문학석사 학위논문

비원어민 한국어 유창성 평가 및 피드백을 위한 기초 연구

2024년 08월

서울대학교 대학원

언어학과 언어학 전공

김 미 경

비원어민 한국어 유창성 평가 및 피드백을 위한 기초 연구

지도 교수 이 호 영

이 논문을 문학석사 학위논문으로 제출함
2024년 08월

서울대학교 대학원
언어학과 언어학 전공
김 미 경

김미경의 문학석사 학위논문을 인준함
2024년 08월

위 원 장 정 민 화 (인)

부위원장 이 호 영 (인)

위 원 이 옥 주 (인)

초 록

이 연구는 말하기 평가 영역에서 핵심적인 평가 항목으로 여겨지는 발화 유창성을 비원어민 한국어 맥락에서 자동으로 평가하기 위한 토대를 다지기 위한 기초 연구이다.

그간 L2 영어를 대상으로 연구되어 온 다양한 유창성 관련 자질들 중 선행 연구에서 핵심적인 자질로 지목되었던 일부 자질들을 분석 대상으로 선정하였다. 이 자질들이 비원어민 한국어 유창성 평가에도 유효한지 살펴보기 위해 발화별로 각 자질의 적용 결과를 자동으로 추출할 수 있는 방법을 제안하고자 한다. 특히, 평가자에 의한 유창성 점수에 대하여 자동 추출된 유창성 자질 값들의 조합이 가지는 설명력을 조사하여 비원어민 한국어 유창성 자동 평가에 어떤 자질들이 얼마나 중요하게 활용될 수 있는지 살펴본다. 이는 비원어민 한국어 유창성 자동 평가와 더불어 피드백 제공 가능성을 탐구하는 것이다.

이 논문에서 제안하는 유창성 자질 값 자동 추출 방식에 대하여, 그 결과 값을 수동으로 음향 분석한 결과와 비교하여 신뢰도를 검증했다. 이러한 방식은 외국인 한국어 학습자에게 매우 유용한 유창성 피드백을 줄 수 있다고 판단된다.

이 연구에서는 AIHUB에서 다운로드 가능한 교육용 비원어민 한국어 음성 데이터를 분석 대상으로 선정하였고, 발화 속도(speech rate), 조음 속도(articulation rate), 발성시간 비율(phonation-time ratio), 분당 연속 발화 평균 길이(mean length of runs), 분당 휴지 평균 개수(mean number of silent pauses), 분당 휴지 평균 길이(mean length of silent pauses) 등 유창성 평가에서 가장 중요한 자질들에 대해 자동 추출 방식을 사용하여 결과 값을 추출했다. 대응표본 t-검정(paired t-test)과 Pearson 상관 분석(Pearson correlation analysis)을 통해 수동 음향 분석 방식과 비교하여 이 자동화 방안의 신뢰성을 입증하였다. 또한, 어떤 유창성 특징들이 비원어민 한국어 유창성 평가에서도 얼마나 유의미한 지표로 기능하는지 살펴보기 위해

진행된 두 가지 단계적 다중 선형회귀 분석 결과, 평가자의 유창성 평가 점수를 설명할 때 유창성 척도만 사용한 첫 번째 다중 회귀 모델의 설명력은 R2=.466에 그친 반면, 문단 읽기 발화 데이터라는 점을 고려하여 발음 평가 점수를 추가 독립변수로 설정한 두 번째 다중 회귀 모델의 설명력은 R2=.686로 상승하였다. 더 나아가, 단계적 회귀 모델이 선택한 자질과 그 자질이 회귀 모델에 미치는 영향력을 살펴보았다. 그 결과, 첫 번째 모델에서 발화 속도(Beta=0.759), 분당 휴지 평균 길이(Beta=0.232) 순서로 유창성 평가 점수를 예측하는 데 유의미한 척도로 밝혀졌으며, 이는 발화 속도와 평균 휴지 길이 자질이 유창성에 대한 인간의 청지각과 높은 상관관계를 보임을 밝혔던 선행 연구와 유사한 결과였다. 이 연구에 사용된 음성 데이터가 문단 읽기 태스크였던 점을 감안하여 발음 평가 점수를 독립변수로 포함시킨 두 번째 모델에서는 발음 점수 (Beta=0.565), 발화 속도(Beta=0.408), 분당 평균 휴지 길이(Beta=0.235) 순으로 회귀 모델에 영향을 미친다는 것을 확인할 수 있었다.

주요어 : 발화 유창성, 비원어민 한국어, 유창성 자질, 유창성 자동 측정, 자동 평가 신뢰성 검토, 유창성 피드백
학 번 : 2019-29853

목 차

표 목차

그림 목차

제 1 장. 서 론

비원어민 말하기 교육 및 평가 분야에서 말하기 유창성(speaking fluency)은 학습자의 전반적인 말하기 능력을 측정하는 데 핵심적인 역할을 한다(Mao et al., 2019). 또한, 유창성은 복잡성(complexity) 및 정확성(accuracy)과 더불어 학습자의 외국어 또는 제2언어(L2) 숙련도를 반영하는 핵심 요소로 간주된다(Chambers, 1997; Housen et al., 2012; Lennon, 1990; Towell, 2012). 따라서 비원어민 말하기 유창성의 본질과 특성을 밝혀 정의하고자 하는 음성학 분야의 연구가 지속적으로 이루어졌다. 특히 유창성에 대한 인간의 청지각 인상을 포착하고 학습자의 말하기 교육 및 평가(teaching & testing)에 평가 지표로 존재하는 유창성을 양적 분석이 가능한 객관적 지표로 설정하고자 하는 연구들이 있었다(Hilton, 2014; Segalowitz, 2010). 이러한 시도들은 주로 특정 유창성 자질들을 측정하여 나온 결과 값과 전문 평가자의 유창성 점수, 더 나아가서는 숙련도 점수와의 상관관계를 살펴봄으로써 어떤 자질들이나 자질들의 집합이 평가 점수를 가장 잘 반영하는지를 보고했다. 하지만 연구 목적과 설정한 가설에 따라 유창성 측면에서 인간의 청지각을 가장 잘 반영하는 자질에 대한 연구자들의 의견은 여전히 상이하다. 이에 덧붙여, 유창성 자질에 대한 대부분의 연구들은 외국어로서의 영어 발화에 초점을 맞추고 있다. 따라서 L2 영어에 초점을 두고 연구한 기존 유창성 자질들이 비원어민 한국어 발화를 평가하는 데에도 유의미한 지표로 기능하는지 검증해야 한다. Baker-Smemoe et al.(2014)는 유창성 자질들이 평가 대상이 되는 L2마다 다를 수 있다고 주장했다. 즉, 각 언어의 특성에 따라 유창성 측정에 필요한 자질들이 달라질 수 있으며, 범언어(cross-linguistic) 자질과 개별 언어(language-specific) 자질의 구분을 바탕으로 비원어민 한국어 발화에 적합한 유창성 자질을 설정해야 한다.

