INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Pronunciation Erroneous Tendency Detection with Language Adversarial Represent Learning

Longfei Yang[†], *Kaiqi Fu*[‡], *Jinsong Zhang*[‡], *Takahiro Shinozaki*[†]

[†]Department of Information and Communication Technology, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan
[‡]Research Institute of International Chinese Language Education, Beijing Language and Culture University, Beijing, China

longfei.yang.cs@gmail.com, kaiq.fu@gmail.com, jinsong.zhang@blcu.edu.cn, shinot@ict.e.titech.ac.jp

Abstract

Pronunciation erroneous tendencies (PETs) are designed to provide instructive feedback to guide non-native language learners to correct their pronunciation errors in language learning thus PET detection plays an important role in computer-aided pronunciation training (CAPT) system. However, PET detection suffers data sparsity problem because non-native data collection and annotation are time-consuming tasks. In this paper, we propose an unsupervised learning framework based on contrastive predictive coding (CPC) to extract knowledge from a large scale of unlabeled speech from two native languages, and then transfer this knowledge to the PET detection task. In this framework, language adversarial training is incorporated to guide the model to align the feature distribution between two languages. In addition, sinc filter is introduced to extract formant-like feature that is considered relevant to some kinds of pronunciation errors. Through the experiment on the Japanese part of BLCU inter-Chinese speech corpus, results show that our proposed language adversarial represent learning is effective to improve the performance of pronunciation erroneous tendency detection for non-native language learners.

Index Terms: Mispronunciation detection, Computer-aided pronunciation training, Pronunciation erroneous tendency, Contrastive predictive coding, language adversarial training

1. Introduction

Increasing demand for learning a second language (L2) in the current environment of globalization makes computer-aided pronunciation training (CAPT) system draw more and more attention in recent years. Comparing to the traditional one-on-one communicative approach between teachers and students, CAPT system saves a lot of resources including but not limited to teaching and administrative staff, and classrooms. It is also more flexible and cheaper. The CAPT system is designed to provide different kinds of informative feedback, like what a language teacher does, mainly including pronunciation scores [1, 2] and mispronunciations [3, 4], to guide the L2 learners in practicing their pronunciations. Pronunciation scores reflect how similar L2 learner's pronunciation is comparing to a native's thus they are suitable for assessment, while mispronunciation information is more useful for learners in terms of correcting pronunciation errors in their future studies.

In conventional approaches, mispronunciation information mainly includes phone substitution, which only tells learners that they have pronunciation errors but does not provide them detailed information on how to correct these errors. Moreover, it may be hard to simply classify some mispronunciations into definite categorical substitutions because these errors are merely slight acoustic deviations from canonical speech [5]. Therefore, we propose pronunciation erroneous tendencies (PETs) in previous works [6]. PETs are designed to represent these deviations from the perspective of articulation manners and placements. Comparing to phone substitution, PETs can describe pronunciation errors across a wider range from acoustic deviations to categorical pronunciation errors, and, thus, are instructive for learners in correcting their mispronunciations.

PET detection plays a crucial role in the CAPT system. With the development of deep learning methods in recent years, PET detection has been evolving with the help of machine learning technologies [7, 8, 9]. Most conventional approaches establish the PET detection component based on supervised learning using only non-native data. However, it is impractical to collect a large amount of non-native speech data. And, annotation is time-consuming since it relies on human speech perception and manual labeling. It is a challenge to handle the data sparsity problem in non-native acoustic modeling.

In recent years, several works have presented promising approaches to obtain audio representation in unsupervised manners [10, 11, 12]. [13] has proposed contrastive predictive coding (CPC) model in which the representation of high-level properties of speech, such as phonetic content and speaker characteristics, can be learned with unsupervised learning using a large amount of unlabeled data. The representation can be employed to improve the performance of downstream tasks, such as automatic speech recognition (ASR). It is more scalable and cheaper.

