7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features

Hyuksu Ryu[1], *Minhwa Chung*[1]

[1]Department of Linguistics, Seoul National University, Seoul, Republic of Korea

oster01@snu.ac.kr, mchung@snu.ac.kr

Abstract

This paper proposes a method to provide an articulatory diagnosis of English produced by Korean learners using articulatory Goodness-Of-Pronunciation (aGOP) features, which are based on the distinctive feature theory in phonology. Previous studies on mispronunciation diagnosis have mainly dealt with pronunciation errors at phone-level. They inform learners of which phone is recognized as a diagnosis, when the corresponding segment is realized as a mispronunciation. However, to provide learners more effective corrective feedback, diagnosis had better be performed at articulatory-level, such as place and manner of articulation, rather than at phone-level. This study aims to provide automatic articulatory diagnosis using articulation-based confidence scores. At first, the speech of learners is forced-aligned and recognized to compute the GOP and aGOPs. When the forced-aligned segment is a consonant, articulatory diagnosis is conducted in three articulatory categories: voicing, place of articulation, and manner of articulation. Otherwise, diagnosis is performed in terms of rounding, height, and backness corresponding to articulatory characteristics of vowels. Experimental results show that F1 scores for voicing, place, and manner corresponding to consonants are 0.828, 0.754, and 0.781, respectively, whereas F1 score for rounding, height, and backness corresponding to vowels are 0.843, 0.782, and 0.824, respectively. These results indicate that the proposed method yields effective articulatory diagnosis.

Index Terms: articulatory Goodness-Of-Pronunciation, mispronunciation diagnosis, CAPT, English produced by Korean learners

1. Introduction

Corrective feedbacks explaining where learners are making pronunciation errors and how to correct them are essential for Computer-Assisted Language Learning (CALL) and Computer-Assisted Pronunciation Training (CAPT) systems [1]. That is, mispronunciation detection and diagnosis of mispronunciation using speech technology are necessary for conducting effective CALL/CAPT.

There have been several studies to detect pronunciation errors of learners [2][3][4][5]. The study of [2] suggested an extended recognition network (ERN), which expands pronunciation dictionaries of learners by predicting frequent erroneous pronunciation sequences. When the erroneous pronunciation sequences are recognized, it is considered that learners made pronunciation errors. However, ERN approach has difficulties in identifying mispronunciation patterns that learners frequently show in terms of each L1-L2 pair. Also, it is difficult to guarantee that ERN covers most of the possible mispronunciations [6].

Another approach for detecting pronunciation errors is using confidence scores such as Goodness-Of-Pronunciation (GOP) [3][4]. Confidence score-based approach has virtues that it has L1/L2 independence and it is easy to compute [7]. However, it is difficult to provide corrective feedback, since learners do not know how to interpret with confidence scores alone and improve their pronunciation with the scores. Diagnosis for the detected pronunciation errors was not provided in these researches.

Several previous studies [6][8][9] conducted diagnosis for mispronunciation as well as detection of pronunciation errors. Li et al. [6] suggested multi-distribution DNN (MD-DNN) by using acoustic features, graphemes, and canonical pronunciation as inputs of DNN to predict actual pronunciation of learners. When the predicted pronunciation is different from the canonical pronunciation, it is considered as a mispronunciation. Wang and Lee [8] proposed hierarchical multi-layer perceptrons (MLPs). First MLP is binary and classifies each frame as correct or incorrect. Then, second MLP classifies each frame identified as incorrect by the first MLP into one of the Error Patterns (EP) as diagnosis. Xie et al. [9] extracted landmark features for nasal codas spoken by learners of Chinese, and detected pronunciation errors by applying SVM. In these studies, diagnosis is performed in a hierarchical way as shown in Figure 1. At first, the pronunciation error detector that they proposed distinguishes between mispronunciation and correct pronunciation. In addition to binary mispronunciation detector, diagnosis is carried out for instances which are correctly detected as mispronunciations (True Rejection in Figure 1). The diagnosis performance is reported by diagnosis error rate (DER), which is defined as the percentage of incorrectly recognized phones among correctly identified as a mispronunciation.

