CLASSIFYING SPEECH INTELLIGIBILITY LEVELS OF CHILDREN IN TWO CONTINUOUS SPEECH STYLES

Yeh-Sheng Lin, *Shu-Chuan Tseng*

Institute of Linguistics, Academia Sinica
yehsheng@gate.sinica.edu.tw, tsengsc@gate.sinica.edu.tw

ABSTRACT

Speech difficulties of children may result from pathological problems. Oral language is normally assessed by expert-directed impressionistic judgments on varying speech types. This paper attempts to construct automatic systems that help detect children with severe speech problems at an early stage. Two continuous speech types, repetitive and storytelling speech, produced by Chinese-speaking hearing and hearing-impaired children are applied to Long Short-Term Memory (LSTM) and Universal Transformer (UT) models. Three approaches to extracting acoustic features are adopted: MFCCs, Mel Spectrogram, and acoustic-phonetic features. Results of leave-one-out cross-validation and models trained by augmented data show that MFCCs are more useful than Mel Spectrogram and acoustic-phonetic features. Respective LSTM and UT models have their own advantages in different settings. Eventually, our model trained on repetitive speech is able to achieve an F1-score of 0.74 for testing on storytelling speech.

Index Terms— Child speech, intelligibility assessment, universal transformer, speech styles

1. INTRODUCTION

1.1. Child speech assessment

Speech difficulty of children can be caused by phonological development problems, communication disorders or diseases [1]. Specific language and speech impairment leads to a variety of linguistic patterns in their speech. In particular, hearing-impaired children may show different difficulties in producing continuous speech. It is important to detect speech problems of hearing and hearing-impaired as early as possible, so that targeted speech therapy and intervention programs can be accordingly initiated [2-4]. A sophisticated assessment of oral language production is subject to multiple linguistic considerations, e.g., segment clarity, lexical use, sentence structure and socio-pragmatic, communicative skills. Among them, speech intelligibility is often evaluated for clarity and fluency of vowels, consonants, tones, rhythm, and intonation [5-7]. Our study focuses on speech intelligibility that involves only segment clarity and lexical prosody. Ideally, continuous speech is the most authentic speech form that reflects the ability of oral language production. In clinical and educational domains, these kinds of assessments are conducted by impressionistic judgements of professional experts [8]. This paper adopts deep learning methods that classify three levels of speech intelligibility, aiming to detect children with severe speech problems. Storytelling data are tested by models solely trained on repetitive data. If we are able to obtain encouraging output performance, it is likely that repetitive speech can serve as an alternative to spontaneous speech, at least for the purpose of early screening.

1.2. Automatic assessment approaches

Approaches based on neural networks, e.g., Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have been recently applied to automatic speech assessment in different fields. CNN has been applied to classification tasks that distinguish speech production of Parkinson patients and healthy people [9] and assess speech produced by aphasia [10, 11] and dysarthria speakers [12]. Different from CNN, RNN is capable of modeling temporal sequences and can deal with variable-length utterances. Long-Short Term Memory (LSTM) [13] and Gated Recurrent Unit (GRU) [14] address vanishing gradient problems around special memory cell units and have the capability of capturing long-term temporal dependencies in acoustic events [15]. GRU is a simplified architecture with an efficiency degree that is comparable to LSTM. These two approaches have been adopted for building automatic speech assessment systems [10, 16-19], e.g., the work done by Korzekwa *et al*. on dysarthric speech [16].

However, the property of sequential computation of RNN-based models in part also hinders parallelization across elements of input sequences. Modern Transformer models [20] allow for more space for parallelization and require less training time on GPU. The architecture of a Transformer model takes into account the global dependencies between input and output with attention mechanisms. Compared to the other deep learning models, Transformer models have achieved surprisingly satisfactory results on complex NLP tasks. But they fail to generalize in some of the simple tasks with which RNN-based models

can successfully handle. Universal Transformer (UT) [21] has thus been proposed to enhance the adaptability of Transformer models.