한편, 그간 음성학 분야의 연구들은 외국어 학습자의 유창성 자질을

측정할 때 수동 음향 분석을 토대로 측정해왔다. 이러한 수동 측정 방식은 다양한 학습자들의 말소리를 담고 있는 대용량 음성 데이터를 대상으로 삼기에 적절하지 않으므로, 실험 대상이 주로 소규모 발화 데이터에 국한된다는 한계가 있다. 특히 언어 시험 환경과 같은 대규모 학습자 음성 데이터를 다루어야 하는 경우, 평가자가 개별적으로 각 유창성 자질별 평가를 진행한다는 것은 불가능에 가깝다. 또한 인간의 지각에 의존하는 수동 음향 분석과 평가는 평가자 개인의 주관과 경험에 따른 편향된 결과를 낳을 수 있다는 가능성을 배제할 수 없을 뿐만 아니라 비용과 시간 소요 관점에서도 비효율적이다. 따라서 객관성과 일관성을 갖춘 고속 분석 도구의 개발이 요구된다. 이러한 도구의 개발은 연구와 평가 양측에서의 효율성과 확장성 향상에 도움을 줄 수 있다. 응용음성학 분야에서는 이를 충족시키기 위해 음성 공학과 더불어 고도화된 최신 머신러닝 기법 및 자동음성인식 기술(Automatic Speech Recognition systems)을 활용하여 유창성 자질 추출부터 평가에 이르는 일련의 과정을 자동화하는 데 주목하고 있다.

최근 주요 유창성 자질을 활용한 말하기 유창성 자동 채점은 말하기 자동평가 시스템을 구성하는 하나의 모듈로 기능하고 있다(Ginther et al., 2010). Versant tests(Pearson, 2009)와 TOEFL iBT Practice test(Zechner et al., 2007) 등은 유창성 자동 채점 모듈을 도입하여 자동 평가를 꾀한 대표적 예시로 볼 수 있다. 특히 TOEFL iBT practice test의 경우, 원어민과 비원어민 발화가 조합된 데이터셋으로 훈련된 자동음성인식기를 기반으로 유창성 관련 특징들을 생성하여 자동 채점을 실시하고 있다. 또한 말하기 유창성 자동 채점은 L2 학습자들을 위한 컴퓨터 보조 발음 훈련(Computer-Assisted Pronunciation Training; CAPT)과 컴퓨터 보조 언어 학습(Computer-Assisted Language Learning; CALL) 분야에서 필수적으로 고려되는 핵심 연구 목표 중 하나로 자리매김하고 있으며, 그 중요성이 부각되고 있다(Detey et al., 2020, Eskenazi, 2009). 그러나 L2 한국어 말하기 자동 평가 연구는 주로 발음 영역에 치중되어 있는 양상을 보인다(Hong et al., 2014; Ryu et

al., 2016; 양승희·정민화, 2017). 또한 선행 연구들은 공통적으로 점수 예측 성능을 보고하는 데에 치우쳐 있어서 사용된 자질들이 얼마나 정확하게 추출되었는지, 그리고 수동 음향 분석과 비교하여 어느 정도의 정확성과 신뢰도를 갖추고 있는지 보고하거나 검증하고 있지 않다.

이 연구는 위와 같은 연구 현황에 근거하여 그간 L2 영어 중심으로 연구되어 온 유창성 자질들을 비원어민 한국어 맥락에서 자동으로 추출하고 수동으로 음향 분석한 결과와 비교하여 비원어민의 한국어 유창성 평가에서 각 자질 별 신뢰성 및 정확성을 검증하고자 한다. 여기에는 유창성 자질 별로 수치화 및 표준화를 통해 비원어민 한국어 학습자를 평가하는 것에 대한 가능성을 검증하는 것을 포함한다. 뿐만 아니라 본 연구는 기존 음성 공학 분야에서 발전되었으며 총점만을 제시하는 유창성 자동 평가 시스템과 달리 자질 별로 결과 값을 추출, 제시함으로써 향후 학습자에게 구체적인 피드백을 제공하는 학습도구로 기능하며, 개인의 필요와 수준에 맞는 학습과 교육이 이루어질 수 있다는 이점을 갖는다. 또한 이러한 방법론은 기존 AI 기반 자동평가 모델과는 달리, 설명 가능한 인공지능(Explainable AI; XAL)으로서 학습자, 교육자, 연구자 모두에게 결과 값의 근거를 해석 가능한 형태로 투명하게 제시할 수 있다는 점에서 그 결과를 이해 가능하고 보다 신뢰할 수 있다는 의의를 가진다.

이 연구는 여기에서 추가적으로 비원어민 한국어 발화에 대한 유창성 평가에서 전문가들에 의해 사전 평가된 유창성 점수에 대해 가장 높은 설명력을 가지는 유창성 자질 혹은 자질 집합을 살펴보고자 한다. 이를 통해 비원어민 한국어 학습자의 말하기 유창성 평가에 있어서 주요 언어적 특성과 영향을 파악하고, 비원어민 한국어를 위한 또 다른 자질의 설정 필요성을 살펴볼 것이다.

이 논문에서 탐구하고자 하는 연구 질문은 다음과 같다:

1. 제안하는 자동 유창성 측정이 수동 측정과 비교했을 때 충분한 신뢰성과 정확성을 담보하는가?
2. 기존 L2 영어를 중심으로 연구되어 온 주요 유창성 자질들이 비원어민 한국어 발화에 대한 전문가의 청지각에 대해서도 유의미한

설명력을 갖추고 있고, 학습 및 평가에 유의미한 지표로 기능하는가?

