ISSN 2982-6578, DOI: 10.23374/jaict.2025.5.3.005
Journal of Artificial Intelligence Convergence Technology

인공지능융합기술연구 Vol. 5, No. 3 : 144~149 September. 2025

아동 음성 분석을 통한 SELSI 기준 표현 언어 발달 단계 분류 모델

이신화[1], 이지수[1], 김지우[1], 김소연[1], 김한섭[2], 김웅식[2*]
[1]건양대학교 의료인공지능학과
[2]건양대학교 인공지능학과

SELSI-Based Classification of Expressive Language Development Stages Using Children's Voices

Sinhwa Lee[1], Jisu Lee[1], Jiwoo Kim[1], Soyeon Kim[1], Hanseob Kim[2], Woongsik Kim[2*]
[1]Department of Medical Artificial Intelligence, Konyang University
[2]Department of Artificial Intelligence, Konyang University

요 약 최근 관련 분야의 연구에 따르면 영유아기의 언어 지연은 학습 장애나 사회성 문제로 이어질 수 있어, 조기 진단과 개입의 중요성이 커지고 있다. 본 연구는 국내 표준 언어 발달 검사도구인 SELSI(Sequenced Language Scale for Infants)를 기반으로 아동의 실제 음성 데이터를 수집하여 표현 언어 능력의 발달 단계를 구분하는 인공지능 모델을 제안한다. 본 연구는 YouTube 및 SNS에서 수집한 아동의 음성을 Mel-Spectrogram으로 변환 후, Swin-Transformer 기반으로 학습하여 분류 성능을 향상시켰다. 실험 결과, 최종적으로 test 정확도 68.42%를 달성하였다. 본 논문에서 제안한 5단계 분류 체계는 실제 발화 기반의 세분화된 언어 발달 평가 가능성을 제시하여 향후 AI Agent를 통한 자동 진단 시스템 개발로 확장하고자 한다.

• 주제어 : 언어 지연, 표현 언어 발달, SELSI, 딥러닝, Mel-Spectrogram, Swin-Transformer, Early Children

Abstract Recent research has demonstrated that language delay in early childhood can lead to learning disabilities and social challenge, highlighting the importance of early diagnosis and intervention. This study proposes a classification model for expressive language development stages by collecting actual speech data from children, based on the SELSI (Sequenced Language Scale for Infants), a standarized Korean language development assessment tool. The research involved converting children's speech, collected from YouTube and social media, into Mel-Spectrograms, and then training a Swin Transformer-based model to improve classification performance. Experimental results achieved a final test accuracy of 68.42%. The proposed 5-stage classification system suggests the possibility of a more detailed language development assessment based on actual vocalizations and provides a foundation for future AI-driven automatic diagnostic systems.

• Key Words : Language delay, Expressive language development, SELSI, Deep Learning, Mel-Spectrogram, Swin-Transformer, Early Children

Received 31 May 2025, Revised 12 June 2025, Accepted 21 June 2025

* **Corresponding Author** Woong-Sik Kim, Dept. of Artificial Intelligence, Konyang University, 158, Gwanjeodong-ro, Seo-gu, Daejeon, Korea. E-mail: wskim@konyang.ac.kr

Ⅰ. 서론

언어 발달은 아동의 사회적 상호작용과 인지 발달에 핵심적인 역할을 한다. 언어 발달 과정에서 나타나는 언어 지연은 8개월 이후의 아동이 나이가 비슷한 아동에 비해서 통계적으로 말이 느린 경우를 의미한다. 이는 언어 발달의 질과 순서는 정상적이지만, 속도가 유의하게 뒤처지는 현상을 의미하며, 단순 언어 장애로 발전할 가능성을 내포하고 있다 [1]. 특히 만 3세경 언어 발달 지연으로 진단된 아동의 약 30%는 8세 이후까지 언어 지연이 지속되며, 만 4세경에 진단되었을 때에는 약 40%가 지속된다 [2]. 이러한 언어 지연은 학습장애나 행동 문제로 이어질 가능성이 높아 조기 진단과 개입이 매우 중요하다 [2].