We apply sequence modeling approaches LSTM and UT to classify speech intelligibility levels of Chinese-speaking hearing and hearing-impaired children in two continuous speech styles. The performance of conventional LSTM models are compared with that of UT models to experiment on the feasibility of establishing an automatic assessment system that is able to detect children with speech difficulties using only repetitive speech as training data. In addition, CNN is used to enhance feature extraction on 2D MFCCs and Mel Spectrogram.

2. SINICA CHILD SPEECH CORPUS

The *Sinica Child Speech Corpus* contains repetitive and storytelling speech data produced by 79 preschool children with normal hearing (NH) and 45 children with hearing impairment [22]. Among them, 30 wore traditional hearing aids (HA, hearing loss degree: mild to profound), and 15 were fitted with a cochlear implant (CI, hearing loss degree: severe to profound). All hearing-impaired children were receiving Auditory-Verbal Therapy (AVT) at the time of recording. Signal-aligned syllable boundary information was obtained by the *ILAS phone aligner* [22] and post-edited by trained phoneticians. Text processing was conducted using the CKIP automatic word segmentation and POS tagging system [23]. Lexical information was integrated with signal-aligned labels to produce multi-layer linguistic annotations. Authorized academic use of the *Sinica Child Speech Corpus* can be granted by the Department of Intellectual Property and Technology Transfer of Academia Sinica.

For repetitive speech collection, 18 sentences containing 99 distinct syllables encompassing all phonemes in Chinese were recorded by a female adult speaker and played one by one for the children to repeat. A word recognition test was conducted by qualified audiologists to HA and CI children to ensure that they were able to hear and understand all content words in the sentences. For storytelling speech collection, picture cards illustrating the content of *The Hare and the Tortoise* were shown in a fixed order to elicit narratives. The content transcription includes paralinguistic phenomena and discourse fillers, particles, and markers (DIS) [22]. Table 1 summarizes the dataset and the articulation rate (AR) of each group in seconds per syllable.

Table 1. Data summary.

Hearing group 79 children (2;11~6;3)		Hearing-impaired group 45 children (3;3~12;5)	
Repetitive	Storytelling	Repetitive	Storytelling
7,511 syll.	9,102 syll. 3,695 DIS	4,064 syll.	7,467 syll. 2,778 DIS
AR: 0.308	0.296	0.299	0.305

Speech intelligibility of the children was rated based on degree of fluency and clarity of segment production at a scale from 1 to 5. Fig. 1 illustrates the sum of the scores given by three phoneticians. As we are mainly concerned with intelligibility levels, the 124 children were divided into three intelligibility groups: LOW (<8), Medium (8~11), and HIGH (>11), each containing 14, 36, and 74 children. The storytelling speech dataset only consists of 119 children's data; thus the numbers of children in LOW, MEDIUM, and HIGH groups are 12, 34, and 73. The main goal of this study is to construct automatic models that can help identify the LOW groups of children, in particular. We used repetitive speech that is on the one hand continuous speech. But on the other hand, it is relatively easy to collect and process for model training. If the models work well on storytelling speech, we may have the option to use repetitive speech instead of spontaneous speech for early screening tasks.

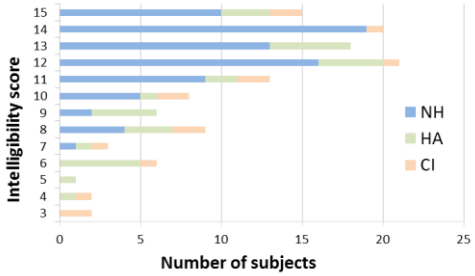

Fig. 1. Subject information (intelligibility ratings).