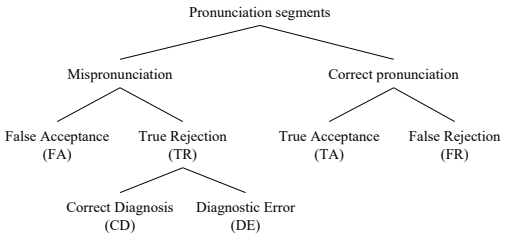

Figure 1: *Hierarchical structure for mispronunciation detection and diagnosis presented in [6][8]*

These hierarchical approaches for diagnosis have a limitation that they provide diagnosis at phone level only. For example, let's assume a learner pronounced a word 'give' /gɪv/ as /gɪb/. When the CAPT system detects pronunciation error at coda position and recognizes the phone as /b/, the system reports a diagnosis of /v/→/b/. However, for more effective corrective feedback or learners, it had better provide diagnosis information at articulatory level such as voicing, place, and

10.21437/SLaTE.2017-12

manner of articulation, not just at phone level. In the example above, it would be more effective if the system tells a learner that there is a diagnosis of fricative→stop at articulatory level of the manner of articulation, rather than diagnosis of /v/→/b/ at phone level. In addition, the diagnosis procedure presented in the existing studies is performed in two steps; detection and recognition. Since mispronunciation detection errors and recognition errors are piled up, diagnosis accuracy could be degraded. For instance, diagnosis will be incorrect when mispronunciation detection shows false acceptance or false reject, as well as when the segment produced by learners is incorrectly recognized, even if the system detects mispronunciation correctly. When considering limitations mentioned above, it will be helpful when mispronunciation diagnosis is performed at articulatory level.

There are several studies utilizing articulatory features for pronunciation assessment and mispronunciation detection. The study of Ryu and Chung [10] proposes articulatory Goodness-Of-Pronunciations (aGOPs) as novel features for pronunciation assessment in English spoken by Korean learners. Furthermore, Li et al. [5] extended GOP into speech attributes to detect mispronunciation of onset consonants in learners' Chinese by decision trees. By using speech attributes, they have shown the possibility to provide corrective feedback to improve pronunciation. However, there is a limitation that they did not present the actual experiments about diagnosis or corrective feedback.

The goal of this paper is to propose a method to provide an articulatory diagnosis in English produced by Korean learners using articulatory Goodness-Of-Pronunciation (aGOP) based on the distinctive feature theory. The remainder of this study is organized as follows. In Section 2, we describe theoretical backgrounds and computational models regarding articulatory features. Section 3 presents the details of methods for diagnosis modeling at articulatory level. In Section 4, we provide quantitative analysis of salient mispronunciation patterns in English by Korean learners. The results of experiments for articulatory diagnosis are presented in Section 5, which is followed by conclusion in Section 6.

2. Articulatory features

Phonemes are differentiated by phonological features [11]. For example, /p/ and /b/ are distinguished by the phonological feature of voicing; /b/ has the feature of voicing ([+voice]), while /p/ does not ([-voice]). In other words, the voicing feature is a phonological criterion which makes the two phonemes 'distinctively' different. The minimum unit that discriminates phonemes in a language is called distinctive features. Therefore, phonemes are represented by a set of distinctive features, which is called natural class [11]. Chomsky and Halle [12] explained that the distinctive features have only two values of presence or absence such as [+voice] or [-voice], and that it is possible to distinguish various phonemes by using a natural class.

