



pISSN 2005-8063
eISSN 2586-5854
2026. 9. 30.
Vol.18 No.3
pp. 83-96

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

<https://doi.org/10.13064/KSSS.2026.18.3.083>



Exploring acoustic features and predictive modeling for automatic fluency assessment in Korean as a foreign language*

Hye-Sook Park¹ · Minhwa Chung² · Sunhee Kim^{1,**}

¹Department of French Language Education, Learning Sciences Research Institute, Seoul National University, Seoul, Korea

²Department of Linguistics, Seoul National University, Seoul, Korea

Abstract

Research on fluency assessment in Korean as a foreign language (KFL) has primarily focused on assessment constructs and manual scoring, whereas studies on automatic fluency assessment remain limited. This study aims to explore effective fluency features for automatic fluency assessment of KFL learners and to compare the performance of different predictive models based on these features. Temporal (T5), acoustic (A8 and A88), and hybrid (H13 and H93) feature sets combining temporal and acoustic features were constructed and applied to tree-based models [Gradient Boosting Classifier (GBC) and XGBoost (XGB)] and transformer-based models for tabular data [FT-Transformer (FTT) and SAINT]. Experiments were conducted on 1,672,398 training and 209,083 validation utterances from the AIHub Korean Learner Speech Corpus. To examine the effect of data imbalance, performance was evaluated under original and balanced score-distribution conditions. Quadratic weighted kappa was the primary metric. Accuracy, Pearson's correlation coefficient (r), root mean squared error, and Macro-F1 were supplementary metrics. Under original conditions, each model's acoustic condition outperformed its temporal and hybrid conditions. A8 was highest for GBC and XGB, A88 for FTT and SAINT, and SAINT-A88 overall. Hybrid sets showed no consistent improvements. Under balanced conditions, patterns varied, but SAINT-A88 remained the best-performing configuration. These findings indicate that performance varied across the feature conditions and predictive models examined.

Keywords: Korean as a foreign language, automatic fluency assessment, acoustic features, predictive modeling

1. 서론

유창성(fluency)은 넓게는 구술 능력 전반을 가리키는 개념으

로 사용되기도 하지만, 좁은 의미에서는 발화가 시간상 얼마나 자연스럽게 이어지는지를 반영하는 수행 특성으로 다루어진다 (Lennon, 1990). Segalowitz(2010, 2016)는 유창성을 인지적 유창

* This research was supported by the Culture, Sports and Tourism R&D Program, funded by the Ministry of Culture, Sports and Tourism, through the Korea Culture Technology Planning and Evaluation Institute, an affiliated institute of the Korea Creative Content Agency in 2025 (Project Name: Development of a Personalized AI Korean Language Learning System for Efficient Korean Language Learning by Foreign Learners, Project Number: RS-2025-02413038, Contribution Rate: 100%).

** sunkim@snu.ac.kr, Corresponding author

Received 14 August 2026; Revised 10 September 2026; Accepted 11 September 2026

© Copyright 2026 Korean Society of Speech Sciences. This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

성(cognitive fluency), 발화 유창성(utterance fluency), 지각된 유창성(perceived fluency)으로 구분하고, 발화 유창성을 인지적 처리 효율성이 발화 산출의 시간적 특성으로 드러난 결과로 설명한다. 이러한 관점에서 모국어 화자의 발화에서 유창성은 비교적 자동화된 언어 처리 과정이 발화 속도(speech rate), 휴지 및 주저 양상(pausing and hesitation patterns), 시간적 연속성(temporal continuity)과 같은 시간적 특성으로 실현된 결과로 이해할 수 있다(Segalowitz, 2010, 2016).

외국어 발화의 유창성 역시 발화의 시간적 양상을 통해 관찰될 수 있지만, 그 양상이 형성되는 조건은 모국어 발화와 다르다. 외국어 학습자는 제한된 언어 지식과 증가한 처리 부담 속에서 발화를 계획하고 산출하므로, 외국어 발화의 유창성은 발화 속도뿐 아니라 산출 과정에서 나타나는 중단(breakdown)과 수정(repair) 양상을 함께 고려하여 설명된다. Skehan(2009)과 Tavakoli & Skehan(2005)은 외국어 발화 유창성을 속도 유창성(speed fluency), 중단 유창성(breakdown fluency), 수정 유창성(repair fluency)의 차원에서 설명하며, 발화 속도, 휴지(pause), 반복(repetition), 자기 수정(self-repair)과 같은 지표가 이러한 하위 차원을 반영한다고 본다. Suzuki et al.(2021)은 발화 속도와 휴지 관련 자질이 청자가 지각하는 유창성 판단과 밀접한 관련이 있다고 보고한다. 따라서 외국어 발화에서 유창성은 발화의 시간적 조직을 반영하는 동시에, 외국어 수행 과정에서 나타나는 산출 부담과 처리 효율성을 설명하는 지표로 볼 수 있다(de Jong, 2018).

외국어 발음 평가 연구에서는 음성 신호를 활용하여 학습자의 발음 수행을 자동으로 평가하려는 연구가 이어져 왔다. Franco et al.(1997)은 자동음성인식 기반 발음 평가의 가능성을 제시하였으며, Neumeyer et al.(2000)은 학습자 발음 평가를 위한 자동 채점 접근을 제안하였다. Cucchiari et al.(2000, 2002)은 자동음성인식 기술을 활용하여 제2언어 학습자의 발음 특성을 분석하였고, Fontan et al.(2018, 2022, 2024)은 음향 자질을 이용한 유창성 예측 연구를 수행하였다. 이러한 연구들은 외국어 발화에서 나타나는 음성 정보를 활용하여 학습자의 발음 및 유창성 수행을 정량적으로 평가하려는 연구 흐름을 보여준다. 이때 발음 평가는 주로 분절음 및 초분절음의 실현을 대상으로 하며, 발화 전반을 포괄하는 말하기 평가 속에서 유창성은 발화의 시간적 조직과 원활성을 판단하는 평가 영역으로 다루어진다.

외국어 학습자 발화의 유창성 자동 평가에서 유창성을 설명하기 위해 반복적으로 다루어져 온 정보 축 중 하나는 발화의 시간적 구조(temporal structure)이다. Skehan(2009)과 Tavakoli & Skehan(2005)은 발화 속도와 휴지를 유창성을 설명하는 주요 요소로 제시하였으며, 발화의 연속성과 끊김이 유창성 판단의 근거가 된다고 설명하였다. de Jong & Wempe(2009)는 음절 핵 검출을 기반으로 발화 속도를 자동 산출하는 방법을 제안하였고, Ginther et al.(2010)은 발화 속도와 휴지 관련 지표가 구술 능력과 연관성이 있다고 보고하였다. Deng et al.(2020) 역시 주저와 같은 비유창성 현상이 자동 유창성 평가에서 활용될 수 있음을 보여주었다.

그러나 발화의 유창성은 발화 속도와 휴지 같은 시간적 구조만으로 충분히 설명되기 어렵다. 유창성 판단은 여러 발화 특성의 결합에 의해 이루어지며(Tavakoli & Skehan, 2005), 음높이, 에너지, 스펙트럼 구조와 같은 음향 정보 역시 발화 수행의 차이를 반영할 수 있다(de Jong, 2018). 이러한 맥락에서 자동 발음 유창성 평가 연구는 시간 자질뿐 아니라 음향 자질을 함께 고려하는 방향으로 확장되었다. Fontan et al.(2018)은 자동음성인식 없이 음향 자질만을 활용하여 제2언어 유창성을 예측할 수 있음을 보여주었으며, Fontan et al.(2022)은 스펙트럼 및 음향 정보를 기반으로 아동 발화의 유창성을 예측하였다. Fontan et al.(2024)은 음향 자질을 활용하여 실어증 화자의 유창성을 자동 측정하는 가능성을 제시하였다. 이러한 연구들은 음향 정보가 발음 유창성 평가에서 시간 자질과 구분되는 독립적인 후보 자질군으로 활용될 수 있음을 보여준다. Eyben et al.(2015)은 음높이, 에너지, 스펙트럼과 같은 정보를 체계적으로 추출하기 위해 GeMAPS(Geneva Minimalistic Acoustic Parameter Set)과 그 확장형인 extended GeMAPS(eGeMAPS)를 제안하였으며, 해당 자질 집합은 자동 평가를 포함한 다양한 음성 분석 연구에서 활용되어 왔다(Schuller et al., 2013).