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 주제인 유창성의 정의에 대해 살펴보고, 유창성에 대한 인간의 청지각을 포착하기 위한 다양한 자질들과 그 측정의 자동화를 논한 선행 연구들을 검토한다. 또한 비원어민 한국어 맥락에서 발화 유창성을 논의한 선행 연구를 살핀다. 3장에서는 본 논문에서 사용하는 데이터와 유창성 자질들을 기술한 후, 해당 자질들의 자동 추출 방식을 제시한다. 더불어 통계 분석을 실시한다. 4장에서는 실험 결과를 요약하고 분석한다. 마지막으로 5장에서는 실험 결과의 의의를 서술하고 이 연구가 가지는 한계 및 후속 연구의 방향성에 대해 서술하며, 결론을 제시한다.

제 2 장. 선행 연구

제 1 절. 유창성의 정의

유창성은 언어 산출(language production)의 여러 측면을 복합적으로 포괄하고 있다. 즉 화자의 L1 배경(De Jong et al., 2015; Derwing et al., 2009), 발화 습관, 심리 기제 등과 같은 개별적 원인(Kormos, 1999), 또는 자유 발화와 낭독 발화와 같은 과제 유형(Foster & Skehan, 1996) 등의 화용론적인 요인들이 복잡한 상호작용을 한다. 따라서 유창성은 단일한 구성체로 파악할 수 없다. 유창성의 본질을 파악하고 정량적으로 측정할 방법을 모색하고자 하는 연구 또한 마찬가지의 이유로 여전히 지속되고 있다.

Fillmore(1979)는 유창성 정의를 시도한 초창기 연구에 해당한다. 그는 유창성을 다음과 같이 (ㄱ) 빈번한 끊김이 없고 적당한 길이로 발화하는 능력, (ㄴ) 논리적으로 조직된 문장을 표현하고 의미론적으로 조밀한 구조를 갖추는 능력, (ㄷ) 다양한 발화 상황과 맥락을 고려하는 능력, 그리고 (ㄹ) 언어 사용에서의 창의성과 생산성의 네 가지 능력으로 정의했다. 그는 전술한 정의를 바탕으로 하여 유창성의 차이가 발생하는 원인을 규명하고자 했다. 또한 형태소, 어휘, 관용구(idiomatic expressions) 및 기타 형식 언어 표현(formulaic language expressions)의 사용 정도의 차이는 절대적 학습량에 기인한다고 주장했다. 이와 동시에 이러한 표현들이 화자들의 머릿속 어휘 사전(mental lexicon)에 존재하더라도 다양한 맥락 속에서 해당 표현들에 무의식적이고 자동적으로 접근하게 되며, 통사적 장치를 활용하는 개개인의 속도로부터 유창성 차이가 발생할 수 있음을 강조했다.

Lennon(1990)은 유창성을 광의와 협의로 구분하여 정의했다. Lennon(1990)에 따르면, 광의의 유창성은 일반적으로 어떤 언어를

“유창하게” 구사한다고 말할 때와 같이, 전반적인 숙련도(proficiency) 수준을 가리키는 포괄적인 의미를 가리킨다. 반면 협의의 유창성은 특히 말하기 평가 측면에서 말하기 능력을 구성하는 한 요소이며, 정확성, 관용성, 관련성, 적절성, 발음, 어휘 범위 등과 같은 요소와 병치되는 단일 척도이다. 또한 그는 원어민의 말하기 속도에 근접하는 숙련도를 표현하는 용어로도 유창성을 정의하고 있다.

그의 연구는 유창성을 다른 구성 요소들과는 달리 언어 수행(linguistic performance)의 영역임을 명확히 했다는 점에서 주목할 만하다. 즉, 어휘나 문법 정확성 등은 언어 능력(linguistic competence)으로서 개인에게 내재된 언어 지식에 따르는 반면, 유창성은 특정 상황에서 실제로 언어를 사용하는 언어 수행을 평가할 수 있는 요소로 정의하였다. 정리하자면, 그의 연구는 유창성에 대한 정의를 보다 구체화하기 시작했고, 협의의 유창성을 구성하는 요소들을 제시함으로써 유창성을 다면적인 개념으로 정의했다는 의의를 가진다. 이 연구에서도 협의의 유창성을 논의의 대상으로 한다.

유창성 평가 및 교육에 관한 대부분의 후속 연구들은 Lennon(1990)이 제시한 협의의 유창성 개념을 기반으로 진행되었고, 더욱 세분화된 유창성 개념이 제시되기 시작했다. Skehan(2003)은 유창성을 속도 유창성(speed fluency), 중단 유창성(breakdown fluency), 수정 유창성(repair fluency)의 세 가지 범주로 구분하였다. 현재에 이르기까지 제안되었던 유창성 자질들은 대부분 이 세 범주 내에 속한다(Tavakoli & Skehan, 2005). 그리고 이러한 연구 틀에 따라 각 연구 목적에 맞게 선택된 자질들이 신뢰할 수 있는 유창성 지표로 기능하는가를 평가하는 방식의 연구가 주류를 이룬다고 할 수 있다. 이 논문 역시 이러한 선행 시도들을 L2 한국어 맥락에서 재현 및 탐구해 보고자 하는 의도를 지닌다.

말하기 유창성의 다차원적 특성을 이해하기 위한 또 다른 대표적 연구에는 Segalowitz(2010)이 있다. 그가 새롭게 제안한 삼원 유창성 모형에 따르면, 유창성은 인지 유창성, 발화 유창성, 그리고 지각

유창성으로 구성된다. 인지 유창성이란 언어 수행에 관여하는 기본적인 정신적 과정이며, 구체적으로는 이러한 과정을 효율적으로 동원하고 조직하는 능력을 지칭한다. 발화 유창성은 실제 발화에서 관찰되는 시간적 특성(temporal features; 발화나 조음의 속도, 휴지의 길이 및 비율 등)으로 측정 가능한 유동성으로 정의하고 있다(Segalowitz, 2016). 최근 연구자들은 시간 특성(temporal features)은 주로 Skehan(2003)의 속도 유창성과 중단 유창성을 묶어 지칭하는 용어로 사용하는 경향이 있다. 이 연구에서도 분석 대상이 되는 유창성 자질로 시간 특성에 초점을 맞출 것이다. 마지막으로 지각 유창성은 인지 및 발화 유창성과는 달리 청자의 관점에서 정의된다는 차이가 있다. 즉, 청자의 감각기관을 통해 인식한 다음 일어나는 화자의 발화에 대한 청자의 주관적 판단이다.