3. METHODOLOGY

3.1. Speech representation

Mel-frequency cepstral coefficients (MFCCs) are commonly used in speech assessment systems for acoustic modeling [10, 17-19, 24] and feature extraction [25, 26]. While deep learning models recently attract intense attentions, Mel Spectrogram is also getting increasingly popular [10, 12, 16]. For our experiments, we implemented MFCCs and Mel Spectrogram feature extraction in Python using the librosa library [27]. For each utterance, 40-dimensional MFCCs and 80-dimensional Mel Spectrogram are derived from short-time frames with a window length of 32 milliseconds.

3.2. Syllable-based acoustic-phonetic features

MFCCs and Mel Spectrogram represent spectrum information. But they do not directly account for linguistic properties. Mandarin Chinese is a syllable-timed, tone language [28]. Syllable-related acoustic information and lexical tones are essential in auditory perception and lexical processing. Thus, we put together acoustic features whose linguistic relevance in continuous speech production has already been confirmed. Eventually, four groups of vocalic and lexical-prosodic features are considered [6, 29]: 1)

vocalic properties: the first three formants, intensity, and F0 value at maximum sonority, 2) *syllable prosody*: duration, the mean, minimum, and maximum of intensity, 3) *pitch-related features*: (a) the initial, final, maximum and minimum F0 values of the voiced region within a syllable, (b) the first three formants and intensity at time points of 3(a), and 4) *tone-related features*: the slopes of the three lines in 3(a), if any. We use the above acoustic-phonetic (AP) features to test whether it is possible for them to serve as alternatives or supplements to MFCCs and Mel Spectrogram features for our LSTM and UT models. It is noteworthy that our AP features do not include those related to onset consonants that are substantially involved in the processes of spoken word recognition and speech intelligibility perception.

3.3. LSTM and Universal Transformer

We use time-domain sequence model LSTM as our baseline model and test the self-attention mechanism UT to explore how well these two models are able to capture the temporal dependency of each syllable. In CNN architecture, we employ only one convolution layer with 32 filters of one stride of 2×2 kernel on MFCCs and 4×4 kernel on Mel Spectrogram. The rectifier linear (RELU) activation function is used in the convolutional layers with a batch normalization layer added after the convolutional layer. The last layer of CNN is a 16-units dense layer with softmax activation function. As in total 135 syllable tokens occur in the 18 sentences, our input sequence is set to be 135-syllable in length. Output sequences of CNN are connected to LSTM and UT to classify the input sequences. Different from 2D MFCCs and Mel Spectrogram, we directly input 1D AP features to LSTM and UT models. LSTM has 256 units with the tanh activation function. UT has 3 depth transformer blocks. We use dropout with probability 0.1 to avoid overfitting and batch normalization and to accelerate the training process. Finally, we combine the speech features (MFCCs and Mel Spectrogram) and AP features by a fully connected layer. All models above are trained by using Adam optimization algorithm with a learning rate of 0.0001 and a batch size of 16. The implementation was conducted by using Keras [30] and Keras-Transformer [31] with default settings except for the above configurations.

3.4. Implementation procedures

MFCCs, Mel Spectrogram and AP are extracted from syllable units by employing the corpus information about syllable boundaries. Each syllable-based feature set is regarded as one data point of the input sequence. We conducted our experiment in two steps. First, we performed the leave-one-out cross-validation on repetitive speech to evaluate LSTM and UT models with the three sets of features. Then we evaluated how the models performed on storytelling speech. For each learning set, the model was trained by 100 epochs. As our training data is small, we augmented our repetitive speech data by first generating 1000 speaker-lists of 18 children (because we have 18 sentences) randomly selected from the respective LOW, MEDIUM, and HIGH groups. Sentences produced by each of the children were concatenated in the order of the speaker-list as input sequences in Fig. 2. Although the speaker-lists contain repeated children, none of the input sequences is identical to the original sentences in the *Sinica Child Speech Corpus*. In the second step, the augmented data were used for training, with a total of 500 epochs for each model. The original sentences of repetitive speech and the complete set of storytelling speech serve as testing data. LSTM and UT models were trained on AP features, while CNN-LSTM and CNN-UT models were constructed by using MFCCs and Mel Spectrogram as well as mixed sets of features.

