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

Automatic Detection of Phonological Errors in Child Speech Using Siamese Recurrent Autoencoder

Si-Ioi Ng, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong
siioing@link.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

Speech sound disorder (SSD) refers to the developmental disorder in which children encounter persistent difficulties in correctly pronouncing words. Assessment of SSD has been relying largely on trained speech and language pathologists (SLPs). With the increasing demand for and long-lasting shortage of SLPs, automated assessment of speech disorder becomes a highly desirable approach to assisting clinical work. This paper describes a study on automatic detection of phonological errors in Cantonese speech of kindergarten children, based on a newly collected large speech corpus. The proposed approach to speech error detection involves the use of a Siamese recurrent autoencoder, which is trained to learn the similarity and discrepancy between phone segments in the embedding space. Training of the model requires only speech data from typically developing (TD) children. To distinguish disordered speech from typical one, cosine distance between the embeddings of the test segment and the reference segment is computed. Different model architectures and training strategies are experimented. Results on detecting the 6 most common consonant errors demonstrate satisfactory performance of the proposed model, with the average precision value from 0.82 to 0.93.

Index Terms: child speech, speech sound disorder, Siamese recurrent auto-encoder

1. Introduction

Children who suffer from speech sound disorder (SSD) commit persistent errors in producing certain speech sounds after the expected age of acquisition. Untreated children with SSD may experience social and academic difficulties, which impact their personal growth in the long term. Currently clinical assessment of SSD is carried out by qualified speech and language pathologists (SLPs) based on perceptual evaluation. The assessment can take various forms, including articulation test, conversation, story telling, etc. The result of each form of test reveals the severity and details of specific speech sound developmental problems. The assessment criteria are established and validated by experts. Timely diagnosis of SSD is crucial to effective treatment and rehabilitation. This is, however, hindered by the significant manpower shortage of SLPs globally. Methods of automatically detecting speech sound errors are highly desired to reduce the pressure on SLPs and benefit a large population of patients.

Child SSD detection is the task of distinguishing abnormal speech sound production from typical ones based on acoustic speech signals. Possible approaches include template matching, statistical modeling and automatic speech recognition (ASR). Given a limited amount of speech data, Yeung et al. [1] investigated an exemplar-based approach to evaluating English rhotic sounds in child speech. With a good amount of data for statistical modeling, Dudy et al. [2][3] improved the goodness of pronunciation (GOP) [4] measure for pronunciation analysis in disordered child speech. Phonetic knowledge about common realizations of target phonemes was applied in the analysis. Similar approaches of knowledge incorporation were found in other works. In [5], assessment of childhood apraxia of speech (CAS) was performed using constrained lattice in an ASR system. The lattice has the advantage that the type of mispronunciation could be beyond a binary decision. For each target word, the lattice was created according to expected mispronunciation rules. This approach was further extended in [6], where phonological error patterns were identified via fine tuning of state transition weights between the correct and mispronounced phone sequences of a target word. However, the nature of being unpredictable in mispronunciation would challenge such systems, which rely on prior knowledge about the concerned errors.

In recent years fixed-dimension representation of speech has been applied widely to speech modeling and classification problems. Such representation encodes the information of variable-length speech segments in low-dimension vectors, which allow different segments to be compared and analyzed in the same embedding space. The similarity between segments can be evaluated by Euclidean distance, cosine distance, or other distance measures. Many approaches have been proposed for extracting embedding from speech. In the present study, the use of sequence-to-sequence auto-encoder (AE) is investigated. It is a neural network model that encompasses an encoder-decoder architecture. The encoder converts the input sequence into a low-dimension embedding while the decoder aims to reconstruct from the embedding an output sequence that is the same as or closely related to the input. The applications of sequence-to-sequence AE are found in unsupervised spoken term discovery, query-by-example spoken term detection and speaker verification, etc. [7][8][9].

A common type of child SSD can be described as the desired phone, typically a consonant, being substituted by another phone. In this study, detection of such phonological errors is formulated as the problem of pairwise contrast between relevant phone segments, based on the embedding representations generated by an AE model. In terms of the network architecture, the AE is combined with a Siamese network, which is jointly trained to contrast the phone segments in the embedding space. Different model setups are evaluated first on test data of "artificial" substitution errors. Subsequently the proposed approach is applied to detect real phonological errors produced by children with SSD.

