INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Child Speech Disorder Detection with Siamese Recurrent Network using Speech Attribute Features

Jiarui Wang, Ying Qin, Zhiyuan Peng, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
`{jiaruiwang, yingqin}@link.cuhk.edu.hk, jerrypeng1937@gmail.com, tanlee@cuhk.edu.hk`

Abstract

Acoustics-based automatic assessment is a highly desirable approach to detecting speech sound disorder (SSD) in children. The performance of an automatic speech assessment system depends greatly on the availability of a good amount of properly annotated disordered speech, which is a critical problem particularly for child speech. This paper presents a novel design of child speech disorder detection system that requires only normal speech for model training. The system is based on a Siamese recurrent network, which is trained to learn the similarity and discrepancy of pronunciations between a pair of phones in the embedding space. For detection of speech sound disorder, the trained network measures a distance that contrasts the test phone to the desired phone and the distance is used to train a binary classifier. Speech attribute features are incorporated to measure the pronunciation quality and provide diagnostic feedback. Experimental results show that Siamese recurrent network with a combination of speech attribute features and phone posterior features could attain an optimal detection accuracy of 0.941.

Index Terms: child speech, speech disorder detection, Siamese network, speech attribute features

1. Introduction

Speech sound disorder (SSD) is a communication disorder in which children have persistent difficulties in articulating speech sounds correctly. Untreated child speech difficulties can lead to a limited ability to effectively participate in social, educational, or occupational activities. To identify individuals who are in need of relevant intervention, speech assessment is required at an early age [1]. Conventional clinical assessment for SSD, typically carried out by a qualified speech therapist (ST), is to perceptually evaluate the production of phonemes in the child's primary language (L1). However, the manpower shortage of experienced STs causes significant delay of assessment and treatment. Acoustics-based assessment systems for automatic detection of speech sound errors has a great potential in reducing the burdens of professional STs and enabling timely assessment for children in need.

The basic idea of child speech disorder detection is to identify inconsistent speech sound production via pronunciation verification. In recent years, there have been extensive studies on pronunciation verification, most commonly in the problem context of second language acquisition [2]. The approaches are generally divided into two categories. One of them is based on automatic speech recognition (ASR) while the other is to directly predict the production of speech sound as correct or not by using a binary classifier.

State-of-the-art ASR systems have demonstrated a high performance level that can be exploited for use in automatic speech assessment. Confidence scores derived based on ASR posterior probabilities provide a way of quantitatively measuring the deviation of mispronounced phonemes from correct ones. The goodness of pronunciation (GOP) score has been widely used to evaluate the pronunciation quality. It is defined as the posterior probability ratio between a canonical phone and a competing phone [3], The effectiveness of GOP score depends highly on the ASR performance. Also, the numerical scores do not provide useful educational feedback about the mispronunciations detected. The lattice-based approaches were proposed to alleviate this problem [4]. The basic idea is to create an extended search lattice for acoustic model decoding, by including the expected error patterns. [5] compares use of extended search lattices with GMM-HMM and DNN-HMM acoustic models in the assessment of childhood apraxia of speech (CAS). This approach can provide diagnostic information related to error patterns. Nonetheless, its performance would be degraded significantly if many unexpected mispronunciations occur.

Detection of mispronunciation has also been achieved by training a binary classifier to learn the decision boundary between correct and erroneous pronunciations. In [6, 7], deep neural network (DNN) models have been applied to improve mispronunciation detection. Despite the high performance demonstrated, the training of DNN classifiers requires a large amount of disordered speech. In practical applications, collection and annotation of disordered speech is a challenging task and particularly difficult for child subjects as children have poor concentration and may not be able to follow instructions [8].

To address the lack of training data for disordered speech, we develop a novel approach to detecting child speech sound disorder based on only normal speech training data. This is achieved with a Siamese network, which measures the similarity between a pair of phones, one being regarded as the canonical phone and the other as a test phone. With the pairwise comparison architecture, a single system can be trained to cover all possible phone-level mispronunciations, instead of handling each error pattern individually. On the other hand, numerous recent studies have shown that articulatory-phonetic features are robust in compensating for acoustic variations related to speakers and speaking style [9, 10]. In our work, a speech attribute-based ASR system is developed to generate articulation-relation speech attribute features. These features are concatenated with phone posterior features to enhance the performance of proposed system. Comprehensive diagnostic feedback can also be retrieved in the aspects of manner, aspiration and place of articulation.