The study of [10] suggested articulatory-based GOPs for pronunciation assessment. The aGOP features are used to compare articulatory characteristics between natives and learners regarding the articulatory attributes. The aGOP features are computed based on phone GOP [4] as the normalized posterior probability for each articulatory attribute as shown in (1).

$$aGOP^k(p) \equiv \log \left[\frac{P(\mathbf{o_p}|q^k)}{\max_i P(\mathbf{o_p}|q_i^k)} \right] /N(p) \qquad (1)$$

, where k and q^k denote the sort of articulatory attribute presented in [12] and the canonical value of the k^{th} articulatory attribute at the position of the forced-aligned target phone p, respectively. $N(p)$ and $P(\mathbf{o_p}|q^k)$ mean the number of frames composing target phone p and the probability of observing $\mathbf{o_p}$ given q^k, respectively. The higher the value of GOP and aGOP, the higher the likelihood that learners utter the canonical phone.

Articulatory features are classified into three categories such as manner, place, and laryngeal [11]. Articulatory features used in this study consist of 24 attributes such as sonorant and continuant (manner), labial and round (place), and voice (laryngeal). The details of articulatory features and the corresponding phones are presented in [10].

The study of [10] used aGOP features as novel predictors for automatic pronunciation assessment of English produced by Korean learners. In addition to well-known features such as Rate of Speech (ROS) in previous studies, [10] showed that the performance was improved by including aGOP features and applying the statistical method such as best subset selection.

The study of [5] also proposed similar articulatory features for mispronunciation detection of Mandarin learners. However, they limited the focus of the study on onset consonants, while this paper deals with every phone including vowels. Also, they performed the articulatory modeling based on categories [13]. For example, the articulatory model of [13] classifies the category of 'place' into multiple attributes, such as bilabial, alveolar, and dental. This kind of modeling has a limitation that it shows low performance when the category has multiple attributes, such as place, as discussed in [13]. On the contrary, aGOP features suggested in [10] and this study are computed based on each attribute, not category. For instance, our model classifies the articulatory attribute of 'alveolar' into a binary value of presence or absence. Furthermore, we specify articulatory attributes in more details by using the phonological theory. Therefore, we can compute more various information and use them for mispronunciation diagnosis.

3. Method

3.1. Corpus and annotation

We use the ETRI (Electronics and Telecommunications Research Institute) English speech corpus produced by Korean speakers. The corpus consists of 21,110 sentences (21 hours) spoken by 151 learners.

Ten Korean annotators participate in phone-level transcription by the procedure in [14]. They are native Koreans who can speak English as L2 and have expertise in phonetics or phonology. Transcribers present 88.13% of the phone-level agreement, which means the annotation is reliable [15]. The result of transcriptions shows 6.32% of overall mispronunciation rate.

3.2. Acoustic model

We use an acoustic model using WSJ corpus of 37,000 sentences spoken by North American native English speakers [16]. The CD-DNN-HMM acoustic model [17] is trained using 39-dim. MFCC+Δ+$\Delta\Delta$ features. In addition to phone acoustic model, we separately train the articulatory attribute models to calculate aGOPs. The parameter configuration and architecture of DNN used in this study is provided by the default configurations of the Kaldi toolkit [18].

3.3. Diagnosis modeling

Mispronunciation diagnosis at the articulatory level suggested in this study is performed as shown in Figure 2. We use GOP [4] and aGOPs [10] as predictors for articulatory diagnosis. Based on forced-alignment, we examine whether the corresponding segment is a consonant or a vowel. When the segment is a consonant, mispronunciation diagnosis is performed in voicing, place of articulation, and manner of articulation, since consonants are phonetically classified in terms of these three dimensions at articulatory level [19]. On the contrary, if the segment is a vowel, mispronunciation diagnosis is carried out at an articulatory level in terms of rounding, height, and backness, corresponding to articulatory characteristics of vowels. Although phonetic transcription of mispronunciation can be converted into articulatory transcription by rules and phone-level mispronunciation can be detected using GOP, rule-based articulatory conversion could be inaccurate, when phone-level mispronunciation detection shows false results. Thus, we perform separate diagnosis modeling at articulatory level using aGOP features as well as GOP.