In this paper, we propose an unsupervised approach to non-native PET detection task. We expect that the knowledge learned from a large scale of two native raw data can help relieve the data sparsity problem in non-native PET detection. In our model, large-scale unlabeled speech from the target language is first fed to the CPC model to learn phonetic properties. The model is trained with language adversarial training using the learner's native language data to align the feature distribution between these two languages. In addition, we introduce sinc filter in the first layer and expect it to capture formant-like feature. Formant is considered relevant to some kinks of PETs from the respect of articulation placements and manners [14]. Then the pretrained model is employed as the feature extractor for the downstream PET detection task. The performance of our proposed method is assessed on the Japanese part of BLCU inter-Chinese speech corpus, which is designed for the learners from Japan who learn Mandarin Chinese as the second language.

http://dx.doi.org/10.21437/Interspeech.2020-2033

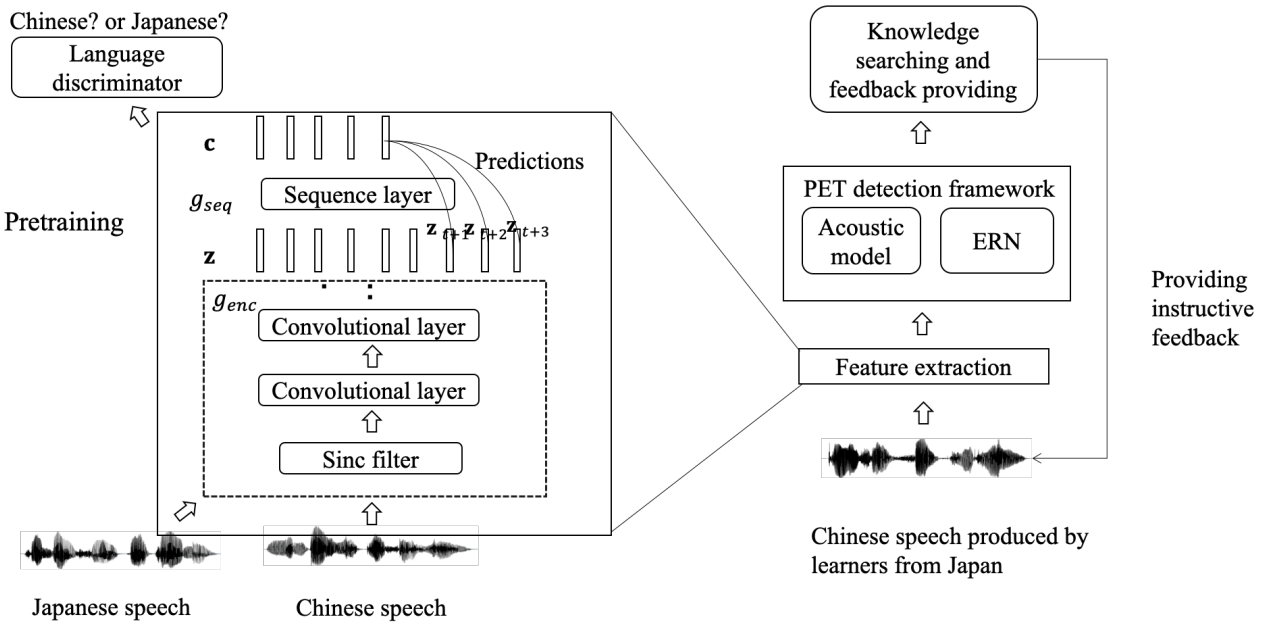

Figure 1: *A demonstration of the proposed model. Pretraining using language adversarial contrastive predictive coding is on the left and pronunciation erroneous tendency detection using the feature extracted by the pretrained model is on the right.*

2. Approach

Our goal is to learn the representation that captures high-level information for non-native acoustic modeling using two native data without human supervision, which can be regarded as a unsupervised process of drawing the distance between these two languages in feature space. We start by elaborating on the pretraining using contrastive predictive coding (CPC) with sinc filter. Then we introduce language adversarial training. Finally, we describe the pronunciation erroneous tendency (PET) detection using the feature extracted by the pretrained model.