상용 말하기 자동 평가 시스템에서도 유창성 평가를 위해 다양한 층위의 음성 정보가 활용된다. ETS SpeechRater는 발화 속도, 평균 침묵 지속시간, 단위 시간당 침묵 빈도, 반복 횟수 및 비 유창성(disfluency) 표지 수와 같은 시간 자질과 음높이 범위(range of normalized pitch), 에너지 변동(standard deviation of power) 등의 음향 자질을 함께 활용하여 학습자의 유창성을 평가한다(Chen et al., 2009, 2018; Evanini et al., 2015; Higgins et al., 2011; Xi et al., 2008; Zechner et al., 2009). Microsoft Pronunciation Assessment 역시 말하기 속도(speaking speed)와 단어 간 묵음 구간(silent breaks between words)뿐 아니라 강세(stress), 억양(intonation), 리듬(rhythm)을 평가에 반영하며(Microsoft Azure, 2023; Nelson & Cardoso, 2023), Pearson Versant 또한 timing, rhythm, intonation과 같은 다양한 음성 정보를 활용한다(Pearson, 2021). 이러한 사례는 발음 평가와 말하기 평가를 포괄하는 자동 평가 체계 안에서도 시간 정보와 음향 정보가 유창성 평가의 후보 자질로 활용될 수 있음을 보여준다.

외국어로서의 한국어에 관한 유창성 연구는 유창성 평가 구인에 관한 연구와 유창성 평가를 다룬 소수의 연구로 이루어져 왔으며, 자동화 연구 역시 최근까지 매우 제한적이었다. Kim & Kim(2017)은 L2 한국어 말하기 유창성 평가의 평가 기준과 채점 신뢰도를 다국면 라쉬 모형으로 분석하였으며, 유창성이 어휘, 문법, 발음, 말속도, 운율과 같은 구성 요소와 함께 평가되는 것을 보여주었다. Kang et al.(2017)은 한국어 학습자의 발화를 평가할 때 말속도(speech rate), f0 범위, 휴지 빈도와 같은 구인이 유창성 판단에 큰 영향을 미친다고 보고하였다. 이후 Liu(2018)와 Liu & Won(2020)은 한국어 학습자의 발화 및 말하기 과제에서 나타나는 유창성 특성을 검토하였고, Lee et al.(2025)은 한국어 학습자의 말하기 유창성 평가를 위한 자동 채점 자질을 설정하고자 하였다. 이러한 연구들은 한국어 학습자의 유창성 평가

에서 시간적 특성과 운율·음향적 특성을 함께 검토할 필요가 있음을 보여준다.

한국어 학습자 발화를 대상으로 한 음성공학 분야의 자동 평가 연구도 제한적이나 지속적으로 이루어져 왔다. Ryu et al.(2016)은 제2언어 한국어 학습자의 발음을 대상으로 음성 정렬과 발화 특성에 기반한 자질을 활용하여 자동 발음평가를 수행하고, 자질 선택을 통해 평가 성능을 비교하였다. 이 연구는 한국어 학습자 발화의 자동 평가에 음향적·음성학적 정보를 활용하였다는 점에서 본 연구와 접점을 가진다. 다만 해당 연구는 학습자의 발음 정확성 평가에 초점을 두고 있어, 발화의 시간적 조직과 음향적 실현 양상을 바탕으로 유창성 수동점수를 예측하는 본 연구와는 평가 대상 및 평가 구인에서 구분된다.

국제적으로는 제2언어 발화의 유창성 또는 운율 수행을 자동으로 평가하기 위해 시간 자질, 음향 자질, 자동음성인식 기반 정보 및 신경망 기반 음성 표현을 활용하는 연구가 축적되어 왔다. Matsuura et al.(2022)은 대화 발화의 특성을 포함한 자질을 활용하여 제2언어 구술 유창성 자동 채점을 수행하였으며, 대화적 자질을 포함한 조건에서 QWK(quadratic weighted kappa) .833을 보고하였다. Dong et al.(2024)은 제2언어 영어 발화에서 음향 자질을 선별하고 음향모델 기반 정보를 결합하여 운율, 유창성 및 발음 정확성 평가를 시도하였다. 그러나 이들 연구는 대상 언어, 발화 과제, 점수 척도, 데이터 규모 및 모델 설정이 본 연구와 상이하므로, 보고된 성능은 직접적인 우열 비교의 기준이라기보다 자동 유창성 평가의 방법론적 참조점으로 해석할 필요가 있다.

이에 본 연구는 외국어로서 한국어 학습자의 발음 유창성 자동 평가를 위한 유의미한 자질을 탐색하고, 이를 기반으로 다양한 예측 모델의 성능을 비교하고자 한다. 구체적으로 시간 자질군, 음향 자질군, 그리고 시간 자질과 음향 자질을 함께 사용한 통합 자질군을 구성하고, 이를 트리 기반 모델과 표형 데이터용 트랜스포머 기반 모델에 적용하여 성능을 비교한다. 이와 같은 자질 기반 유창성 연구는 유창성에 대한 평가 성능을 향상할 수 있을 뿐 아니라 한국어를 외국어로 학습하는 학습자들에게 유창성에 대한 피드백을 제공하는 데 있어 매우 유용할 것으로 예상된다.

본 논문의 구성은 다음과 같다. 2장에서는 발음 유창성 자질 탐색과 예측 모델링을 포함한 방법론을 제시한다. 3장에서는 실험 데이터, 선별 음향 자질의 선택 절차와 결과, 유창성 예측 모델의 설정 및 성능 평가 결과를 제시한다. 4장에서는 앞선 실험 결과를 바탕으로 자질군 구성, 모델 구조 및 점수 분포 조건에 따른 성능 양상을 논의한다. 마지막으로 5장에서는 연구 결과를 종합하고 연구의 의의와 한계, 향후 과제를 제시한다.

2. 방법론

2.1. 유창성 자질군 탐색

2.1.1. 시간 자질

본 연구에서 유창성 자질로서 먼저 시간 자질군은 선행연구

를 기반으로 하여 다섯 개를 제안한다. 시간 자질은 읽기 기반 발음 과제 음성에서 원음성 신호만으로 안정적으로 산출할 수 있는 핵심 자질이다. 발화 속도와 조음 속도가 유창성 및 구술 능력과 관련된 대표적인 시간 지표로 다루어져 왔으며(Ginther et al., 2010; Skehan, 2009; Tavakoli & Skehan, 2005), 휴지와 중단은 발화의 연속성과 끊김을 설명하는 주요 지표로 활용되어 왔다(Deng et al., 2020; Skehan, 2003, 2009). 본 연구에서는 이러한 시간적 특성을 반영하기 위해 전체 발화 시간(speech duration, SD), 발화 속도(speaking rate, SR), 조음 속도(articulation rate, AR), 발화 중단(voice breaks, VB), 발화 중단 비율(voice break ratio, VBR)의 다섯 개 자질을 사용한다(T5). 이때 SD는 발화 속도와 발화 중단 비율을 산출하기 위한 기준 시간 단위이자, 읽기 과제에서 발화 산출에 소요된 전체 수행 시간을 나타내는 기초 시간 자질로 사용하였다.

본 연구에서는 이들 자질을 원음성 신호의 RMS(root mean square) 에너지 변화를 이용하여 자동 산출하였다. 음성 신호는 25 ms의 프레임 길이와 10 ms의 프레임 이동 간격으로 분할하였으며, 각 프레임에서 RMS 에너지를 계산하였다. RMS 에너지는 프레임 내 음성 표본의 제곱 평균에 대한 제곱근으로 정의되며, 본 연구에서는 수식 (1)과 같이 산출하였다(O'Shaughnessy, 2000).

$$RMS(t) = \sqrt{\frac{1}{N} \sum_{n=1}^N \chi_t(n)^2} \quad (1)$$

수식 (1)에서 $\chi_t(n)$ 은 t 번째 프레임의 n 번째 표본값이며, N 은 해당 프레임의 표본 수이다. 산출된 RMS 시계열은 이동평균으로 평활화하였고, 평균 RMS를 기준으로 발화 구간과 휴지 구간을 구분하였다. 또한 RMS 곡선의 국소 피크를 이용하여 음절 수를 추정하였다.

