INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-guided and Data-driven Decision Trees

Wei Li[1], Kehuang Li[1], Sabato Marco Siniscalchi[1, 2], Nancy F. Chen[3], and Chin-Hui Lee[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[2]Department of Telematics, Kore University of Enna, Enna, Italy
[3]Institute for Infocomm Research, Singapore

{lee.wei, kehle}@gatech.edu, marco.siniscalchi@unikore.it, nfychen@i2r.a-star.edu.sg, chl@ece.gatech.edu

Abstract

We propose a novel decision tree based framework to detect phonetic mispronunciations produced by L2 learners caused by using inaccurate speech attributes, such as manner and place of articulation. Compared with conventional score-based CAPT (computer assisted pronunciation training) systems, our proposed framework has three advantages: (1) each mispronunciation in a tree can be interpreted and communicated to the L2 learners by traversing the corresponding path from a leaf node to the root node; (2) corrective feedback based on speech attribute features, which are directly used to describe how consonants and vowels are produced using related articulators, can be provided to the L2 learners; and (3) by building the phone-dependent decision tree, the relative importance of the speech attribute features of a target phone can be automatically learned and used to distinguish itself from other phones. This information can provide L2 learners speech attribute feedback that is ranked in order of importance. In addition to the abovementioned advantages, experimental results confirm that the proposed approach can detect most pronunciation errors and provide accurate diagnostic feedback.

Index Terms: mispronunciation detection and diagnosis, decision tree, deep neural network (DNN), automatic speech attribute transcription (ASAT), CAPT

1. Introduction

Nowadays, the need to acquire a second language (L2) is gaining increasing importance, and computer assisted language learning (CALL) systems can make second language (L2) learning and teaching more efficient. It is known that the L2 learning process is heavily affected by a well-established habitual perception of sounds and articulatory motions in the learners' primary language (L1), which often causes mistakes and imprecisions in speech production of the L2 learners, e.g., a negative language transfer [1]. Therefore, an important component of CALL is the CAPT subsystem [2], which can be employed to automatically assess L2 learners' pronunciation quality and provide corrective feedbacks.

Automatic speech recognition (ASR) systems can be used in the CAPT module to define ad-hoc confidence scores and provide pronunciation scores to the end learners. For example, the log-likelihood ratio (LLR) was adopted in [3] as a confidence score to measure the difference between native-like and non-native acoustic phone models. Subsequently, "Goodness of Pronunciation (GOP)" [4] and its variants [5, 6, 7] were also proposed to assess the quality of learners' pronunciation. However, when facing lower confidence scores, L2 learners are more likely to feel helpless, because they do not know what is wrong with their pronunciation and how to improve it with only numeric scores.

In [8], it was shown that L2 learners can improve their production of the targeted phones by receiving the feedback about the mispronunciation error at phone level. Nowadays, more and more research work has thus focused on how to use automatic mechanisms to generate finer detection results and corrective information. For example, an extended recognition network (ERN) [9, 10] containing canonical phones and frequent erroneous patterns was proposed to provide diagnostic feedback related to phone substitutions, i.e., phone /A/ is mispronounced as phone /B/. Nonetheless, a major assumption made by providing phone-level feedback is that learners are aware of which articulatory movements (e.g., manner and place of articulation [11, 12]) have to be corrected in order to restore the pronunciation of the canonical phone. Unfortunately, that is a challenging task for L2 beginners, as discussed in [13]. Moreover, Yoon *et al*. pointed out in [14] that there are many "distortion errors", i.e., the erroneous sound is always between two canonical phones, rather than an absolute phoneme substitution. In short, phone-level feedback is not sufficient to give direct corrective instructions and deal with such distortion errors.