이러한 세 가지 측면 중 발화 유창성은 관측이 가능하다는 점으로 인해 양적 연구에 적합하다(Foster, 2020; Tavakoli et al., 2020). 그간 연구되어 온 넓은 범위의 유창성 자질들은 발화 유창성을 측정하기 위한 일종의 도구라고 할 수 있다. 이 도구를 바탕으로 유창성 점수와의 상관관계를 살펴보는 연구들은 청취자의 주관적인 지각 유창성을 측정 가능한 방식으로 정의하여 표준화 방안을 강구하기 위한 노력으로 해석할 수 있다. 이 연구에서는 이러한 노력을 비원어민 한국어 맥락에서 재현함과 동시에 발화 유창성의 다면적 성격을 밝혀 객관적으로 측정 및 평가하는 데에 일조하고자 한다.

제 2 절. 인간의 청지각과 유창성 측정

전문가의 유창성 평가 점수나 포괄적인 말하기 숙련도 점수로 표상되는 인간의 청지각 능력과 다양한 유창성 자질들 간의 상관관계를 밝힘으로써 유창성을 정의하고자 하는 논의는 오랜 기간 지속되어 왔다. 그러나 2.1.절에서 확인한 유창성의 다면적 특성은 유창성에 대한

인간의 청지각적 상관 자질을 찾기 어렵게 만든다. 이를 극복하려면 다양한 범위의 유창성 자질을 도입하여 전문가의 유창성 평가 점수와 상관관계가 가장 높은 자질을 밝혀야 하고, 선행 연구들도 그러한 방식으로 진행되었다. 그럼에도 불구하고 학자들마다 유창성을 가장 잘 측정해 주는 자질에 대한 견해는 여전히 상이하다.

Kormos and Dénes (2004)는 16명의 헝가리인 영어 학습자들의 발화를 대상으로 유창한 발화와 그렇지 않은 발화를 구분하고 평가자들의 유창성 지각에 영향을 미치는 유창성 변수를 식별하는 데 초점을 두었다. 그 결과, 해당 연구는 발화 속도(speech rate), 분당 연속 발화 평균 길이(mean length of runs), 발성 시간 비율(phonation-time ratio), 그리고 분당 강세단어 수(pace)가 유창성 점수의 강력한 예측 변수임을 보여주었다. 이는 유창성의 시간적 특성이 중요하다는 사실을 내포한다. 또한 분당 강세단어 수(pace) 자질이 중요한 이유는 영어가 강세 박자 언어(stress-timed language)이기 때문이다. 반면 공백 휴지와 비공백 휴지[①], 기타 말더듬 현상은 유창성 지각에 유의미한 영향을 미치지 않는 것으로 보고했다.

한편, 말하기 숙련도 점수라는 큰 맥락에서 유창성 자질들이 미치는 영향력을 살핀 연구들이 많이 있다. 대표적으로 Iwashita et al.(2008)은 TOEFL iBT 시험에서 200개의 발화 샘플을 대상으로, 다섯 레벨로 구별된 말하기 숙련도 평가에서 유창성 자질 집합을 포함한 다양한 범주의 자질들이 기여하는 정도를 조사하는 대규모 연구를 실시했다. 말하기 숙련도에 미치는 유창성 자질들은 발화 속도(speech rate), 분당 공백 휴지 횟수(number of unfilled pauses), 그리고 총 휴지 시간(total pausing time)이었으며, 이는 높은 수준의 학습자들이 더 빠르게 발화하고, 더 적게 멈추며, 휴지 구간이 적다는 것을 보여준다.

Ginther et al.(2010)은 유창성 관련 시간 자질들과 OEPT(the Oral English Proficiency Test)의 종합 점수 간의 상관관계를 탐구했다.

① unfilled pauses를 공백 휴지, filled pauses를 비공백 휴지로 번역하였으며, 이하 내용에서는 편의를 위해 unfilled pauses는 휴지라는 표현으로 사용한다.

150명의 참가자들의 발화를 대상으로 OEPT 점수와 가장 강한 상관관계를 보인 특성은 발화 속도(speech rate), 조음 속도(articulation rate), 분당 연속 발화 평균 길이(mean length of run)였으며, 휴지 시간(silent pause time)[②]의 경우 상대적으로 약한 음의 상관관계를 보임을 밝혔다.

Baker-Smemoe et al.(2014)는 다양한 L2를 대상으로 발화 유창성 자질들의 범언어적 유효성을 탐구하고자 했다. 이 연구에서는 86명의 영어 원어민 학습자가 L2로 프랑스어, 독일어, 일본어, 아랍어 또는 러시아어를 발화한 126개의 OPIs(ACTEL Oral Proficiency Interviews) 발화 데이터를 사용했으며, 언어별 가능성을 살폈다는 점에서 주목할 만한 결과를 보여주었다. 그들은 특정 유창성 측정 자질을 사용하여 말하기 능력을 추정하는 것이 실용적일 수 있으나, 말하기 수준이 높은 학습자들을 대상으로 유창성 평가를 진행할 시 세분화된 하위 자질들을 범언어적으로 적용하는 것에 주의가 필요함을 강조했다. 이런 이유로 비원어민 한국어 발화의 유창성 평가에는 어떤 자질들이 유의미하게 작용하는지 확인해 볼 필요가 있다.

비교적 최근 연구인 Tavakoli et al.(2020)에서는 영국문화원의 Aptis Speaking Test에서 서로 다른 네 가지 숙련도 수준을 가진 것으로 평가된 32명의 화자의 발화를 대상으로 이들의 말하기 숙련도 수준을 구분할 수 있는 유창성 측정 자질이 무엇인지 파악하기 위한 실험을 진행했다. 연구자들은 기존에 Skehan(2003)이 제시한 세 범주에 복합 유창성(composite fluency) 범주를 추가하여 유창성 자질을 미시적으로 분석하였다. 그리고 그 결과 속도 유창성(speed fluency)과 복합 유창성(composite fluency) 범주에 속하는 자질 목록들이 하급에서 상급 중간 수준(A2-B2) 학습자들의 유창성 수준을 일관성 있게 구별하고, 중단 유창성(breakdown fluency)의 경우 가장 낮은 수준(A2)을 나머지와 구별하는 정도에 그쳤음을 확인하였다.

[②] silent pause는 unfilled pause와 같은 자질을 의미한다.

