

DEEP SEGMENTAL PHONETIC POSTERIOR-GRAMS BASED DISCOVERY OF NON-CATEGORIES IN L2 ENGLISH SPEECH

Xu Li, Xixin Wu, Xunying Liu, Helen Meng

Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong

{xuli, wuxx, xyliu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Second language (L2) speech is often labeled with the native, phone categories. However, in many cases, it is difficult to decide on a categorical phone that an L2 segment belongs to. These segments are regarded as non-categories. Most existing approaches for Mispronunciation Detection and Diagnosis (MDD) are only concerned with categorical errors, i.e. a phone category is inserted, deleted or substituted by another. However, non-categorical errors are not considered. To model these non-categorical errors, this work aims at exploring non-categorical patterns to extend the categorical phone set. We apply a phonetic segment classifier to generate segmental phonetic posterior-grams (SPPGs) to represent phone segment-level information. And then we explore the non-categories by looking for the SPPGs with more than one peak. Compared with the baseline system, this approach explores more non-categorical patterns, and also perceptual experimental results show that the explored non-categories are more accurate with increased confusion degree by 7.3% and 7.5% under two different measures. Finally, we preliminarily analyze the reason behind those non-categories.

Index Terms— Computer-aided pronunciation training, mispronunciation detection and diagnosis, segmental phonetic posterior-grams, non-categorical patterns

1. INTRODUCTION

Mispronunciation Detection and Diagnosis (MDD) is one of the core technology in Computer-Aided Pronunciation Training (CAPT) systems. It aims to detect the mispronunciation in a second language (L2) learner's speech, and further diagnose the error type and give learners effective feedback. There are several ways to address this task: The methods based on pronunciation scoring [1–4] are very popular and achieve a promising performance on mispronunciation detection. However, this kind of methods can not deal with mispronunciation diagnosis. Alternative methods, such as extended recognition network (ERN) [5–7] and acoustic-phonemic model (APM) [8], also perform well. ERN incorporates manually

Word	n o r t h		
Canonical Transcription	n a o r t h		
Real Pronunciation	l n a o r t h		
Traditional Annotation	n a o r t h	Detection	Diagnosis
Recognition Result 1	l a o r t h	Correctly Detected	Wrong
Recognition Result 2	n a o r t h	Missed	Wrong

Fig. 1. An example for how non-categorical mispronunciations are wrongly treated in traditional MDD

designed or data-derived phonological rules to generate possible phone paths in a word, including a canonical phonemic path and common mispronunciation paths. If a decoded path involves a mispronunciation path, then the mispronunciation can be detected and also diagnosed. Different from traditional acoustic models, APM feeds in not only acoustic features, but also phone context information for better performance.

Based on the categorical phone set, most existing approaches to model L2 speech only target categorical phone errors, but ignore non-categorical errors (i.e., the segments for which it is difficult to label as a single phone category). For example, L2 English speech uttered by native Cantonese speakers shows that the phone [n] is often pronounced as a sound that bears resemblance to both [l] and [n] ([l.n] in Fig. 1). In current MDD proposals, they are often coarsely labeled as one of the approximate phones. Fig. 1 shows an example where the canonical annotation for "north" should be [n a o r t h], but in face of the non-categorical pronunciation that resembles both [l] and [n], it may be recognized as either [l a o r t h] (as in Recognition Result 1) which enables mispronunciation detection but inaccurate diagnosis. Alternatively, if the non-categorical segment is recognized as [n a o r t h] (as in Recognition Result 2), it will fail to neither detect nor diagnose the mispronunciation. From the learner's perspective, the unawareness of non-categorical errors resulting from the training system will greatly decrease the efficiency of their learning, and they may even pronounce with their own accents through the training.