현재 국내에서는 SELSI(영유아 언어발달 검사지) [3]를 통해 아동의 언어 발달 수준을 진단하고 있다. 실제 임상에서 SELSI를 활용한 조기 진단과 개입이 아동의 언어 발달 향상에 유의미한 효과를 나타낸 연구가 보고되고 있다 [4]. 연구에 따르면, 생후 18개월 영유아 건강검진에서 발달 지연이 의심된 만 3세 아동을 대상으로 전반적인 발달 평가를 실시한 결과, 언어 발달 연령이 생활 연령에 비해 약 1년 3개월 지연된 것으로 확인되었다. 이후 8주간 총 16회기의 언어치료를 통해 SELSI 표현 언어 점수가 46점에서 56점으로 향상되는 변화를 보였다. 이러한 사례는 SELSI 기반의 조기 진단과 개입이 언어 발달에 실질적인 효과를 가져올 수 있음을 보여주는 임상적 근거이며, 조기 진단을 통한 개입이 발달 지연 아동의 치료적 지원에 있어 핵심적인 요소임을 시사한다.

그러나 SELSI와 같은 문서 기반 검사 도구는 전문가의 해석과 평가 과정이 필요하기 때문에 시간과 비용이 많이 소요되는 한계가 있다. 따라서 보다 신속하고 자동화된 진단 도구에 대한 필요성이 제기되고 있다.

Ⅱ. 관련 연구

최근에는 인공지능 기술을 활용한 아동 언어 분석 연구가 등장하고 있다 [5]. 특히 자연어 처리(Natural Language Processing) 기술을 기반으로 아동의 발화를 텍스트로 변환한 후, 문법적 정확성과 언어 구조를 분석하여 발달 단계를 분류하는 딥러닝 기반 연구들이 시도되고 있다 [5]. 이는 실제 음성 신호에 내재된 억양, 피치, 발화 속도, 멜주파수(MFCC) 등의 음향 특성이나 발화 패턴을 충분히 반영하지 못하는 한계가 있다.

Ⅲ. 연구내용과 방법

3.1 제안하는 연구 내용

본 연구는 이러한 한계를 극복하기 위해 아동의 실제 음성 데이터를 기반으로 SELSI 항목에 따라 표현 언어 발달 단계를 분류하는 방법을 제안한다. 아동의 자연 발화를 기반으로 음성 특성을 정량화하고, 딥러닝 모델을 활용하여 언어 발달 단계를 자동 분류함으로써, 조기 선별 가능성과 실용성을 동시에 갖춘 인공지능 기반 진단 시스템을 구축하고자 한다.

3.2 데이터 수집

데이터 수집의 기준으로 활용된 SELSI는 생후 4개월부터 35개월까지의 영유아를 대상으로 언어의 전반적인 영역을 통합적으로 평가할 수 있는 국내 표준화된 검사 도구이다 [6]. 이는 임상 현장 [7]에서 널리 사용되며 말을 이해하고 알아듣는 수용 언어 능력과 자신의 생각, 의도를 전달하는 표현 언어 능력의 두 가지 평가 영역으로 구성되어 있다.

Table 1. SELSI Utilization Question Example

No.	Months	Question
27	16~17	말하는 억양이 문장처럼 들린다(정확한 낱말이나 문장이 아니어도 됨).
34	20~21	자신의 감정이나 느낌을 표현할 수 있다(예: “싫어”, “예뻐” 등).

Table 2. SELSI Delete Question Example

No.	Months	Question
24	14~15	유아가 의미를 알고 정확하게 사용하는 낱말이 4개 이상 있다.
35	20~21	말할 수 있는 낱말이 적어도 10~20개가 있다.

데이터는 표현 언어 항목을 활용하여 수집하였다. Table 1은 표현 언어 항목의 예시이다. 또한SELSI 문항에서 Table 2와 같이 단편적인 모습만으로 판단하기 어려운 문항들을 제거하여 총 56개 문항 중 실제 활용이 가능한 37개 문항을 선정하였다.