2. Background & Speech Database

2.1. Speech acquisition by Cantonese-speaking children

The present study is focused on Cantonese, a major Chinese dialect that is widely spoken in Hong Kong, Macau, Guangdong and Guangxi Provinces of Mainland China, as well as overseas

http://dx.doi.org/10.21437/Interspeech.2020-2145

Chinese communities. Cantonese is a monosyllabic and tonal language. Each Chinese character is pronounced as a single syllable carrying a lexical tone. A Cantonese syllable can be divided into an Initial part and a Final part. The Initial is a consonant while the Final could be a diphthong or comprise a vowel nucleus followed by a consonant coda (final consonant). There are a total of 19 consonants, 11 vowels and 11 diphthongs in Cantonese. The present-day Cantonese uses over 700 legitimate syllables (Initial-Final combinations). If the tone difference is taken into account, the number of distinct syllables exceeds $1,600$ [10][11]. In this study, we focus on Cantonese spoken in Hong Kong. The target group of speakers is pre-school children in Hong Kong.

In [12], So and Dodd examined speech sounds of typically developing (TD) and Cantonese-speaking pre-school children. It was shown that children were able to acquire tones, most of the vowels and diphthongs by the age of 2;0 (years;months). The acquisition of final consonants and initial consonants was achieved by the age of 4;6 and 5;0 respectively. To at el. [13] investigated acquisition of Hong Kong Cantonese by children aged 2;4 to 12;4. The study revealed a longer time required for speech sound acquisition. Vowels and diphthongs were acquired by 5;0 and 4;0 respectively, and all initial consonants were acquired by 6;0. In the process of speech sound acquisition, children may try to simplify a target speech sound by substituting it with other sounds. This is mainly due to the undeveloped motor skills for speech sound production. TD children gradually stop using the substitution sounds and return to typical pronunciation when they grow up. Nonetheless, some children would persist the substitution errors beyond the expected age of acquisition. The symptoms are referred to as phonological disorder and disordered children are recommended to seek treatment offered by SLPs.

Table 1: *Statistics of speakers in available speech data*

Age (years;months)	3;0-3;11	4;0-4;11	5;0-5;11	6;0-6;11
Male, healthy	7	26	31	14
Female, healthy	13	33	35	20
Male, atypical	9	9	5	1
Female, atypical	6	9	6	0

2.2. Child Speech Database: CUCHILD

A Cantonese child speech corpus named CUCHILD is used in the present study [14]. The corpus contains speech data collected from $1,986$ kindergarten children (aged 3;3-6;11) in Hong Kong. All speakers use Cantonese as their first language (L1). CUCHILD is designed to support acoustic modeling of Cantonese child speech and research on automatic assessment of SSD [15][16]. The speech material consists of a total of 130 Cantonese words of 1 to 4 syllables in length, covering the 19 consonants and 11 most commonly used vowels. Speech recording was carried out in classrooms provided by the kindergartens. A digital recorder was located at 20-50 centimeters in front of the children's mouth. Yet environmental noise such as reverberation, school bells, people walking around, etc. was unavoidable. To minimize effects of background noise, the gain and the position of recorders were adjusted manually. Child speech was elicited via a picture naming task. Each word was also accompanied by a pictorial illustration. A research assistant showed the pictures one by one and guided the child to speak the intended words.

All participants were assessed with the Hong Kong Cantonese Articulation Test (HKCAT) [17]. The HKCAT is a standardized test for children which reflects the severity of developmental delay and the types of speech sound errors. Among all participants, 230 children were found to have SSD.

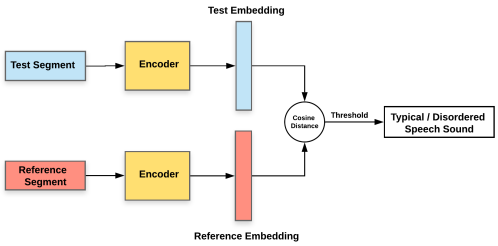

Figure 1: *Speech sound disorder (SSD) detection system.*

The speech data were collected recently and detailed work of data processing and annotation are still ongoing, The present study makes use of a subset of the whole corpus, which covers the recordings from 233 child speakers. The data was manually annotated and segmented into child speech and research assistants' speech. Spoken words manifesting SSD were labelled manually by SLPs. The syllable-level orthographic transcriptions were manually verified. Table 1 summarizes the speaker information in our dataset.