In [13, 15, 16], the authors have investigated articulatory-level feedback to overcome the limitations of phone-level feedback. Indeed, it has been reported that L2 learners prefer receiving direct instruction on how to correct mispronunciation at the articulatory level [17, 18]. Unlike conventional acoustic features (e.g., MFCC) used in [15, 16], the speech attribute features describing articulatory characteristic are proposed [13] to directly measure pronunciation quality and give corrective feedbacks based on articulation manner and place. However, the decision boundary of each speech attribute feature in [13] is optimized for its own classification purpose, not directly related to phone-level mispronunciation detection. In this paper we aim to optimize the decision boundaries of each attribute feature for different target phones. Moreover, in order to inspect how inaccurate speech attribute features could lead to mispronunciations, a white-box and interpretable classifier is investigated in this work, instead of relying only on a fully black-box approach, as we did in [13].

http://dx.doi.org/10.21437/Interspeech.2016-517

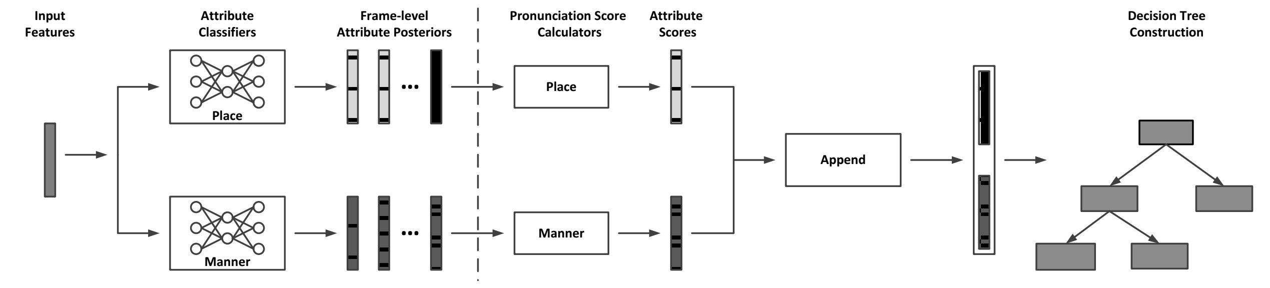

Figure 1: *Overview of mispronunciation detection framework*

In this paper, speech attribute features, such as voicing, aspiration, and manner and place of articulation, are used to construct two types of decision trees [19] to model articulatory characteristics of correct and incorrect phone-level pronunciations. The first type is a knowledge-guided decision tree in which the input uses only speech attribute features that belong to the target phone. The other is a data-driven decision tree, which is built by automatically selected "optimal" speech attribute features. Both decision trees are "readable" models that allows each phone-level mispronunciation to become interpretable by traversing the corresponding path from a leaf node to the root node. We can thus find and analyze how speech attribute features result in mispronunciations. Subsequently, pertinent articulatory-level feedbacks could be formulated to help the L2 learners improve their pronunciations. Finally, through constructing the decision trees, we can automatically learn which speech attributes are more important for distinguishing one phone from others. This information helps rank the corrective feedback by importance. Detailed analysis and examples of the abovementioned decision trees will be given in our experimental section. It should be mentioned that a decision tree, which is able to inspect and analyze how speech features affect mispronunciations, was used with success to detect prosody-level mispronunciations in [20].

Table 1. *Speech attributes and their associated Pinyin initials*

Category	Attribute	Phone set
Place	Labial	B,P,M,F
	Alveolar	D,L,N,T, C,S,Z,
	Retroflex	ZH,CH,SH,R
	Palatal	J,Q,X
	Velar	G,H,K,NG
	N/A	VOWELS
Manner	Stop	B,P,D,T,G,K
	Fricative	F,S,SH,X,H
	Affricative	Z,ZH,C,CH,J,Q
	Nasal	M,N,NG
	Liquid	L, R
	N/A	VOWELS
Aspiration	Aspirated	P,T,K,C,CH,Q
	Unaspirated	B,D,G,Z,ZH,J
	N/A	F,H,L,M,N,R,S,SH,X,NG, VOWELS
Voicing	Voiced	M,N,L,R,NG, VOWELS
	Unvoiced	B,P,M,F,D,T,N,L,G,K,H,J,Q,X ZH,CH,SH,R,Z,C,S
Silence	Silence	SIL

2. Mandarin Phones & Speech Attributes

We focus on European learners of Chinese in this study. So this section will describe Mandarin phones and their corresponding speech attribute features. Each Chinese character corresponds to one spoken syllable, consisting of an initial, usually a consonant, and a final, usually a vowel(s) or vowel(s) followed by a nasal. There are a total of 21 syllable initials and 38 syllable finals. As a preliminary study, we are concerned with mispronunciation of 21 syllable initials, because initial errors are more prone to cause miscommunication in Mandarin when compared to finals [21].