Concerned with the above problem, there are several ap-

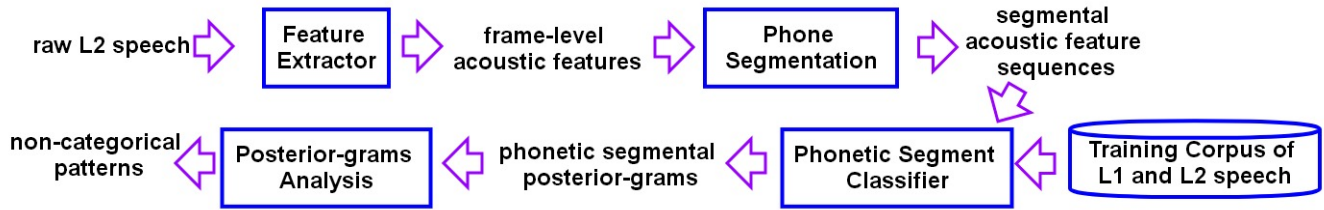


Fig. 2. Non-categorical patterns acquisition framework

proaches [9–13] proposed to capture the L2 pronunciation deviations from categorical phones. [9, 12, 13] try to capture each L2 error patterns of a given categorical phone, while [11] focuses on extending a native phone set to include the non-categorical patterns in L2 English speech. Actually, this work is an extension of [11]. The distinction lies in 1) this approach extracts segment-level features to model the phonetic information while [11] uses non-segmental frame-level features and their average, thereby the accumulated errors are reduced and the discovered non-categories are more accurate with higher confusion degree; 2) this work uses a more simple but effective method to explore non-categories, while not involving k-means clustering method used in [11]. This prevents a data imbalance problem and high-dimension data issues in k-means clustering, and hence discovers more non-categories.

The contributions of this work include 1) introduce a novel non-categorical patterns acquisition framework; 2) preliminarily analyze the reasons behind these non-categories.

The rest of this paper is as follows. In Section 2, we describe the framework and approaches. Experiment setups and results are presented in Section 3 and some analysis for these discovered non-categories are conducted in Section 4. Finally, Section 5 concludes the paper.

2. FRAMEWORK AND APPROACHES

The framework in this work is shown in Fig. 2. We first extract acoustic features of the raw speech on frame level. And then, since the text transcription is available, forced alignment is performed to concatenate acoustic features into segment-level features corresponding to phonemes. Feed on these segmental acoustic feature sequences, a phonetic segment classifier is trained on the corpus of both L1 and L2 speech. We use output of the softmax layer, i.e. segmental phonetic posterior-grams (SPPGs) to represent phone segment-level information. This representation is widely used in the L2 English acoustic-phonemic space [9, 12, 14, 15] due to its speaker independence, and more environment-robust properties. Based on SPPGs, an effective method is used to explore non-categories.

2.1. Phonetic segment classifier

This phonetic segment classifier consists of convolutional layers, recurrent layers and dense layers, as shown in Fig. 3. This unified architecture [16] takes advantages of the complementarity of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Deep Neural Networks (DNNs), and it is widely used in Automatic Speech Recognition (ASR) tasks [17–19]. In this work, we applied 2D CNN upon the input segmental acoustic feature sequences. The output of the final CNN layer is fed to the RNN layer. Since Gated Recurrent Unit (GRU) [20] has a performance on par with Long Short-Term Memory (LSTM) [21] while is more efficient in computation, we adopt GRU layer in our model. The final state of the GRU layer is fed to the FC layers and the softmax layer. Within this structure, CNNs serve as the speaker-independent feature extractor, which is robust to spectral variation. RNNs can model the temporal characteristics within a phone segment. DNNs are used to transform the final hidden features into a space that makes the output easier to classify.

To train this phonetic segment classifier, we need L1 (native) speech to implicitly guide the model to learn what canonical pronunciation of a phone should be. At the same time, since second language (L2) space has some deviations from the native space, we also need L2 speech to adapt the model parameters to fit it well. Hence both L1 and L2 speech corpus are used for training this classifier. In Section 3, the segment-level phone recognition rates are shown under different training corpus setups, and it indicates the importance and correctness of using both L1 and L2 speech to train the model.

2.2. Non-categories exploration

The output of the classifier is segmental phonetic posterior-grams, which is a probability vector where each element corresponds to a phone category. The higher probability one element preserves, the more likely the segment belongs to the phone category which the element corresponds to. According to this, for a segment of canonical pronunciation, the classifier will output a vector with a high probability (confidence) on that canonical phone element. However, for the segments of L2 non-categories, the classifier will be ambiguous on more than one element, which results in a probability