3. The Proposed System of SSD Detection

3.1. SSD detection system

In clinical assessment of SSD, the child is guided to speak a list of test words. The responsible speech pathologist observes the speech production and decides if the child makes errors on specific parts of the words. The judgement depends highly on the clinician's experience in differentiating atypical speech sounds from typical ones.

Towards automated assessment of SSD, the proposed system aims to determine whether a phonological error occurs in a test speech segment. The test segment contains a specific phoneme as part of a test word spoken by the child. To detect the error, we may choose one or multiple reference segments of the expected speech sound to compare with the test segment in a pairwise manner. Using multiple reference segments is preferred as they can represent the deviation of the expected speech sound. As illustrated in Figure 1, the comparison is in an embedding space and all embeddings are extracted by the encoder obtained from the trained Siamese RAE model. The cosine distance is computed for each pair of embedding. The binary decision is based on a pre-defined threshold. If the score is above the threshold, the test segment is classified as typical pronunciation. Otherwise it is a disordered pronunciation.

The present study is focused on a set of initial consonants in Cantonese, which are considered as reliable markers for child speech acquisition. Details of the model are described in the following sections.

3.2. Recurrent autoencoder

A recurrent autoencoder (RAE) model is used to generate a compact representation of phone segment. This representation is referred to as the embedding. The RAE converts variable-length phone segments into fixed-dimensional embedding vectors, on which distance or similarity measure could be applied straightforwardly. The RAE has three components. The encoder receives an input sequence. The hidden state of the encoder's last layer reaches the linear layer and generates the embedding, which is passed to the decoder to construct the

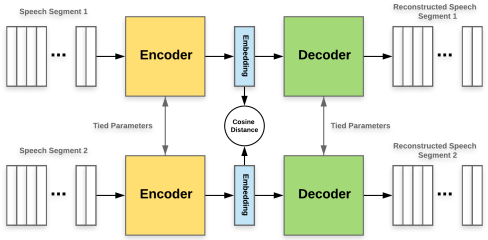

Figure 2: *Siamese network architecture.*

output sequence. The RAE is trained such that the embedding is adequate for reconstructing a certain type of target output. One common choice of the target output sequence is to make it equal to the input sequence. This can be achieved by minimizing the mean squared error (MSE) loss in the training of encoder and decoder networks. For the input sequence $S = [\, \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, ..., \boldsymbol{x_T} \,]$, the MSE loss is given as,

$$L_{mse} = \sum_{t=1}^{T} \|\boldsymbol{x_t} - D_t(E(S))\|^2 \tag{1}$$

where $D_t(\cdot)$ refers to the decoder output at time step t and $E(\cdot)$ denotes the last hidden layer output of the encoder, while T is the length of input sequence.

The RAE model is also commonly applied with a weakened input-output relation, i.e., without requiring the decoder to perform exact recovery of the input sequence. Such design aims at sharing mutual information between non-identical but closely related training segments [7]. This type of RAE is known as the correspondence RAE (Cor-RAE). In this work, the Cor-RAE model is trained using speech segments carrying the same phoneme. Consider a pair of segments $S_1 = [\, \boldsymbol{x_1^{(1)}}, \boldsymbol{x_2^{(1)}}, \boldsymbol{x_3^{(1)}}, ..., \boldsymbol{x_{T_1}^{(1)}} \,]$ and $S_2 = [\, \boldsymbol{x_1^{(2)}}, \boldsymbol{x_2^{(2)}}, \boldsymbol{x_3^{(2)}}, ..., \boldsymbol{x_{T_2}^{(2)}} \,]$ from the same phoneme category, the MSE loss for the training of Cor-RAE is,

$$L_{mse} = \sum_{t=1}^{T_2} \|\boldsymbol{x_t^{(2)}} - D_t(E(S_1))\|^2 \tag{2}$$

where T_1 and T_2 denote the lengths of S_1 and S_2 respectively.

3.3. Siamese recurrent autoencoder

As discussed earlier, the task of phonological error detection is formulated as a process of contrasting a test segment against the target phonemes. This process is realized with a Siamese network. It consists of two identical neural networks with shared parameters, which process two input representations in parallel. By inserting the Siamese loss in training, the network parameters are optimized to learn the similarity between the input representations. Our implementation of the Siamese RAE follows the work in [18], where the loss is computed with a pair of embeddings extracted from the RAE, as shown in Figure 2.