Yan, et al.(2021)은 말하기 숙련도와 유창성 자질들 사이의 관계에 유의하여 유창성 자질들을 종합적 및 세부적 범주로 나누어 정의하고 오디오 기반 측정과 수동 전사 결과물 기반 측정이라는 두 가지 방법론을 바탕으로 삼아 실험을 진행했다. 또한 실험에 사용한 발화는 Aptis 발화 코퍼스에서 추출한, 영어를 제2언어로 학습한 사람들에 의해 발화된 발화 샘플 125개였다. 이 연구 결과는 일반적으로 사용되는 종합적 유창성 변수인 발화 속도(speech rate), 조음 속도(articulation rate) 및 분당 휴지 횟수(number of silent pauses)가 말하기 숙련도 점수와 강한 상관 관계를 가진다는 사실을 보여주었다. 또한, 세부적 유창성 자질로 분류된 특징(분당 연속 발화 평균 길이(mean length of runs), 끊기 휴지 비율(juncture pause rate), 수정 성공률(repair success rate))이 발화 유창성과 더 강한 상관 관계를 가진다는 사실을 보여주었다. 다르게 말하자면, 화자의 말하기 수준이 높을수록 휴지의 위치가 자연스러우며 말더듬이 발생하여도 효과적으로 수정하는 능력을 보여주었다.

이상의 논의 내용을 L2, 음성 표본 수, 그리고 사용된 유창성 자질들에 따라 <표 2-1>과 같이 정리하였다.

위에서 검토한 연구들이 가지는 한계는 다음과 같다. 먼저, 이들은 공통적으로 수동 음향 분석을 통해 진행되었기 때문에 충분하지 않은 수의 발화 샘플을 연구 대상으로 삼고 있다. 발화 샘플의 수가 충분하지 않다면 다양한 맥락에서의 유창성의 복합성과 변동성을 포착하기 어렵다. 뿐만 아니라 이러한 수동 분석은 대용량 데이터를 대상으로 하기에는 불가능에 가까우므로 자동화 방안을 모색할 필요가 있다. 다음으로, 대부분의 선행 연구들은 주로 비원어민의 영어 발화를 분석 대상으로 하고 있다. 이들은 범언어적(cross-linguistic)인 유창성에 대한 주목할 만한 통찰을 제시한다고 볼 수 있으나, 개별 언어(language-specific) 측면에서의 비원어민 한국어 유창성 평가를 위해서는 언어 고유의 특성을 고려하여 연구되어 온 자질들에 대한 검증을 거쳐야 할 것이다. 마지막으로, 선행 연구마다 인간의 청지각에 높은 상관관계를 보이는

자질 목록과 그들의 효과에 대한 논의가 통일되어 있지 않다.

이러한 점들로 미루어 보아, 비원어민 한국어의 유창성을 평가하기 위해선 기존의 유창성 지표가 비원어민 한국어 유창성에 대한 전문가의 평가를 예측하는 데에 어느 정도의 신뢰성과 타당성을 가지는지 필수적으로 검증해야 한다. 이러한 검증 절차는 비원어민 한국어 유창성 개념에 대한 이해를 도모하고 비원어민 한국어 학습자의 유창성을 평가하는 객관적 기준을 확립하는 데 도움을 줄 수 있을 것이며, 학습자들의 한국어 유창성을 향상시키기 위한 지침으로 삼을 수 있을 것이다.

<표 2-1> 선행 연구에서 사용된 유창성 자질 목록

선행 연구	음성 표본 수	L2	유창성 자질
Kormos and Dénes (2004)	16	영어	**발화 속도**, 조음 속도, **발성시간 비율**, **분당 연속 발화 평균 길이,** 분당 휴지 횟수, 분당 휴지 평균 길이, 분당 비공백 휴지 횟수, 분당 비유창성 특징 횟수, **Pace**[③], Space[④],
Iwashita et al.(2008)	200	영어	분당 비공백/**공백 휴지 횟수**, 분당 말수정 행위 횟수, **전체 휴지 길이**, **발화 속도**, 분당 연속 발화 평균 길이

[③] Pace는 분당 강세 단어 횟수를 의미하는 용어로 사용되었다.
[④] Space는 전체 단어 개수 중 강세 단어의 개수의 비율을 의미하는 용어이다.

Ginther et al.(2010)	150	영어	응답 소요 시간, 음성발화 시간 비율, 음절/휴지/비공백 휴지 수, **발화 속도, 조음 속도**, **분당 연속 발화당 평균 음절 수**, 총 휴지 시간, 분당 휴지 시간, 휴지 비율, 비공백 휴지 시간, 분당 평균 비공백 휴지 시간, 비공백 휴지 비율
Baker-Smemoe et al.(2014)	126	불어, 독일어, 일본어, 아랍어, 노어	**발화 속도**, 휴지 길이와 횟수, 분당 연속 발화 평균 길이, 망설임 횟수, 시작 오류 횟수
Tavakoli et al.(2020)	32	영어	**발화 속도, 조음 속도,** **분당 연속 발화 평균 길이,** 발성시간 비율, 모든 종류의 휴지(공백/비공백)의 평균 길이 및 횟수, **분당 공백 휴지 평균 횟수 및 길이,** 문장 내 위치(중간/끝)에 따른 휴지의 평균 길이 및 횟수, 문장 내 위치(중간/끝)에 따른 비공백 휴지의 평균 길이 및 횟수, 비공백 휴지의 횟수, 수정 행위/시작 오류/ 문장 재구성 행위/반복 횟수
Yan et al.(2021)	125	영어	**발화 속도,** **조음 속도,** **분당 연속 발화 평균 길이,** **휴지 횟수,** **끊기 휴지 비율,** **수정 성공률**

참고. 굵은 글씨체는 각 연구에서 보고된 가장 예측력 있는 지표를 말한다.

제 3 절. 유창성 측정과 평가의 자동화

이 절에서는 유창성 관련 자질들을 자동으로 추출하는 방법을 제안한 연구들을 검토한다. 유창성 측정 자동화를 논의하는 연구의 대부분은 유창성 자동 평가나 말하기 자동 평가라는 맥락에서 진행되었으며, 궁극적으로는 전문가의 수동 평가 대체제로서의 초석을 다지기 위한 노력의 일환이었다는 사실을 이해할 필요가 있다. 따라서 유창성 자동 평가 관련 연구 현황도 함께 살펴보아야 한다.