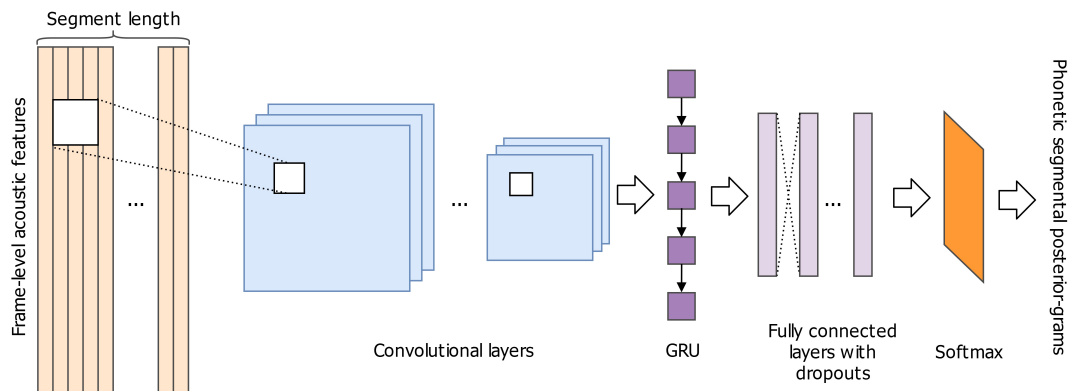


Fig. 3. Phonetic segment classifier architecture

Table 1. Experimental data setup

	training set	evaluation set
TIMIT (L1)	3.0 hours (144K segments)	0.4 hours (18K segments)
CU-CHLOE-C (L2)	5.5 hours (139K segments)	0.7 hours (19K segments)

Table 2. Segment-level phone recognition rate

training set	training set	evaluation set (L1)	evaluation set (L2)
L1	74.27%	72.92%	32.12%
L2	80.08%	27.24%	71.02%
L1+L2	75.24%	72.06%	70.29%

vector with more than one peak of comparable probabilities. From this point, we can set a threshold and look for the segments where the probability vector has more than one element higher than this threshold. These segments can be regarded as non-categories. After this exploration, we can use an identical name to denote all the segments with the same element set above the threshold. Then we have non-categorical patterns. For example, we can use l_n to represent the non-categorical segments with an element set of $\{l, n\}$ above a given threshold. Section 3 gives the perceptual test results on these explored non-categories.

3. EXPERIMENTS

3.1. Experimental setup

Our experiments are conducted on two speech corpora: 1) the CU-CHLOE-C (Chinese University-Chinese Learners of English - Cantonese speakers) data set [22] as the L2 speech corpus; and 2) the TIMIT data set as the L1 speech corpus. Both L1 and L2 corpus are divided by speakers into training set and evaluation set, as shown in Table 1.

The acoustic feature adopted in this work is 13 dimensional Mel-Frequency Cepstrum Coefficient (MFCC). There are 3 CNN layers with 64 output channels for each. The kernel size is 3 by 3, and no pooling function is used. The GRU layer has 128 hidden units. Its final state is transformed to output by 3 dense layers with 512 hidden units per layer, and the dropout [23] rate is set as 0.2 to prevent over-fitting prob-

lems. The implementation software used in this work is Pytorch 0.4.1 [24]. We use the combined training set of L1 and L2 to train the model, and use the evaluation set to test the performance on each corpus. During the training, we split 10% data from training set as the validation set. Adam [25] is used as the optimizer, and the initial learning rate is set as 0.001. The probability threshold used for exploring non-categories is set to 0.4.

3.2. Experiment results

Table 2 shows the phone recognition performance of our system trained on different training sets. If only L1 corpus involved in training, the system can have a phone recognition rate of 72.92% and 32.12% on L1 and L2 evaluation set respectively. This implies there is some deviation from L2 space to L1 space. On the counterpart, if only L2 corpus involved in training, the system has a phone recognition rate of 27.24% and 71.02% on L1 and L2 evaluation set respectively. However, if a combined training set of both L1 and L2 is used, the performance on L1 and L2 evaluation set can almost respectively reach the phone recognition rate when merely using the corresponding corpus as the training set. This implies even though there is some deviation between L1 and L2 space, a model trained on a combination training set can project them into a common space and still preserve the recognition performance. This system will be used for non-categories exploration.

Table 3. Average proportion of four options in each non-categories (The previous results [11] are shown in parentheses). The higher score of Option 3 and lower score gap between Option 1 and 2 are preferable.

