Detection of Mispronunciation in Non-native Speech Using Acoustic Model and Convolutional Recurrent Neural Networks

Yufei Huang1, *, Yulin Huang2
1School of Intelligence and Computing, Tianjin University, Tianjin, 300350
2School of English, Beijing Foreign Studies University, Beijing, 100089

*Corresponding author: yufeihuang666@tju.edu.cn

Abstract. Computer-Aided Language Learning (CALL) frameworks have gained much popularity because of their adaptability, enabling students to refine their language abilities at their own pace. Much research has been done to help improve CALL systems and dig out the more suitable features for targeted Mandarin mispronunciation detection is still an open research area. The acoustic model of convolutional recurrent neural networks (CRNN) (CNN + LSTM) + connectivity time series classification (CTC) model is used to convert acoustic signals into pinyin label sequences. As many Chinese speech data sets were trained using this model, the initial and vowel pronunciation error rates were finally obtained. Therefore, through linguistic classification, mainly four pronunciation errors related to the native Spanish pronunciation habits are further discovered. Moreover, based on the final analysis obtained, some corresponding instructive suggestions for further international Chinese teaching from different aspects are also put forward. Apart from proposing practical suggestions from different perspectives for further Mandarin CALL international teaching according to the experiment results' evaluations, this system still has room for further improvement.

Keywords: CALL; Acoustic model; Mispronunciation; Mandarin; Spanish.

1. Introduction
With the “mandarin fever” in recent years, learning Chinese has gained increasing popularity worldwide. Moreover, as Chinese tone is challenging for most foreigners to pronounce correctly, to correct mispronunciation error is thus an essential part of learning native mandarin should be prioritized by instructors. Therefore, the mispronunciation detection of Mandarin has become more significant, especially for those who attempt to get the hang of this language through the computer-assisted tutorial. However, compared to other languages such as English, Dutch, and Japanese, less work has been done on Chinese phonemes mispronunciation detection. Furthermore, as shown in the literature review, even though much work was done on Chinese mispronunciation detection for non-native Chinese speech using different methods, existing approaches still work on confusing phonemes of Mandarin instead of all phonemes. As a result, digging out the more suitable features for all mandarin mispronunciation detection is still an open research area.
In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery (Valueva 235). They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics (Zhang 32). Nowadays, as computer-Aided Language Learning (CALL) frameworks have gained much more popularity because of their adaptability, enabling students to refine their language abilities at their own pace. There is much research that has been done to help improve CALL systems. Moreover, a more particular CALL sub-region called Computer-Aided Pronunciation Training (CAPT) emphasizes areas such as recognizing an error in non-native speech. And there have been lots techniques of mispronunciation detection proposed by many former researchers regarding the deep Convolutional Neural Network-based methods. Recently, Deep Neural Networks (DNNs), which attempt to model high-level abstractions in data, have greatly improved the discrimination of acoustic models in speech recognition. Qian et al. (2012) applied the application of the Deep Trust Network (DBN) to mispronunciation detection and diagnosis in L2 English. Then the DNN-HMM framework is further proposed for error detection by Mohamed et al. (5060), and three types of acoustic features are used, such as MFCC, PLP, and filter bands. And in the work of Hu et al. 2013 (1886), they first used the acoustic model trained by DNN to score the quality of English pronunciation. They recommend using DNN and transferring learning to the task of detecting phone-level pronunciation errors. Then in 2015, in the DNN-based acoustic model training, a multi-layer neural network was trained as a non-linear basic function to represent the speech signal, and the top layer of the network was trained differently to enhance the senone of the posterior probability (Hu et al. 2015). More importantly, Joshi (697) proposed a deep neural network acoustic model training to detect vowel pronunciation errors. In addition, Li et al. (193) proposed an acoustic phonological model that uses a multi-distributed deep belief network with acoustic features and equivalent canonical pronunciation to detect mispronunciation in English corpus. In addition, many other authors use the new concepts of fuzzy approximation (Zhao et al. 3847), adaptive neural output (Wang et al. 3658), and improved stability criteria (Yin et al. 3472) to perform mispronunciation detection.

The above literature shows that different techniques/methods have been used for mispronunciation detection and a lot of work has been done on mispronunciation detection for different languages (English, Dutch, and Japanese), but little work is done on mandarin mispronunciation detection for particularly non-native speakers from Spanish. Consequently, this paper uses the acoustic model of CRNN (CNN + LSTM) + CTC to convert acoustic signals into pinyin label sequences. By using a large number of Chinese speech data sets, the statistical result of the pronunciation error rate would be finally obtained. Therefore, the contributions of this paper are:

1. We develop an acoustic model for efficient detection of mispronunciation in Chinese Mandarin phonemes.
2. We use phonetics knowledge to summarize the experiment results as four types of the most error-prone points for Spanish learning Chinese.
3. Based on the final analysis obtained, we give corresponding instructive suggestions for further international Chinese teaching from different aspects.