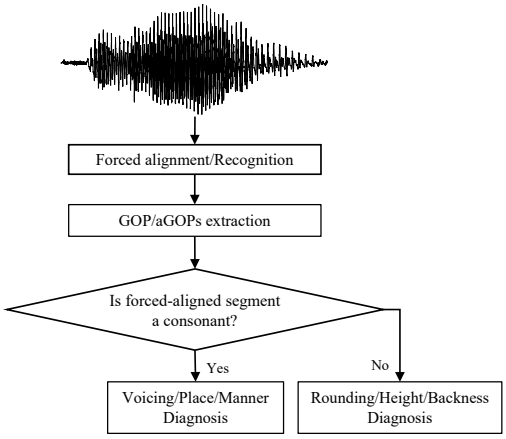

Figure 2: *The general scheme for mispronunciation diagnosis at articulatory level*

For consonants, mispronunciation diagnosis at articulatory level is carried out separately in terms of voicing, place, and manner as demonstrated in Figure 3. As stated above, GOP and 24 aGOPs for the associated articulatory attributes are used as predictors for diagnosis modeling. The binary value of correctness or incorrectness is used for the response variable of each diagnosis model by comparing the canonical pronunciation and the actual realization in terms of articulatory levels. For example, let's assume an observation of /θ/→/s/, which means the canonical pronunciation is /θ/ and the actual realization is /s/. At articulatory level, the canonical pronunciation /θ/ is voiceless dental fricative, while the actual realization /s/ is voiceless alveolar fricative. In this case, the value of 'correctness' is assigned to the articulatory level of voicing and manner of articulation, since these articulatory levels of the canonical pronunciation and the actual realization are matched as voiceless and fricative, respectively. On the other hand, in terms of place of articulation, the canonical and the actual are dental and alveolar, respectively, so the value of incorrectness is given to the articulatory level of place.

Feed-Forward Neural Network (FFNN) is applied to each diagnosis modeling of mispronunciation at articulatory level. All the diagnosis models are implemented using TensorFlow [20]. The FFNNs are configured with fully connected hidden layers and a softmax output layer. We tune the number of hidden layers between 3 and 7 and the number of nodes in the range [128, 256, 512, 1024]. Each layer is followed by an Exponential Linear Unit (ELU) [21], which is known to show better performance compared to Rectified Linear Unit (ReLU). We also apply batch normalization [22] and dropout with rate of 0.5. Weights are initialized in the form of He initialization [23]. We use a learning rate of 0.005 with the Adam optimizer [24] which uses cross-entropy for the loss function. The FFNNs are trained for 10,000 epochs with a batch size of 100 and early-stopping is performed based on the accuracy of the validation set.

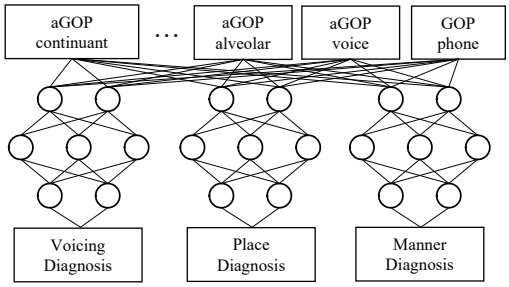

Figure 3: *Mispronunciation diagnosis modeling for consonants at articulatory level*

For vowels, mispronunciation diagnosis at articulatory level is separately performed in terms of rounding, height, and backness as shown in Figure 4. For instance, when there is an observation of /ɑ/→/o/, the canonical pronunciation /ɑ/ is an unrounded low back vowel, while the actual realization /o/ is a rounded mid back vowel. Therefore, 'incorrectness' is assigned to rounding and height where the value of the canonical (unrounded low) does not match with that of the actual realization (rounded mid). On the contrary, the articulatory level of backness has 'correctness', because both the canonical and the actual are back vowel. The detailed configurations of diagnosis modeling are identical to that of consonants as mentioned above.