	[ax_er]	[aa_ao]	[ey_ih]	[ae_ay]	[eh_ey]	[d_t]	[l_n]	[b_p]	[f_v]	[m_n]
Option 1	38.9% (46.0%)	35.2% (16.1%)	56.5% (46.6%)	25.9% (25.0%)	14.7% (17.7%)	49.0% (57.8%)	24.1% (56.6%)	27.8% (18.8%)	51.9% (37.2%)	41.1% (55.6%)
Option 2	37.0% (30.6%)	48.1% (50.0%)	25.0% (27.8%)	44.4% (47.5%)	63.7% (60.4%)	28.7% (18.0%)	43.5% (21.7%)	35.2% (35.2%)	29.6% (45.7%)	36.3% (27.5%)
Option 3	13.0% (3.2%)	9.3% (9.7%)	13.9% (4.4%)	14.9% (7.3%)	19.6% (3.1%)	16.7% (7.0%)	12.0% (8.0%)	15.7% (8.6%)	14.8% (11.0%)	15.7% (8.8%)
Option 4	11.1% (20.2%)	7.4% (24.2%)	4.6% (21.2%)	14.8% (20.2%)	2.0% (18.8%)	5.6% (17.2%)	20.4% (13.7%)	21.3% (37.4%)	3.7% (6.2%)	6.9% (8.1%)
Score gap ¹	1.9% (15.4%)	12.9% (33.9%)	31.5% (18.8%)	18.5% (22.5%)	49.0% (42.7%)	20.3% (39.8%)	19.4% (34.9%)	7.4% (16.4%)	22.3% (8.5%)	4.8% (8.1%)

Option 1: More similar to P_1 ; Option 2: More similar to P_2 ;
Option 3: Equal similarity to P_1 and P_2 ; Option 4: Not similar to either P_1 or P_2 .

3.3. Perceptual tests

To verify the explored non-categories indeed perceptually different from their corresponding categorical phones, we designed perceptual tests on series of non-categorical confusion groups. Each non-categorical confusion group consists of audios from the non-category, and also the corresponding categorical phones. In our case, since the threshold to explore non-categories is set to 0.4, we only consider the non-categories that resemble between two categories. For simplified writing, we mark the non-category that resemble two categories P_1 and P_2 as $P_1_P_2$.

In each group, we randomly select 12 audios (6 audios from the non-category, 3 audios from each of two categories). For the audios in each category, we select the segments where the SPPGs have a high probability (confidence) of 0.9 at that phone element.

These perceptual tests are conducted in the same way as [11]. During the tests, the listeners are asked to give their perceptual choice among 4 options after carefully listening each segment with short contexts: 1) More similar to P_1 ; 2) More similar to P_2 ; 3) Equal similarity to P_1 and P_2 ; 4) Not similar to either P_1 or P_2 . 25 listeners majoring in linguistics or English were invited to participate in these perceptual tests.

Fig. 4 shows the average proportion of 4 options in three sample non-categorical confusion groups. The complete confusion scores are provided in Supplemental Material A. From these pie charts, we observe that listeners almost have an agreement on native phone categories, and give the corresponding phones as their choices. However, for non-categories, there is no majority perceptual vote among the given categorical phones. This implies these non-categories are indeed perceptually different from the corresponding categories, and bear a resemblance to both of them.

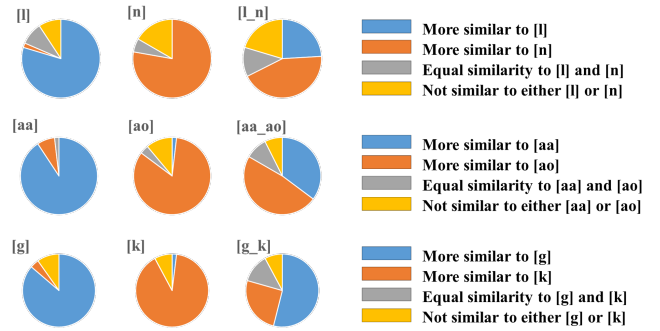


Fig. 4. Statistic perceptual results of each non-categorical confusion group

To compare the confusion degree of the explored non-categories in [11] and this work, we display the averaged confusion scores of the non-categories that are discovered in both [11] and this work. Table 3 shows the average proportion of each option in such non-categories. The averaged confusion scores in this work are shown in the first line of each option, while the previous results are shown within parentheses in the second line. The score gap between Option 1 and 2 is worked out and displayed in the bottom part. We observe that for most non-categories, such as [ax_er], [aa_ao] and [d_t], the score gap of the first two options is much smaller in this work. From the average sense, this score gap is decreased from 26.1% to 18.8%. Moreover, the averaged score of Option 3 is also obviously increased from 7.1% to 14.6%. These two observations verify that the non-categories explored in this approach are more accurate with higher confusion degree.

Table 4 shows the explored non-category set compared with that in [11]. There are 9 additional and 10 existing non-categories. To explore the possible reason, the k-means clustering algorithm used in [11] may be subjected to a data im-

¹The score gap between Option 1 and 2

[n] with [l] in colloquial Cantonese, for e.g. (you) [nei] pronounced as (logic) [lei] [26]. Hence, Cantonese learners often unconsciously perform these substitutions in English, resulting in the non-category [l_n].