2. Background

2.1. Problem statement

Cantonese is a major Chinese dialect widely spoken in Guangdong and Guangxi Provinces of Mainland China, Hong Kong,

http://dx.doi.org/10.21437/Interspeech.2019-2320

Macau and many overseas Chinese communities. It is a monosyllabic language. Each Chinese character is associated to one spoken syllable, which is composed of an Initial unit and a Final unit. The Initial is typically a consonant and the Final comprises a vowel nucleus, followed by optional nasal and stop coda. For subword acoustic modeling in Cantonese ASR, we use a phone set that consists of 19 consonants and 13 vowels.

Clinical assessment of SSD is based on subjective evaluation of the production of specific speech sounds. A child is considered to have SSD, if he/she is not able to pronounce certain speech sounds correctly beyond the expected age. Based on previous research of child speech development, a child should be able to master all vowel sounds by the age of 3, and the consonants should be correctly pronounced by the age of 6 [11]. For children aged from $2-6$, speech sound errors occur more frequently in the initial consonants than vowels. In this work, we focus on the sound disorder related to the initial consonants. The most common error patterns, as suggested by clinical observation, include fronting, backing and deaffrication, which are caused by incorrect place and/or manner of articulation. This clinical knowledge is incorporated into the proposed design of SSD detection system. As shown in Table 1, each Cantonese initial consonant can be characterized by a set of speech attributes that describe the manner, place of articulation and the aspiration status [12].

2.2. Speech corpus

This work is carried out on CUChild127, which is a newly collected large-scale database of child speech. The database is developed to support research on automatic assessment of SSD in Cantonese-speaking pre-school children [8]. It contains the speech collected from $1,500$ children (aged $3-6$) in Hong Kong. Among them, 310 children were found to have SSD based on subjective assessment with the Hong Kong Cantonese Articulation Test (HKCAT) [13]. During data collection, each child subject was asked to read a list of 127 Cantonese words (1 to 4 syllables in length), which cover all of the 19 consonants and 13 vowels in Cantonese.

3. ASR-based posterior features

3.1. Phone posterior features

Phone posterior features are generated using an ASR system with a phone-based acoustic model. The modeling units cover the 19 consonants and 13 vowels. The audio recordings of 150 speakers selected from CUChild127 are used as training data. For each speaker, there are 127 word utterances, accompanied by their syllable-level transcriptions. A pronunciation lexicon that contains 689 syllables is used to convert the syllable transcription into a phone sequence. For example, the Cantonese syllable /tau/ is decomposed into three phones, “t”, “a”, “u:”.

The acoustic features used for acoustic model training are 40-dimensional MFCCs. The acoustic model is built based on a time-delay neural network (TDNN), which contains 6 hidden layers with 1024 neurons per layer. The context configuration for each layer of TDNN is $\{[-2,2],\{0\},\{-1,2\},\{-3,3\},\{-3,3\},\{-7,2\}\}$. Rectified linear unit (ReLU) followed by re-normalization is applied to each hidden layer. The softmax activation function is applied in the output layer. 1408 neurons of the output layer correspond to the tied tri-phone states (senones). Time alignment of the state-level senones are obtained from a context-dependent GMM-HMM (CD-GMM-HMM) acoustic model. For the TDNN acoustic model, a syllable error rate of 16.43% was attained in the syllable recognition task assuming all Cantonese syllables have equal probability.

Table 1: *Speech attributes and their corresponding Cantonese initial consonants labeled with the Jyut-Ping scheme*

Category	Attribute	Phone Set
Manner	Plosive	[b] [p] [g] [k] [d] [t] [gw] [kw]
	Nasal	[m] [n] [ng]
	Affricate	[z] [c]
	Fricative	[s] [f] [h]
	Glide	[j] [w]
	Liquid	[l]
Aspiration	Aspirated	[p] [t] [k] [kw] [c]
	Unaspirated	[b] [d] [g] [gw] [z]
	N/A	[s] [f] [h] [j] [w] [l] [m] [n] [ng]
Place	Alveolar	[d] [t] [z][c] [s] [j]
	Lateral	[l]
	Labial	[b] [p] [w] [m]
	Velar	[g] [k] [ng]
	Velar-labial	[gw] [kw]
	Dental-labial	[f]
	Vocal	[h]

With this ASR system, we extract frame-level phone posterior features. Each of modeled phones is associated with a set of senones in the softmax layer. Phone posterior probability can be calculated by summing up the senone posteriors associated to the same phone from the softmax layer. As a result, a 33-dimensional (including the silence model) frame-level phone posterior feature vector is obtained from the 1408 senones.