The following part begins by presenting the detail of the proposed methodology. Then, the results and suggestions obtained from the experiment are further present as follows.

2. Methodology

2.1. A: Pronunciation error detection system

1. Detection framework based on automatic speech recognition

To detect incorrect pronunciation, we use a detection framework based on statistical speech recognition to detect pronunciation errors automatically. The entire detection framework is shown in Figure 1. First, the audio of the sentence read by the learner is converted into a two-dimensional
spectral image signal, and then the acoustic model is matched. We use the CRNN (CNN + LSTM) + CTC acoustic model. The acoustic signal is converted into a pinyin label sequence, and then the identified factor sequence is compared with the correct pronunciation sequence of the read sentence. The statistical result of the mispronunciation is obtained through a certain statistical method. The process is demonstrated as follows:

![Figure 1 Frame diagram of the detection system](image)

3. CRNN-CTC acoustic model
The entire training process of the CRNN-CTC acoustic model is roughly divided into two stages. The convolutional recurrent neural network (Shi et al.) (CRNN) structure is shown in Figure 2. First, the 200-dimensional spectrogram data feature value sequence is passed to the relevant convolutional neural network, and the convolutional layer and the maximum pooling layer are connected for feature extraction and classification. The convolution kernel size is 3x3, and the pooling window size is 2, and then build a long and short-term memory (Hochreiter 1750) (Gers et al. 138) (LSTM) network, perform a fully connected linear calculation, and input it into the output layer, which uses softmax as the activation function. The entire network is trained using Stochastic Gradient Descent (SGD). The backpropagation algorithm calculates the gradient. In particular, in the convolutional pooling layer, the architecture of the convolutional layer is based on the architecture of VGG (Simonyan), and in the transcription layer, as described in (Graves et al. 5), errors are backpropagated using a forward algorithm. In the loop layer, backpropagation over time (BPTT) is applied to calculate the error.
Finally, we use the conditional probability defined in the connection time classification (CTC) layer proposed by Graves et al. (5). When using the negative log-likelihood of this probability as the objective function of training the network, we only need the image and its corresponding label sequence, which avoids the labor of labeling the position of a single character and can merge a large number of consecutive repeated symbols to obtain the final actual phonetic symbol sequence.

4. Experiment

4.1. A: Experimental configuration
The CRNN-CTC acoustic model was trained and tested on the Tsinghua University’s THCHS30 Chinese speech data set and AIShell-1 open-source data set. After 70 epochs of training, we obtained a speech with an accuracy of 80% on the test set model.

This experiment collected forty speech samples from eight native Spanish speakers from Colombia, Argentina, and others. On average, they are all learning Chinese at a level higher than L1. Among the forty speech samples, thirty were obtained by six persons reading the same five sentences in Chinese, and the remaining ten samples were obtained by randomly selecting sentences practiced by two persons. As a result, our database has been formed below (Figure 3).
Corpus statistics

|                |       |
|----------------|-------|
| text           | 16    |
| Recorder       | 8     |
| Total number of sentences | 41    |
| Total phonemes | 3904  |
| Average number of phonemes | 95    |

Figure 3 Corpus statistics

Then we input each voice sample, split it into initials and vowels through the detected pinyin sequence. By counting the total number of initials and vowels detected and the number of pronunciation errors, the initial and vowel pronunciation error rates are finally obtained. The experiment chooses to convert the ordinary wav speech signal into the two-dimensional spectral image signal required by the neural network through operations such as framing and windowing, that is, the spectrogram. Furthermore, pass them to the convolutional neural network for the feature extraction process.

4.2. B: Evaluation Index

In this experiment, according to our test process and goal, we used two common evaluation indicators to measure the effect of the experiment, namely the initial error rate (IER) and final error rate (CER).

\[
\text{IER} = \frac{\text{Number of errors}}{\text{Total number of initial detections}} \\
\text{CER} = \frac{\text{Number of errors}}{\text{Total number of vowel detections}}
\]

4.3. Results and evaluations

Detailed results are shown in the Figure 4 and 5.

Figure 4 Initial detection error rate
Figure 5 Vowel detection error rate

Firstly, it can be seen from the figures above that z has the highest pronunciation error rate among the initials, with an error rate of 58.14%, exceeding 50%, followed by c(46.15%), zh(40.48%), ch(39.53%), s(38.10%), and the correct rate of m is as high as 100%, including the pronunciation of h, w, g, and d consonants. The correct rate is more than 90%.