5. CONCLUSIONS

This work introduces a novel non-categorical phone acquisition framework, based on SPPGs learned via deep neural networks. A phonetic segment classifier is utilized to extract SPPGs, and we apply a simple but effective method to explore the potential non-categories. This approach not only helps increase the confusion degree of the explored non-categories, thereby improving the accuracy of non-categories exploration, but also discovers more non-categorical patterns. Possible reasons behind the non-categories are also given based on the phone inventories comparison between L1 and L2 languages. Since this framework has not language specific constraint, it can be smoothly adapted to other language pairs. In the future, how to facilitate MDD performance with these non-categories will be investigated.

6. REFERENCES

- [1] H. Franco, L. Neumeyer, and et al., “Automatic pronunciation scoring for language instruction,” in *Proc. ICASSP*. IEEE, 1997, vol. 2, pp. 1471–1474.
- [2] 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.
- [3] J. Zheng, C. Huang, and et al., “Generalized segment posterior probability for automatic mandarin pronunciation evaluation,” in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. IV–201.
- [4] 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, pp. 1886–1890.
- [5] A. M. Harrison, W. Lo, and et al., “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.
- [6] W. Lo, S. Zhang, and H. Meng, “Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system,” in *International Speech Communication Association*, 2010.
- [7] X. Qian, F. K. Soong, and H. Meng, “Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (capt),” in *International Speech Communication Association*, 2010.
- [8] K. Li, X. Qian, and H. Meng, “Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks,” *IEEE/ACM TASLP*, vol. 25, no. 1, pp. 193–207, 2017.
- [9] Y. Wang and L. Lee, “Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning,” *IEEE/ACM TASLP*, vol. 23, no. 3, pp. 564–579, 2015.
- [10] A. Lee, N. F. Chen, and J. Glass, “Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery,” in *Proc. ICASSP*. IEEE, 2016, pp. 6145–6149.
- [11] S. Mao, X. Li, and et al., “Unsupervised discovery of an extended phoneme set in L2 English speech for mispronunciation detection and diagnosis,” in *Proc. ICASSP*. IEEE, 2018, pp. 6244–6248.
- [12] Y. Wang and L. Lee, “Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning,” in *Proc. ICASSP*. IEEE, 2013, pp. 8232–8236.
- [13] X. Li, S. Mao, and et al., “Unsupervised discovery of non-native phonetic patterns in L2 English speech for mispronunciation detection and diagnosis,” *Interspeech*, pp. 2554–2558, 2018.
- [14] L. Sun, K. Li, and et al., “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in *ICME*. IEEE, 2016, pp. 1–6.
- [15] A. Lee, Y. Zhang, and J. Glass, “Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams,” in *Proc. ICASSP*. IEEE, 2013, pp. 8227–8231.
- [16] T. N. Sainath, O. Vinyals, and et al., “Convolutional, long short-term memory, fully connected deep neural networks,” in *Proc. ICASSP*. IEEE, 2015, pp. 4580–4584.
- [17] D. Amodei, S. Ananthanarayanan, and et al., “Deep speech 2: End-to-end speech recognition in English and Mandarin,” in *ICML*, 2016, pp. 173–182.
- [18] W. Xiong, L. Wu, and et al., “The Microsoft 2017 conversational speech recognition system,” in *Proc. ICASSP*. IEEE, 2018, pp. 5934–5938.
- [19] G. Trigeorgis, F. Ringeval, and et al., “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proc. ICASSP*. IEEE, 2016, pp. 5200–5204.
- [20] K. Cho, B. Van Merriënboer, and et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *Proc. EMNLP*, 2014.
- [21] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [22] H. Meng, W. Lo, and et al., “Development of automatic speech recognition and synthesis technologies to support Chinese learners of English: The CUHK experience,” *Proc. AP-SIPA ASC*, pp. 811–820, 2010.
- [23] N. Srivastava, G. Hinton, and et al., “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [24] N. Ketkar, *Introduction to PyTorch*, pp. 195–208, Apress, Berkeley, CA, 2017.

- [25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *ICLR*, 2015.
- [26] H. Meng, Y. Y. Lo, and et al., “Deriving salient learners mispronunciations from cross-language phonological comparisons,” in *ASRU. IEEE Workshop on*. IEEE, 2007, pp. 437–442.
- [27] R. Lado, “Linguistics across cultures: Applied linguistics for language teachers.,” 1957.
- [28] E. Zee, “Chinese (hong kong cantonese),” *Journal of the International Phonetic Association*, vol. 21, no. 1, pp. 46–48, 1991.
- [29] D. M. Decker et al., *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999.