본 연구에서는 SELSI의 표현 언어 항목을 기반으로 아동의 개월 수가 명시된 콘텐츠를 선별하였다. 일반인에게 공개된 영상 플랫폼(YouTube, SNS)에서 수집한 음성 데이터를 활용하였으며, 수집된 모든 연구용 데이터는 원 음성으로부터 Mel-Spectrogram [8] 형태로 변환되어 비식별화 과정을 거쳤기 때문에 윤리적 문제를 해결하였다. 따라서 데이터 수집 목적은 학술적 용도에 한정되며, 상업적 활용은 일체 배제하였다.

아동의 자연발화를 수집하는 과정에서, 부모의 언어적 격려나 감정적 반응 등 상호작용 요소가 함께 포함된 사례가 일부 존재하였다. 데이터 전처리 과정에서 이러한 비아동 음성을 최대한 제거하고자 하였으나, 아동의 발화와 동시적으로 발생한 부모 음성은 기술적 한계로 인해 완전히 제거하지는 못하였다.

Table 3. Dataset Components

Class	Number of Samples	Avg length(s)
0-11	147	2.06
12-17	151	1.44
18-23	150	1.46
24-29	150	2.33
30-35	150	2.48

본 연구에서는 ASHA(미국 언어 청각 협회) [9]에서 정의한 개월별 분류 기준을 참고하여 Table 3과 같이 0~11개월, 12~17개월, 18~23개월, 24~29개월, 30~35개월 총 5개의 클래스로 구분하였으며, 클래스별 약 150개의 데이터를 수집해 총 748개의 데이터셋을 Table 3과 같이 구축하였다. Avg length는 수집한 데이터셋의 클래스별 평균 음성 시간 길이이며, 아동 발달에 따른 언어 능력 향상으로 인해 음성 길이가 점차 길어지는 경향을 확인할 수 있다.

3.3 데이터 전처리

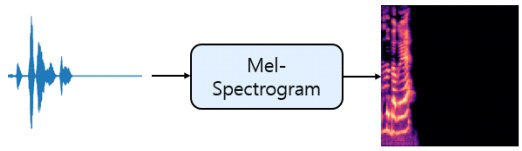

Fig. 1 Spectrogram Generation Procedure

음성 데이터를 정제하기 위해 noisereduce 라이브러리를 사용하여 음성 신호에 포함된 외부 소음을 감소시켰다. 또한, 음성 데이터를 5초로 통일하여 학습에 활용하기 위해 padding과 trimming 기법을 적용하였다. 5초보다 짧은 음성은 padding을 통해 무음 신호로 보완하였고, 5초보다 긴 음성은 trimming을 통해 일정 길이로 잘라내었다. 아동의 옹알이와 같은 특수한 음성 데이터는 명확한 음소 구분이 어렵고, 주파수 분포가 불규칙한 특성을 가진다 [10]. 이러한 비정형적 음성 데이터는 일반적인 음성 처리 방법으로는 효과적인 분류가 어렵다. 따라서 본 연구에서는 전처리된 음성 데이터를 Fig. 1과 같이 Mel-Spectroram으로 변환하여, 음향적 특성을 보다 정밀하게 2D 이미지 기반 딥러닝 모델의 입력으로 반영할 수 있도록 하였다. 이후 발달체계에 대한 다중 분류를 진행하기 위해 레이블에 대해 one-hot encoding을 적용하였으며 train, test 비율을 9대 1로 분할 하였다.

3.4 ResNet-34 기반 모델 학습

ResNet-34 [11]는 깊은 신경망에서 발생할 수 있는 기울기 소실 문제를 해결하기 위해 skip connection을 도입한 잔차 구조를 기반으로 하며, 이미지 분류 분야에서 널리 활용되는 대표적인 모델이다. 본 연구에서는 사전학습된 가중치를 적용한 ResNet-34를 사용하여 학습을 수행하였다. 하이퍼파라미터를 조정하여 학습을 수행한 결과, Test 데이터에서 63.29%의 정확도를 기록하였다.