유창성 자동 평가 방식은 수동 음향 분석을 토대로 하는 응용 음성학 분야의 연구에서 최신 기술과의 융합에 이르기까지 지속적인 발전을 이루어 왔다. 수동 음향 분석은 2.2.절에서 언급되었던 연구들이 주로 채택했던 방법론에 해당한다. 그러나 수동 음향 분석은 연구자나 전문가가 직접 개입한다는 점에서 시간과 비용의 한계가 분명히 존재한다. 수동 음향 분석은 평가자의 풍부한 경험을 바탕으로 한 청지각 인상과 직관을 통해 다양한 맥락을 고려한 유동적 평가가 가능하다는 이점이 있다. 하지만 평가자 개개인의 주관이 개입되거나 평가자의 피로감으로 인해 일관성이 저해될 여지가 있으며, 대량 분석이나 평가에는 평가자 내/간 일치도 문제에 직면한다.

이렇듯 수동 유창성 분석 및 평가의 방법론적 한계를 극복할 필요성과 맞물려, 최근 빅데이터 구축 기술과 자동 음성 인식과 같은 컴퓨터 알고리즘의 발전으로 말하기 평가를 자동화하고자 하는 움직임이 활발하게 일어나고 있다. 특히 유창성 측정과 평가의 자동화에 대한 연구는 전통적인 특성 기반 접근법과 종단간 접근법(E2E: end-to-end)이라는 두 가지 주요 접근 유형으로 구분할 수 있다. 전자는 자동 음성 인식(Automatic Speech Recognition; ASR)과 강제 정렬 시스템으로부터 얻은 전사(transcriptions)의 타임스탬프를 이용해 간접적으로 특징을 계산해 내거나, 원시 음성 신호로부터 직접 유창성 특성들을 추출하고 계산한다. 그런 다음, 유창성 점수를 예측하기 위해 훈련된 기계 학습 모델이 사용되며, 그 성능은 인간 평가와의 상관 관계

분석을 통해 최종적으로 평가된다. 반면에 후자의 종단간 접근법에서는 음성 신호라는 원시 데이터(입력)만을 사용해 유창성 점수(출력)를 결과값으로 도출하는 하나의 통합된 계산 모델을 제공하려는 목표를 설정한다.

Cucchiarini et al.(2000)은 자동 음성 인식(ASR) 시스템을 사용한 특성 기반 접근법을 취한 연구에 해당한다. 이 연구에서는 강제 정렬로 얻은 전사물에 대해, 유창성 관련 변수를 자동으로 계산하고 얻은 유창성의 양적 측정 값이 전문가의 유창성 평가를 예측하는 데 잠정적으로 사용될 수 있는지 조사했다. 구체적으로, 먼저 연속 음성 인식 시스템의 구성 모듈인 39 Hidden Markov Models 기반 음향 모델(acoustic model), 언어 모델(language model), 그리고 어휘 사전(lexicon)을 80명의 원어민과 비원어민이 낭독한 네덜란드어 발화 데이터로 훈련시켰다. 이후 강제 정렬 알고리즘을 사용하여 음성과 그 음소 단위 인식 결과물을 매핑함으로써 음소마다 시작과 끝 경계와 같은 시간 정보를 추적할 수 있도록 전사하고, 발화 및 비발화 부분의 위치 정보가 출력되도록 설계했다. 그리고 이러한 전사 정보를 기반으로 선택된 시간 자질들(temporal features)에 대한 양적 측정 값을 자동으로 계산하는 방식을 제안하였다. 그 결과 발화 속도, 발성시간 비율, 조음 속도, 휴지 수, 휴지부 총 길이, 분당 연속 발화 평균 길이의 자동 측정 값이 유창성 평가와 높은 상관관계를 보이는 것으로 보고되었다. 특히 발화 속도가 가장 강력한 예측 변수로 나타났으며, 평균 휴지 길이의 경우 유의미한 상관 관계를 보이지 않는다는 결론이 도출되었다. 또한 단계적 다중 회귀 분석을 통해 발화 속도와 휴지 수를 변수로 포함한 모델이 유창성 평가 점수에 대해 가장 좋은 설명력(R=0.94)을 가지는 것으로 보고했다. 이와 같은 결과를 바탕으로, 해당 연구에서는 자동 음성 인식 기술을 통해 계산된 유창성의 시간적 특성에 대한 측정 값들이 낭독 발화의 유창성에 대한 객관적 평가 도구를 개발하는 데 기여할 수 있음을 보였다.

이와 같은 연구는 자동 음성 인식 기술을 사용하여 유창성 평가의

자동화 잠재성을 살펴보는 연구의 시초가 되었다는 점에서 큰 의의를 지닌다. 그러나 강제 정렬의 정확도에 대한 검증이 소량의 음성 샘플에 대해서만 이루어졌다고 밝혔으며, 대부분의 결과물이 높은 정확도를 보였으나 음소의 시작과 끝 경계 즉, 위치 정보(타임스탬프)가 사람이 분석하는 것만큼의 정확성을 보이지는 않는다는 검증 결과를 보고하였다. 위치 정보를 직접적으로 사용하지 않는다고 할지라도 이러한 추상적인 분석은 신뢰도를 낮출 수 있으며, 분석의 완전성에 대한 의문이 발생할 수 있다. 대부분의 유창성 자질들에 대한 계산은 모두 휴지 인식과 휴지 구간의 길이가 핵심 기준으로 이루어지기 때문이다. 따라서 이러한 연구는 유창성 평가 측면에서 성능이 보장된다고 할지라도 학습자에게 이해가능한 구체적 피드백을 주기에는 적절하지 않다는 한계를 가진다.

Deshmukh, et al.(2009)은 IBM사에서 제작했고 개인화된 음성 언어 평가 및 학습 관리 서비스를 제공하는 일종의 컴퓨터 보조 언어 학습(CALL) 시스템인 Sensi(Chandel et al., 2007)에 영어 유창성 평가 모듈을 통합하고자 하는 목적으로 진행된 연구이다. 이 연구에서는 시간 특성(temporal features)을 설정하여 유창성 측정 값을 도출하는 연구들과 달리 주로 비유창성을 정의하는 자질들, 즉 수정 유창성(repair fluency) 자질들의 자동 추출 방식을 고민했다는 점이 특징적이다. 구체적으로, 자동 음성인식기의 결과물에 수동 전사가 더해진 전사물을 대상으로 자연언어처리(Natural Language Processing; NLP) 기술을 접목하여 어휘 자질(lexical features)을 분석 및 추출하고 원시 음성 데이터로부터 직접 얻은 운율 자질(prosodic features)을 결합했다. 그러나 어휘 자질을 분석하기 위해 수동 태깅 절차를 요구하므로 완전히 자동화된 평가 방식이라고 볼 수 없다. 연구진들 역시 이를 극복하기 위해 자동 음성 인식기의 높은 정확도가 확보되어야 함을 강조하였다. 한편, 운율 자질은 제1포먼트와 제2포먼트를 추적하는 알고리즘을 설계하여 측정하였다. 발화자가 다음으로 생각을 형성하거나 생각을 전달할 단어를 선택하는 데 지연이 발생할 때, 휴지(unfilled pauses) 또는 비공백 휴지(filled pauses;