시간 자질 산출을 위해 음성 구간의 총 지속시간 T_s , 휴지 구간의 총 지속시간 T_p , 추정 음절 수 N_{syll} , 휴지 또는 음성 단절 구간의 수 N_p 를 정의하였다. 이들은 T5 산출을 위한 중간 변수이며, 최종 입력 자질은 SD, SR, AR, VB, VBR이다. SD는 수식 (2)와 같이 음성 구간과 휴지 구간의 합으로 계산된다.

$$SD = T_s + T_p \quad (2)$$

SR은 SD 대비 산출된 음절 수로 수식 (3)과 같다.

$$SR = \frac{N_{syll}}{SD} \quad (3)$$

AR은 휴지 구간을 제외한 실제 발화 구간 대비 음절 수로 계산하며, 수식 (4)와 같다.

$$AR = \frac{N_{syll}}{T_s} \quad (4)$$

VB는 발화 내에서 탐지된 휴지 구간의 개수로 정의하며, 수식 (5)와 같다.

$$VB = N_p \quad (5)$$

VBR은 SD에서 휴지 구간이 차지하는 비율로 계산하며, 수식 (6)과 같다.

$$VBR = \frac{T_p}{SD} \times 100 \quad (6)$$

표 1은 T5에 포함된 다섯 개 시간 자질과 각 자질의 조작적 정의를 제시한 것이다.

표 1. 시간 자질
Table 1. Temporal feature set

Equation No.	Feature	Operational definition
(2)	Speech suration (SD)	Total duration of the utterance
(3)	Speaking rate (SR)	Number of syllables per speech duration
(4)	Articulation rate (AR)	Number of syllables per speech duration excluding pauses
(5)	Voice breaks (VB)	Number of detected pause intervals
(6)	Voice break ratio (VBR)	Proportion of pause duration in speech duration

2.1.2. 음향 자질

음향 자질은 기본적으로 openSMILE에서 제공하는 eGeMAPS를 기반으로 추출하였다(Eyben et al., 2010, 2015). 그리고 각 발화에서 추출한 eGeMAPS 88개 자질 전체를 사용한 음향 자질군(A88) 이외에, 이 88개 가운데 유창성 수동점수 예측에 상대적으로 높은 기여도를 보인 8개 자질을 선별하였다(A8). 따라서 A88은 전체 음향 자질군, A8은 선별 음향 자질군에 해당하며, 두 조건 모두 시간 자질을 포함하지 않는 음향 자질 단독 조건이다.

A8의 자질 선별을 위하여 AIHub 교육용 한국어 학습자 발화 코퍼스(AIHub & NIA, 2022a, 2022b, 2022c, 2022d)에서 일부를 사용하여 별도의 데이터를 구성하였다. 영어권(AIHub & NIA, 2022a), 유럽어권(AIHub & NIA, 2022b), 중국 및 일본어권(AIHub & NIA, 2022c), 아시아어권(AIHub & NIA, 2022d)의 네 개 모국어권 학습자 발화 가운데 유창성 수동 점수가 부여되고 eGeMAPS 자질을 추출할 수 있는 발음 과제 발화를 각 집단에서 11,735개씩 선정하여, 총 46,940개 발화를 사용하였다.

eGeMAPS를 기반으로 추출한 88개 자질 가운데 1차 자질 선별은 Gradient Boosting Classifier를 사용하여 교차검증 기반 재귀적 자질 제거(recursive feature elimination with cross-validation) 방식으로 수행하였다. 5겹 교차검증(five-fold cross-validation)을 적용하였고, 평가 지표로는 예측 점수와 수동점수 간 평균절대오차로 설정하였다. 자질 선택 모델의 주요 설정은 n_estimators=100, learning_rate=0.1, max_depth=3, loss=log_loss, random_state=42였다. 이를 통해 유창성 점수 예측에 상대적으로 기여하는 후보 자질을 우선 추출하였다.

이후 후보 자질을 중요도 순으로 정렬하고, 상위 10, 15, 20,

25, 30, 40, 60, 80개 자질 조건의 성능을 비교하였다. 이 단계에서는 random_state=42로 설정한 GBR(gradient boosting regressor)를 사용하여 후보 자질의 중요도를 산출하였다. 각 자질 수 조건의 성능 평가는 중앙값 대체와 GBR로 구성한 파이프라인에 5겹 교차검증(five-fold cross-validation)을 적용하여 수행하였다(n_splits=5, shuffle=True, random_state=42). 평가지표는 유창성 수동점수와 예측값 간 평균절대오차(mean absolute error)로 설정하였으며, 교차검증의 각 폴드(fold)에서 산출된 평균절대오차를 바탕으로 조건별 평균과 표준편차를 계산하였다. 후보 자질 수에 따른 5겹 교차검증 평균절대오차 비교 결과는 표 2에 제시하였다.

표 2. 후보 자질 수에 따른 5겹 교차검증 평균절대오차
Table 2. Five-fold cross-validated mean absolute error by candidate feature set size

Number of features	Mean MAE	SD
10	.8449	.0037
15	.8408	.0041
20	.8391	.0040
25	.8378	.0042
30	.8375	.0036
40	.8373	.0034
60	.8375	.0033
80	.8381	.0035

가장 낮은 평균절대오차는 40개 자질 조건에서 나타났다. 본 연구에서는 최저 평균절대오차와의 차이가 0.01 이하인 조건을 예측 성능 차이가 제한적인 후보군으로 설정하였다. 이는 유창성 수동점수가 0점에서 5점까지 1점 간격으로 부여되는 척도임을 고려하여 설정한 운영 기준이다. 이 기준에 따라 최저 평균절대오차에 0.01을 더한 .8473 이하의 조건 가운데 가장 적은 수의 자질을 포함한 10개 자질 조건을 최종 후보로 선택하였다. 이와 같이 예측 성능이 유사한 후보 조건 사이에서 더 간명한 조건을 선택하는 방식은 불필요한 자질 수 증가를 줄이고 해석 가능성을 확보하기 위한 모형 선택 원칙과 맥을 같이한다(Hastie et al., 2009).

최종 10개 후보에는 시간 관련 자질인 음절 수(syllable count)와 휴지 시간(pause duration)이 포함되어 있었으므로 제외하였다. 이에 따라 A8은 eGeMAPS 기반 후보 자질 가운데 선별된 8개의 음향 자질로 구성하였다. 표 3은 최종 선별된 음향자질(A8)에 포함된 자질과 각 자질의 음향 영역 및 조작적 정의를 제시한 것이다.

표 3. 선별 음향 자질군(A8)

Table 3. Selected acoustic feature set (A8)

No	Feature	Acoustic domain	Operational definition
1	loudnessPeaksPerSec	Loudness	Rate of loudness peaks per second
2	loudness_sma3_meanFallingSlope		Mean falling slope of the smoothed loudness contour
3	loudness_sma3_stddevNorm		Normalized standard deviation of the smoothed loudness contour
4	StddevUnvoicedSegmentLength	Voicing	Standard deviation of unvoiced segment duration
5	F0semitoneFrom27.5Hz_sma3nz_pctlrang0-2	Pitch	Percentile range of F0 in semitones
6	F0semitoneFrom27.5Hz_sma3nz_percentile20.0		20th percentile of F0 in semitones
7	F1amplitudeLogRelF0_sma3nz_amean	Formant	Arithmetic mean of F1 amplitude relative to F0
8	loudness_sma3_percentile80.0	Loudness	80th percentile of the smoothed loudness contour

2.1.3. 유창성 자질군

다음 표 4는 본 연구에서 제안한 시간 자질군(T5)과 두 음향 자질군(A8, A88), 그리고 이들을 통합한 통합 자질군(hybrid feature set)으로 시간 자질(T5)과 음향 자질(A8, A88)을 각각 통합한 통합 자질군(H13, H93)이다. 시간 자질군은 발화 유창성의 시간적 조직을, 음향 자질군은 음성 신호의 실현 양상을 반영하며, 통합 자질군은 두 정보 층위를 하나의 입력 공간에서 함께 사용하기 위한 조건으로 설정하였다.

표 4. 유창성 자질군

Table 4. Fluency feature sets

Feature set	Type	Components	Main purpose
T5	Temporal	5 RMS-based temporal features	Temporal fluency representation
A8	Acoustic	Selected eGeMAPS acoustic features	Selected acoustic representation
A88		Full eGeMAPS acoustic features	Full acoustic representation
H13	Hybrid	T5 + A8	Selected temporal-acoustic integration
H93		T5 + A88	Full temporal-acoustic integration

eGeMAPS, extended Geneva Minimalistic Acoustic Parameter Set.