Two types of Siamese loss are considered and compared in this work. The first one is the contrastive loss, which is expressed as,

$$L_c = \frac{1}{2} y * d + \frac{1}{2}(1-y) * max(0, m-d), \tag{3}$$

where $d = 1 - cos(\boldsymbol{z_1}, \boldsymbol{z_2})$, and $\boldsymbol{z_1}, \boldsymbol{z_2}$ are the pair of embeddings representing two input speech segments. Both embeddings are generated by the encoder in the Siamese RAE.

The other type of loss function is the triplet loss defined as,

$$L_t = max(0, m + d_{ap} - d_{an}), \tag{4}$$

The loss function involves three embeddings as input, which include an anchor z_a, a positive sample z_p and a negative sample z_n. d_{ap} and d_{an} are the cosine distances of the anchor-positive pair and the anchor-negative pair respectively, and $d_{ap} = 0.5 * (1 - cos(\boldsymbol{z_a}, \boldsymbol{z_p}))$.

The overall objective function for Siamese RAE training combines the MSE loss and the contrastive/triplet loss as,

$$L_{mse,c} = (1-w) * L_c + w * \frac{L_{mse1} + Loss_{mse2}}{2} \tag{5}$$

$$L_{mse,t} = (1-w) * L_t + w * \frac{L_{mse\_a} + L_{mse\_p} + L_{mse\_n}}{3} \tag{6}$$

where w is a scalar weight to balance the reconstruction loss and similarity loss.

4. Experiments and Results

4.1. Data pre-processing

Consonant segments in TD and atypical child speech were extracted automatically by forced alignment with GMM-HMM triphone models. The triphone models were trained with speech data from 80 TD children of age $5;0$ - $6;11$ among the 233 speakers as summarized in Table 1. TD children in this age range are expected to make few mistakes in speech production and their speech are considered to be free of SSD problems. Acoustic features for GMM-HMM training consist of 13-dimensional Mel-frequency cepstral coefficients (MFCC) and their first- and second-order derivatives extracted every 0.01 second. For triphone model training, linear discriminant analysis (LDA), semi-tied covariance (STC) transform and feature space Maximum Likelihood Linear Regression (fMLLR) were applied [19][20][21]. With a basic syllable pronunciation dictionary, an error rate of 17.35% was achieved on the task of free-loop syllable recognition with test speech from 15 unseen TD children in the same age range. Forced alignment was applied to the speech data shown in Table 1 according to the canonical pronunciations of the 130 test words. Feature extraction, acoustic model training and forced alignment were all carried out with the Kaldi speech recognition toolkit [22]. As a result, a pool of consonant segments were extracted and they were divided into different subsets as shown in Table 2.

Table 2: *Summary of data (phone segments) for training and evaluation of the RAE model.*

Name of subset	Clinical group	Age range	No. of segments
Training	TD	5;0 - 6;11	17400
Reference	TD	5;0 - 6;11	4000
Development	TD	5;0 - 6;11	3500
Test1	TD	3;0 - 4;11	21000
Test2	TD & Disordered	3;0 - 6;11	706 & 726

4.2. Training of the Siamese RAE

In this study, the gated recurrent units (GRU) are adopted as the recurrent neural network architecture in the Siamese RAE [23]. The input representations are 40 dimensional Filter-bank features, with mean and variance being globally normalized. The training phone segments are paired randomly. A training target '1' is assigned to the pairs of same-class segments , and '0'