‘음’, ‘아’ 등의 간투사) 현상이 빈번히 나타난다. 그리고 이 구간에서 조음 기관의 움직임이 발생하지 않는다는 가정을 기반으로 알고리즘이 설계되었다. 그 결과, 이들의 연구는 어휘 자질(R=0.598)이 운율 자질(R=0.546)보다 인간의 유창성 평가 점수와 상관관계가 더 높다고 보고했다. 특히 고유 단어(unique words)의 사용이 풍부할수록 유창성 점수가 높음을 보였다.

상술한 두 연구는 자동 음성인식 기술과 강제 정렬 알고리즘을 접목한 방법으로 유창성 자질들의 추출을 (반)자동화하고, 전문가 평가 점수와의 상관분석을 통해 높은 상관관계가 있음을 증명하며 자동 평가의 가능성을 살핀 연구이다. 그러나 자동 음성인식 기술을 활용하는 방법은 사전 훈련된 개별 언어로만 제한되고, L2 학습자의 말하기 숙련도가 낮을수록 인식기의 성능이 급격히 저하되기 때문에 신뢰성이 저하될 수 있다는 문제가 발생한다.

이러한 배경을 바탕에 두고서, Fontan, et al.(2018)은 컴퓨터 보조 발음 훈련(CAPT) 소프트웨어에 통합 가능한 유창성 평가 도구를 개발하기 위한 기초 연구를 수행하였다. 이 연구는 총 48분 분량(252개 문장)의 일본인 학습자 L2 프랑스어 낭독 발화를 대상으로 자동 음성 인식 시스템을 사용하지 않는 유창성 평가 자동화 방안을 논의했다. 구체적인 방법은 다음과 같았다. 원시 음성 신호(low-level signal)에 대해 순방향-역방향 발산 분할(Forward-Backward Divergence Segmentation; FBDS) 알고리즘을 사용하여 음향적으로 동질한 단위로 분절음과 휴지부 등을 분할했다. 그리고 이러한 방식을 통해 총 여섯 가지 유창성 자질에 대해 분석하였다. 먼저, 주어진 발화 시간 대비 FBDS에 의해 생성된 분절음의 수가 클수록 발화 속도(speech rate)가 빠르다고 보았다. 또한 새로운 유창성 자질들을 도입했다. FBDS 알고리즘에 의해 발견된 분절음의 길이의 표준편차가 클수록 긴 휴지부나 망설임이 많은 것으로 가정한 유창성 자질을 발화 속도의 규칙성(regularity of speech rate)이라 칭하고 그 결과를 측정했다. 유사한 방식으로 휴지의 분당 평균 길이와 그 표준편차, 발화

비율(percentage of speech)을 측정했다. 마지막으로 화자의 발화 계획이 원활하게 이루어지지 않는 유창성이 낮은 화자일수록 제1포먼트의 급작스러운 변동이 빈번히 발생한다고 가정한 다음, 포먼트 추적 알고리즘을 사용하여 발화 변동성(speech fluidity)을 측정하였다. 이 자질들의 결과 값을 사용하여 단계별 다중 선형 회귀 분석을 수행했다. 그 결과 다중 회귀 모델의 설명력은 'R=0.82' 였으며, 발화 속도(speech rate)의 예측력이 가장 높았으며, 발화 속도의 규칙성 > 발화 비율 > 발화 변동성 순의 기여도를 보였다. 이처럼 이 연구는 이전까지의 연구들에서 제안되지 않은 새로운 자질들을 도입했다는 점에서 주목할 만하다. 또한 수동 전사나 음성 데이터를 활용한 훈련 절차를 요구하는 음성인식기의 사용 없이 저수준(low-level) 음향 신호에서 직접적으로 특성 분석을 진행했다. 즉, 목표하는 L2 언어에 제약이 없기 때문에 확장성 측면에서 큰 가치를 가진다. 하지만 연구진은 망설임 등에 의해 발생하는 간투사의 존재는 고려되지 않았음을 언급했다. 이 지점에서 FBDS 알고리즘 방식이 가진 한계를 도출할 수 있다. 대상이 되는 L2와 학습자의 숙련도에 따라 변동성이 크다는 ASR 기반 유창성 자동 평가의 한계를 극복하고자 했으나 FBDS 알고리즘은 일종의 상쇄효과(trade-off)가지고 있다. ASR system은 훈련 목적에 따라 간투사 등을 고려할 수 있는 반면, FBDS는 저수준(low-level) 음향 신호를 분석하기 때문에 어떤 음절 혹은 음소가 어떻게 분석되었는지 구체적으로 알 수 없기 때문이다. 즉, 이러한 알고리즘을 사용한 방식 역시 각 자질에 대해서 신뢰성을 검증하기 어려우며, 여전히 개개인의 학습자들을 위한 자질 별 피드백을 주기에는 적절하지 않다.

Liu et al.(2023)은 종단 간 접근 방법을 통해 유창성 자동 평가를 시도한 연구이다. 이 연구에서는 음향, 언어 모델, 어휘 사전이라는 세 모듈로 이루어진 자동 음성인식 시스템 대신 wav2vec2.0을 사용하였다. 또한 음성 표본에서 직접적으로 프레임 수준의 음성 특성을 추출하기 위해 자기 지도 학습(self-supervised learning; SSL)을 활용하였다.

