Zero-shot Mispronunciation Detection by Knowledge-based Data Augmentation

Zhenghai You
SAI@BUPT
Beijing, China
youzhenghai@bupt.edu.cn

Mewlude Nijat
SISE@XJU
Wulumuqi, China
mewludenijat@stu.xju.edu.cn

Ying Shi, Chen Chen, Wenqiang Du
BNRist, CSLT@Tsinghua University
Beijing, China
shiying,cchen,duwq@cslt.org

Askar Hamdulla
SISE@XJU
Wulumuqi, China
askar@xju.edu.cn

Dong Wang
BNRist, CSLT@Tsinghua University
Beijing, China
wangdong99@mails.tsinghua.edu.cn

***Abstract*—We propose a zero-shot mispronunciation detection approach that does not require any non-native data for model training. Central to our method is a knowledge-based data augmentation process. This process synthesizes mispronunciations by taking into account the typical error patterns of the target group, subsequently using this synthesized data to train an SVM as the detection model. To validate our approach, we constructed a new L2 speech dataset named UY/CH-CHILD, which comprises L2 Chinese speech samples from Uyghur children. Experimental findings suggest that our knowledge-based augmentation strategy proficiently identifies pronunciation mistakes made by non-native children. Interestingly, with such a zero-shot learning, the performance of the detection system is on par with that of native human annotators. The dataset and code will be available online. The code is now available online: https://github.com/youzhenghai/knownledge-gop**

***Index Terms*—Pronunciation error detection, Data augmentation, GOP**

I. Introduction

As an indispensable part of Computer Assisted Language Learning (CALL), Computer Assisted Pronunciation Training (CAPT) equipped with pronunciation assessment technology has been widely applied in foreign language learning [1]. We focus on mispronunciation detection (MD) in this paper, a core task in CAPT with the aim to identify mispronounced phones given the canonical pronunciation [2].

Most of the present MD research is based on automatic speech recognition (ASR). Among these methods, Goodness of Pronunciation (GOP) [3] is perhaps the most widely used, due to its simplicity and good generalizability. In principle, GOP collects frame-level phone posteriors and integrates them into a phone-level score, and the score can be used to make MD decisions by comparing a pre-defined threshold. Several variants of GOP variants were proposed. For instance, Shi et al.[4] considered the confidence of each frame when integrating the frame-level posteriors. Sudhakara et al. [5] refined the GOP by considering HMM transitions.

The paper was partly supported by the National Science Foundation of China (NSFC) with NO. 62171250.

While GOP offers utility, its singular score might fall short in mirroring the multifaceted quality of pronunciation. In pursuit of a more comprehensive MD technique, Hu et al. [6] presented a classifier approach. Specifically, they derived a set of *features* from the frame-level posteriors and used the features to train a classifier to discriminate correct and incorrect pronunciations [6]. The classification approach offered significant performance improvement compared to the simple GOP approach and has been recognized as a strong baseline for MD. Note that the classifier-based approach is flexible: it can be used to involve multiple acoustic and phonetic features, and can be used to collect segmental features to produce utterance-level scores [7, 8].

The key strength of the classification approach is that the classifier can learn pronunciation patterns of L2 learners from non-native speech data. However, this requires a large amount of non-native speech with high-quality annotation. Unfortunately, annotating L2 speech is always challenging as assessing pronunciations requires expert knowledge and the assessment is highly subjective. This partly explains why public L2 speech databases are rare.

To address the problem, we propose a zero-shot learning approach without human-annotated L2 data but can still train a reasonable MD classifier (support vector machine (SVM) in our study) using only L1 data, e.g., the data used to train ASR systems. Since L1 data do not contain pronunciation errors, we synthesize mispronunciations by perturbing the phoneme labels. Most importantly, the perturbation is based on the knowledge of pronunciation errors of the target population, so that the classifier can learn the mistakes that the L2 learners tend to make. We call the approach knowledge-based data augmentation. It should be highlighted that our proposal is a zero-shot approach, as the real speech of L2 learners is not required at all, making it applicable in any scenario where mispronunciation annotation is impossible.

To test our proposed approach, we constructed a new dataset, UY/CH-CHILD. The dataset involves 32021 utterances in Chinese (mostly single words) spoken by 113 Uyghur

children aged from 4-8 years. Experiments show that the SVM MD system trained with knowledge-based data augmentation can effectively detect pronunciation errors made by non-native children, and the MD system works not worse than human annotators.