2.2. 유창성 예측 모델

본 연구에서는 시간 자질과 음향 자질의 예측적 유용성을 동일한 평가 조건에서 비교하기 위해 네 개의 유창성 예측 모델을 사용하였다. 모델은 트리 기반 모델(tree-based model)과 표형 데이터용 트랜스포머 기반 모델(transformer-based tabular model)로 구분하였다.

트리 기반 모델로는 GBC(gradient boosting classifier; Friedman,

2001)와 XGB(extreme gradient boosting; Chen & Guestrin, 2016)을 사용하였다. GBC는 순차적으로 약한 학습기를 추가하여 이전 단계의 오차를 보정하는 그라디언트 부스팅(gradient boosting) 계열의 분류 모델이다. XGB는 정규화와 병렬 처리 구조를 강화한 확장 그라디언트 부스팅 모델로, 본 연구에서는 회귀 기반 모델로 구성하였다.

표형 데이터용 트랜스포머 기반 모델로는 특징 토큰나이저 트랜스포머(feature tokenizer transformer, FTT; Gorishniy et al., 2021)와 셀프어텐션 및 샘플 간 어텐션 트랜스포머(self-attention and intersample attention transformer, SAINT; Somepalli et al., 2021)를 사용하였다. FTT는 표형 데이터의 각 입력 자질을 토큰화한 뒤 셀프어텐션(self-attention) 구조를 통해 자질 간 관계를 학습하는 모델이다. SAINT는 셀프어텐션뿐 아니라 표본 간 어텐션(intersample attention)을 함께 활용하여 자질 간 관계와 표본 간 관계를 동시에 학습하도록 설계된 모델이다.

GBC의 모델 예측값은 각 유창성 점수 범주에 대한 예측 확률을 산출하므로, 연속 예측 점수는 각 점수 범주 k 의 예측 확률 $p(k|x)$ 에 해당 점수를 가중하여 합산한 확률가중 기대점수로 계산하였다. GBC의 예측 점수 $\hat{y}$ 는 수식 (7)과 같다. 여기서 $p(k|x)$ 는 각 발화에서 추출한 입력 자질 벡터 x 가 주어졌을 때 해당 발화가 유창성 점수 범주 k 에 속할 예측 확률을 의미한다.

$$\hat{y} = \sum_{k=0}^5 kp(kvertx) \quad (7)$$

XGB는 회귀 기반 모델로서 0점에서 5점 범위의 연속 점수를 직접 산출하였다. FTT와 SAINT 역시 평균제곱오차를 목적함수로 하여 연속 점수를 직접 출력하였다. 따라서 네 모델의 예측값은 모두 연속 점수로 통일하여 Pearson's r과 RMSE를 산출하였다.

2.3. 모델 평가 방법

모델 성능 지표로는 이차 가중 카파 계수(quadratic weighted kappa, QWK), 정확도(accuracy, ACC), 피어슨 상관계수(Pearson correlation coefficient, Pearson's r), 평균제곱근오차(root mean squared error, RMSE)를 사용하였다. QWK는 예측 점수와 수동 점수 간 일치도를 점수 간 거리까지 고려하여 산출하는 지표이므로, 0점에서 5점까지의 순서형 유창성 점수 평가에 적합하다. ACC는 예측 점수와 수동점수가 정확히 일치한 비율을, r은 두 점수 간 선형 상관을, RMSE는 예측 오차의 크기를 나타낸다. 본 연구에서는 QWK를 주 지표로 삼고, ACC, r, RMSE를 보조 지표로 함께 제시하였다.

또한 점수 범주별 예측 성능을 확인하기 위해 정밀도(Precision)와 재현율(Recall)의 조화평균인 F1 점수(F1-score)를 점수별로 산출하고, 그 평균값인 Macro-F1을 함께 제시하였다. 정밀도는 특정 점수로 예측한 사례 가운데 실제로 해당 점수에 속한 사례의 비율을, 재현율은 실제 해당 점수에 속한 사례 가운데 모델이 해당 점수로 정확히 예측한 사례의 비율을 의미한다.

다. Macro-F1은 각 점수 범주의 F1 점수를 단순 평균한 값으로, 특정 점수 범주의 표본 수가 전체 성능 해석에 미치는 영향을 완화하여 확인하기 위한 지표이다.

Pearson's r 과 RMSE는 연속 예측값을 기준으로 산출하였다. 반면 QWK, ACC, 점수별 F1 및 Macro-F1은 예측값을 유창성 점수 범위(0점에서 5점 사이)로 제한한 뒤 가장 가까운 정수로 반올림하여 산출하였다. 또한 동일 검증 발화에서 산출된 조건 간 예측 오차 차이를 비교하기 위해 쌍체 절대오차(paired absolute error)를 기준으로 Wilcoxon 부호순위 검정을 수행하였고, 다중 비교에는 Holm 보정(Holm, 1979)을 적용하였다.

3. 실험 결과

3.1. 데이터

본 연구는 AIHub에서 구축한 교육용 한국어 학습자 발화 코퍼스(AIHub & NIA, 2022a, 2022b, 2022c, 2022d)를 활용하였다. 해당 코퍼스는 발음 과제와 말하기 과제로 구분되며, 본 연구에서는 발음 유창성 자동 평가에 필요한 시간 자질과 음향 자질을 산출할 수 있고, 전문가 유창성 수동점수를 기준값으로 활용할 수 있는 발음 과제 발화를 선정하였다. AIHub 원천 데이터는 영어권 1,601,367개, 1,029.8시간(AIHub & NIA, 2022a), 유럽어권 2,434,822개, 1,538.5시간(AIHub & NIA, 2022b), 중국 및 일본어권 1,331,180개, 1,006.2시간(AIHub & NIA, 2022c), 아시아어권 2,355,685개, 1,510.6시간(AIHub & NIA, 2022d)의 네 개 언어권 데이터로 구성되어 있다. 원천 데이터는 음성 파일, 발화 구간 정보가 포함된 강제 정렬 데이터, 발화 및 화자 정보를 포함한 메타데이터를 제공한다.

본 연구에서는 전체 코퍼스 가운데 시간 자질과 음향 자질을 산출할 수 있고, 유창성 수동점수가 함께 확보된 발화만을 실제 분석에 사용하였다. 최종 실험 데이터는 AIHub 원천 데이터의 훈련·검증 구분을 유지하여 구성하였으며, 훈련 데이터는 1,672,398개, 검증 데이터는 209,083개 발화로 이루어졌다. 표 5는 최종 선정 데이터의 제1언어권 분포를 제시한 것이다. 훈련 및 검증 분할에서 언어권별 구성 비율은 거의 동일하게 유지되었으며, 이를 통해 이후 성능 비교에서 특정 제1언어권의 과대 또는 과소 표집이 개입할 가능성을 줄였다.

표 5. 선정 데이터의 분할에 따른 제1언어권 분포
Table 5. L1 group distribution of the selected data by split

L1 Group	Training (N)	Training (%)	Validation (N)	Validation (%)
Asia	508,383	30.40	63,558	30.40
English	347,195	20.76	43,406	20.76
Europe	530,306	31.71	66,299	31.71
China & Japan	286,514	17.13	35,820	17.13
Total	1,672,398	100	209,083	100

각 발화에는 0점부터 5점까지 1점 간격의 유창성 수동점수가 부여되었다. 이 점수는 범주형 표지이면서 동시에 낮은 점수에서 높은 점수로 이어지는 순서형 척도이므로, 본 연구의 예측 과정에는 단순 다중 분류가 아니라 점수 간 순서와 거리 정보를 함께 고려해야 하는 발음 유창성 평가 문제로 설정된다. 표 6은 선정 데이터의 유창성 수동점수 분포를 나타낸다.

표 6. 선정 데이터의 유창성 수동점수 분포
Table 6. Distribution of manual fluency scores in the selected data

Fluency score	Train (N)	%	Validation (N)	%
0	1,266,447	75.73	158,362	75.74
1	17,570	1.05	2,276	1.09
2	71,645	4.28	8,904	4.26
3	136,879	8.18	17,315	8.28
4	138,542	8.28	17,115	8.19
5	41,315	2.47	5,111	2.44
Total	1,672,398	100	209,083	100

표 6에서 보듯이 선정 데이터는 훈련 및 검증 분할 모두에서 강한 점수 불균형을 보인다. 특히 0점 발화가 전체의 약 4분의 3을 차지하는 반면, 1점과 5점 발화는 상대적으로 적다. 이러한 점수 분포가 예측 성능에 미치는 영향을 함께 검토하기 위해 ‘원천 점수 분포 유지 조건’과 ‘균형 점수 분포 조건’으로 구성하였다. 원천 점수 분포 유지 조건에서는 3.1절에서 제시한 훈련 발화 1,672,398개와 검증 발화 209,083개를 원래의 점수 분포 그대로 사용하였다. 이 조건은 실제 발음 과제 데이터에서 나타나는 점수 분포를 반영한 상태에서 모델 예측 성능을 확인하기 위한 것이다.