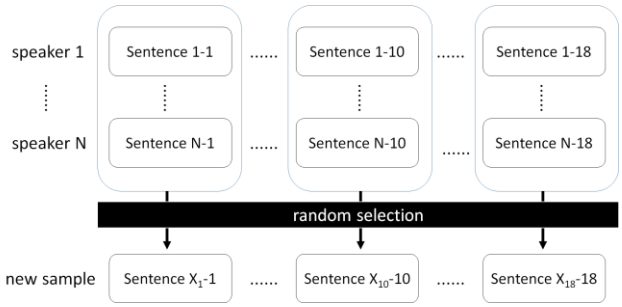

Fig. 2. New sample.

4. DATA EXPERIMENTAL RESULTS

4.1. Model performance in two speech styles

Our main aim is to identify speech intelligibility level of the LOW groups of children. The leave-one-out cross-validation results on repetitive speech in Table 2 show that LSTM models were more successful than UT models. But UT models outperformed LSTM models with augmented training data. The size of training data and the demand for complex calculation may be attributed to this difference. To our surprise, AP features work relatively well, even though the features merely take into account vocalic and lexical prosodic properties. Both of our CNN-LSTM and CNN-UT models performed better with MFCCs than with Mel Spectrogram and AP features. In the case of storytelling speech, F1 scores of leave-one-out cross-validation are consistently low. The best result we were able to achieve by CNN-LSTM model using MFCCs is 0.48 for the LOW group.

Using augmented training data, the models worked better for both repetitive data and storytelling data. CNN-LSTM and CNN-UT models with MFCCs still outperformed those with Mel Spectrogram and AP features. Model performance dropped dramatically with AP features on storytelling

Table 2. Results of model performance.

Model	Feature	Leave-one-out cross-validation						Training by augmented data					
		F1 (repetitive)			F1 (storytelling)			F1 (repetitive)			F1 (storytelling)		
		L	M	H	L	M	H	L	M	H	L	M	H
LSTM	AP	0.77	0.97	0.97	0.24	0.43	0.67	0.87	0.87	0.93	0.25	0.51	0.54
CNN-LSTM	MFCC	**0.92**	**0.99**	**0.98**	**0.48**	**0.39**	**0.77**	0.76	0.93	0.94	0.70	0.60	0.80
	MelS	0.92	0.99	0.99	0.24	0.49	0.72	0.74	0.90	0.91	0.58	0.56	0.76
UT	AP	0.4	0.58	0.85	0.2	0.25	0.67	0.81	0.81	0.86	0.32	0.39	0.76
CNN-UT	MFCC	0.52	0.69	0.9	0.36	0.38	0.73	**0.93**	**0.94**	**0.95**	**0.71**	**0.62**	**0.81**
	MelS	0.38	0.67	0.91	0.04	0.37	0.72	0.64	0.88	0.91	0.35	0.46	0.79

Table 3. Results of models using mixed features.

Model	Feature	F1 (repetitive)			F1 (storytelling)		
		L	M	H	L	M	H
CNN-LSTM	MFCC+AP	0.92	0.88	0.94	0.63	0.61	0.84
	MelS+AP	0.68	0.96	0.88	0.33	0.47	0.69
CNN-UT	MFCC+AP	0.90	0.94	0.94	0.69	0.61	0.81
	MelS+AP	0.81	0.84	0.89	0.45	0.48	0.77

Table 4. Best results in subject groups.