II. Related work

Mispronunciation detection has been approached in three ways. The first approach is pronunciation scoring [9], which collects information from ASR output and composes a phone-level score to reflect the pronunciation quality. GOP is the most representative in this category [3]. This approach is simple and effective, but requires a powerful acoustic model, ideally adapted to the L2 data. The second MD approach uses an extended recognition network (ERN) [10], i.e., augmenting the decoding graph with mispronunciation paths. This approach cannot deal with unknown mispronunciations, and the performance often degrades when the number of mispronunciations is large. The third approach is based on phone ASR, and MD is performed by comparing the ASR output with the canonical phone sequence directly. This approach attracted more attention recently, due to the thriving of the end-to-end ASR techniques [11–15]. A key shortage of this approach is that it requires a large amount of L2 training data, making it unsuitable in many situations. Our proposed approach is based on pronunciation scoring, though it involves a classifier to perform MD, and the classifier is trained in a zero-shot manner.

The idea of data augmentation is not new in MD. For instance, Fu et al. [12] synthesized mispronounced labels to deal with the data imbalance problem when training end-to-end MD models. Recently, [16] treated data augmentation as the main tool to solve the data sparsity problem when training end-to-end MD models. Our work is related, but differs from theirs significantly: (1) We use pronunciation knowledge when synthesizing mispronunciations; (2) We use the synthesized data to train a simple SVM MD classifier rather than the heavy end-to-end MD model.

III. Method

We will briefly describe the ERN, the GOP, and the classification approach, and then introduce the zero-shot learning with knowledge-based data augmentation.

A. Extended Recognition Network

ERN is a popular approach for MD. It constructs an extended decoding graph called ERN for each single test utterance, and the ERN involves the paths of the canonical pronunciation as well as potential mispronunciations, hence representing how L2 learners with different native language backgrounds might mispronounce. The potential mispronunciations can be regarded as prior knowledge, and can be obtained from either linguistic experts or an L2 dataset. When testing an utterance, construct an ERN using its transcription and the prior knowledge, and then conduct speech recognition constrained by the constructed ERN. If the resultant optimal path is not the one corresponding to the canonical pronunciation, a mispronunciation is detected. Fig. 1 shows an example of ERNs where the canonical pronunciation is 'da4 cong1'.

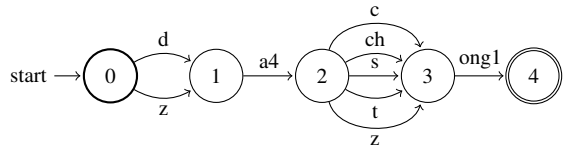


Fig. 1. An extended recognition network (ERN) that involves the canonical pronunciation 'da4 cong1' and potential mispronunciations such as 'za4 chong1', 'da4 song1'.

B. GOP scoring

Denote $p(q_i|o_t)$ the frame-wise posterior over phone q_i at time t, the log posterior probability (LPP) of speech segment $o[t_s, t_e]$ is defined as follows:

$$\begin{aligned} LPP(q_i) &= \log p\left(q_i \mid o; t_s, t_e\right) \\ &\approx \frac{1}{t_e - t_s + 1} \sum_{t=l_s}^{t_e} \log p\left(q_i \mid o_t\right) \end{aligned} \tag{1}$$

The GOP score of a particular phone q_i on this segment is then computed as the log ratio of LPP:

$$GOP(q_i) = \log \frac{LPP(q_i)}{\max_{q_j \in Q} LPP(q_j)} \tag{2}$$

where Q is the phone set. In practice, for each utterance the canonical phone sequence is known, which can be used to segment the speech into phone segments by forced alignment. In our experiments, we use a well-trained DNN as the acoustic model to perform the forced alignment and compute the frame-wise phone posterior $p(q_i|o_t)$.

Given a threshold, it is easy to judge whether a phone q in the canonical transcription is well pronounced, by comparing $GOP(q)$ and the threshold.