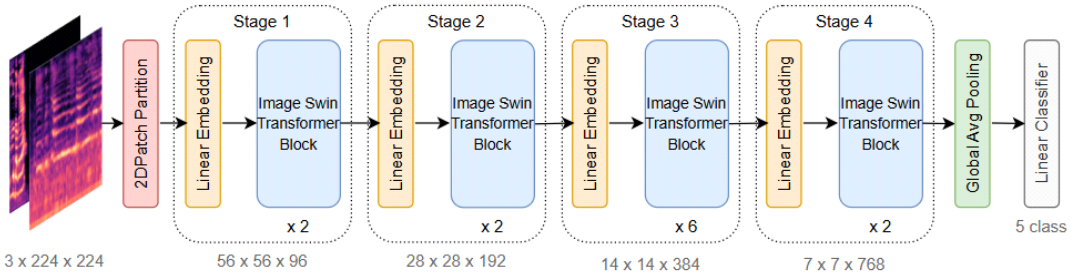

Fig. 2 Swin-Transformer Architecture

3.5 Swin-Transformer 기반 모델 학습

본 연구에서는 Fig. 2 구조의 Swin-Transformer[12]를 백본 네트워크로 사용하였다. Swin-Transformer는 ViT(Vision Transformer)를 기반으로 하며, Shifted Window 방식을 통해 다양한 스케일의 시각 정보를 효과적으로 처리할 수 있다. 먼저 입력 이미지에 Patch Partition과 선형 임베딩 과정을 적용하여 입력 텐서의 채널 수를 조정한 후 Swin-Transformer 블록에서는 W-MSA

(Window-based Multi-head Self-Attention)와 SW-MSA(Shifted Window MSA)를 적용하여 Self-Attention을 수행하였다. 이를 통해 Patch Merging을 통해 계층적으로 축소된 피처 맵을 생성하여 특징을 효과적으로 추출하고자 하였다 [13].

Ⅳ. 결과

Table 4. ResNet-34, Swin-Transformer results

Model	Accuracy(%)	Precision(%)
ResNet-34	63.29	64.76
Swin-Transformer	68.42	69.07

Table 5. Swin-Transformer final hyperparameters

epoch	25
learning_rate	0.0001
batch_size	32
optimizer	AdamW
scheduler	CosineAnnealingLR

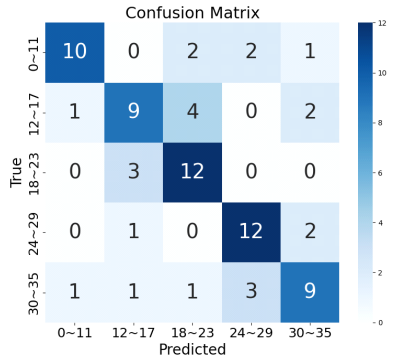

Fig. 3 Swin-Transformer Confusion Matrix

Swin-Transformer는 ResNet-34에 비해 파라미터 수가 많고 모델이 상대적으로 무겁지만, 다양한 스케일의 정보를 학습할 수 있다는 장점이 있다. 이러한 구조적 특성을 바탕으로 Swin-Transformer의 하이퍼파라미터를 조정하여 학습을 수행한 결과, Test 데이터에서 68.42%의 정확도로 Table 4와 같이 ResNet-34보다 약 5% 상승된 결과를 보인다. Table 5는 최종 모델에 적용된 하이퍼파라미터이며, Fig. 3은 테스트 결과의 Confusion Matrix를 나타낸다.

본 연구의 test 정확도는 68.42%, 정밀도는 69.07%이다. 정밀도가 높은 모델은 불필요한 오진을 줄이고 실제 위험군 아동을 보다 효과적으로 선별하는 데 기여할 수 있다[14,15].