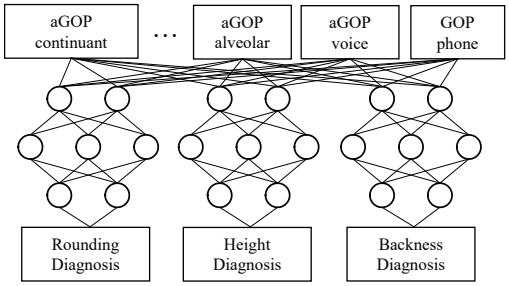

Figure 4: *Mispronunciation diagnosis modeling for vowels at articulatory level*

4. Quantitative analysis of salient mispronunciations

Prior to experiments for mispronunciation diagnosis at the articulatory level, we quantitatively analyze the tendency of mispronunciation patterns in English produced by Korean learners using the corpus mentioned in 3.1. By the result of corpus anal-

ysis, there are entirely 602,810 phones. Among them, 38,100 phones are marked as incorrect, which shows 6.32% of variation rate. We choose phones with the variation rate above the overall variation rate (6.32%) and the entire frequency over 500 instances as salient phones. Details of salient phones and the corresponding variation frequency and variation rates are presented in Table 1. The sum of variation frequencies of nine salient phones is 26,553 instances, occupying around 70% of entire variation frequency (38,100 instances).

Table 1: *Frequency and rate of salient phones*

Category	Phone	Entire Freq.	Variation Freq.	Variation rate
Consonant	/z/	12,603	3,425	27.18%
	/ð/	14,967	3,488	23.30%
	/θ/	3,392	613	18.07%
	/v/	9,492	1,434	15.11%
	/d/	23,814	2,999	12.59%
	/t/	49,804	4,206	8.45%
Vowel	/ɑ/	11,327	2,381	21.02%
	/ɔ/	10,204	1,690	16.56%
	/ʌ/	44,490	6,317	14.20%

To determine the most noticeable variations which appear only in the learners' speech, we choose variations which are more frequent than in native speech [25] among salient phones in Table 1. The most noticeable variations are determined by adopting the analysis of [25], and we consider the most noticeable variations as salient mispronunciation patterns. Salient variations for /d/ and /t/ are deletion in consonant clusters (ex. 'just' /ʤʌsṯ/→/ʤʌs/) and flapping (ex. 'body' /bɑdi/→/bɑṛi/). Because such variations frequently appear in natives' speech [25], they are not included in the list of salient mispronunciation patterns. Details of salient mispronunciations in consonants and vowels are shown in Table 2 and Table 3, respectively. The parenthesis in the Canonical column (Canon.) denotes the corresponding frequency of the canonical phones where variations occur.

Table 2: *Salient mispronunciation patterns in consonants in terms of articulatory level*

Level	Canon.	Act.	Example	Freq.	Ratio
Voicing	/z/ (3,425)	/s/	does /dʌz̲/→/dʌs̲/	2,935	85.69%
	/v/ (1,434)	/f/	love to /lʌv̲ tʊ/→/lʌf̲ tʊ/	305	21.27%
Place	/ð/ (3,488)	/d/	this /ð̲ɪs/→/d̲ɪs/	3,235	92.75%
	/θ/ (613)	/s/	thing /θ̲ɪŋ/→/s̲ɪŋ/	213	34.75%
		/t/	thank /θ̲æŋk/→/t̲æŋk/	331	54.00%
Manner	/ð/ (3,488)	/d/	this /ð̲ɪs/→/d̲ɪs/	3,235	92.75%
	/θ/ (613)	/s/	thing /θ̲ɪŋ/→/s̲ɪŋ/	213	34.75%
	/v/ (1,434)	/b/	give /gɪv̲/→/gɪb̲/	766	53.42%