C. SVM as MD classifier

As mentioned, a classifier trained with correct and incorrect pronunciations is superior to the naive GOP scoring and thresholding. We have chosen to use SVM as the classifier due to its theoretical robustness. Following [6], the feature vector of the SVM model is composed as follows:

$$\begin{aligned} &[LPP(q_1), LPP(q_2), \ldots, LPP(q_M), \\ &LPR(q_1 \mid q_i), LPR(q_2 \mid q_i), \ldots, LPR(q_M \mid q_i)] \end{aligned} \tag{3}$$

where $LPP(q_i)$ is defined in Eq. 1 and $LPR(q_i|q_j)$ is the log posterior ratio between q_j and q_i, defined as follows:

$$\begin{aligned} LPR\left(q_j \mid q_i\right) = &\log p\left(q_j \mid o; t_s, t_e\right) \\ &- \log p\left(q_i \mid o; t_s, t_e\right) \end{aligned} \tag{4}$$

The SVM approach offers advantages over the GOP scoring. Not only can it integrate more diverse information than the

GOP, but it also allows for the contribution of each information factor to be learned from the target L2 data. This learning is especially important when the acoustic model does not match the test data. Unfortunately, in many situations L2 data are not available. We will develop a data augmentation approach to solve the problem.

D. Knowledge-based data Augmentation

We can utilize L1 data in place of L2 data to train the MD classifier. Given that L1 data does not contain mispronunciations, we can adopt a synthesis approach. Specifically, we randomly change the phone label of a speech segment s from r to r' and use the resultant pair (s, r') as a 'positive sample' of mispronunciations (labeled as 'F'). The original correctly labeled pair (s, r) is used as a 'negative sample' (labeled as 'T'). It should be emphasized that the segmentation information (starting and ending time) is obtained from the forced alignment with the correct labels. Note that in L2-ARCTIC[17], three pronunciation errors were defined: insertion, substitution and deletion. However in our experiments, our task is to detect mispronunciations in isolation words, for which deletion and insertion can be ignored. Moreover, tone substitution is simulated, as this type of error is typical for Chinese L2 learners.

Fig. 2 shows an example of the data augmentation process, where phone c is substituted for s, and the MD label corresponding to c changes from 'T' to 'F'. The newly labeled samples are used to train the SVM MD classifier.

a3	a3	a3	a3	ch	ch	ch	ch	ch	u2	u2	u2	u2
T	T	T	T	T	T	T	T	T	T	T	T	T

↓

a3	a3	a3	a3	sh	sh	sh	sh	sh	u2	u2	u2	u2
T	T	T	T	F	F	F	F	F	T	T	T	T

Fig. 2. Data Augmentation in Force Alignment

The random data augmentation has been employed in previous studies, e.g., [16]. As shown in the next section, this approach does not lead to expected performance on mismatched test data. A hypothesized reason is that the random substitution cannot well represent the error patterns of pronunciations of the target population, which is Uyghur children in our experiments. For example, tone is important in Chinese but does not exist in Uyghur. Therefore for Uyghur speakers, tone error is a main source of mispronunciation, especially when they started to learn Chinese in childhood. Unfortunately, random sampling cannot produce sufficient tone substitutions. The knowledge about how the speakers may make pronunciation errors is crucial information and can be used to guide the synthesis process to produce the most important samples for the SVM classifier.

A simple way to collect this knowledge is to compute the probability that a canonical phone is pronounced as another phone by Uyghur children, forming a confusion matrix as shown in Fig. 3. At each location of the matrix, the color represents the probability that the phone on the x-axis is pronounced as the phone on the y-axis. In our study, the UY/CH-CHILD dataset was used to derive the confusion matrix.

Utilizing the confusion matrix and guided by expert knowledge, we can systematically synthesize relevant mispronunciations. Specifically, if the probability at location (q_i, q_j) is 6.2, then whenever q_i is encountered, it will be altered to q_j with a likelihood of 6.2%. The MD label is then set to 'F'. Additionally, to enhance the robustness of the model during training, we still introduced a certain amount of random errors.

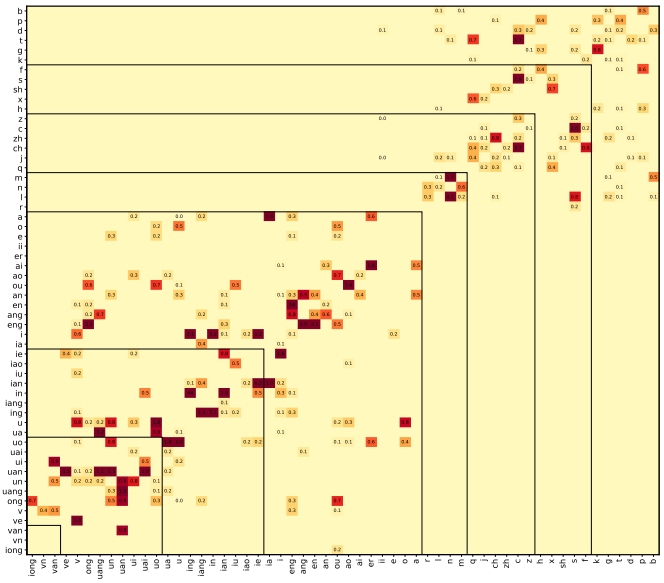


Fig. 3. The confusion matrix estimated from UY/CH-CHILD. The x-axis represents the canonical phone, and the y-axis represents the phone really pronounced.