3.2. Speech attribute features

To obtain speech attribute features, we train three different acoustic models that model the variation of articulation manner, articulation place and aspiration respectively. The three acoustic models have the same architecture as the phone-based acoustic model. The training data are also the same. The only difference is that the phone sequence representation is changed to speech attribute sequence. For consonants, the phone-attribute mapping is based on Table 1. All vowels are grouped together, being represented by a single modeling unit. For example, the syllable /tau/ is represented by the model sequence of “plosive”-“vowel” in terms of articulation manner.

Similar to the phone-based model, frame-level speech attribute posterior feature vectors are obtained by summing up the neuron outputs corresponding to the same speech attribute. The dimensions of speech attribute posterior feature vectors are 14, 16 and 8 for articulation manner, articulation place and aspiration respectively.

4. SSD detection with Siamese network

In [14], the Siamese network was successfully applied to assess semantic similarity between sentences. Motivated by this work, we propose a child SSD detection system, which measures the similarity between a pair of phones using a Siamese network. Specifically, a sound disorder is detected if the claimed phone is classified as being distinct from the canonical phone (pronounced correctly) according to the measured similarity. Figure 1 illustrates the proposed system design. The system takes two phone segments as input and computes frame-level ASR posterior features for each of them as described above. A binary classifier based on Siamese network is used to compare the two resulted feature sequences and decide on whether they are the same phone or not.

4.1. Feature representation

Each utterance in our speech corpus contains a spoken word (one of the 127 test words). As mentioned earlier, child SSD typically concerns a specific initial consonant in each test word. Phone-level forced alignment is performed using a GMM-HMM acoustic model to locate and extract the designated phone segment. Subsequently frame-level posterior features are computed from the segment, as described in Section 3. As a result, a variable-length sequence of feature vectors is obtained. Each of the feature vectors is formed by appending four types of posterior features, namely phone posteriors, articulation manner posteriors, articulation place posteriors and aspiration posteriors.

4.2. Siamese network classifier

As shown in Figure 1, the Siamese network classifier takes in two input feature sequences, which may have different lengths. The classifier gives a positive output if the two input segments are from the same phone category and negative output if they are different phones. A detailed architecture of the classifier is shown as Figure 2. The network contains two identical neural networks with shared parameters [15].

GRU Encoder: Recurrent neural network (RNN) is commonly adopted to deal with sequence classification problem. Similar to long short-term memory (LSTM), the gated recurrent unit (GRU) is an extension of conventional RNN to alleviate the gradient vanishing problem. As a modification of LSTM, GRU can speed up training with simpler structure [16]. In our work, GRU is used to embed a sequence of input data into a new feature representation of fixed size. With the feature transformation, it is expected that a positive pair of phones (same phone category) are pulled closer and a negative pair of phones are pushed farther away from each other. The GRU has 2 hidden layers with 200 memory cells per layer. Bi-directional GRU (BiGRU) is attempted for further improvement of the performance.

Similarity Computation Layer: The similarity is measured by the element-wise absolute difference between the embedding features, i.e., $|v_1 - v_2|$, which is a 200 dimensional feature vector. It is used as the input for subsequent binary classification.

Fully-Connected Layer: At the top of the architecture there are two fully connected layers. ReLU is used as the activation function in these layers. The Sigmoid function is used in the output layer, which generates a binary decision. The Sigmoid output value approaching 1 represents the segments in a pair are more likely to be in the same phone category. Otherwise the Sigmoid output value is approaching 0.

