P-44

Usefulness of Automatic Speech Recognition for Evaluating Children with Speech Sound Disorders

Dae-Hyun Jang, M.D, PhD, Jae won Kim M.D.*

Department of Rehabilitation Medicine, Incheon Saint Mary`s Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

Objective

In this study, an **automatic speech recognition (ASR) for speech sound disorder evaluation** was developed to detect articulation errors in children

Participants & Methods

The study targeted children under 18 seeking rehabilitation for articulation issues, excluding those with intellectual disabilities, autism spectrum disorders, motor speech disorders, severe speech intelligibility problems, extensive time abroad, or severe speech impediments

➢Automatic speech recognition model

This is an end-to-end model, pre-trained using **Mel-frequency Cepstral Coefficients (MFCC)**. The original training dataset consisted of **436,000 hours of adult voice** databases. The model was further trained using **137 hours** of speech data from typically developing children and **93.6 minutes** of speech data (6,935 words) from children with speech sound disorders.

➢ Evaluation on the model performance

Two Korean standardized speech sound disorder tests, **APAC** (Assessment of Phonology and Articulation for Children) and **U-TAP** (Urinal Test of Articulation and Phonology), were used in the study. The participants' responses were recorded using iPhone 10. The resulting transcriptions of the ASR model were compared with those made by speech-language pathologists (SLPs).

Results

A total of 30 children with speech sound disorder, including 10 females, aged 3-7 years, took part in the tests. **The reliability between the SLPs and ASR model** for both the percentage of consonants correct (PCC) and the percentage of vowels correct (PVC) was assessed as **'very good reliability (0.81~1.00)'** for both APAC and UTAP.

Table 1. Reliability of children's percentage of consonants correct and percentage of vowels correct assessed by SLPs and ASR models

	Test	**APAC**		**UTAP**			**Test**	**APAC**		**UTAP**	
	Evaluator	**SLPs**	**ASR**	**SLPs**	**ASR**		**Evaluator**	**SLPs**	**ASR**	**SLPs**	**ASR**
Average Percentage of consonants correct	M (±SD)	74.76% (±15.21%)	76.71% (±15.20%)	73.88% (±16.13%)	76.05% (±15.26%)	Average Percentage of vowels Correct	M (±SD)	84.17% (±10.91%)	85.17% (±10.38%)	79.67% (±10.33%)	80.33% (±11.59%)
	ICC (95%CI)	0.984 (CI: .953- .994)		0.978 (CI: .941 - .990)			ICC (95%CI)	0.929 (CI: .853- .966)		0.838 (CI: .659 - .923)	

The **phoneme error rate (PER)** was **11.5% for APAC**, and **12.22% for UTAP** which represents the percentage of instances where the transcription of the ASR model and SLPs were differed at the phoneme level.

The total number of ASR recognition disagreements transcribed as correct pronunciations by SLPs averaged **2.37 occurrences** per child in APAC and **2.7 occurrences** per child in UTAP.

On the other hand, the total number of ASR recognition disagreements transcribed as incorrect pronunciations by SLPs averaged **7.8 occurrences** per child in APAC and **7 occurrences** per child in UTAP.

Table 2. Common ASR disagreements transcribed as correct pronunciations by SLPs

	APAC				UTAP			
No.	Target word [IPA]	Phoneme in error*	ASR transcript-ion	Frequency (%)	Target word [IPA]*	Phoneme in error [IPA]**	ASR transcript-ion	Frequency (%)
1	화장실 [hwadzaŋɕil]	ㄹ [l]	Omission	4 (5.26%)	짹짹 [tsʼɛk̚tsʼɛk̚]	ㄱ [k̚]	Omission	4 (1.90%)
2	눈사람 [nunsʼaram]	ㄴ [n]	Omission	3 (3.94%)	싸움 [sʼaum]	ㅁ [m]	Omission	3 (1.42%)
3	눈사람 [nunsʼaram]	ㅁ [m]	Omission	3 (3.94%)	그림 [kɯrim]	ㄱ [k]	Omission	3 (1.42%)
4	화장실 [hwadzaŋɕil]	ㅈ [dz]	ㄷ [d]	3 (3.94%)	그림 [kɯrim]	ㅁ [m]	Omission	3 (1.42%)
5	컵 [kʰʌp̚]	ㅂ [p̚]	Omission	3 (3.94%)	귀 [kwi]	ㄱ [k]	ㅈ [dz]	3 (1.42%)

Table 3. Common ASR disagreements transcribed as incorrect pronunciations by SLPs

	APAC					UTAP				
No.	Target word [IPA]	Phoneme in error*	SLPs Transcripti-on	ASR transcripti-on	Frequen-cy (%)	Target word [IPA]	Phoneme in error**	SLPs Transcript-ion	ASR transcripti-on	Frequen-cy (%)
1	딸기 [tʼalgi]	ㄹ [l]	Omission	ㄹ [l]	6 (2.56%)	동물원 [doŋmurwʌn]	ㅇ [ŋ]	ㅁ [m]	ㅇ [ŋ]	6 (2.85%)
2	이빨 [ipʼal]	ㄹ [l]	Omission	ㄹ [l]	5 (2.14%)	괴물 [kwemul]	ㄹ [l]	Omission	ㄹ [l]	5 (2.38%)
3	딸기 [tʼalgi]	ㄱ [g]	ㄲ [kʼ]	ㄱ [g]	5 (2.14%)	눈썹 [nunsʼʌp̚]	ㅂ [p̚]	Omission	ㅂ [p̚]	4 (1.90%)
4	단추 [tantsʰu]	ㅊ [tsʰ]	ㅉ [tsʼ]	ㅊ [tsʰ]	4 (1.71%)	단추 [tantsʰu]	ㅊ [tsʰ]	ㅉ [tsʼ]	ㅊ [tsʰ]	4 (1.90%)
5	눈사람 [nunsʼaram]	ㄴ [n]	Omission	ㄴ [n]	4 (1.71%)	짹짹 [tsʼɛk̚tsʼɛk̚]	ㄱ [k̚]	Omission	ㄱ [k̚]	4 (1.90%)

* The phoneme in error refers to the phoneme in which the result of ASR differs from that of SLPs. The SLPs transcription matches the phoneme in error;

APAC: Assessment of Phonology and Articulation for Children; UTAP: Urinal Test of Articulation and Phonology; SLPs: Speech Language Pathologists; ASR: Automatic Speech Recognition model; IPA: International Phonetic Alphabet

Conclusion

The model had **reliability over 90% in agreement** with SLP transcriptions. This suggests using such a model in speech-language pathology is promising.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022R1C1C1008337)