%. This result suggests that the characteristics of phone boundaries with respect to speaker accent are stress-dependent. For example, a phone boundary following a stressed vowel may bear more accent information than following an unstressed vowel.

4.3. Confusion matrix results

Table 3 is the confusion matrix of the best classification results on the test data (with stress and phones & boundaries information). We can see that dialect regions R1 and R2 were more confused by the classifier, as well as R3 and R4. The pattern is consistent with Chinese dialectology. If we merged R1 with R2, and R3 with R4, and perform three-way dialect classification, we found we can achieve a 1-best accuracy of 92.4% and 2-best accuracy 99.9%.

Table 3. Confusion matrix of classification results.

	R1	R2	R3	R4	R5
R1	234	23	8	0	2
R2	62	103	45	4	6
R3	3	6	432	4	4
R4	0	4	33	54	3
R5	0	0	16	1	207

4.4. Importance of features

LightGBM provides two metrics to measure the importance of features in a model: *split* – the number of times a feature is used to split the data; *gain* – the average information gain a feature has when used to split the data.

Figure 3 compares the overall importance of vowels, consonances, and boundaries in the best model above (5 dialect regions, stress, phone & boundaries) on the two metrics respectively, in which the totals of each category are shown.

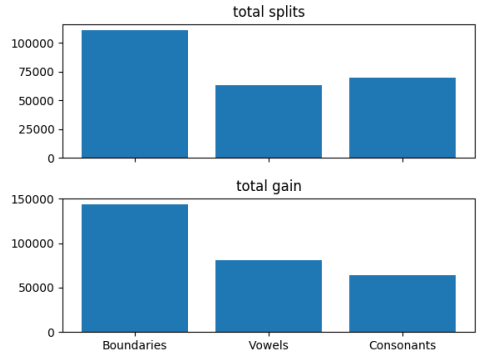

Figure 3. The total importance of vowels, consonants, and boundaries on *split* and *gain* in the model.

We can see that overall, phone boundaries are more important than vowels and consonants in terms of tree splits and information gain. We should also note that there were more phone boundaries (85) than vowels (43) and consonants (62) in each word sequence.

The 10 most important phonemes/boundaries in the model on *split* and *gain* are listed in Table 4 respectively. For every phone and boundary, the *split* and *gain* values on all feature dimensions of the phone/boundary were totaled. If a phoneme/boundary had more than one token, we took the average of those tokens. Among the 20 most important ones shown in Table 4, 15 of them are phone boundaries (marked with a '-' between two phones), suggesting that phone boundaries are more useful in this task.

Table 4. Phonemes/boundaries with most splits and gain.

Phonemes/ boundaries	Number of splits	Phonemes/ boundaries	Information gain
UW1-S	4549	B-AW1	13410.1
SH-UH1	4242	UW1-S	10208.9
N-T	2812	V-R	8701.1
B-AW1	2808	SH-UH1	8325.8
EY1	2791	EY1	7965.5
W	2760	UH1-R	6368.7
AO1-NG	2540	AY1	5371.7
M-AO1	2451	N-T	4127.8
UW1-Z	2426	F-EY1	3826.2
UH1-R	2351	ER0	3521.1

4.5. Robustness of LightGBM

During the study we found that LightGBM was extremely robust to overfitting in our task. To test its robustness, we trained a LightGBM model (on the feature set with stress and phone boundaries, 13,090 dimensions) until no more leaves met the split requirements, with the same parameter settings as in Section 4.2.

Figure 4 shows the training and test loss for the first 5,000 trees/iterations. A number of representative loss values and classification accuracies are listed in Table 5. We can see that the test loss reached the minimum at 494 trees where the accuracy was 77.3%. After that, although the test loss was increasing, the test accuracy kept going up until stabilized at around 82.0%, which was significantly better than the accuracy corresponding to the minimum loss. The training terminated after 19,033 trees, when no leaves could be further split, with an accuracy of 82.0%.

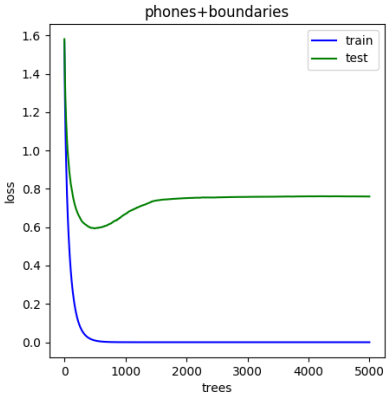

Figure 4. Training and test losses.

Table 5. Loss values and accuracies on test set.