In consonants, salient mispronunciation patterns are classified into voicing, place, and manner at articulatory level. As presented in Table 2, in terms of voicing, learners show devoicing patterns such as /z/→/s/ and /v/→/f/. Devoicing on /z/ mainly occurs at word final, while devoicing on /v/ is mostly caused by regressive assimilation. In terms of place of articulation, salient mispronunciation pattern is shown as a variation from dental to alveolar. Finally, in terms of manner of articulation, Korean learners have difficulties of producing English fricatives. These mispronunciation patterns are the cases where the learners failed to produce fricatives which do not exist in native language and substituted them with their corresponding stops.

Table 3: *Salient mispronunciation patterns in vowels in terms of articulatory level*

Level	Canon.	Act.	Example	Freq.	Ratio
Round	/ɑ/ (2,381)	/oʊ/	project /prɑ̲ʤɛkt/→ /proʊ̲ʤɛkt/	295	12.39%
Height	/ɑ/ (2,381)	/oʊ/	project /prɑ̲ʤɛkt/→ /proʊ̲ʤɛkt/	295	12.39%
	/ɔ/ (1,690)	/oʊ/	law /lɔ̲ː/→/loʊ̲/	735	43.49%
	/ʌ/ (6,317)	/ɑ/	another /ənʌ̲ðɚ/→ /ənɑ̲ðɚ/	1,106	17.51%
		/æ/	and /ʌ̲nd/→/æ̲nd/	1,030	16.31%
Backness	/ʌ/ (6,317)	/æ/	and /ʌ̲nd/→/æ̲nd/	1,030	16.31%
		/ɛ/	Helen /hɛlə̲n/→/hɛlɛ̲n/	654	10.35%

In vowels, salient mispronunciation patterns are classified into rounding, height, and backness at articulatory level. As presented in Table 3, in terms of rounding, learners tend to realize unrounded /ɑ/ as rounded vowel /oʊ/. In terms of height, there are two variations, one is raising from low to mid vowels, and the other is lowering from mid to low vowels. In backness, all salient patterns occur as fronting from back to front vowels.

Substitution of /ɔ/ for /oʊ/ occurs because /ɔ/ does not exist in native language and learners tend to replace it with the most similar phoneme /oʊ/ which exists in L1. The remaining mispronunciation patterns in Table 3 are explained as orthographic interference, in which inappropriate inference from the spelling influences the learners' pronunciation [25]. For instance, the learners substituted /oʊ/ for /ɑ/, as in /proʊʤɛkt/ for the word 'project' /prɑʤɛkt/, where the variation is influenced from the grapheme 'o' for /ɑ/.

5. Experiments

5.1. Experimental setup

Based on the quantitative analysis of salient mispronunciation patterns in Section 4, we perform an articulatory diagnosis experiment on the seven salient phones. As can be seen in Section 4, the number of the correctly pronounced phone is much larger than the number of mispronounced, which could make trained models biased. To prevent the bias problem, we adopt other phones' correctly pronounced observations as mispro-

nounced samples of the target phone as much as the difference between the number of correct instances and the number of incorrect instances to make a balance [5].

We split balanced observations of each phone into the training and the test set with a ratio of 8:2. Furthermore, 20% of the training set is assigned as a validation set to determine hyperparameters of FFNN, such as the number of layers and the number of nodes. We do not include artificially augmented instances as mentioned above in the test set, maintaining the balance between the number of correct and incorrect instances. The details of the training, validation, and test sets are presented in Table 4.