균형 점수 분포 조건에서는 점수 범주별 표본 수의 차이가 평가 결과에 미치는 영향을 통제하기 위해, 훈련 데이터와 검증 데이터에서 각 점수 범주의 표본 수를 동일하게 구성하였다. 검증 데이터는 가장 적은 표본 수를 보인 1점 범주의 2,276개를 기준으로, 0점부터 5점까지 각 점수 범주에서 동일한 수의 발화를 추출하여 총 13,656개 발화로 구성하였다. 훈련 데이터도 1점 범주의 17,570개를 기준으로 각 점수 범주에서 동일한 수의 발화를 추출하여 총 105,420개 발화로 구성하였다. 표본 추출은 기존 훈련·검증 분할 속에서 각각 독립적으로 수행하였으며, 난수 시드는 42로 고정하였다. 두 점수 분포 조건은 조건 간 성능의 직접적인 우열을 검증하기 위한 것이 아니라, 실제 점수 분포를 유지한 평가와 점수 범주별 표본 수를 통제된 평가에서 자질군 및 모델별 성능 양상이 어떻게 달라지는지를 함께 확인하기 위한 보완적 조건으로 설정하였다.

3.2. 모델 설정

본 연구에서는 2.2절에서 제시한 트리 기반 모델인 GBC와 XGB, 그리고 표형 데이터용 트랜스포머 기반 모델인 FTT와 SAINT를 사용하였다. 2.1.3절에서 구성한 자질군은 각 모델의 입력 조건으로 설정하였다. 구체적으로 GBC와 XGB에는 시간

자질군 T5, 선별 음향 자질군 A8, 그리고 T5와 A8을 결합한 통합 자질군 H13을 적용하였다. FTT와 SAINT에는 시간 자질군 T5, 전체 음향 자질군 A88, 그리고 T5와 A88을 결합한 통합 자질군 H93을 적용하였다. 이에 따라 각 모델 내에서 시간 자질 조건, 해당 음향 자질 조건, 그리고 통합 자질 조건의 예측 성능을 비교하였다.

자질군의 배치는 각 모델 계열에서 검토하고자 한 음향 정보의 범위를 달리하여 설정하였다. 트리 기반 모델에는 자질 선별 절차를 통해 도출된 A8을 적용하여, 유창성 수동점수 예측에 기여하는 핵심 음향 자질을 축소된 입력 조건에서 검토하고자 하였다. 이는 예측 결과를 개별 자질 수준에서 해석하고, 향후 학습자 피드백으로 연계할 수 있는 자질 구성을 고려한 것이다. 반면 표형 데이터용 트랜스포머 기반 모델에는 eGeMAPS 88개 자질 전체로 구성된 A88을 적용하여, 자질 토큰 간 관계를 학습하는 모델에 전체 음향 정보를 제공하고자 하였다. 이를 통해 선별 과정에서 제외된 정보를 포함하는 전체 음향 자질 조건에서의 예측 성능을 검토하였다.

GBC는 HistGradientBoostingClassifier를 사용하였고, 주요 설정은 learning_rate=0.05, max_depth=3, max_iter=300, random_state=0이었다. XGB는 XGBoost 기반 회귀 모델(XGBRegressor)을 사용하였다. 주요 설정은 n_estimators=600, max_depth=6, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8, reg_lambda=1.0, random_state=42였다.

FT와 SAIN는 동일한 입력 자질 조건에서 비교하였으며, 두 모델 모두 난수 시드는 42로 고정하였다. 주요 설정은 d_token=24, n_blocks=1, n_heads=4, d_ff=48, dropout=0.10, learning_rate=3e-4, weight_decay=1e-였다. 최적화에는 AdamW를 사용하였다(Loshchilov & Hutter, 2017). 손실 함수는 평균제곱오차로 설정하였고, 학습 에포크는 T5 조건에서 20회, A88 및 H93 조건에서 15회로 설정하였다.

3.3. 실험 결과

표 7은 원천 데이터의 유창성 점수 분포를 유지한 검증 데이터에서 모델 및 자질군별 전체 성능을 제시한 것이다.

표 7. 원천 분포 유지 조건에서 모델 및 자질군별 전체 성능

Table 7. Overall performance by model and feature set under the original data score distribution

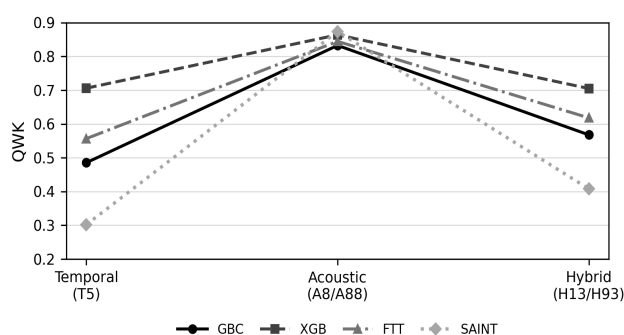
Model	Feature Set	QWK	ACC	r	RMSE	Macro-F1
GBC	T5	.485	.543	.583	1.701	.268
GBC	A8	.833	.765	.853	.777	.317
GBC	H13	.568	.542	.649	1.520	.268
XGB	T5	.706	.410	.801	1.068	.261
XGB	A8	.864	.791	.878	.713	.331
XGB	H13	.705	.349	.797	1.152	.247
FTT	T5	.557	.457	.659	1.514	.244
FTT	A88	.845	.785	.859	.763	.322
FTT	H93	.618	.543	.704	1.413	.247
SAINT	T5	.302	.432	.415	2.237	.222
SAINT	A88	.874	.813	.893	.676	.341
SAINT	H93	.408	.534	.513	2.018	.241

QWK, quadratic weighted kappa; ACC, accuracy; RMSE, root mean squared error; GBC, gradient boosting classifier; XGB, XGBoost; FTT, FT-Transformer; SAINT, self-attention and intersample attention transformer.

원천 분포 유지 조건에서는 네 모델 모두에서 A8 또는 A88이 가장 높은 성능을 보였다. GBC와 XGB에서는 A8 조건이, FTT와 SAINT에서는 A88 조건이 각 모델에 적용된 조건 가운데 가장 높은 성능을 나타냈다. 전체 조건 중에서는 SAINT-A88이 QWK, ACC, r, Macro-F1에서 가장 높고, RMSE에서 가장 낮았다.

T5는 모든 모델에서 해당 모델에 적용된 음향 자질 조건보다 낮은 성능을 보였다. 통합 자질 조건인 H13과 H93은 일부 모델에서 T5보다 높은 성능을 보였으나, 각 모델에 적용된 A8 또는 A88의 성능을 넘어서지는 못하였다. 예를 들어 GBC에서는 H13이 T5보다 QWK, ACC, Pearson's r 및 RMSE에서 높은 성능을 보였으나 A8보다 낮았고, FTT와 SAINT에서도 H93은 T5보다 높은 성능을 보인 일부 지표가 있었으나 A88의 성능에는 이르지 못하였다.

원천 점수 분포 유지 조건의 검증 데이터에서 0점 발화는 전체의 75.74%를 차지하므로, 모든 발화를 0점으로 예측하는 최빈 범주 기준 정확도는 .757이다. 각 모델에서 가장 높은 ACC를 보인 조건은 GBC-A8(.765), XGB-A8(.791), FTT-A88(.785), SAINT-A88(.813)이며, 이는 최빈 범주 기준 정확도보다 각각 .008, .034, .028, .056 높았다.



GBC, gradient boosting classifier; XGB, extreme gradient boosting; FTT, FT-Transformer; SAINT, self-attention and intersample attention transformer; QWK, quadratic weighted kappa.

그림 1. 원천 분포 유지 조건에서 자질군 구성에 따른 QWK 변화

Figure 1. QWK trends across temporal, acoustic, and hybrid feature sets under the original data score distribution

그림 1은 원천 분포 유지 조건에서 자질군 구성에 따른 QWK 변화로서, 여기에서도 표 7과 동일한 양상이 확인된다. 모든 모델에서 각 모델에 적용된 A8 또는 A88의 QWK가 가장 높게 나타났다. T5는 이들보다 낮은 QWK를 보였다. H13과 H93은 GBC, FTT, SAINT에서 T5보다 높은 QWK를 보였으나, 모든 모델에서 A8 또는 A88의 QWK를 넘어서지는 못했다. GBC와 XGB에서는 A8이, FTT와 SAINT에서는 A88이 각 모델에 적용된 조건 가운데 가장 높은 QWK를 보였다.