Each initial's articulatory characteristic can be described using its corresponding speech attribute features [11, 12]. For example, when people pronounce the initial "B", the airflow from the lungs is blocked by the place of articulation "labial", causing a pressure difference to build up. Once the closure is opened, the released airflow produces a sudden impulse causing an audible sound, or burst [22]. This whole process is called the articulation manner "stop". As articulation place and manner, speech attribute features "labial" and "stop" are used to describe how "B" is produced. In addition to manner and place of articulation, we also consider voicing and aspiration. When a phone is pronounced, voicing is used to describe if the vocal cords vibrates; whereas, aspiration is used to describe whether there is a brief puff of air after an obstruction is released. Table 1 lists the mapping table between speech attribute and Mandarin initials denoted in Pinyin [23, 24, 25].

3. Overview of Detection Framework

Figure 1 shows the proposed mispronunciation detection framework, which consists of three modules: (i) the speech attribute feature extraction module, which has also been used in the automatic speech attribute transcription (ASAT) paradigm [26]; (ii) the segmental pronunciation score computation module for each speech attribute; and (iii) the phone-dependent decision tree training module based on speech attribute scores.

3.1. Attribute Feature Extraction

Following the ASAT framework [26], speech attributes are extracted using a bank of speech attribute detectors. A context dependent DNN-based classifier is separately trained for each articulatory-motivated attribute category described in Table 1. A window of 11 speech frames centered on the current frame is fed into each DNN classifier, which in turn generates a set of confidence scores in terms of posterior probabilities that the current frame pertains to each possible attribute within the target category. Finally, the frame-level attribute posteriors are sent to the segmental-level pronunciation scoring module.

3.2. Attribute Pronunciation Score Calculation

In this module, regarding to each speech attribute listed in Table 1, the Eq. (1) below is used to calculate pronunciation scores at a segment level by summing up frame-level log posteriors. The higher the pronunciation sore, the more likely the corresponding speech attribute exists in the current segment. For example, when people pronounce initial "B" (also denoted as /B/ for it is also a phone), if the speech attribute "labial" with a very high pronunciation score is observed, we might conclude that the learners' place of articulation of this initial is correct. Namely, the L2 learners correctly use their labial articulators (lips) to block the airflow from the lungs.

$$\log P(p|\boldsymbol{o}; t_s, t_e) = \frac{1}{t_e - t_s} \sum_{t=t_s}^{t_e} \log \sum_{s \epsilon p} P(s|\boldsymbol{o_t}), \quad (1)$$

where unit p is our target speech attribute, $\boldsymbol{o_t}$ is the input feature at frame t; t_s and t_e are the start and end times of unit p, obtained by forced-alignment. $P(s|\boldsymbol{o_t})$ is the frame-level posterior; s is the senone label; $\{s \in p\}$ is the set of senones, corresponding to unit p.

3.3. Phone-dependent Decision Tree Construction

Next, we build a phone-dependent decision tree with the C4.5 algorithm [27, 28] according to the annotated phone label and the calculated pronunciation scores of the speech attributes. Each decision tree was iteratively constructed by using C4.5 algorithm selecting the speech attribute with the highest normalized information gain to split the current node set of samples into subsets. The resulting leaf nodes classify each phone segment into either the correct or incorrect (mispronounced) categories. With respect to the mispronounced category, we can traverse the corresponding path from the current leaf node to the root node to know how this mispronunciation has occurred. Since each non-leaf node of the decision tree is associated with one speech attribute and its corresponding splitting value, we can easily give quantitative and qualitative corrective feedback.

4. Experiments

4.1. Speech Corpora

Two speech corpus, (i) a native speech corpus from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [29], and (ii) a non-native speech corpus a subset of iCALL [30], are mixed to train our speech attribute classifiers. More details can be found in [13].

Our non-native testing set consists of 1662 utterances spoken by 30 leaners from iCALL. There were no speaker overlap between the training and test sets. Furthermore, those 30 learners came from six different countries with six different L1s, including English, French, Spanish, Italian, and Russian. Such L1 diversity makes mispronunciation detection challenging, because the error types made by different L2 learners are influenced by their L1s, as discussed in Section 1.