Among the vowels, the highest pronunciation error rate is v, with an error rate of 71.43%, followed by ang, with an error rate of 61.76%, and the finals iu, eng, en, un, ei, in, with an error rate of 52.94%, respectively. 51.85%, 50.00%, 46.43%, 45.45%, 38.89%, 37.50%. The lowest error rate is er. No mispronunciation is detected in the statistical samples. The error rate of finals a, i is also low, but by comparison, the average error rate of vowel pronunciation is higher than the error rate of consonant pronunciation, and the composition of vowel pronunciation is even more complicated.

Therefore, to classify all the mispronunciations obtained above according to the pronunciation parts, the following four categories are mainly obtained, namely:

a Type of pre-and post-tongue or pre-nasal deviation: When the vowel is pronounced too far forward or backward, the front nasal sounds are similar to the back nasal sounds, or the back nasal sounds are similar to the front nasal sounds. In Figure 6, three of the top five with the highest error rate belong to this type, namely ang [ɑŋ], eng [ɤŋ], and ing [iŋ], all of which are prone to be pre-nasal sounds.

b (Tongue pronunciation) Shorten the type of bias: the type of bias where the aspiration time is insufficient when sending the breath sound. This type includes three aspirated and aspirated sounds before, during and after the tip of the tongue (z [ts], c [tsʰ] and t [tʰ], d [t], etc.) and p [pʰ] and b [p] It is easy to confuse each other.

c Types of tongue-leafed errors: Japanese students are prone to tongue-leafed errors when they pronounce zh, ch, and sh. In Figure 5, zh [tʂ] and ch [tʂʰ] ranked third and fourth in error rate. According to our statistics, these two are easy to pronounce and easy to pronounce z, c, s, and other non-tongue sounds.

d. The type of lip-shaped roundness and spreading error: The lip shape is somewhat stretched when the lips are rounded or rounded when the lip sounds are developed. For example, among the vowels, the round labial sound o [o] with a high error rate is easy to be pronounced, and u [u] is easy to pronounce as i [i] and e [ɤ]. Similarly, the labial sound e [ɤ] is easy to round and is pronounced like ou [ou] and other round labial sounds.
In addition, we also analyzed the types of mispronunciation of these easy-to-read mispronunciations, using the pronunciation method as the standard, and summarized three types of typical mispronunciations, as follows:

a. Diphthong: In Figure 6, iu [joʊ] and ei [eɪ] with higher error rates are both diphthongs. Because the pronunciation of diphthongs is more complicated than the monophonic sound, it is not only the type error rate of diphthongs higher. The recognition result is also more prosperous. Since it is often impossible to read both vowels to the standard, it will lead to a single vowel and may be recognized as other polysyllables or other diphthongs. For example, in all the mispronunciations that we have counted, iu [joʊ] is easily mispronounced as i+ao, ing, u, i, e, etc.

b. Unvoiced and voiced sounds: In the pronunciation of many consonants, unvoiced and voiced sounds are prone to appear. For example, l [l] is recognized as d [t], etc., due to errors.

c. Affricate: Among the consonants, the pronunciation of plosives, fricatives, and affricates is easy to confuse. For example, the affricate q [tɕʰ] is easily mispronounced as x [ɕ] and so on.

4.4. D. Comparison between Chinese and Spanish pronunciation of alphabet systems

In learning the second language pronunciation, the interference of the mother tongue is the main reason for the pronunciation error. As Linguists have classified the world’s languages into what they call ‘language families’, the difficulty of a new language that we want to learn is very relative and is significantly related to how similar it is to our mother tongue(s) or other languages we speak well(Fixmer). Especially in the primary stage, learners generally naturally follow the original pronunciation habits of the mother tongue to pronounce the Chinese initials and phonemes, so that the bottom sound of the mother tongue In the entire process of pronunciation acquisition, a deep imprint has been burned, which has become what we call "foreign accents."

Therefore, to evaluate these results, it is important to be aware of some major differences between the two languages to understand better these common mistakes in pronouncing Mandarin Chinese. The very basic thing is that English is part of the Indo-European language family, and Chinese is part of the Sino-Tibetan one. As there are 4 basic tones in Chinese for each word in a sentence, if you pronounce the same syllable in different tones, the corresponding meanings would vary distinctly. Therefore, tones are a unique concept for most foreign speakers. More importantly, Spanish uses the Roman alphabet for the phonemes, which means the phonemic and phonological difference in the foundation. As shown in Figure 5 and Figure 6, while Spanish has only five vowel sounds, Mandarin Chinese has 24 ones which are hard to handle. Moreover, Spanish also does not have the following consonant sounds, such as zh, ch, sh. Besides, the flat, curled, and tongue-twisting sounds of Chinese are also difficult for the Spanish.