IV. Experiment

A. Dataset

Two public datasets were used in our experiments: AIShell-1 [18] was used to train the SVM classifier, and THCHS30 [19] was used to test the performance of random data augmentation. The datasets are both in native Chinese and recorded in silent environments with a sampling rate 16k Hz and the precision is 16 bits.

To test the zero-shot situation, we also constructed a Chinese L2 dataset UY/CH-CHILD. The data set involves 113 children (4-8 years old) in child gardens or primary schools in Yining, Xinjiang Uygur Autonomous Region of China. All the children are native Uyghur speakers and were learning Chinese. The children were presented a picture of an object, e.g., apple or washing machine, and were asked to pronounce the name of the object. At each recording session, a supervisor guided the child to pronounce the correct name of each object, and

demonstrated the pronunciation if necessary, until the object was recognized.

The whole session was recorded, including the speech of the supervisor and the child. In the post-processing phase, only the names of the objects pronounced by the children were kept and all the rest speech signals were removed. Therefore, each utterance in the UY/CH-CHILD dataset contains a single Chinese word, typically two or three characters.

Two kinds of labels were annotated for the utterances: (1) the canonical phones and the truly pronounced phones; (2) whether the pronunciation is correct (T) or not (F). The phone labels (more precisely, initials and finals in Chinese) were used to extract the confusion matrix shown in Fig. 3, and the T/F labels were used to evaluate the MD performance. Note that the T/F labels are not always consistent with the phone labels: when the pronunciation quality is poor, it could be labeled as 'F' even if the phone is correctly pronounced.

Considering the high inter-rater variation, we invited 24 native speakers to perform the T/F annotation. There are 3 annotators for each utterance, and a phone segment is finally labeled as T (correct) only if more than two annotators think it is correct. In total, we identified 1,212 mispronunciations from 21,162 phone pronunciations.

B. System configuration

We experimented with three types of systems: an ERN system, a GOP system, and an SVM system. All the three systems make use of the same acoustic model. The backbone is a C-TDNN architecture composed of 2 Convolution block 7 TDNN blocks, Each block incorporates ReLU and BatchNorm. The model was trained on 11k Chinese speech data, including a large variety in speakers and noise conditions. Accents were included but with a limited amount. Very few children's speech were included. The training utilized frame-wise cross-entropy as the loss function.

The SVM model uses a linear kernel, and was trained on samples excerpted and synthesized from AIShell-1, by either random data augmentation (SVM (Random)) or knowledge-based data augmentation (SVM (Knowledge)).

C. Simulation Test

The first experiment is a simulation test with THCHS30. The mispronunciations on the test set were simulated by random data augmentation. For SVM, we only report the results with random data augmentation as it fully matches the test data. Note that the ERN system does not apply in this simulation test as the phone substitution is fully random, leading to a trivial and weak decoding network, as reported in [11].

We use Pearson correlation ($Corr.$) and the F1 measurement as the main metrics, and the precision and recall are also presented. For SVM, the decisions from the SVM were used to perform MD. For GOP, we selected a decision threshold that yielded the same precision as the SVM system to simplify the comparison. The results are shown in Table I. The results indicate that the SVM classifier notably outperforms the GOP, demonstrating that mispronunciation can be effectively detected with simple data augmentation.

TABLE I
RESULTS ON THCHS30.

	Corr.	Precision	Recall	F1
GOP	0.59	0.61	0.60	0.60
SVM(Random)	0.66	0.61	0.77	0.68

D. Results on UY/CH-CHILD

The results on UY/CH-CHILD are shown in Table II, where the SVM model was trained with either random or knowledge-based data augmentation. As in our previous experiments, the decision point for the GOP system was set to match the precision of the SVM model trained with random augmentation. The ERN determined by the pronunciation rules under three different error probability thresholds is also reported. The same phonetic confusion matrix as in the knowledge-based augmentation was used to construct the extended decoding graph, by adding the high-confusing phones into the decoding graph as alternative pronunciations. Three thresholds on the substitution probability p in the confusion matrix were used to identify the high-confusing phones: $p > 0$, $p > 0.01$, $p > 0.02$, denoted by ERN(0), ERN(0.01), and ERN(0.02), respectively.