4.2. Speech Attribute Classification System Setup

The input feature (see Figure 1) is a window of 11 speech frames, each includes a 39-dim MFCC+Δ+ΔΔ vector. After forced-alignment with context dependent (CD) attribute labels, we used these labels to separately train the set of corresponding DNNs. Each DNN has 6 hidden layers each with 2048 sigmoid units. Softmax was employed at the output layer. Except for the CD attribute labels, the DNN configuration used in our study is that provided in the default settings of the Kaldi toolkit [31]. At evaluation time, we used those DNNs to map the input feature vectors into frame-level attribute posteriors. Finally, we computed the pronunciation scores for each speech attribute with Eq. (1).

4.3. Decision Tree Setup

Speech attribute pronunciation scores are first concatenated into a feature vector, and the phone-dependent decision tree is then built according to the phone level labels (correct or incorrect) and corresponding feature vectors. Before constructing a phone-dependent decision tree implemented by WEKA [30], a data imbalance problem had to be addressed. For one target phone, the number of correct samples is much higher than that of the incorrect samples, leading to a biased decision tree with a high precision rate, but a low recall rate. In order to resolve this issue, we use other phones' correctly pronounced samples as target phone's incorrect samples to increase the number of incorrect samples, as is done in [31]. Although there are nearly 20 different speech attributes in our feature vector, the decision tree can automatically select the more important attributes for distinguishing target phone from others. This is called data-driven based decision tree (DDBDT). In addition to DDBDT, this paper also investigates knowledge-guided based decision tree (KGBDT), where the feature vector only contains speech attributes related to the target phone. For example, if our target phone is /B/, only four attributes (labial, stop, unaspirated and unvoiced) are used to construct feature vector, which is used for training KGBDT.

4.4. Evaluation Metrics

As in [15, 16, 34], the following three metrics are used: false acceptance rate (FAR, the ratio between the number of mispronounced phones that are misclassified by the system as correct and the number of all the incorrect phonemes), false rejection rate (FRR, the proportion between the number of correct phones that are misclassified by the system as incorrect and the number of all the correct pronounced phones) and diagnostic accuracy (DA, the percentage of the provided diagnostic feedbacks that are the same as human annotations).

4.5. Experimental Results

KGBDT and DDBDT were separately constructed for each phone. The average mispronunciation detection performance of 21 different phones is summarized in Table 2. In order to give an insight into how a specific decision tree was used to detect mispronunciations and give corrective feedback, we first give two phone-dependent KGBDT examples, as shown in Figures 2 and 3. Next, an additional example is illustrated in Figure 4 to compare and contrast the different mispronunciation detection mechanisms of DDBDT and KGBDT (shown in Figure 3).

Table 2. *Mispronunciation detection performance of two different decision trees on testing set*