2.1. Pretraining

2.1.1. Contrastive predictive coding with sinc filter

Contrastive predictive coding (CPC) is an unsupervised represent learning framework [10]. The objective is designed to extract the feature that makes long-term predictions about future observations while retaining properties or structures of the input by maximizing the mutual information of these feature with those extracted from future timesteps. Predictions on different timescales will capture different levels of information. A quickly varying representation can be considered to be of local structure, and a slowly varying one can be higher abstractions or global structures, such as phonemes and words in speech signal [15]. An overview of the CPC model is illustrated in the left part of Figure 1. To be detailed, let $\mathbf{x} = \{x_1, x_2, ..., x_l\}, x_i \in \mathbb{R}$, denotes an raw speech signal in L discrete time steps where x_i is the acoustic amplitude at time i. First, an encoder g_{enc} encodes the signal into the embedding vector $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_t\}, \mathbf{z}_i \in \mathbb{R}^{d_z}$.

$$\mathbf{z} = g_{enc}(x_1, x_2, ..., x_l), \tag{1}$$

In our model, the main difference from the original CPC is that we employ sinc filter [16] in the first layer, which is reported can extract the formant-like feature [16]. Formant is relevant to PETs in terms of articulation manners and placements [14], such as the position of tongue and the shape of the lip. The operation of sinc filter can be written as follows:

$$y[n] = x[n] * g[n, \theta] \tag{2}$$

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{3}$$

where sinc filter is defined as $sinc(x) = sin(x)/x$ and f_1 and f_2 are the learned low and high cutoff frequencies respectively.

Then a sequence model g_{seq} summarizes the past representations and produces context-aware embeddings, which can be denoted as $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_t\}, \mathbf{c}_i \in \mathbb{R}^{d_c}$, i.e.,

$$\mathbf{c} = g_{seq}(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_t), \tag{4}$$

The representations are learned by minimizing the InfoNCE loss [10], which is a loss function based on noise contrastive estimation as the lower-bound on the mutual information between the context aware embedding $\mathbf{c}_t$ at time t and future latent representations $\mathbf{z}_{t+k}$ for $k \in \{1, ..., K\}$. Given a set $Z = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$ which contains one positive sample from $p(\mathbf{z}_{t+k}|\mathbf{c}_t)$ and $N-1$ negative samples from "noise" distribution $p(\mathbf{z})$. The loss function for each step t can be donated as follows:

$$\mathcal{L}_{tk}^N = -\mathbb{E}_Z \left[\log \frac{f_k(\mathbf{c}_t, \mathbf{z}_{t+k})}{\frac{1}{N} \sum_{\tilde{\mathbf{z}} \in Z} f_k(\mathbf{c}_t, \tilde{\mathbf{z}})} \right] \tag{5}$$

where $f_k(\mathbf{c}_t, \mathbf{z}_{t+k})$ is a scoring function that can be a log-bilinear model:

$$f_k(\mathbf{c}_t, \mathbf{z}_{t+k}) = \exp(\mathbf{c}_t^\intercal \mathbf{W}_k \mathbf{z}_{t+k})) \tag{6}$$

where $\mathbf{W}_k$ is the parameter in each model for every k. The total loss to be minimized is a sum of the loss for each step:

$$\mathcal{L}^N = \sum_t \sum_k \mathcal{L}_{tk}^N \tag{7}$$

in which negative samples are sampled uniformly from representations in the same speech signal in this work.

2.1.2. Language adversarial training

The aim of incorporating language adversarial learning is to align the feature distribution between the learner's target and native languages. It is similar to the process of language learning as the learner's non-native speech may be affected by their native language, which is referred to as L1 transfer. The proposed unified framework is shown in Figure 1. In our setup, the CPC model is paired with another output that shares the internal representations of the input and tries to discriminate whether the input speech signal comes from the target language or from the native to represent the language discrepancy. In this experiment, we focus on the speech in Chinese produced by language learners from Japan. Nevertheless, the proposed approach is not limited to Chinese-Japanese condition. The training process of this language discriminator is adversarial with respect to the shared hidden layers by using gradient reversal to maximize the language adversarial loss rather than minimize it to confuse the discriminator. The language discriminator is optimized using negative log-probability as the language adversarial loss:

$$\mathcal{L}^{LA} = -l \log y_w - (1-l)\log(1-y_w) \quad (8)$$

$$y_w = p(l=1|h;\theta_{LA}) = softmax(W_l h + b) \quad (9)$$

where $l \in \{0,1\}$ denotes the ground-truth language label of input (0 for Chinese and 1 for Japanese), y_w and $W_l, b \in \theta_{LA}$ are the outputs and weights of the final layer, and h is the representation generated by the CPC model. In our experiments, we make a comparison between using the output of convolutional layers and that of sequence layers as the input of the language discriminator.