Table 6. Comparison with Precedent Study

	Prior (Byoung-Doo Oh et al.[5])	Ours
Test Accuracy	78%	68.42%
Class	3	5
Data Type	Text	Children's Speech
Model	koBERT embedding->CNN-> Siamese Network	Swin -Transformer
Classification Method	Sentence- Similarity	Speech Pattern
Collect Instrument	Self-Development	SELSI

이는 서론에서 언급된 논문 [5]의 test 정확도 78%보다 낮은 수치를 보였다. 그러나 Table 6과 같이 실험 설계와 분류 조건의 차이를 고려할 때, 단순한 성능 수치만으로 비교하기에는 무리이다. 언어 발달은 연속적인 특성을 가지며 인접한 개월 수 간의 발화 특징이 명확히 구분되지 않기 때문에 본 연구에서 설정한 5개 연령 구간 간에는 클래스 간 중첩이 존재할 수밖에 없다. 반면, 관련 연구 2세, 4세, 6세의 세 구간만을 대상으로 하여 상대적으로 분류 간격이 넓고 클래스 간 모호성이 낮은 구조를 갖는다. 또한 본 연구가 음향 패턴 기반으로 분류를 진행한 것과 달리, 비교 논문은 문장 유사도 기반 접근 방식을 사용하여 분류 조건에도 차이가 있다[16,17,18].

이러한 차이들을 감안하면 두 연구 간의 성능을 동일 선상에서 비교하는 것은 한계가 있으며 본 연구는 보다 복잡하고 현실적인 분류 조건 하에서 의미 있는 성능을 도출하였다는 점에서 의의가 있다.

Ⅴ. 결론 및 향후 연구

본 논문은 SELSI를 기반으로 아동 음성 데이터를 분석하여 언어 발달 단계를 분류하는 모델을 개발하였다. Mel-Spectrogram을 활용하여 Swin-Transformer 기반으로 학습한 결과, 68.42%의 정확도를 도출하였다. 실험 결과를 바탕으로, 본 연구에서 제안한 5-class 분류 체계는 더욱 세분화된 언어 발달 단계를 구분하려는 시도로서 유의미한 접근으로 해석할 수 있다. 향후 연구에서는 LLM을 활용한 접근이 아닌, 본 연구에서 제안한 음성 분류 모델을 AI Agent 형태로 구성하여 반복적 피드백 학습 및 자기 강화 학습 구조를 도입함으로써, 보다 정밀하고 실용적인 자동 진단 체계를 구축하고자 한다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업(2024-0-00047) 지원과 대학혁신지원사업의 캡스톤디자인 교과목 지원을 받아 수행되었음.

REFERENCES

[1] Mi-Ae Jang et al., A Study on the Effectiveness of Parent-Child Interaction Therapy (PCIT) for Children with Language Delay, pp. 1-73, Hanyang University, 2019.

[2] Jae-Yeon Bae et al., "Speech Characteristics of Infants Aged 1 to 18 Months According to the Speech Development Model", Phonetics and Speech Sciences, vol. 2, no. 2, pp. 27-36, 2010.

[3] Hye-Ryun Yoon et al., "A Study on the Characteristics of Language Development in Infants and Toddlers According to Gender: Focusing on SELSI", Communication Sciences and Disorders, vol. 9, no. 1, pp. 30-44, 2004.

[4] Hye-Won Oh et al., "The Effects of Early Intervention on Postural Control and Language Ability in a 3-Year-Old Child with Developmental Delay: A Case Study", Journal of Korean Academy of Neurocognitive Rehabilitation Therapy, vol. 16, no. 1, pp. 77-87, 2024.

[5] Byoung-Doo Oh et al., "Deep Learning-Based End-to-End Language Development Screening for Children Using Linguistic Knowledge", Applied Sciences, vol. 12, no. 9, 2022, article 4651.

[6] Hye-Ryeon Yoon et al., "A Study on the Characteristics of Language Development in Infants and Toddlers According to Gender: Focusing on SELSI", Communication Sciences and Disorders, vol. 9, no. 1, pp. 30-44, 2004.