또한 2.3에서 제시한 방식에 따라 동일 검증 발화에 대한 쌍체 절대오차를 기준으로 Wilcoxon 부호순위 검정을 수행한 결과, 원천 분포 유지 조건에서 각 모델에 적용된 음향 자질 조건은 시간 자질 조건 및 통합 자질 조건보다 유의하게 작은 예측 오차를 보였다(Holm 보정 $p < .001$). SAINT-A88은 XGB-A8보다도 유의하게 작은 오차를 보였다(Holm 보정 $p < .001$).

이와 같이 원천 분포 유지 조건에서 각 모델에 적용된 A8 또는 A88은 T5보다 높은 성능을 보였다. H13은 GBC에서 T5보다 QWK와 Pearson's r 이 높고 RMSE가 낮았으며, H93은 FTT와 SAINT에서 모든 평가 지표에서 T5보다 높은 성능을 보였다. 그러나 H13과 H93은 해당 모델에 적용된 음향 자질 조건의 성능을 넘어서지는 못했다.

표 8은 원천 분포 유지 조건에서 각 모델의 QWK가 가장 높은 자질 조건을 기준으로 산출한 점수별 F1을 제시한 것이다. 이에 따라 GBC와 XGB에서는 A8을, FTT와 SAINT에서는 A88을 적용한 결과를 제시하였다.

표 8. 원천 분포 유지 조건에서 모델별 최고 QWK 자질 조건의 점수별 F1
Table 8. Class-wise F1 scores of the highest-QWK feature condition for each model under the original data score distribution

Model	Feature set	F1-0	F1-1	F1-2	F1-3	F1-4	F1-5	Macro-F1
GBC	A8	.945	.035	.206	.443	.275	.000	.317
XGB	A8	.961	.033	.206	.451	.336	.000	.331
FTT	A88	.955	.027	.183	.458	.307	.000	.322
SAINT	A88	.974	.047	.252	.468	.302	.000	.341

GBC, gradient boosting classifier; XGB, extreme gradient boosting; FTT, FT-Transformer; SAINT, self-attention and intersample attention transformer; QWK, quadratic weighted kappa.

점수별 성능에서는 전체 성능과 구별되는 양상이 나타났다. 네 모델에서 QWK가 가장 높은 자질 조건은 모두 0점 구간에서 가장 높은 F1을 보였다. 반면 1점 구간에서는 모든 조건에서 낮은 성능을 보였고, 5점 구간에서는 네 조건 모두 F1이 .000이었다. GBC-A8은 5점으로 1개 발화를 예측하였으나 정분류는 없었고, XGB-A8, FTT-A88 및 SAINT-A88은 5점으로 예측한 발화가 없었다. 2점, 3점, 4점 구간에서는 모델별 차이가 나타났으나, 전반적으로 0점 구간에 비해 낮은 수준의 성능을 보였다. 2점과 3점 구간에서는 SAINT-A88이 가장 높은 F1을 보였고, 4점 구간에서는 XGB-A8이 가장 높은 F1을 보였다. Macro-F1은 SAINT-A88(.341), XGB-A8(.331), FTT-A88(.322), GBC-A8(.317) 순으로 높게 나타났다.

원천 점수 분포 유지 조건에서 확인된 성능이 특정 점수 범주의 높은 비율에 영향을 받는지를 확인하기 위하여, 점수 범주별 표본 수를 동일하게 구성한 균형 점수 분포 조건에서도 성능을 평가하였다. 표 9는 균형 점수 분포 조건에서의 모델 및 자질군별 전체 성능을 제시한 것이다.

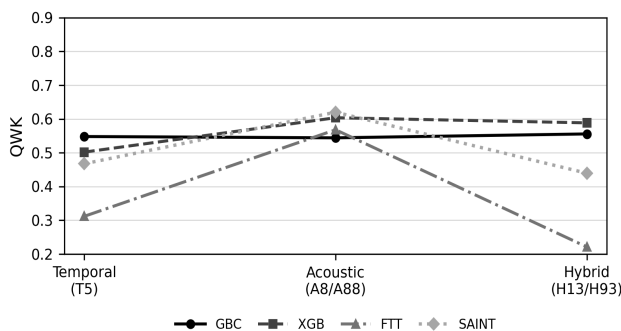
표 9. 균형 점수 분포 조건에서 모델 및 자질군별 전체 성능
Table 9. Overall performance by model and feature set under the balanced data score distribution

Model	Feature set	QWK	ACC	r	RMSE	Macro-F1
GBC	T5	.548	.291	.644	1.370	.266
GBC	A8	.544	.267	.635	1.322	.243
GBC	H13	.556	.297	.652	1.374	.281
XGB	T5	.502	.308	.588	1.460	.298
XGB	A8	.604	.302	.672	1.265	.285
XGB	H13	.589	.304	.665	1.332	.299
FTT	T5	.312	.264	.435	1.773	.214
FTT	A88	.568	.301	.631	1.326	.276
FTT	H93	.221	.247	.317	1.926	.189
SAINT	T5	.467	.286	.574	1.523	.247
SAINT	A88	.620	.326	.685	1.249	.310
SAINT	H93	.439	.288	.547	1.624	.271

QWK, quadratic weighted kappa; ACC, accuracy; RMSE, root mean squared error; GBC, gradient boosting classifier; XGB, XGBoost; FTT, FT-Transformer; SAINT, self-attention and intersample attention transformer.

균형 점수 분포 조건에서는 원천 점수 분포 유지 조건과 비교하여 모델 및 자질군별 성능 양상이 다르게 나타났다. 그럼에도 SAINT-A88은 전체 조건 중 가장 높은 QWK, ACC, Pearson's r, Macro-F1과 가장 낮은 RMSE를 보였다. XGB-A8도 QWK, Pearson's r, RMSE에서 상대적으로 높은 수준을 보였다.

자질군 구성에 따른 성능 양상은 모델별로 다르게 나타났다. GBC에서는 H13이 QWK, ACC, Pearson's r, Macro-F1에서 가장 높은 값을 보였으나, RMSE는 A8에서 가장 낮았다. XGB에서는 A8과 H13의 차이가 원천 점수 분포 유지 조건보다 작아졌으며, A8은 QWK, Pearson's r, RMSE에서, H13은 ACC와 Macro-F1에서 상대적으로 큰 값을 보였다. FTT와 SAINT에서는 A88이 T5와 H93보다 높은 성능을 보였다.



GBC, gradient boosting classifier; XGB, extreme gradient boosting; FTT, FT-Transformer; SAINT, self-attention and intersample attention transformer; QWK, quadratic weighted kappa.

그림 2. 균형 점수 분포 조건에서 자질군 구성에 따른 QWK 변화
Figure 2. QWK trends across temporal, acoustic, and hybrid feature sets under the balanced score distribution

그림 2는 균형 점수 분포 조건에서 자질군 구성에 따른 QWK 변화로서, 균형 점수 분포 조건에서 모델 간 QWK 차이가 원천 점수 분포 유지 조건보다 축소된 양상이 확인된다. XGB에서는 A8이, SAINT에서는 A88이 각 모델에 적용된 조건 가운데 가장 높은 QWK를 보였고, FTT에서도 A88이 T5와 H93보다 높은 QWK를 보였다. 반면 GBC에서는 H13이 가장 높은 QWK를 보였다. 그러나 H13과 H93은 모든 모델에서 일관된 최적 조건으로 나타나지 않았다.