We optimize the weighted-sum of two losses using a hyperparameters λ. The overall training objective of the composite model can be written as follows:

$$\mathcal{L} = \lambda \sum_{N} \mathcal{L}^{N} - (1-\lambda) \sum_{M+N} \mathcal{L}^{LA} \quad (10)$$

where N and M are the numbers of samples of two languages ($M = N$ in this work).

We look for parameters that satisfy a min-max optimization criterion as follows:

$$\min_{\theta_{cpc}} \max_{\theta_{LA},\theta_s} \mathcal{L} \quad (11)$$

where θ_{cpc}, θ_{LA} and θ_s denote parameters of the CPC model, language discriminator and shared part respectively. Such optimization will involve a maximization with respect to the language discriminator and a minimization with respect to the CPC model.

Algorithm 1: Language adversarial training

Input: Chinese data $\mathbf{x}_C$, Japanese data $\mathbf{x}_J$, batch size b

Output: learned model parameters

1. Initialize model parameters;
2. **repeat**
   (1) Randomly sample $\frac{b}{2}$ examples from $\mathbf{x}_C$
   (2) Randomly sample $\frac{b}{2}$ examples from $\mathbf{x}_J$
   (3) Compute $\mathcal{L}^N$ and $\mathcal{L}^{LA}$
   (4) Take a gradient step for $\lambda\frac{2}{b}\nabla_{\theta_{cpc}}\mathcal{L}^N$
   (5) Take a gradient step for $(1-\lambda)\frac{2}{b}\nabla_{\theta_{LA}}\mathcal{L}^{LA}$
   `// Gradient reversal`
   (6) Take a gradient step for $-(1-\lambda)\frac{2}{b}\nabla_{\theta_s}\mathcal{L}^{LA}$

until convergence;

Algorithm 1 presents the pseudocode for the language adversarial training to train the model. The parameters are initialized first. Then we create minibatches by randomly sampling $b/2$ samples from $\mathbf{x}_C$ and $b/2$ from $\mathbf{x}_J$. Note that, Chinese samples take part in calculating both $\mathcal{L}^N$ and $\mathcal{L}^{LA}$ while Japanese data only participate in computing $\mathcal{L}^{LA}$.

2.2. PET detection

Once the pretraining is finished, the pretrained model is employed as the feature extractor for the PET detection framework. The non-native speech is first fed into our pretrained model to generate resulting context vectors (c) as the feature. Then these feature are sent to the acoustic model to output phone-level transcription with the extended pronunciation network (ERN). Then the system evaluates the speech according to the difference between these recognized transcriptions and canonical ones. Finally, the system searches the knowledge from the PET database and provide instructive feedback to learners. During training on non-native data, the parameters of the acoustic model for PET are trained while those of the pretrained model remain frozen.

3. Experimental setting

3.1. Datasets

AISHELL corpus is employed as the Chinese source, which is an open-source Mandarin Chinese speech corpus [17], and Corpus of Spontaneous Japanese (CSJ) is used as the Japanese source, which is a database containing a large collection of Japanese spoken language data [18]. We randomly choose 150 hours of data from two corpora above as our training set for pretraining.

For PET detection, BLCU inter-Chinese speech corpus, which is collected for language learners who learn Mandarin Chinese as their second language [6], is employed as our non-native dataset for PET detection. All ground-truth labels for PETs in this corpus are annotated by well-trained phoneticians. In this work, we focus on the top 16 kinds of pronunciation errors. Around 90% of this corpus is used as the training set and the rest for testing. There is no overlap of speakers between the training and testing set.