Table 4: *Details of training, validation, and test sets in terms of salient phones*

Cat.	Phone	Training (Validation)	Test	Total
consonant	/z/	14,685 (2,937)	3,671	18,356
	/ð/	18,367 (3,673)	4,591	22,958
	/θ/	4,447 (889)	1,111	5,558
	/v/	12,893 (2,578)	3,223	16,116
vowel	/ɑ/	14,314 (2,862)	3,578	17,892
	/ɔ/	13,624 (2,724)	3,405	17,029
	/ʌ/	61,077 (12,215)	15,269	76,346

5.2. Experimental results

Table 5 shows the performance of mispronunciation diagnosis for salient consonants in terms of articulatory levels for the test set. By the results, mispronunciation diagnosis at articulatory level using articulatory features shows more than 70% accuracy and 0.75 F1 scores for all articulatory levels in average. The results denote that the proposed method using articulatory features is effective for articulatory diagnosis for consonants.

Although the proposed method shows high accuracy in average at articulatory levels, the articulatory level of place for /ð/ and /θ/ presents slightly lower performance than average. These phones are the only dental sounds at the articulatory level of place. Since they are inter-dental fricative, they all have relatively small amount of amplitude, which make them difficult for them to distinguish mispronunciation. These factors appear to affect the performance.

Table 5: *Diagnosis performance for salient consonants in terms of articulatory levels*

Phone	Level	Accuracy	Precision	Recall	F1
/z/	voicing	70.14%	0.683	0.890	0.773
	place	85.57%	0.857	0.877	0.867
	manner	79.38%	0.821	0.825	0.823
/ð/	voicing	83.60%	0.837	0.898	0.866
	place	60.50%	0.623	0.670	0.646
	manner	62.13%	0.632	0.852	0.726
/θ/	voicing	79.68%	0.814	0.857	0.835
	place	65.83%	0.672	0.697	0.684
	manner	71.76%	0.761	0.830	0.794
/v/	voicing	80.18%	0.821	0.859	0.840
	place	75.43%	0.795	0.842	0.818
	manner	71.40%	0.751	0.815	0.782
average	voicing	78.40%	0.789	0.876	0.828
	place	71.83%	0.737	0.772	0.754
	manner	71.17%	0.741	0.831	0.781

Table 6 presents a performance of mispronunciation diagnosis for salient vowels in terms of articulatory levels for the test set. As well as consonants, vowels also show high average articulatory diagnosis performance, except in height. Other levels have more than 70% accuracy, but height shows 65% accuracy in average. In the case of vowels, the training sets contain variations from salient phones to diphthongs, such as /ɑ/→/aɪ/. Articulatory characteristics of diphthongs vary during pronunciation of a vowel. For example, the tongue in /aɪ/ moves from low to high in terms of height. Also, /ɔɪ/ shows changes of articulation in height (mid→high) and backness (back→front). Thus, such drastic changes in the articulation of a vowel could affect low performance.

Table 6: *Diagnosis performance for salient vowels in terms of articulatory levels*

Phone	Level	Accuracy	Precision	Recall	F1
/ɑ/	rounding	83.79%	0.839	0.888	0.863
	height	59.60%	0.658	0.853	0.743
	backness	80.94%	0.811	0.896	0.851
/ɔ/	rounding	70.43%	0.731	0.855	0.788
	height	70.93%	0.759	0.898	0.823
	backness	75.98%	0.753	0.887	0.815
/ʌ/	rounding	88.25%	0.883	0.865	0.874
	height	65.16%	0.693	0.887	0.778
	backness	65.59%	0.761	0.859	0.807
average	rounding	80.82%	0.818	0.869	0.843
	height	65.23%	0.703	0.879	0.782
	backness	74.17%	0.775	0.881	0.824