Type	Group	L	M	H
Repetitive	NH	1.00	0.98	0.99
	HA	1.00	0.87	0.86
	CI	0.80	0.92	0.57
Storytelling	NH	-	0.65	0.86
	HA	0.67	0.44	0.50
	CI	0.80	0.80	0.75

speech, suggesting that our AP features do not represent sufficient information that is relevant to intelligibility assessment. In contrast, there are only marginal differences between repetitive and storytelling speech with MFCCs and Mel Spectrogram. Concerning the LOW group, CNN-UT model with MFCCs is able to achieve an F1 of 0.93 on repetitive speech and an F1 of 0.71 on storytelling speech. In spite of the fact that the model may have seen part of the repetitive speech in the training phase, the overall model performance on storytelling speech is quite positive.

We also tested whether CNN-LSTM and CNN-UT models can achieve better results with mixed feature sets. Results are summarized in Table 3. Only small increases in F1 scores in some subject groups were achieved. Mixed features did not improve the performance significantly. In addition to model enhancement, we will also experiment on integrating information about onset consonants and feature attributes that have been proved useful for assessing pathological speech into our AP features [19].

4.2. Model performance in subject groups

We analyzed the classification output produced by the best CNN-UT model with MFCCs. Results are presented in Table 4 in terms of NH, HA, and CI subgroups. For repetitive speech, our models are in principle successful with all F1 scores higher than 0.8 except for the CI HIGH group. Some of these children produced very clear segments, but their sentences are less fluent in prosody [6]. It seems that none of the three feature extraction methods is able to account for this complexity.

For storytelling speech, F1 scores of the CI subgroups are quite satisfactory (LOW 0.8, MEDIUM 0.8 and HIGH 0.75). But our best models did not perform well on data produced by HA children. Moreover, the discrepancy in model performance between repetitive and storytelling data of HA children is also the greatest. As our HA and CI children were receiving the AVT training program, their spoken language ability can be individually very different and does not necessarily correlate with hearing loss degree or sensory aid types. Nevertheless, our results suggest a wider range of speech variability in HA children.

5. CONCLUSIONS

Our models preliminarily achieved positive results in assessing intelligibility levels tested on repetitive and storytelling speech produced by different groups of children. More training data are required to account for speech variability in continuous speech. We are currently preparing a large-scale speech dataset of developing children. Models and acoustic feature extraction methods used in this study will be applied to the forthcoming data to construct sophisticated models that are able to identify speech intelligibility levels and hopefully also able to specify speech difficulties of children with hearing or speech impairment.

6. ACKNOWLEDGEMENTS

This study was financially supported by the Children's Hearing Foundation, the National Science Council of Taiwan (under grant NSC 99-2410-H-001-097), and the Ministry of Science and Technology of Taiwan (under grant MOST 106-2410-H-001-045-MY2).

7. REFERENCES

[1] H. Zhu and B. Dodd, "The phonological acquisition of Putonghua (Modern Standard Chinese)," *Journal of child language*, 2000, vol. 27, pp. 3-42.

[2] S.-C. Tseng. "Speech Production of Mandarin-speaking Children with Hearing Impairment and Normal Hearing," in *Proc. ICPhS*, 2011, pp. 2030-2033.

[3] M.-E. Bouchard, M.-T. Normand, and H. Cohen, "Production of consonants by prelingually deaf children with cochlear implants," *Clinical linguistics & phonetics*, 2007, vol. 21, pp. 875-84.

[4] D. Dornan, L. Hickson, B. Murdoch, and T. Housston, "Outcomes of an Auditory-Verbal Program for Children With Hearing Loss: A Comparative Study With a Matched Group of Children With Normal Hearing," *Volta Rev*, 2007, vol. 107.

[5] S.-C. Peng, A. Weiss, H. Cheung, and Y.-S. Lin, "Consonant Production and Language Skills in Mandarin-Speaking Children With Cochlear Implants," *Archives of otolaryngology--head & neck surgery*, 2004, vol. 130, pp. 592-7.

[6] S.-C. Tseng, K. Kuei, and P.-C. Tsou, "Acoustic characteristics of vowels and plosives/affricates of Mandarin-speaking hearing-impaired children," *Clinical linguistics & phonetics*, 2011, vol. 25, pp. 784-803.