3.2. Model setup

The baseline is the original CPC model. The encoder contains five 1-dimensional convolutional layers with a 160 downsampling factor thus there is a feature vector for every 10ms of speech, which keeps consistent with the rate of phoneme sequence labels obtained with Kaldi. For convolutional layers, the size of filters are [10, 8, 4, 4, 4], the strides are [5, 4, 2, 2, 2] and the paddings are [3, 2, 1, 1, 1]. 512 hidden units of each layer are with ReLU activation. Batch normalization is employed following each convolutional layer. A recurrent neural network with gated recurrent units (GRU RNN) with 256 dimensional hidden states is employed as the sequence part of the model. The output of GRU at every timestep is used as the context $\mathbf{c}$ to predict 12 timesteps in the future. In each training iteration, a segment containing 20480 data points (around 1.28s) is randomly selected from every utterance. Adam optimizer with a learning rate of 2e-4 is used to train the model with a minibatch of size 8. We simply mix data from two languages into a large dataset for training the original CPC model.

For the language adversarial training, the language discriminator contains two layers with hidden units whose size is [64,

2], and the final layer outputs the one-hot representations to distinguish two languages. λ in Eq (10) is set to 0.5. To get the language adversarial loss, we make a comparison between two approaches. In the first approach, we take the output of the last convolutional layer as the input into the language discriminator; in the second, we take the output of the sequence layers as the input. The detailed analysis can be found in Section 4.

For the PET detection part, we use uni- and bi-directional recurrent neural networks with gated recurrent units (GRU-RNN) for acoustic modeling. First, we directly establish modeling using the surface feature 40-dimensional MFCC extracted from non-native data only and feed them into three-layered uni-GRU where each layer has 512units and five-layered bi-GRU, where each layer has 550 units. Then we employ our pretrained model as the feature extractor. The subsequent acoustic model uses uni-GRU with three layers in which each layer has 512 units, and the bi-GRU has one layer with 550 units. Batch normalization and a dropout of 0.2 are performed following each layer. RmsProp is employed with a batch size of 16.

4. Results and Discussion

For overall performance for phone recognition, as shown in Table 1, we can find that the feature learned from two native speech through the unsupervised framework are effective for non-native spoken language processing, but the original CPC model does not perform well without language information. The performance by using the output of the sequence model DLA-CPC as the input of the language discriminator is better than SLA-CPC using that of the encoder. We think it may be because the sequence layer in SLA-CPC does not receive the guidance from language information. Our proposed framework DLA-CPC improves the overall phone error rate (PER) from 11.94% to 10.89% for the uni-GRU model, and from 10.73% to 9.85% for the bi-GRU model, which is better than the modeling using only non-native data and the original CPC model.

Table 1: *The detection performance of different approaches. The results with "uni" and "bi" mean the model of our downstream PET detection. "L2-only" denotes that the model is trained with only non-native data, and "-CPC" uses the pretrained model as the feature extractor. "SLA-" denotes the output of encoders is used as the input fed into the language discriminator, and "DLA-" means we use the output of the sequence model for it. The best results are marked with bold fonts.*

Model	Recall	Precision	F1 score	PER
L2-only, uni	38.32%	51.61%	43.98	11.94%
CPC, uni	35.33%	50.43%	41.55	12.13%
Sinc-CPC, uni	31.74%	50.48%	38.97	12.27%
SLA-CPC, uni	33.53%	45.9%	38.75	13.45%
DLA-CPC, uni	**41.92%**	51.09%	**46.05**	10.89%
DLA-Sinc-CPC, uni	39.52%	**55.0%**	45.99	**10.54%**
L2-only, bi	40.12%	54.92%	46.37	10.73%
CPC, bi	34.73%	54.2%	42.33	10.78%
Sinc-CPC, bi	37.13%	58.49%	45.42	10.43%
SLA-CPC, bi	37.72%	51.79%	43.64	11.03%
DLA-CPC, bi	**44.91%**	58.14%	50.68	**9.85%**
DLA-Sinc-CPC, bi	44.14%	**60.22%**	**50.94**	9.9%

For PET detection, it can be found that our proposed model improves the F1 score from 43.98 to 46.05 for the uni-model ,and from 46.37 to 50.68 for the bi-model when comparing to L2-only, which is better than original the CPC as well. The improvement of the $recall$ means the feature extracted by our model can help the PET detector find out more mispronunciations for language learners. We also notice that even though introducing sinc filter can help the model improve the $precision$, which measures how many "mispronunciation" detected by the machine are truly errors, but the $recall$ decreases a little.