4. 논의

본 연구의 결과는 원천 점수 분포 유지 조건에서 각 모델에 적용된 음향 자질 조건이 시간 자질 조건과 통합 자질 조건보다 높은 성능을 보였다는 점으로 정리할 수 있다. 원천 점수 분포 유지 조건에서는 모든 모델에서 A8 또는 A88이 T5보다 높은 성능을 보였다. GBC와 XGB에서는 A8이, FTT와 SAINT에서는 A88이 각 모델에 적용된 조건 가운데 가장 높은 QWK를 보였다. 균형 점수 분포 조건에서도 XGB-A8과 SAINT-A88은 상대적으로 높은 성능을 유지하였다. 반면 H13과 H93은 두 점수 분포 조

건에서 일관된 최적 조건으로 나타나지 않았다. 이러한 자질군의 배치는 트리 기반 모델에서 선별 자질 조건을 통해 핵심 음향 자질의 예측적 유용성과 해석 가능성을 검토하고, 표형 데이터용 트랜스포머 기반 모델에서 전체 음향 자질을 통해 자질 간 관계를 학습하도록 한 분석 목적에 따른 것이다. 따라서 A8과 A88의 결과는 두 자질군의 일반적 우열을 모델 구조와 독립적으로 비교한 결과라기보다, 각 모델 계열에 설정된 시간 자질, 음향 자질 및 통합 자질 조건의 성능 양상을 제시하는 것으로 해석할 수 있다. 외국어 유창성 연구가 발화 속도와 휴지 등 시간적 수행 특성을 주요 지표로 다루어 온 것(de Jong, 2018; Skehan, 2009; Tavakoli & Skehan, 2005)과 비교하면, 본 결과는 한국어 학습자 발화의 유창성 수동점수 예측에서 시간 자질과 구분되는 음향 자질을 함께 검토하게 한다.

A8은 음량(loudness), 음높이(pitch), 음성성(voicing), 포먼트(formant) 영역의 자질로 구성되며, 이 가운데 음량 관련 자질의 비중이 높다. 이 자질군은 GBC와 XGB에서 각 모델의 가장 높은 QWK를 보였다. Kang et al.(2017)은 한국어 학습자 발화의 유창성 판단에서 말속도, 기본주파수(F0) 범위, 휴지 빈도를 다루었다. 본 연구의 A8은 시간 자질을 포함하지 않는 조건에서 기본주파수 및 음량 관련 음향 정보를 함께 사용하였다. 이에 따라 본 연구에서는 한국어 학습자 유창성에서 논의되어 온 기본주파수 관련 특성을 시간 자질과 구분되는 음향 자질 조건에서 검토하였다. 또한 음향 자질을 사용한 제2언어 유창성 예측을 제시한 Fontan et al.(2018)과 비교할 때, 본 연구는 한국어 학습자 발화에서 선별 음향 자질군과 전체 음향 자질군을 구분하고 이를 여러 예측 모델에 적용하였다.

시간 자질을 결합한 H13과 H93이 음향 자질 단독 조건보다 일관되게 높은 성능을 보이지 않은 결과는, 시간 자질의 결합이 본 실험에서 설정한 모든 모델 조건에서 성능 향상으로 이어지지 않았음을 보여준다.

점수 분포 조건에 따른 결과는 유창성 자동 평가의 성능을 해석하는 기준과 관련된다. 원천 점수 분포 유지 조건에서 각 모델의 최적 자질 조건은 전체 QWK, ACC, Pearson's r, RMSE에서 높은 성능을 보였으나, 점수별 F1에서는 0점 구간과 1점·5점 구간 사이에 큰 차이가 나타났다. 특히 5점 범주에서는 네 조건 모두 정분류가 없었으며, GBC-A8에서만 5점 예측 1건이 산출되었고 나머지 세 조건에서는 5점 예측이 산출되지 않았다. 균형 점수 분포 조건에서는 자질군과 모델별 성능 양상이 원천 점수 분포 유지 조건과 다르게 나타났으며, 일부 모델에서는 자질군별 상대적 성능도 달라졌다. 따라서 표 7의 전체 성능, 표 8의 점수별 F1, 표 9의 균형 점수 분포 조건 결과는 유창성 수동점수 예측의 서로 다른 측면을 제시한다. 발음 과제 발화에 부여된 유창성 수동점수를 대상으로 한 본 연구에서, 전체 지표는 전반적 예측 성능을, 점수별 F1은 각 점수 범주의 예측 양상을 제시한다.

5. 결론

본 연구는 외국어로서의 한국어 학습자의 유창성 자동 평가

를 위한 음향 자질 탐색 및 예측 모델링을 목적으로 수행되었다. 한국어 학습자 발화의 유창성을 자동으로 평가하는 연구가 제한적인 상황에서, 본 연구는 발음 과제 발화에 부여된 유창성 수동점수를 예측 대상으로 설정하였다. 이를 위해 시간 자질(T5), 선별 음향 자질(A8), 전체 음향 자질(A88), 통합 자질(H13, H93)을 구성하고, 트리 기반 모델과 표형 데이터용 트랜스포머 기반 모델의 성능을 원천 점수 분포 유지 조건 및 균형 점수 분포 조건에서 비교하였다.

실험 결과, 원천 점수 분포 유지 조건에서는 각 모델에 적용된 음향 자질 조건이 시간 자질 조건보다 높은 성능을 보였다. GBC와 XGB에서는 A8이, FTT와 SAINT에서는 A88이 각 모델에 적용된 조건 가운데 높은 성능을 보였으며, 전체 조건에서는 SAINT-A88이 가장 높은 성능을 나타냈다. 통합 자질 조건은 음향 자질 단독 조건보다 일관되게 높은 성능을 보이지 않았고, 점수 분포 조건에 따라 전체 성능과 자질군별 상대적 성능이 달라졌다. 이러한 결과는 한국어 학습자 발화의 유창성 수동점수 예측에서 음향 자질을 탐색하고, 자질군과 예측 모델의 조합에 따른 성능을 비교한다는 연구 목적에 직접적으로 답한다.

본 연구의 기여는 한국어 학습자 발화의 유창성 자동 평가에서 시간 자질과 음향 자질을 구분하고, 선별 음향 자질군과 전체 음향 자질군을 서로 다른 모델 계열의 입력 조건으로 구성하여 성능을 검토한 데 있다. 다만 실험 대상이 발음 과제 발화에 한정되었고, 점수 분포의 불균형이 컸으며, 통합 자질군은 단순 결합 방식으로 구성되었다. 후속 연구에서는 발화 유형을 확장한 데이터에서 자질-모델 조합의 성능을 비교하고, 소수 점수 범주의 예측 양상과 시간-음향 자질의 결합 방식을 추가로 검토할 수 있다.

References

- AIHub, & National Information Society Agency (NIA). (2022a). Korean speech data from native English speakers for educational purposes. Retrieved from <https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71469>
- AIHub, & National Information Society Agency (NIA). (2022b). Korean speech data from native European speakers for educational purposes. Retrieved from <https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71489>
- AIHub, & National Information Society Agency (NIA). (2022c). Korean speech data from native Chinese and Japanese speakers for educational purposes. Retrieved from <https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71490>
- AIHub, & National Information Society Agency (NIA). (2022d). Korean speech data from native Asian speakers for educational purposes. Retrieved from <https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71479>
- Chen, L., Zechner, K., & Xi, X. (2009, May). Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 442-449). Princeton, NJ.
- Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., ... Gyawali, B. (2018). Automated scoring of nonnative speech using the *SpeechRater*sm v. 5.0 engine. *ETS Research Report Series*, 2018(1), 1-31.
- Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). San Francisco, CA.
- Cucchiari, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107(2), 989-999.
- Cucchiari, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, 111(6), 2862-2873.
- de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, 15(3), 237-254.
- de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385-390.
- Deng, H., Lin, Y., Utsuro, T., Kobayashi, A., Nishizaki, H., & Hoshino, J. (2020, May). Automatic fluency evaluation of spontaneous speech using disfluency-based features. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 9239-9243). Barcelona, Spain.
- Dong, W., van Hout, R., Cucchiari, C., & Strik, H. (2024, July). L2 prosody assessment by combining acoustic and neural model features. *Proceedings of the Speech Prosody 2024* (pp. 165-169). Leiden, The Netherlands.
- Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the *TOEFL Junior*[®] comprehensive writing and speaking test. *ETS Research Report Series*, 2015(1), 1-11.
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2), 190-202.
- Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on*