We also present the Precision-Recall (PR) curve for the GOP system, elucidating the relationship between precision and recall, as shown in Fig. 4. The positions of the two SVM systems and the ERN (0.01) system are also illustrated in the picture. Note that both the SVM and ERN systems are decisive, for which PR curves are not straightforward.

The results indicate that ERN performs well in terms of recall, but its precision score is low. The overall performance (F1 in Table II and the position in Fig. 4) is lower than the GOP system. This is not surprising and demonstrates a key shortage of the ERN approach: controlling the alternative pronunciations is crucial, and to cover as much as mispronunciations more false detections are inevitable.

Random augmentation does not prove effective as well, resulting in performance that is inferior to that of the GOP. The knowledge-based augmentation, on the other hand, performs significantly better, manifesting a marked performance improvement compared with the GOP system. This underscores the importance of incorporating prior knowledge about pronunciation errors from actual speech during the training of a mispronunciation classifier.

Overall, the above experiments demonstrated that the classifier approach accompanied by knowledge-based data augmentation is effective for zero-shot mispronunciation detection, and it is a better way to use pronunciation knowledge than the ERN approach.

E. Discussion

Compared to the results on THCHS30, the MD performance on UY/CH-CHILD looks poor. An apparent reason is that

TABLE II
RESULTS ON UY/CH-CHILD.

	Corr.	Precision	Recall	F1
GOP	0.22	0.23	0.33	0.27
ERN(0)	0.14	0.11	0.58	0.19
ERN(0.01)	0.16	0.14	0.43	0.21
ERN(0.02)	0.15	0.14	0.49	0.21
SVM (Random)	0.16	0.23	0.18	0.20
SVM (Knowledge)	0.25	0.20	0.52	0.29

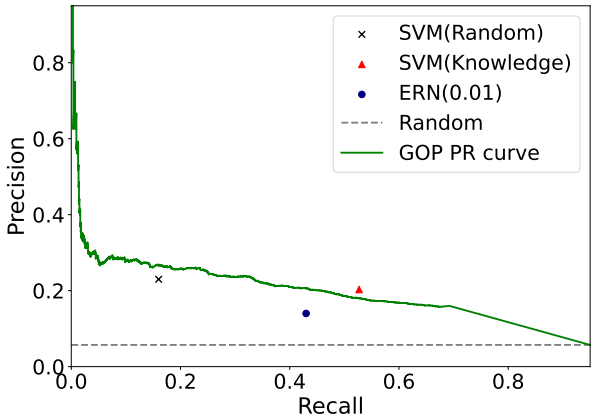
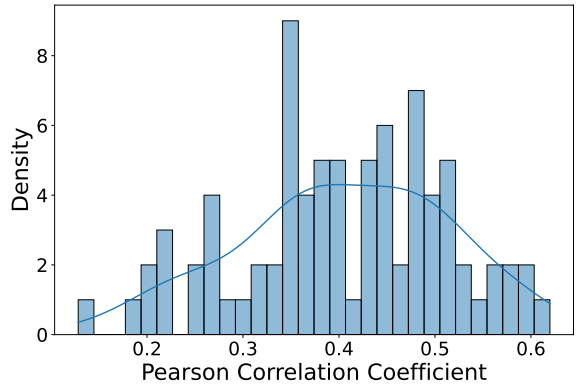
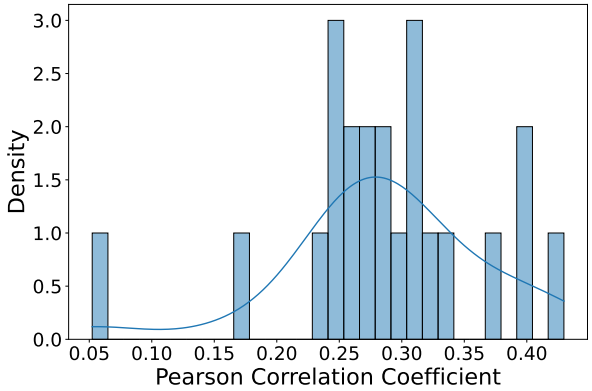


Fig. 4. The Precision-Recall (PR) curve of the GOP system when tested on UY/CH-CHILD is compared with a random classification curve. The decision points of two SVM models and ERN(0.01) are also indicated. The performance of a random classifier is also illustrated.