6. Conclusion

In this paper, we proposed a method to provide an articulatory diagnosis of English produced by Korean learners using articulatory Goodness-Of-Pronunciation (aGOP) features, which are based on the distinctive feature theory in phonology. So far, previous studies regarding mispronunciation diagnosis have limitations that they have carried out diagnosis at phone level. To provide effective corrective feedback, mispronunciation diagnosis had better be performed at the articulatory level, rather than at phone level. We applied different models of articulatory diagnosis depending on the consonants and vowels considering articulatory characteristics. Mispronunciation diagnosis for consonants was conducted in three articulatory levels: voicing, place of articulation, and manner of articulation. On the other hand, articulatory diagnosis for vowels was performed in terms of rounding, height, and backness. Furthermore, we quantitatively analyzed salient mispronunciation patterns in English produced by Korean learners using corpus-based analysis. By the results, the proposed method for articulatory diagnosis presented more than 70% accuracy and 0.75 of F1 scores in average for all articulatory levels except height in vowels. It is noteworthy that these results indicate the proposed method yields effective mispronunciation diagnosis at articulatory level.

However, there is a limitation that the proposed method only decides that the pronunciation is correct or not at the articulatory level and does not provide corrective feedback on how to correct the pronunciation at articulatory level. Thus, in future work, we will extend the experiment to provide corrective feedback at articulatory level as well as mispronunciation diagnosis.

7. Acknowledgements

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2016-0-00464) supervised by the IITP (Institute for Information & communications Technology Promotion).

8. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] A. M. Harrison, W.-K. Lo, X. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Proc. Speech and Language Technology in Education (SLaTE 2009)*, Birmingham, UK, pp. 45–48, 2009.

[3] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 2, Münich, Germany, pp. 1471–1474, 1997.

[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] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees," in *Proc. Interspeech 2016*, San Fransisco, CA, pp. 3127–3131, 2016.

[6] K. Li, X. J. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[7] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2016)*, Jeju, Republic of Korea, pp. 1–7, 2016.

[8] Y. B. Wang and L. S. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *Ieee-Acm Transactions on Audio Speech and Language Processing*, vol. 23, no. 3, pp. 564–579, 2015.

[9] Y. Xie, M. Hasegawa-Johnson, L. Qu, and J. Zhang, "Landmark of Mandarin nasal codas and its application in pronunciation error detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, pp. 5370–5374, 2016.

[10] H. Ryu and M. Chung, "Automatic pronunciation assessment of English produced by Korean learners using articulatory features," *Phonetics and Speech Sciences*, vol. 8, no. 4, pp. 103–113, 2016.

[11] B. Hayes, *Introductory phonology*. Malden, MA; Oxford: Wiley-Blackwell, 2009.

[12] N. Chomsky and M. Halle, *The sound pattern of English*. New York: Harper & Row, 1968.

[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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, pp. 6135–6139, 2016.

[14] H. Ryu, K. Lee, S. Kim, and M. Chung, "Improving transcription agreement of non-native English speech data transcribed by non-natives," in *Proc. Speech and Language Technology in Education (SLaTE 2011)*, Venice, Italy, pp. 61–64, 2011.

[15] H. Ryu, S. Kim, and M. Chung, "Comparing transcription agreement on non-native English speech corpus between native and non-native annotators," in *Proc. Interspeech 2012*, Portland, OR, pp. 2366–2369, 2012.

[16] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," Philadelphia: Linguistic Data Consortium, 1993.

[17] G. E. Dahl, Y. Dong, D. Li, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[18] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, Big island, HI, 2011.

[19] P. Ladefoged and M. Halle, "Some major features of the international phonetic alphabet," *Language*, vol. 64, no. 3, pp. 577–582, 1988.

[20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: http://www.tensorflow.org

[21] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, 2016.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. the 32nd International Conference on Machine Learning (ICML 2015)*, Lille, France, pp. 448–456, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE International Conference on Computer Vision (ICCV 2015)*, Santiago, Chile, pp. 1026–1034, 2015.

[24] J. Ba and D. Kingma, "Adam: A method for stochastic optimization," in *Proc. International Conference for Learning Representations (ICLR 2015)*, San Diego, CA, 2015.

[25] H. Hong, S. Kim, and M. Chung, "A corpus-based analysis of english segments produced by korean learners," *Journal of Phonetics*, vol. 46, pp. 52–67, 2014.