[7] Daegu Metropolitan City, Daegu Metropolitan City Welfare Portal, Welfare Policy Division. Accessed May 27, 2025. Availabl

e: https://www.daegu.go.kr/welf/index.do?menu_id=00000518
[8] Nian Shao et al., "CleanMel: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR", arXiv preprint, arXiv:2502.20040, 2025.
[9] S. Lim, "Trends in Automation Technologies for Computer Graphics Using Artificial Intelligence" Journal of Artificial Intelligence Convergence Technology, vol. 5, no. 1, pp. 37-42, Mar. 2025. Doi: 10.23374/jaict.2025.5.1.006
[10] Pamela R. Mitchell et al., "Phonetic variation in multisyllable babbling", Journal of Child Language, vol. 17, no. 2, pp. 247-265, 1990.
[11] S. M. Jo, "A study on technical analysis of efficient recommendation systems," Journal of Artificial Intelligence Convergence Technology, vol. 5, no. 1, pp. 17-22, Mar. 2025. Doi: 10.23374/jaict.2025.5.1.003
[12] Hyung-Seok Lee et al., "Estrus Detection in Hanwoo Cattle Using the Swin Transformer Model", Proceedings of the Korean Institute of Information Scientists and Engineers Conference, pp. 609-611, 2023.
[13] Kwang-Ho Yoon et al., Automatic Corrosion Segmentation in Ship Inspection Images Using Swin Transformer, pp. 1-45, Busan University, 2023.
[14] Steven A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence", Scientific Reports, vol. 12, article 5979, 2022.
[15] S. M. Jo, "A Study on Generalization Performance Analysis of Artificial Intelligence Data Learning Techniques," Journal of Artificial Intelligence Convergence Technology, vol. 5, no. 2, pp. 55-60, Jun. 2025. doi: 10.23374/jaict.2025.5.2.001
[16] S. K. Kang, "Deep Learning-based Image Analysis for Moving Object Detection," Journal of Artificial Intelligence Convergence Technology, vol. 5, no. 2, pp. 83-88, Jun. 2025. doi: 10.23374/jaict.2025.5.2.005
[17] Shihao Cai et al., "Agentic feedback loop modeling improves recommendation and user simulation", arXiv preprint, arXiv: 2410.20027, 2025.
[18] M. A. Kang, "Structure and Development Trends of YOLO Object Detection Models," Journal of Artificial Intelligence Convergence Technology, vol. 5, no. 1, pp. 43-48, Mar. 2025. doi: 10.23374/jaict.2025.5.1.007

저자소개

이 신 화 (Sinhwa Lee)
2022년 3월~현재 : 건양대학교 의료인공지능학과 학사과정
관심분야 : 딥러닝, 의료인공지능, 영상처리, 음성인식, HCI, XAI

이 지 수 (Jisu Lee)
2022년 3월~현재 : 건양대학교 의료인공지능학과 학사과정
관심분야 : 머신러닝, 딥러닝, 빅데이터, ICT, 데이터마이닝, 컴퓨터 비전

김 지 우 (Jiwoo Kim)
2022년 3월~현재 : 건양대학교 의료인공지능학과 학사과정
관심분야 : 인공지능, 딥러닝, 머신러닝, 영상처리, HCI, 음성인식

김 소 연 (Soyeon Kim)
2022년 3월~현재 : 건양대학교 의료인공지능학과 학사과정
관심분야 : 인공지능, 머신러닝, 천문학, 과학, 지구과학, 과학수사

김 한 섭 (Hanseob Kim)
2019.02 조선대학교 컴퓨터공학과 (공학사)
2021.08 고려대학교 컴퓨터학과 (공학석사)
2025.02 고려대학교 컴퓨터학과 (공학박사)
2019.01~2024.01 한국과학기술연구원 인공지능연구단 인턴연구원
2025.03~현재 건양대학교 인공지능학과 조교수
관심분야 : 가상현실, 증강현실, HCI, 감성컴퓨팅, 가상인간, 인공지능

김 웅 식 (Woong-Sik Kim)
1989년 2월 : 인하대학교 정보공학과(공학석사)
2007년 2월 : 인하대학교 컴퓨터공학과(공학박사)
2006년 3월~현재 : 건양대학교 인공지능학과 교수
관심분야 : 인공지능, 의료공학, 임베디드, 뇌파