(a) Human-Human



(b) Machine-Human

Fig. 5. Distributions of the Pearson correlation coefficients (a) between the human annotators, and (b) between the SVM MD system and the human annotators.

the ASR model does not match the test data that are non-native and from children. Another contributing factor is the inherent difficulty in labeling the dataset. This is due to the high subjectivity in determining whether a pronunciation is correct, and in most situations, the pronunciation is just slightly different from the canonical way, leading to significant uncertainty even for native speakers, not to say MD systems. In comparison, the random substitution in the simulation test produces a large proportion of mispronunciations that are easy to test, e.g., replacing s to g, t to a. This means the real dataset is much more difficult than the simulation data, not only for MD systems, but also for humans.

To test this hypothesis, we plot the distribution of the Pearson correlations (1) between any two human annotators; (2) between the knowledge-based SVM system and human all annotators. It can be seen from Fig. 5 that the inter-human correlation is widely distributed, ranging from 0.1 to 0.6. In comparison, the machine-human correlation is from 0.05 to 0.4, mostly in the range of inter-human correlations. This means that the SVM MD system works not worse than a human annotator on average.

V. CONCLUSIONS

In this study, we introduced a zero-shot mispronunciation detection method that does not need any L2 speech data. The core of our approach is knowledge-based data augmentation. By leveraging knowledge of the target population's error patterns, we synthesize mispronunciations and subsequently train an SVM on this synthesized data to detect mispronunciations. Our approach was evaluated on the UY/CH-CHILD dataset, which comprises L2 Chinese speech recordings from Uyghur children. Our results indicate that knowledge-based data augmentation is rather effective, and it outperforms the GOP and ERN, two most popular approaches. We also found that prior knowledge of mispronunciation patterns is important for training an effective MD classifier — random data augmentation cannot lead to a reasonable performance. Finally, our experiments demonstrated that the knowledge-based SVM system rivals native human annotators.

Although promising results have been achieved with our approach, zero-shot MD remains a very challenging task. Future work includes deep investigation on the variation of phone posteriors with L2 speech, test with speech of L1 children, and focus on tone change, the main source of mispronunciations of Chinese learners.

REFERENCES

[1] Bin Zou, *Computer-Assisted Foreign Language Teaching and Learning: Technological Advances: Technological Advances*, IGI Global, 2013.

[2] Diane Litman, Helmer Strik, and Gad S Lim, "Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities," *Language Assessment Quarterly*, vol. 15, no. 3, pp. 294–309, 2018.

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

[4] Jiatong Shi, Nan Huo, and Qin Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," 2020.

[5] Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities," in *Interspeech 2019*, 2019.

[6] Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[7] Bin Su, Shaoguang Mao, Frank Soong, Yan Xia, Jonathan Tien, and Zhiyong Wu, "Improving pronunciation assessment via ordinal regression with anchored reference samples," in *International Conference on Acoustics, Speech, and Signal Processing*, 2021.

[8] Yuan Gong, Ziyi Chen, Iek Heng Chu, Peng Chang, and James Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," 2022.

[9] Yoon Kim, Horacio Franco, and Leonardo Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Eurospeech 1997*, 1997, pp. 645–648.

[10] Alissa M Harrison, Wai-Kit Lo, Xiao-jun Qian, and Helen Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[11] Yiqing Feng, Guanyu Fu, Qingcai Chen, and Kai Chen, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *ICASSP*. IEEE, 2020, pp. 3492–3496.

[12] Kaiqi Fu, Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, and Binghuai Lin, "A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques," *arXiv preprint arXiv:2104.08428*, 2021.

[13] Wai-Kim Leung, Xunying Liu, and Helen Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *ICASSP*. IEEE, 2019, pp. 8132–8136.

[14] Daniel Korzekwa, Jaime Lorenzo-Trueba, Szymon Zaporowski, Shira Calamaro, Thomas Drugman, and Bozena Kostek, "Mispronunciation detection in non-native (l2) english with uncertainty modeling," in *ICASSP*. IEEE, 2021, pp. 7738–7742.

[15] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, "Transformer based end-to-end mispronunciation detection and diagnosis.," in *Interspeech*, 2021, pp. 3954–3958.

[16] Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek, "Computer-assisted pronunciation training—speech synthesis is almost all you need," *Speech Communication*, vol. 142, pp. 22–33, 2022.

[17] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, and John Levis, "L2-arctic: A non-native english speech corpus," in *Interspeech*, 2018.

[18] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *O-COCOSDA*. IEEE, 2017, pp. 1–5.

[19] Dong Wang and Xuewei Zhang, "ThCHS-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.