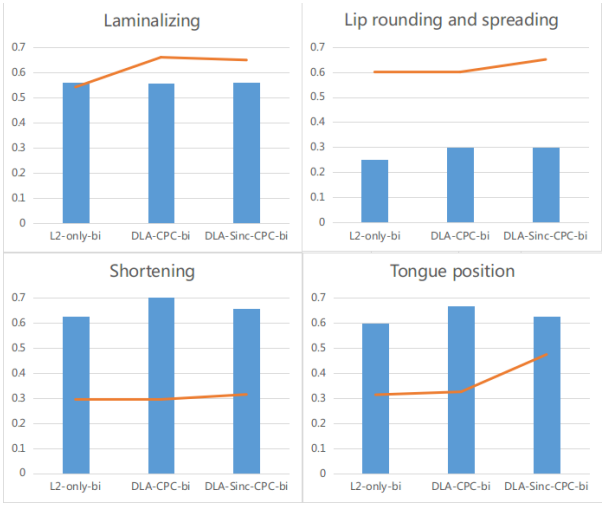

Figure 2: *Results of four groups of PETs. Column is Recall and line is Precision*

To make a detailed analysis, we group the employed PETs into four groups and divide the results of the bi-model, as shown in Figure 2. Four groups are: whether the shape of lip is rounding or spreading, the position of tongue is advancing or backing, the aspiration or constriction is sufficient or not, and laminalizing means that some balade-palatal phonemes are pronounced like Japanese lamina-alveolar. We notice that our model improves the $recall$ for all four groups. It also can be found that the $precision$ for detecting PETs about laminalizing is significantly improved by introducing language adversarial learning. The performance for PETs about tongue position and the shape of lip is improved by incorporating sinc filter. We think it may be because the first formant (F1) and the second formant (F2) formant are sensitive to the tongue position (high and low to F1, front and rear to F2), and the third formant (F3) is considered related to the shape of lip (round or spread) [14], which is in line with our expectations.

5. Conclusions

In this paper, we propose an unsupervised approach to learn representations from a large amount of Chinese and Japanese raw speech data for non-native acoustic modeling of PET detection. In our model, contrastive predictive coding is employed to learn phonetic structures from Chinese speech, and sinc filter is incorporated to extract the formant-like feature that is relevant to some kinds of PETs. The model is trained with language adversarial learning using Japanese speech to align the feature distribution between two languages. The experimental results show that our proposed model is effective for non-native PET detection. In future works, we're going to incorporate sequence learning and other criteria to further improve the performance with our pretrained model.

6. References

[1] W. Hu, Y. Qian, and F. K. Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call)," in *INTERSPEECH*, 2013.

[2] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W.-p. Ye, "Generalized segment posterior probability for automatic mandarin pronunciation evaluation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–201.

[3] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5049–5052.

[4] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. 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.

[5] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[6] W. Cao, D. Wang, J. Zhang, and Z. Xiong, "Developing a chinese l2 speech database of japanese learners with narrow-phonetic labels for computer assisted pronunciation training," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[7] R. Duan, J. Zhang, W. Cao, and Y. Xie, "A preliminary study on asr-based detection of chinese mispronunciation by japanese learners," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Y. Gao, Y. Xie, W. Cao, and J. Zhang, "A study on robust detection of pronunciation erroneous tendency based on deep neural network," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] L. Yang, Y. Xie, Y. Gao, and J. Zhang, "Improving pronunciation erroneous tendency detection with convolutional long short-term memory," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 52–56.

[10] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[11] A. Hyvarinen and H. Morioka, "Unsupervised feature extraction by time-contrastive learning and nonlinear ica," in *Advances in Neural Information Processing Systems*, 2016, pp. 3765–3773.

[12] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," *arXiv preprint arXiv:2002.02848*, 2020.

[13] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[14] Z. Wu and M. Lin, in *Summary of Experimental Phonetics*, 1989.

[15] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[16] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028.

[17] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[18] K. Maekawa, H. Kikuchi, and W. Tsukahara, "Corpus of spontaneous japanese: design, annotation and xml representation," *Reproduction*, vol. 16, no. 16, pp. 5–5, 2004.