- Multimedia* (pp. 1459-1462). Firenze, Italy.
- Fontan, L., Kim, S., De Fino, V., & Detey, S. (2022, November). Predicting speech fluency in children using automatic acoustic features. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1085-1090). Chiang Mai, Thailand.
- Fontan, L., Le Coz, M., & Detey, S. (2018, September). Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of French. *Interspeech 2018* (pp. 2544-2548). Hyderabad, India.
- Fontan, L., Prince, T., Nowakowska, A., Sahraoui, H., & Martinez-Ferreiro, S. (2024). Automatically measuring speech fluency in people with aphasia: First achievements using read-speech data. *Aphasiology*, 38(5), 939-956.
- Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (1997, April). Automatic pronunciation scoring for language instruction. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1471-1474). Munich, Germany.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.
- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379-399.
- Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932-18943.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
- Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2), 282-306.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.
- Kang, S., Ahn, H., Hong, E., Min, B., Cho, S., Lee, L., & Park, H. (2017). A study of Korean production and fluency as a foreign language. *The Korean Society of Bilingualism*, 67, 1-29.
- Kim, N., & Kim, Y. (2017). Scoring reliability verification of L2 Korean speaking fluency assessment: Use of many-faceted Rasch model. *Studies in Linguistics*, 45, 483-524.
- Lee, H. W., Oh, J., Lee, J., Choi, E. J., Lee, S., & Kim, H. (2025). A study on the development of automatic scoring features for assessing Korean language learners' speaking fluency. *Korean Education*, (143), 443-475.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387-417.
- Liu, T. (2018). An analysis of fluency characteristics in Korean learners' speech. *Proceedings of the Korea Grammar Education Circle*, 2018(1), 89-99.
- Liu, T., & Won, M. J. (2020). Analysis of fluency in Korean learners' speaking tasks. *Teaching Korean as a Foreign Language*, 56, 265-284.
- Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv. <https://doi.org/10.48550/arXiv.1711.05101>
- Matsuura, R., Suzuki, S., Saeki, M., Ogawa, T., & Matsuyama, Y. (2022, November). Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1312-1320). Chiang Mai, Thailand.
- Microsoft Azure. (2023). Speech service ignite update: New enhancement features for pronunciation assessment. Retrieved from <https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/speech-service-ignite-update-%E2%80%93new-enhancement-features-for-pronunciation-assess/3978093>
- Nelson, C., & Cardoso, W. (2023, August). Evaluating the effectiveness of Microsoft Transcribe for automating the assessment of pronunciation in language proficiency tests. *EuroCALL 2023. CALL for All Languages* (pp. 117-122). Reykjavik, Iceland.
- Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30(2-3), 83-93.
- O'Shaughnessy, D. (2000). *Speech communications: Human and machine* (2nd ed). New York, NY: Institute of Electrical and Electronics Engineers.
- Pearson. (2021). Versant English test: Official guide for test-takers. Retrieved from <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/pearson-languages/en-gb/pdfs/versant-resources/official-test-guide-versant-english-test.pdf>
- Ryu, H. H., Hong, H., Kim, S., & Chung, M. (2016, December). Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1-6). Jeju, Korea.
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language – State-of-the-art and the challenge. *Computer Speech & Language*, 27(1), 4-39.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. England, UK: Routledge.
- Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, 54(2), 79-95.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*,

36(1), 1-14.

- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510-532.
- Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training (version 1). arXiv. <https://doi.org/10.48550/arXiv.2106.01342>
- Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, 105(2), 435-463.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. *Planning and Task Performance in a Second Language*, 11, 239-273.
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SPEECHRATERsm V1.0. *ETS Research Report Series*, 2008(2), i-102.
- Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10), 883-895.

• 박혜숙 (Hye-Sook Park)

서울대학교 불어교육과 강사, 학습과학연구소 객원 연구원
서울시 관악구 관악로 1
Tel: 02-880-7690
Email: cielcine@snu.ac.kr
관심분야: 프랑스어 교정음성학, 발음 자동평가, 발음 교수법

• 정민화 (Minhwa Chung)

서울대학교 언어학과 교수
서울특별시 관악구 관악로 1
Tel: 02-880-9195
Email: mchung@snu.ac.kr
관심분야: 음성인식, 음성언어처리, 컴퓨터 기반 언어교육

• 김선희 (Sunhee Kim) 교신저자

서울대학교 불어교육과 교수, 학습과학연구소 겸무 연구원
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: sunhkim@snu.ac.kr
관심분야: 프랑스어 음성학, 음성언어처리

외국어로서의 한국어 학습자의 유창성 자동 평가를 위한 음향 자질 탐색 및 예측 모델링*

박 혜 숙¹ · 정 민 화² · 김 선 희¹

¹서울대학교 불어교육과·학습과학연구소, ²서울대학교 언어학과

국문초록

외국어로서의 한국어 유창성 평가 연구는 주로 유창성 평가 구인이나 수동 평가를 다루어 왔으며, 유창성 자동 평가 연구는 제한적이다. 본 연구는 외국어로서의 한국어 학습자의 유창성 자동 평가를 위한 효과적인 자질을 탐색하고, 이들 자질에 기반한 서로 다른 예측 모델의 성능을 비교하고자 한다. 이를 위해 시간 자질군(T5), 음향 자질군(A8, A88), 그리고 시간 자질과 음향 자질을 결합한 통합 자질군(H13, H93)을 구성하여 트리 기반 모델[GBC(gradient boosting classifier), XGB(XGBoost)]과 표형 데이터용 트랜스포머 기반 모델[FTT(FT-Transformer), SAINT(self-attention and intersample attention transformer)]에 적용하였다. 데이터로는 AIHub 교육용 한국어 학습자 발화 코퍼스의 발음 과제에서 추출한 1,672,398개의 훈련 발화와 209,083개의 검증 발화를 사용하였다. 데이터 불균형의 영향을 분석하기 위해 원천 점수 분포 유지 조건과 균형 점수 분포 조건으로 구분하여 성능을 평가하였다. 평가 지표는 이차가중카파(quadratic weighted kappa)를 주 지표로 하고, 정확도, 피어슨 상관관계수(Pearson's r), 평균제곱근오차, Macro-F1을 보조 지표로 사용하였다. 원천 점수 분포 유지 조건에서는 각 모델에 적용된 음향 자질 조건이 시간 자질군과 통합 자질군보다 높은 성능을 보였다. GBC와 XGB에서는 A8이, FTT와 SAINT에서는 A88이 각 모델에 적용된 조건 가운데 가장 높은 성능을 나타냈으며, 전체 조건에서는 SAINT-A88이 가장 우수한 성능을 보였다. 통합 자질군은 일관된 성능 향상을 보이지 않았으며, 균형 점수 분포 조건에서는 자질군과 모델에 따라 성능 양상이 달라졌지만 SAINT-A88이 해당 조건에서도 가장 우수한 성능을 유지하였다. 이러한 결과는 본 연구에서 설정한 자질 조건과 예측 모델의 조합에 따라 유창성 자동 평가 성능이 달라질 수 있음을 시사한다.

핵심어: 외국어로서의 한국어, 유창성 자동 평가, 음향 자질, 예측 모델링

참고문헌

- 강석한, 안현기, 홍은실, 민병곤, 조수진, 이성준, 박현정(2017). 외국어로서의 한국어 발음과 유창성 연구. *이중언어학*, 67, 1-29.
- 김나미, 김영주(2017). L2 한국어 말하기 유창성 평가의 신뢰도 검증: 다국면 라쉬 모형 활용. *언어학 연구*, (45), 483-524.
- 이향원, 오재혁, 이정란, 최은지, 이선영, 김형민(2025). 한국어 학습자의 말하기 유창성 평가를 위한 자동 채점 자질 설정 연구. *새국어교육*, (143), 443-475.
- 유천연(2018). 한국어 학습자의 발화에서 나타난 유창성 특징 분석. *한국문법교육학회 학술발표논문집*, 2018(1), 89-99.
- 유천연, 원미진(2020). 한국어 학습자의 말하기 과제에 나타난 유창성 분석. *외국어로서의 한국어교육*, 56, 265-284.
- AIHub, 한국지능정보사회진흥원(2022a). 교육용 영어 모국어 사용자의 한국어 음성 데이터. <https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71469>
- AIHub, 한국지능정보사회진흥원(2022b). 교육용 유럽어 모국어 사용자의 한국어 음성 데이터. <https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71489>
- AIHub, 한국지능정보사회진흥원(2022c). 교육용 중·일어 모국어 사용자의 한국어 음성 데이터. <https://www.aihub.or.kr/>

* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원 부설 한국문화기술기획평가원의 2025년도 문화기술 연구개발사업으로 수행되었음(과제명: 외국인의 효율적인 한국어 말하기 학습을 위한 AI 활용 한국어 진단 시스템 개발, 과제번호: RS-2025-02413038, 기여도율: 100%).

[aihubdata/data/view.do?dataSetSn=71490](https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71490)

AIHub, 한국지능정보사회진흥원(2022d). 교육용 아시아어(중·일
어 제외) 사용자의 한국어 음성 데이터. [https://www.aihub.or.kr/
aihubdata/data/view.do?dataSetSn=71479](https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71479)