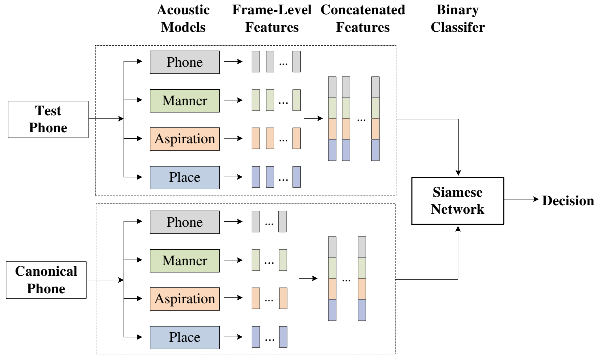

Figure 1: *The framework of child speech disorder detection system*

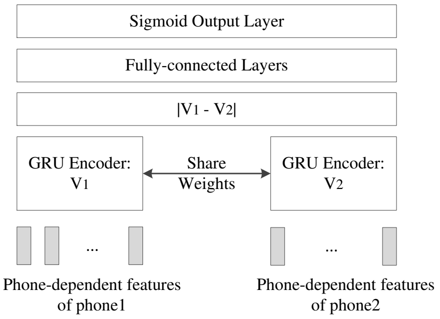

Figure 2: *The structure of the binary classifier based on Siamese network*

4.3. Classifier training

Training the Siamese network classifier in Figure 2 requires pairs of phone segments. In our work, the training segments are obtained by applying forced alignment on the spoken word utterances from 150 normal child speakers in the CUChild127 database (see Section 2). Each segment of initial consonant is paired with one segment of the same phone and four segments of other phones. This arrangement leads to a total of $67,800$ positive training pairs and $271,200$ negative training pairs. For better generalization in learning, vowel segments are also included in the negative pairs. In supervised training of the Siamese network, the positive training pairs are labeled with target output 1 while the negative pairs are labeled with 0. The GRU is first initialized by orthogonal initialization. The binary cross entropy is used as loss function which is optimized by the stochastic gradient descent (SGD) algorithm. We fix the mini-batch size of 128 with learning rate of 10^{-2}.

5. Experiments

5.1. Preparation for test data

The proposed system is tested on the task of detecting typical initial consonant production errors. Similar to the processing of training data, the test consonant segments are obtained by forced alignment with the same acoustic model. The test data comprises word utterances of normal speech from 50 child speakers and word utterances of disordered speech from another 35 speakers in CUChild127. However, the number of correctly produced initial consonants is much higher than that of mispronounced initial consonants. In order to augment disordered speech, artificial errors are created from normal speech by replacing the claimed initial consonant with another one. The replacement rule is a random process. As a result, there are totally $2,630$ correctly produced initial consonants, $2,630$ artificial consonant errors and 640 real consonant errors.

5.2. Experimental setup

Experiments are carried out separately in detecting artificial errors and real errors as described in Section 5.1. Each test seg-

Table 2: *Classification performance on detecting artificial consonant errors of two Siamese models, using filter-bank features, phone posterior features and Phone + Attribute posterior features*

Model	**Filter-bank**			**Phone posterior**			**Phone+Attribute**		
	Accuracy	AUC	Precision/Recall	Accuracy	AUC	Precision/Recall	Accuracy	AUC	Precision/Recall
Siamese-GRU	64.4%	0.697	0.613/0.768	91.3%	0.961	0.941/0.856	93.3%	0.962	0.937/0.900
Siamese-BiGRU	80.7%	0.918	0.889/0.702	92.7%	0.963	0.940/0.883	**94.1%**	**0.965**	**0.942/0.910**

ment is labeled with a claimed phone category. The Siamese network determines whether the real production of the test segment matches the claimed phone category. This is implemented by pairwise comparison of the test segment to all canonically pronounced training segments corresponding to this claimed phone category. A set of Sigmoid outputs are thus obtained. The binary decision on whether the test segment of initial consonant is disordered depends on the average of these Sigmoid outputs. If the average value is lower than the pre-defined threshold, the test segment is classified as a disorder, otherwise it is classified as a correct pronunciation. In this work, threshold is tuned to obtain the highest accuracy. The Siamese network with uni-directional GRU (Siamese-GRU) and bi-directional GRU (Siamese-BiGRU) are compared to see the capacity of feature embedding. Moreover, 40-dimension filter-bank features, 33-dimension phone posterior features and 71-dimension (33 + 14 + 16 + 8) composite posterior features (Phone + Attribute features) are evaluated with the proposed system.

Four quantitative metrics, namely accuracy, precision, recall and Area Under receiver operating characteristics Curve (AUC) [17] are used for performance evaluation on the proposed system. A good system should optimize all these metrics simultaneously.

5.3. Experimental results and discussion

Table 2 shows the performance evaluation results of detecting the artificial initial consonant errors. The Siamese-BiGRU based system and the Siamese-GRU based system are compared. Three different types of input features are adopted, which are filter-bank features, phone posterior features and Phone+Attribute features. Test data is composed of $2,630$ correctly produced initial consonants and $2,630$ artificial consonant errors, which cover all the initial consonant categories except the consonant [n]. For each consonant category, the correctly produced samples and the mispronounced samples are equal. As shown in Table 2, the Siamese network with BiGRU encoder outperforms that with GRU encoder when using the same type of input features. It is expected that the BiGRU encoder has better ability to represent the production of phones than the GRU encoder. On the other hand, using Phone+Attribute features makes further improvements to the detection results compared to filter-bank features and phone posterior features. This demonstrates that using phone posterior features appended with articulatory-related information is more robust to detect the consonant errors. Finally, the Siamese-BiGRU system using Phone+Attribute features achieves the best results, with the accuracy and the AUC score being 94.1% and 0.965 respectively. This implies that our proposed system based on the Siamese network has potential in distinguishing the disordered speech from the normal speech.