Consequently, three of the most distinct differences between Mandarin alphabet systems and Spanish ones are that, first of all, as mentioned above, compared with Western phonemes, Chinese has much more ones, which means that there are some “vacancies.” The most typical one is the vowel: the pronunciation error rate of v[y] is shown in Figure 6. It is as high as 71.43%. It is not difficult to observe that the same or similar pronunciation does not exist in the phoneme system of Spanish, so this kind of round-labial sound is very unfamiliar to Hispanic learners. Secondly, the unvoiced and voiced consonants in Spanish correspond one-to-one, but there are fewer voiced consonants in Chinese, only five, and the more unvoiced consonants are divided into aspirated and non-aspirated consonants.

Last but not least, the two pronunciation systems have essential differences in the pronunciation of many phonemes. Although the writing forms are the same, the sound quality and even the pronunciation parts in the mouth are completely different. For example, r [ɻ] is a large tongue trill in Spanish (Ma 105), but it belongs to the tip of the tongue in Chinese pinyin. It makes Spanish learners susceptible to the influence of their mother tongue when they pronounce the "r" sound. Consciously transfer the large and small tongue sounds in the mother tongue, and the curling degree of the tongue tip does not meet the pronunciation requirements of the Chinese standard initial consonant "r," which makes it difficult to pronounce the accurate Chinese initial consonant "r." Besides, j, q, v, c, z, and so on all have comparably different pronunciations.
5. Conclusion and further work

In this paper, we proposed an acoustic model for mispronunciation detection of Mandarin phonemes that could aid in the CAPT area. According to the pronunciation error trend caused by the learner's inaccurate pronunciation and pronunciation, this paper uses deep neural network (DNN) for acoustic modeling and is based on the statistical speech recognition framework for pronunciation error trend detection. Besides, the acoustic feature detection system constructed in this article can more effectively capture specific types of errors. As our model achieves an accuracy of 80%, the results show that z has the highest pronunciation error rate among the initials, with an error rate of 58.14%, exceeding 50%, followed by c(46.15%), zh(40.48%), ch(39.53%), s(38.10%), and the correct rate of m is as high as 100%. Moreover, among the vowels, the highest pronunciation error rate is v, with an error rate of 71.43%, followed by ang, with an error rate of 61.76%, and the finals iu, eng, o, ing, en, un, ei, in, with an error rate of respectively 52.94%, 51.85%, 50.00%, 46.43%, 45.45%, 38.89%, 37.50%. Therefore, through linguistic classification, we discovered that there are mainly four pronunciation errors related to the native Spanish pronunciation habits. Consequently, this paper proposes practical suggestions from different perspectives for further Mandarin CALL international teaching according to the experiment results' evaluations.

1) Innovate teaching design, and create an exclusive corpus

Only through making good use of a variety of online teaching resources, creating high-quality courses, MOOCs, micro-classes, teaching corpora, teaching databases, and teaching cases databases, electronic reference books, etc., especially the entire corpus for Spanish native speakers, the resources for Spanish learners could be more comprehensively improved and utilized.

2) Enrich the types of online courses and improve customized teaching materials

Teachers need to adjust the teaching materials in advance to prepare for the training of key pronunciation and intonation to overcome the weaknesses in pronunciation of Hispanic learners to achieve better pronunciation correction effect and improve teaching efficiency.

3) Develop exclusive teaching software and improve the resource sharing platform

Even though many language learning apps have emerged in recent years, the typical ones for people speaking specific native languages (such as Spanish) are rare. Therefore, to make better use of the resource-sharing platform, it is necessary to develop exclusive teaching software for different languages that could help better adapt to the international education of Chinese Mandarin.

However, this system still has room for improvement, which can be improved from the following aspects: Firstly, specific distinguishing features (such as VOT, formant, etc.) or particular classifiers are introduced to detect specific types of pronunciation errors. In addition, it can be based on two. The statistical construction of corpus phoneme-level tagging can reflect the expanded recognition network.
of different phoneme probabilities (Luo et al.). Besides, the scale of training data must be increased to improve the acoustic model further.