	FAR	FRR	DA
KGBDT	17.72%	7.82%	96.8%
DDBDT	17.24%	8.32%	96.2%

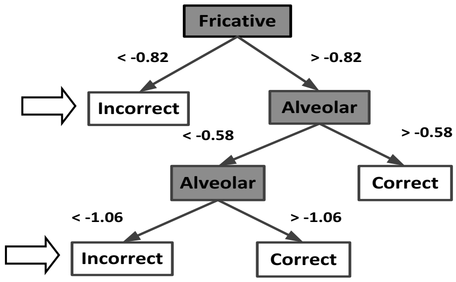

Figure 2: *The KGBDT of phone S*

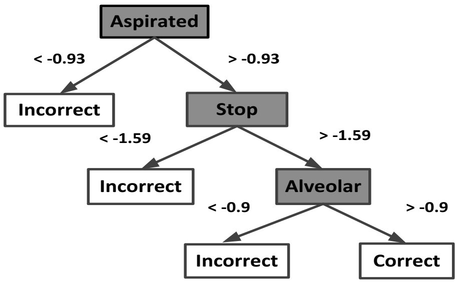

Figure 3: *The KGBDT of phone T*

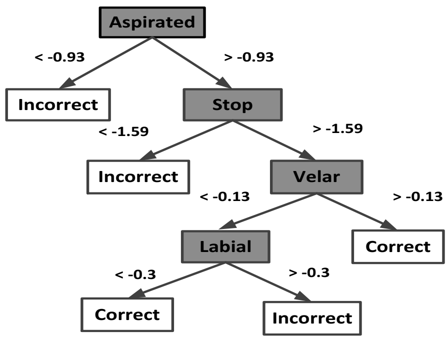

Figure 4: *The DDBDT of phone T*

4.6. Analysis of Results

A 10-fold cross validation was conducted to prune the decision tree to avoid overfitting so that only the most important speech attributes were selected. Take phone /S/ in Figure 2 for example, only "fricative" and "alveolar" were used to build the decision tree. This observation is consistent with Table 1, where the intersection between "fricative" and "alveolar" sets only contains the phone /S/. Moreover, we can see that the higher a node is, the more important the speech attribute associated with this node is for distinguishing /S/ from other phones. Therefore, the speech attribute priority for phone /S/ is: fricative (manner) > alveolar (place). Facing phone /S/ mispronunciation caused by a combination of inaccurate manner and place, manner related feedback should be given priority.

Each non-leaf node in the decision tree is associated with one speech attribute and its corresponding splitting value. Thus how each input feature leads to a mispronunciation can be quantitatively and qualitatively analyzed. Take phone /T/ for example, the decision tree first checks the pronunciation score of ***aspiration***, if the score is smaller than -0.93 (this splitting value or decision boundary is automatically learned by using C4.5 algorithm and optimized for each target phone), we can conclude that the lack of aspiration contributes to a poor pronunciation. Obviously, the speech attribute associated with each tree node gives a qualitative description about which speech attribute results in a mispronunciation. Moreover, the pronunciation score of each speech attribute itself gives a quantitative description of the mispronunciation tendencies, i.e., the lower the attribute score is, the more likely this attribute is not well pronounced.

By traversing a corresponding path from a leaf node to the root node, each mispronunciation can be interpreted, i.e., the reason why this mispronunciation occurs and how to correct it can be communicated to the L2 learner. KGBDT classifies the phone segment as an incorrect category when target phone-related attributes' pronunciation scores are small, which shows that expected articulatory manner or place is not observed in the current phone segment. In Figure 2, there are two possible error patterns, marked by arrows. The upper arrow points to phone /S/ being mispronounced due to articulation manner, i.e., the frication is not pronounced well. Consequently, an articulation manner-based feedback can be given. The second arrow shows that phone /S/ is mispronounced due to the articulation place: the alveolar attribute has a low score although its articulation manner is correct. Therefore, articulation place-based feedback can be given, e.g., your tongue could be closer to the superior alveolar ridge.

Through comparing Figures 3 and 4, we found that speech attributes, such as velar and labial, were used to construct DDBDT for phone /T/. These speech attributes do not belong to phone /T/, as shown in Table 1. DDBDT thus classifies the phone segment into the incorrect category when these attribute scores are high, which indicates that unexpected articulatory manner or place is observed in the current phone segment.

From Table 2, we can find that both DDBDT and KGBDT can detect most pronunciation errors and provide accurate diagnostic feedback. Although FAR is higher than FRR, as reported in previous work [15, 16, 34], higher FAR and lower FRR will bring about more confidence to L2 learners to study foreign languages.

5. Conclusion

In this paper, speech attribute based decision tree was proposed to detect phonetic segmental mispronunciations and provide articulatory level feedback based on manner and place of articulation. Compared with conventional score-based systems, our approach can tell the L2 learners why their mispronunciations occur and how to correct them. While giving more intuitive and instructive feedback, our system also achieves a high mispronunciation detection performance. For future work, the combination of DDBDT and KGBDT will be investigated. Moreover, current attribute-level pronunciation score is just calculated using simple summation of frame-level posteriors. In the future, advanced method such as dynamic programing [35] and articulatory landmark mechanism [36], will also be studied. Finally, due to the limited page size, this preliminary work only reports the average mispronunciation detection performance and analyzes only a few trees in Figures 2-4, more individual phone-dependent detection performance will be analyzed in the near future with a full report.

6. Acknowledgment

The first author was partially supported by a grant from the China Scholarship Council. The authors would like to thank Zhen Huang for his valuable discussions.

7. References

[1] H. Meng, “Developing speech recognition and synthesis technologies to support computer-aided pronunciation training for Chinese learners of English,” In *Proc. 23rd Pacific Asia Conference on Language, Information and Computation*, 2009..