[7] W.-Y.C. Tiago H. Falk, Fraser Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, 2012, vol. 54(5), pp. 622 - 631.

[8] S. McLeod and S. Verdon, "A Review of 30 Speech Assessments in 19 Languages Other Than English," *American journal of speech-language pathology / American Speech-Language-Hearing Association*, 2014, vol. 23.

[9] J. Vasquez, J.R. Orozco, and E. Noeth. "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease," in *Proc. Interspeech*, 2017, pp. 314-318.

[10] Y. Qin, Y. Wu, T. Lee, and A.P.H. Kong, "An End-to-End Approach to Automatic Speech Assessment for Cantonese-speaking People with Aphasia," *Journal of Signal Processing Systems*, 2020.

[11] Y. Qin, T. Lee, and A.P.H. Kong. "Automatic Assessment of Language Impairment Based on Raw ASR Output," in *Proc. Interspeech*, 2019, pp. 3078-3082.

[12] J.C.-V. Correa, T. Arias, J.R. Orozco-Arroyave, and E. Nöth. "A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease," in *Proc. Interspeech*, 2018, pp. 456-460.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997, vol. 9(8), pp. 1735-1780.

[14] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proc. SSST-8*, 2014, pp. 103-111.

[15] H. Sak, A. Senior, and F. Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338-342.

[16] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak. "Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech," in *Proc. Interspeech*, 2019, pp. 3890-3894.

[17] Y. Qin, T. Lee, and A.P.H. Kong, "Automatic Assessment of Speech Impairment in Cantonese-Speaking People with Aphasia," *IEEE Journal of Selected Topics in Signal Processing*, 2020, vol. 14(2), pp. 331-345.

[18] Y. Qin, T. Lee, S. Feng, and A.P.-H. Kong. "Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning," in *Proc. Interspeech*, 2018.

[19] J. Wang, Y. Qin, Z. Peng, and T. Lee. "Child Speech Disorder Detection with Siamese Recurrent Network using Speech Attribute Features," in *Proc. Interspeech*, 2019, pp. 3885-3889.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin "Attention Is All You Need," *arXiv:1706.03762*, 2017.

[21] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser "Universal Transformers," *arXiv e-prints arXiv:1807.03819*, 2018.

[22] S.-C. Tseng. "ILAS Chinese Spoken Language Resources," in *Proc. LPSS*, 2019, pp. 13-20.

[23] K.-j. Chen, C.-R. Huang, L. Chang, and H.-L. Hsu, "SINICA CORPUS: Design methodology for balanced corpora". Vol. 167. 1996.

[24] Y. Liu, Y. Qin, S. Feng, T. Lee, and P.C. Ching. "Disordered Speech Assessment Using Kullback-Leibler Divergence Features with Multi-Task Acoustic Modeling," in *Proc. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2018, pp. 61-65.

[25] S. Gillespie, y.-y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel. "Cross-Database Models for the Classification of Dysarthria Presence," in *Proc. Interspeech*, 2017, pp. 3127-3131.

[26] M.S. Paja and T. Falk. "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *Proc. Interspeech*, 2012, pp. 62-65.

[27] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. "librosa: Audio and music signal analysis in python," in *Proc. the 14th python in science conference*, 2015, pp. 18-25.

[28] S. Duanmu, "The phonology of standard Chinese". 2007: OUP Oxford.

[29] P. Ladefoged and K. Johnson, "A Course in Phonetics (5th)", in *Thomson Wadsworth*. 2006. p. Chap. 6-9, pp. 133-236.

[30] F. Chollet. *keras*. GitHub repository 2015; Available from: https://github.com/fchollet/keras.

[31] K. Mavreshko. *keras-Transformer*. GitHub repository 2018; Available from: https://github.com/kpot/keras-transformer.