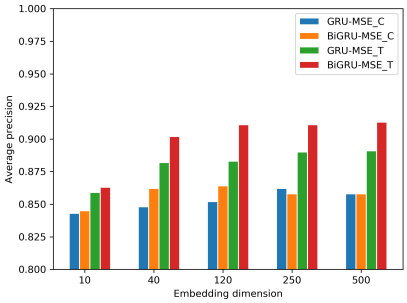

Figure 3: *Performance on the development set.*

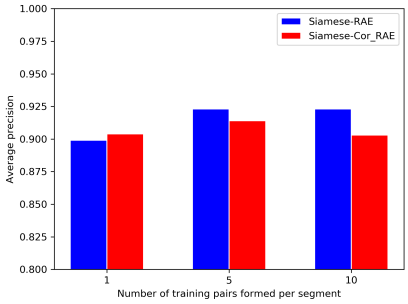

Figure 4: *Performance on artificial errors.*

assigned to pairs of segments from different classes. The encoder and decoder networks both consist of 3 hidden layers and 400 hidden units. The embeddings are L2-normalized. Both the Siamese RAE and Siamese Cor-RAE models are trained by the Adam optimizer [24] with a batch size of 256, a learning rate of 10^{-4}, weight decay of 10^{-5} and for 50 epochs. Training of the Siamese Cor-RAE starts with a pre-trained standard Siamese RAE model. The margin is 0.9 for the contrastive loss and 0.25 for the triplet loss. A loss weight of 0.5 is applied to both loss functions. The training processes are implemented with PyTorch [25].

Different embedding sizes, loss functions (contrastive vs. triplet) and Siamese RAE model designs are evaluated on the development set. The evaluation is carried out on the same-different discriminability task as described in [26]. Given a pair of test segments $(p_i, p_j), i \neq j$, p_i and p_j are declared to contain the same phoneme if the embedding distance $d \leq \tau$, where τ is the decision threshold. Each development segment is randomly paired with a segment from the reference dataset. The segment pairs assigned with '0' are regarded as artificial substitution errors, in which the target phone is substituted by another phone. The cosine distance is computed for each segment pair. The average precision (AP) is used as the evaluation metric of system performance. The value of AP is obtained from the precision-recall (PR) curve, which portrays the system performance across varying decision thresholds. The results in terms of AP are shown as in Figure 3. Overall, using the triplet loss and bi-directional network structure (BiGRU-MSE_T) leads to better performance on the same-different task. In the following experiments, this setting with an embedding size of 120 is used.

4.3. Performance evaluation on artificial errors

In this part, the *Test1* dataset in Table 2 is used to evaluate the performance of the Siamese RAE and Siamese Cor-RAE. Each test segment in *Test1* is paired up with a segment randomly selected from the reference set. The results in terms of AP are reported as in Figure 4. In the figure we also compare different training strategies in which each segment in the training dataset is used to form 1, 5 and 10 training pairs with other training segments. It can be seen that the conventional Siamese RAE consistently outperforms the Siamese Cor-RAE. The change of training pairs shows a noticeable impact on the performance level. The results imply that it is beneficial to use more training pairs. However, a large number of training pairs would not yield further improvement, in particular for the Siamese Cor-RAE. It seems the learning of mutual information shared across short segments from the same phoneme in Siamese Cor-RAE does not work as successfully as where sub-word and word level speech units are used [7][27]. This could be caused by the short duration of speech units or hyperparameter settings of the model. More works are required to draw definitive conclusions.

4.4. Performance evaluation on real errors

The Siamese RAE with Bi-GRU trained with 5 training pairs, i.e., the best performing model shown in Figure 4, is evaluated on the task of detecting real phonological errors with the *Test2* dataset. The test data cover the 6 most common error patterns, which concern the Cantonese consonants /f/, /k/, /s/, /k^h/, /t^h/ and /p^h/. The errors are made on the phonological processes of stopping (e.g. /f/ to /p/), fronting (e.g. /k/ to /t/), deaspirtation (e.g. /k^h/ to /k/) and affrication (e.g. /s/ to /ts/) etc. They are caused mainly by incorrect place or manner of articulation. It should be noted that children with SSD often had incomplete speech sound inventories and the errors were not limited to these common patterns.

Table 3: *AP on real consonant errors.*

Consonant	Error Pattern	No. of Consonant Segments {Disordered, Typical, Reference}	AP
/f/	Stopping	{65,93,153}	0.898
/k/	Fronting	{52,137,409}	0.824
/s/	Affrication, Stopping	{178,183,294}	0.917
/k^h/	Deaspiration, Fronting	{141,102,169}	0.938
/t^h/	Deaspiration, Backing	{205,115,220}	0.861
/p^h/	Deaspiration	{85,76,126}	0.921

Each test segment is compared with all reference segments carrying the same consonant. The average cosine distance is computed. The results in terms of AP are shown as in Table 3. The highest AP is achieved on the detection of atypical aspirated consonant /k^h/ sound, while the performance of detecting unaspirated consonant /k/ is the worst among all test patterns. This suggests that unaspirated consonants may not be reliably detected in automatic assessment of SSD. It was noted that the Siamese Cor-RAE did not yield good performance. The use of the correspondence model for SSD detection with higher-level speech units (e.g. syllable or word level) will be investigated in our future work.