[2] J.-M. Kim, C. Wang, M. Peabody, and S. Seneff, “An interactive English pronunciation dictionary for Korean learners.,” in *Proc. Interspeech, 2004.*

[3] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,” in *Proc. Eurospeech*, 1999.

[4] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[5] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W. Ye, “Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation,” in *Proc. ICASSP*, 2007.

[6] S. Wei, G. Hu, Y. Hu, R.H. Wang, “A new method for mispronunciation detection using support vector machine based on pronunciation space models,” *Speech Communications*, vol. 51, no. 10, pp. 896–905, 2009.

[7] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, “Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers,” *Speech Communication*, 67, pp. 154-166, 2015.

[8] A. Neri, C. Cucchiarini ,and H. Strik, “ASR-based corrective feedback on pronunciation: does it really work?,” in Proc. Interspeech, 2006.

[9] H. Meng, Y. Lo, L. Wang, and W. Yiu, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *Proc. ASRU*, 2007.

[10] W. K. Lo, S. Zhang and H. Meng, “Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System,” in *Proc. Interspeech*, 2010.

[11] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, *MIT Press*, 2000.

[12] G. Fant, *Speech Sounds and Features*. Cambridge, MA, MIT Press, 1973.

[13] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, “Improving Non-native Mispronunciation Detection And Enriching Diagnostic Feedback With DNN-BASED Speech Attribute Modeling”, in *Proc. ICASSP*, 2016.

[14] S. Yoon, M. Hasegawa-Johnson, and R. Sproat, “Landmark Based Automated Pronunciation Error Detection”, in *Proc. Interspeech*, 2010.

[15] R. Duan, *et al*, “A Preliminary Study on ASR-based Detection of Chinese Mispronunciation by Japanese Learners,” in *Proc. Interspeech*, 2014.

[16] Y. Gao, *et al*, “A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network,” in *Proc. Interspeech*, 2015.

[17] O. Bälter, O. Engwall, A. Öter, and H. Sidenbladh-Kjellström, “Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction,” in *Proc. 7th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 36–43, Oct. 2005.

[18] O. Engwall, “Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher,” *Computer Assisted Language Learning*, 25:37–64, 2012.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*, Wadsworth,1984.

[20] Y. H Guan, J. C. Chen, H. C. Liao, *et al*, “Decision Tree Based Tone Modeling with Corrective Feedbacks for Automatic Mandarin Tone Assessment,” in *Proc. Interspeech*, 2010.

[21] J.-S. Zhang, W. Li, et al, “A Study On Functional Loads of Phonetic Contrasts Under Context Based On Mutual Information of Chinese Text And Phonemes,” in *Proc. ISCSLP*, 2010.

[22] https://en.wikipedia.org/wiki/Stop_consonant.

[23] https://en.wikipedia.org/wiki/Pinyin.

[24] 林焘, 王理嘉, 语音学教程[M]. 北京大学出版社, 2013

[25] 张家騄, 汉语人机语音通信基础[M]. 上海科学技术出版社, 2010.

[26] C.-H. Lee and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089– 1115, 2013.

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[28] T. Hastie, R. Tibshirani and J. Friedman, *Elements of Statistical Learning*, Springer, 2009.

[29] S. Gao, *et al*, “Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR,” in *Proc. ICSLP*, 2000.

[30] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, “iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent,” in *Proc. Interspeech*, 2015.

[31] 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,” in *Proc. ASRU*, 2011.

[32] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (2nd Edition), Morgan Kaufmann Publishers, 2005.

[33] J. Jiang and B. Xu, “Exploring The Automatic Mispronunciation Detection of Confusable Phones for Mandarin,” in *Proc. ICASSP*, 2009.

[34] K. Li and H. Meng, “Mispronunciation Detection and Diagnosis in L2 English Speech Using Multi-Distribution Deep Neural Networks,” in *Proc. ISCSLP*, 2014.

[35] H. Li, S. Wang *et al*, “High Performance Automatic Mispronunciation Detection Method Based on Neural Network and TRAP Features,” in *Proc. Interspeech*, 2009.

[36] Y. Xie, M. Hasegawa-Johnson, L. Qu, and J. Zhang, “Landmark OF Mandarin Nasal Codas And Its Application In Pronuncaition Error Detection,” in *Proc. ICASSP*, 2016.