The experimental results of detecting the real consonant errors from child disordered speech are shown in Table 3. The Siamese-BiGRU based system using Phone+Attribute features is applied in this experiment. Five common types of error patterns in SSD are summarized from the real consonant errors, namely fronting, backing, deaspiration, deaffrication and stopping. For each type of error, the number of error samples and the number of correctly produced samples are identical. From Table 3, the following observations can be made:

Table 3: *Detection results of real disordered speech: The errors are described as the form of [claimed phone] - [real production]*

Error Patterns	**Speech Attibute**	**Errors**	**Number**	**AUC**
Fronting	Place	[k] - [t]	18	0.972
		[t] - [k]	63	0.909
Backing		[g] - [d]	50	0.849
		[d] - [g]	59	0.837
Deaspiration	Aspiration	[t] - [d]	143	0.971
		[k] - [g]	112	0.936
		[c] - [z]	83	0.914
Deaffrication	Manner + Aspiration	[c] - [s]	38	0.889
Deaffrication+Stopping		[c] - [d]	74	0.977

- The highest AUC score is achived by detecting the error [c] - [d] that is related to two error patterns:deaffrication and stopping. It is therefore reasonable to assume that our proposed system performs better in detecting the errors involved more error patterns.
- The detection accuracy of the consonant pairs of [d] - [g] and [g] - [d] are worse than that of [t] - [k] and [k] - [t]. This suggests that the system should be improved to detect the mispronunciation between two voiced consonants.

6. Conclusions

We propose a child speech disorder detection system based on Siamese recurrent network to detect initial consonant errors. With help of the pair-based structure of Siamese network, this work mitigates the lack of properly annotated disordered speech in three aspects: 1) The system is trained only with normal speech; 2) Pairwise training naturally augment the amount of training data; 3) The disorder detection of all the phone categories can be implemented in one system, without training multiple systems for each phone category individually. In addition, speech attribute feature is adopted to provide further improvements and comprehensive diagnostic feedback. The results are promising since careful selection of positive pairs and negative pairs could further improve the performance of the system.

7. Acknowledgements

This research is partially supported by the Major Program of National Social Science Fund of China (Ref: 13&ZD189), the Hear Talk Foundation, a GRF project grant (Ref: CUHK 14227216) from the Hong Kong Research Grants Council, and a direct grant from CUHK Research Committee.

8. References

[1] H. M. Sharp and K. Hillenbrand, “Speech and language development and disorders in children,” *Pediatric Clinics of North America*, vol. 55, no. 5, pp. 1159–1173, 2008.

[2] 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.

[3] 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.

[4] S. M. Abdou, S. E. Hamid, M. Rashwan, A. Samir, O. Abdel-Hamid, M. Shahin, and W. Nazih, “Computer aided pronunciation learning system using speech recognition techniques,” in *Proc. ICSLP*, 2006.

[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. INTERSPEECH*, 2014.

[6] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling,” in *Proc. ICASSP*. IEEE, 2016, pp. 6135–6139.

[7] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, “Improving mispronunciation detection for non-native learners with multisource information and LSTM-based deep models.” in *INTERSPEECH*, 2017, pp. 2759–2763.

[8] 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. ISCSLP*. IEEE, 2018, pp. 1–4.

[9] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. R. Rabiner, “An overview on automatic speech attribute transcription (ASAT),” in *Proc. INTERSPEECH*, 2007.

[10] C.-H. Lee and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[11] J. Law, J. Boyle, F. Harris, A. Harkness, and C. Nye, “The feasibility of universal screening for primary speech and language delay: Findings from a systematic review of the literature,” *Developmental medicine and child neurology*, vol. 42, no. 3, pp. 190–200, 2000.

[12] 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.

[13] P. Cheung, A. Ng, and C. To, “Hong Kong Cantonese articulation test,” *Hong Kong SAR, Peoples Republic of China: Language Information Sciences Research Centre, City University of Hong Kong.*, 2006.

[14] J. Mueller and A. Thyagarajan, “Siamese recurrent architectures for learning sentence similarity,” in *Proceedings of Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[15] S. Chopra, R. Hadsell, Y. LeCun *et al.*, “Learning a similarity metric discriminatively, with application to face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitin (CVPR (1))*, 2005, pp. 539–546.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.

[17] J. Huang and C. X. Ling, “Using AUC and accuracy in evaluating learning algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.