5. Conclusion

An approach to automatic detection of phonological errors in child speech has been investigated and evaluated with both artificial and real speech sound errors. It has been shown that the proposed Siamese Recurrent Auto-encoder model is able to learn compact representations from variable-length speech segments, which are effective in distinguishing erroneous segments from correct ones. Specifically, for the 5 most common consonant errors in Cantonese, the achieved values of average precision range from 0.82 to 0.93. These results reveal the good potential of applying the proposed approach to automatic assessment of speech sound disorder in real-world settings. Future work will include the incorporation of clinical knowledge in the model design and the discovery of domain knowledge through acoustical analysis of child speech.

6. References

[1] G. Yeung, A. Afshan, K. E. Ozgun, C. Kaewtip, S. M. Lulich, and A. Alwan, "Predicting clinical evaluations of children's speech with limited data using exemplar word template references." in *Proc. of SLaTE*, 2017, pp. 161–166.

[2] S. Dudy, M. Asgari, and A. Kain, "Pronunciation analysis for children with speech sound disorders," in *Proc. of EMBC*, 2015, pp. 5573–5576.

[3] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer speech & language*, vol. 50, pp. 62–84, 2018.

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

[5] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A comparison of gmm-hmm and dnn-hmm based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Proc. of INTERSPEECH*, 2014, pp. 1583–1587.

[6] L. Ward, A. Stefani, D. Smith, A. Duenser, J. Freyne, B. Dodd, and A. Morgan, "Automated screening of speech development issues in children by identifying phonological error patterns," in *Proc. of INTERSPEECH*, 2016, pp. 2661–2665.

[7] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *Proc. of ICASSP*, 2019, pp. 6535–3539.

[8] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *Proc. of INTERSPEECH*, 2016, pp. 765–769.

[9] H.-S. Lee, Y.-D. Lu, C.-C. Hsu, Y. Tsao, H.-M. Wang, and S.-K. Jeng, "Discriminative autoencoders for speaker verification," in *Proc. ICASSP*, 2017, pp. 5375–5379.

[10] R. S. Bauer and P. K. Benedict, *Modern cantonese phonology*. Walter de Gruyter, 2011, vol. 102.

[11] T. Lee, W. K. Lo, P. Ching, and H. Meng, "Spoken language resources for cantonese speech processing," *Speech Communication*, vol. 36, no. 3-4, pp. 327–342, 2002.

[12] L. K. So and B. J. Dodd, "The acquisition of phonology by cantonese-speaking children," *Journal of child language*, vol. 22, no. 3, pp. 473–495, 1995.

[13] C. K. To, P. S. Cheung, and S. McLeod, "A population study of children's acquisition of hong kong cantonese consonants, vowels, and tones," *Journal of Speech, Language, and Hearing Research*, 2013.

[14] S.-I. Ng, C. W.-Y. NG, J. Wang, T. Lee, K. Y.-S. Lee, and M. C.-F. Tong, "Cuchild: A large-scale cantonese corpus of child speech for phonology and articulation assessment," in *Accepted to INTERSPEECH*, 2020.

[15] J. Wang, S. I. Ng, D. Tao, W. Y. Ng, and T. Lee, "A study on acoustic modeling for child speech based on multi-task learning," in *Proc. of ISCSLP*, 2018, pp. 389–393.

[16] S. I. Ng, D. Tao, J. Wang, Y. Jiang, W. Y. Ng, and T. Lee, "An automated assessment tool for child speech disorders," in *Proc. of ISCSLP*, 2018, pp. 493–494.

[17] P. Cheung, A. Ng, and C. To, "Hong kong cantonese articulation test," *Hong Kong: Language Information Sciences & Research Centre*, 2006.

[18] Z. Zhu, Z. Wu, R. Li, H. Meng, and L. Cai, "Siamese recurrent auto-encoder representation for query-by-example spoken term detection." in *Proc. of Interspeech*, 2018, pp. 102–106.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[20] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE transactions on speech and audio processing*, vol. 7, no. 3, pp. 272–281, 1999.

[21] ——, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[23] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end asr-free keyword search from speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1351–1359, 2017.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[26] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. of INTERSPEECH*, 2011, pp. 821–824.

[27] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proc. of INTERSPEECH*, 2015, pp. 3199–3203.