Number of trees	1-best accuracy	2-best accuracy	Test loss	Train loss
200	75.4%	93.0%	0.6888	0.1420
494	77.3%	93.5%	0.5943	0.0096
800	78.2%	93.9%	0.6281	0.0007
2000	80.9%	95.0%	0.7519	8.1e-07
5000	82.1%	95.7%	0.7603	2.8e-07
19033	82.0%	96.3%	0.7616	1.8e-07

The robustness of LightGBM on this task, including the substantial discrepancy between log loss and classification accuracy, deserves further investigation.

5. CONCLUSIONS

We applied LightGBM on softmax-based features extracted from deep neural networks to classify Chinese dialect regions from L2 English speech in a text-dependent manner. The results demonstrated that phone boundaries and vowel stress were useful on this task, with phone boundaries being more useful than vowels and consonants. Our system achieved 68% accuracy on five dialect regions using one sentence and 82% accuracy using 23 words. The results represent a nearly 50% error reduction over a baseline system based on HMM/GMM and forced alignment. We found that the LightGBM classifier was extremely robust for this task.

6. REFERENCES

[1] J. Flege, M. Munro, and D. MacKay, "Factors affecting strength of perceived foreign accent in a second language," *Journal of the Acoustical Society of America,* 97, pp. 3125–3134, 1995.

[2] J. Bloem, M. Wieling, and J. Nerbonne, "Automatically identifying characteristic features of non-native English accents," In M. Côté, R. Knooihuizen, and J. Nerbonne (eds.), *The future of dialects*, pp. 155–173, 2016.

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

[4] R. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," *Speech Communication*, 42, pp. 109–123, 2004.

[5] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for shanghai-accented mandarin," *Interspeech 2005*, pp. 217–220.

[6] M. H. Bahari, R. Saeidi, H. Van hamme, and D. Van Leeuwen, "Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech," *ICASSP 2013*, pp. 7344–7348.

[7] H. Behravan, V. Hautamaki, S. M. Siniscalchi, T. Kinnunen, and C. Lee, "i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24, pp. 29–41, 2016.

[8] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features," *Interspeech 2016*, pp. 2388–2392.

[9] Y. Qian, K. Evanini, P. L. Lange, R. A. Pugh, R. Ubale, and F. K. Soong, "Improving native language (L1) identification with better VAD and TDNN trained separately on native and non-native English corpora," *ASRU 2017*, pp. 606–613.

[10] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *Journal of the Acoustical Society of America,* 89, pp. 874–886, 1991.

[11] J. Sereno, L. Lammers, and A. Jongman, "The relative contribution of segments and intonation to the perception of foreign-accented speech," *Applied Psycholinguistics*, 37, pp. 303–322, 2016.

[12] T. Cho and P. Ladefoged, "Variation and Universals in VOT: Evidence from 18 Languages," *Journal of Phonetics*, 27, pp. 207-229, 1999.

[13] J. Flege, "Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language," *Journal of the Acoustical Society of America*, 89, pp. 395-411, 1991.

[14] J. Yuan and M. Liberman, "Phoneme, Phone Boundary, and Tone in Automatic Scoring of Mandarin Proficiency", *Interspeech 2016*, pp. 2145-2149.

[15] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *Interspeech 2013*, pp. 2306–2310.

[16] Y. Zhang, S. L. Nissen, and A. L. Francis, "Acoustic Characteristics of English Lexical Stress produced by Native Mandarin Speakers," *Journal of the Acoustic Society of America*, 123, pp. 4498-4513, 2008.

[17] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks," *Speech Communication*, 96, pp. 28-36, 2018.

[18] E. Byers and M. Yavas, "Vowel reduction in word-final position by early and late Spanish-English bilinguals," *PloS One*, vol. 12, no. 4, p. e0175226, 2017.

[19] A. Sluijter and V. van Heuven, "Spectral Balance as an Acoustic Correlate of Linguistic Stress". *Journal of the Acoustical Society of America*, 100, pp. 2471- 2485, 1996.

[20] F. Rallo, "Can nonnative speakers reduce English vowels in a native-like fashion? Evidence from L1-Spanish L2-English bilinguals," *Phonetica*, 72, pp. 162–81, 2015.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," In *Advances in Neural Information Processing Systems*, pp. 3149–3157, 2017.

[22] X. Wen and Y. Jia, "Joint Effect of Dialect and Mandarin on English Vowel Production: A Case Study in Changsha EFL Learners," *Interspeech 2016*, pp. 185-189.

[23] HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk .

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," *ASRU 2011*.