References
[1] Bahl, L.R. Brown, P.F. de Souza, P.V. and Mercer, R.L. “Maximum mutual information estimation of hidden Markov model parameters for speech recognition.” 1986.
[2] Fixmer, E. “Why Is the Chinese Languages so Difficult to Learn?” Rosetta Translation, 25 Mar. 2015.
[3] Gao, Y. Xie, Y. Cao, W. and Zhang, J. “A study on robust detection of pronunciation erroneous tendency based on deep neural network”, Speech Commun. Assoc. 2015, pp. 693-696.
[4] Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. “Learning precise timing with LSTM recurrent networks”. JMLR, 2002, pp. 115–143.
[5] Graves, A.S. et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. ICML, 2006.
[6] Hu, W. Qian, Y. and Soong, F.K. “A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL).” Proc. InterSpeech-ISCA 2013, pp. 1886-1890.
[7] Hu, W. Qian, Y. Soong, F.K. and Wang, Y. “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers”, Speech Commun., vol. 67, 2015, pp. 154-166.
[8] Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning,” Sixth European Conference on Speech Communication and Technology, 1999.
[9] Joshi, S. Deo, N. and Rao, P. “Vowel mispronunciation detection using DNN acoustic models with cross-lingual training”, Speech Commun. Assoc. 2015, pp. 697-701.
[10] Kewley-Port, D. Watson, C. Maki, D. and Reed, D. “Speaker-dependent speech recognition as the basis for a speech training aid,” ICASSP, 1987, vol. 12, pp. 372-375.
[11] Li, H. Wang, S. Liang, J. Huang, S. and Xu, B. “High performance automatic mispronunciation detection method based on neural network and TRAP features”, Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. 2009, pp. 1911-1914.
[12] Li, K. Qian, X. and Meng, H. “Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks”, IEEE/ACM Trans. Audio Speech Language Process. 2017, vol. 25, no. 1, pp. 193-207.
[13] Luo, D; Yang, X and Wang, L. “Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus.” Twelfth Conference of the International Speech Communication Association, 2011.
[14] Ma, Xiaoyun, et al. “An Analysis of the Pronunciation Errors of the Chinese Pinyin "r" for International Students whose Native Language is Indo-European Languages”. Jingchu Academics, 2017 Issue 3 (Total Issue 10). Ed. Overseas Returnees Think Tank (Wuhan) Strategic Investment Management Co., Ltd. Company., 2017, 103-108.
[15] Mohamed, A.-R. Sainath, T. N. Dahl, G. Ramabhadran, B. Hinton, G. E. and Picheny, M. A. “Deep belief networks using discriminative features for phone recognition”, Speech Signal Process. (ICASSP), 2011, pp. 5060-5063.
[16] Muscariello, A. Gravier, G. and Bimbot, F. “Towards robust word discovery by self-similarity matrix comparison,” Proc. ICASSP, 2011.
[17] Qian, X. Meng, H.M. Soong, F.K. “The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training.” Proc. InterSpeech-2012, ISCA, 2012.
[18] Shi, Baoguang.,et al. “An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition,” ArXiv, 2015.
[19] Simonyan, Karen, Zisserman, Andrew. “Very Deep Convolutional Networks for Large-Scale Image Recognition,” ArXiv, 2014.
[20] Valueva, M.V., et al. “Application of the Residue Number System to Reduce Hardware Costs of the Convolutional Neural Network Implementation.” *Mathematics and Computers in Simulation*, 2020, pp. 232–243, doi:10.1016/j.matcom.2020.04.031.

[21] Wang, H. Liu, P. X. Li, S. and Wang, D. “Adaptive neural output-feedback control for a class of nonlower triangular nonlinear systems with unmodeled dynamics”, *IEEE Trans. Neural Netw.* 2018, vol. 29, no. 8, pp. 3658-3668.

[22] Wang, Y.-B. Lee, L.-S. “Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training.” *ICASSP*, 2012.

[23] Yin, Y. Zong, G. and Zhao, X. “Improved stability criteria for switched positive linear systems with average dwell time switching”, *J. Franklin Inst.* 2017, vol. 354, no. 8, pp. 3472-3484.

[24] Zhang, Wei. “Shift-invariant pattern recognition neural network and its optical architecture”. *Proceedings of Annual Conference of the Japan Society of Applied Physics*. 1988.

[25] Zhang, Y. Salakhutdinov, R. Chang, H. and Glass, J. “Resource configurable spoken query detection using deep Boltzmann machines,” *Proc. ICASSP*, 2012.

[26] Zhao, X. Wang, X. Zong, G. and Li, H. “Fuzzy-approximation-based adaptive output-feedback control for uncertain nonsmooth nonlinear systems”, *IEEE Trans. Fuzzy Syst.* 2018, vol. 26, no. 6, pp. 3847-3859.