이후 추출한 프레임에 강제 정렬을 적용하여 음소 단위의 정보를 얻는 대신, K-평균 군집(K-means clustering) 알고리즘을 사용하여 각 프레임에 의사 레이블(즉, 클러스터 인덱스)을 할당했다. 이 군집화 작업은 유창성을 위한 고유 특성을 학습하는 데 사용된다. 마지막 단계로서, 양방향 장단기 메모리 (Bidirectional Long Short-Term Memory; BLSTM) 모델이 프레임 단위 SSL 특성과 클러스터 인덱스를 사용하여 발화 수준에서 유창성 점수를 예측하도록 훈련했다. 이를 바탕으로 비원어민의 영어 낭독 발화 데이터를 대상으로 진행된 실험 결과는 다른 방법을 이전 방법들을 적용한 연구 결과와 유사한 성능을 달성했다고 보고하였다. 종단 간 방식은 음성 신호에서 유창성 관련 정보를 직접 얻도록 설계되므로 음성 전사나 시간 정보가 필요하지 않다. 따라서 음성 인식기의 오류나 강제 정렬 알고리즘의 정확도에 영향을 받지 않는다. 또한 모델의 구조가 비교적 간결하며, 입력과 출력 간의 관계를 직접 학습하기 때문에 다양한 L2 맥락에서의 유창성 평가가 용이하다는 장점을 지닌다. 그러나 전체 과정이 하나의 블랙박스로 간주되기 때문에 내부 작동 원리를 이해하기 어려워, 평가 결과에 대해 학습자와 교수자에게 설명 가능한 방식의 피드백을 제공하는 데에 한계가 있다.

요약하자면, 선행 연구에서는 자동 유창성 자질 추출 및 평가 모델의 성능을 평가하기 위해 다양한 컴퓨터 알고리즘을 사용하여 결과값을 도출하였다. 또한 2.2절 응용 음성학 분야 연구들과 마찬가지로 인간이 매긴 유창성 점수와의 상관관계 분석이나 회귀 분석을 통해 모델의 성능을 보고하였고, 그 결과 역시 지금까지 상당한 진전을 보였다. 그러나 이들 연구에는 주목할 만한 연구 공백이 존재한다. 선행 연구들은 공통적으로 성능 보고에 치우쳐 있다. 이로 인해 유창성을 이루는 각 자질이 얼마나 정확하게 측정되고 있는지에 대한 보고가 부족하며, 전문가의 수동 측정과 비교한 성능 차이의 정도에 대한 검증이 소홀했다. 자질 별 결과 값을 제공하고 있지 않기 때문에 평가 점수를 도출하더라도 피드백을 제공하기 어려우며, 유창성을

향상시키고자 하는 L2 학습자들이 실제로 활용할 수 있는 학습 도구로써 기능하기에는 어려움이 있는 실정이다. 따라서 독립적 유창성 요소의 정확도를 조사하고 검증하며, 인간 평가와의 비교 성능을 살펴봄으로써 신뢰성을 확보할 필요가 있다. 이외에도 다소 무거운 컴퓨터 알고리즘이 활용되기 때문에 언어 학습, 교육 및 평가 분야나 음성학 관련 연구자들이 접근하기 쉽지 않다는 문제점도 존재한다. 따라서 보다 접근이 용이하며 설명 가능한 결과를 도출할 수 있는 방법론의 개발이 요구된다.

제 4 절. 비원어민 한국어 말하기 유창성

이 절에서는 비원어민 한국어 맥락에서 발화 유창성을 논의한 연구들을 검토한다.

제2언어(L2) 또는 외국어로서의 한국어 교육 분야의 연구자들은 비원어민 한국어 말하기 평가에서 유창성은 필수 평가 요소라는 점에 동의하고 있다(김나미·김영주, 2018; 김태경·박초롱, 2015).

대표적인 한국어 능력 평가인 한국어능력시험(TOPIK)의 경우, 말하기 평가 영역이 도입되어야 한다는 필요성에 대한 논의(정원기, 2015; 한국언어문화교육학회, 2016)가 지속된 끝에 2022년 11월 말하기 평가가 최초로 실시되었다. 하지만 말하기 평가의 핵심 평가 요소인 유창성에 대해 '발화 속도가 자연스러운가?[⑤]' 라는 다소 단순한 내용으로 기술하고 있다. 유창성이 말하기 평가의 필수 요인이라는 점에 학자들은 모두 동의하고 있는 만큼 그 내용과 평가 준거 역시 보다 구체화될 필요가 있다. 또한 전세계 각지의 한국어 학습 수요와 외국인 유학생의 수가 지속적으로 증가하고 있는 경향성과 더불어 한국어 시험을 치르는 학습자들의 수 역시 증가하는 추세로 미루어 볼 때, 대규모 말하기 평가를 위한 자동 채점 시스템의 도입은 불가피할 것으로 예상된다.

비단 평가에서만 아니라 해외에서 한국어를 학습하고자 하는 이들에게도 개인화된 한국어 말하기 학습 도구에 일정 수준 이상의 학습자들을 위한 유창성 피드백을 줄 수 있다면 이들의 전반적인 말하기 숙련도 향상에 큰 도움이 될 것이다.

이러한 배경을 근거로 비원어민 한국어 맥락에서 논의되어 온 발화 유창성에 대한 정의를 검토하고, 관련 자질들을 자동 추출하고 평가하는 과정에 대한 논의 현황을 살펴보고자 한다.

강석한 외.(2017)은 외국인 학습자들의 발음 및 유창성 채점 결과에 대해, 선정된 유창성 평가 루브릭들이 어느 정도의 영향을

[⑤]출처 https://www.topik.go.kr/

미치는지 검토하였다. 한국어능력시험에서 4~6급 사이의 중급 및 고급 수준인 60명의 비원어민 한국어 학습자를 대상으로 하고 있다. 이 학습자들은 특정 주제에 대한 경험에 대해 1분 30초 내에 답변하는 과제를 수행했고, 총 1시간 30분 분량의 음성 샘플을 확보했다. 그 다음, 발화 속도, 휴지 빈도, F0 범위, 발화 길이라는 네 가지 음성학적 자질의 측정값이 평가자가 채점한 발음과 유창성 점수[⑥]에 어떤 영향을 주는지 회귀 분석을 실시했다. 그 결과, 발화 속도, F0 범위, 휴지 빈도가 유의미한 영향을 미친다고 보고하며 비원어민의 한국어 유창성의 습득이 범언어적 흐름과 유사한 양상을 보인다고 보고했다. 이들의 연구는 억양과 관련된 F0 자질을 설정함으로써 F0 범위가 상대적으로 좁다고 알려진 음절 언어(syllable-timed language)인 한국어도 F0 범위가 넓은 학습자가 유창성 점수가 높다는 새로운 통찰을 제시했다는 점에서 주목할 만하다. 그러나 유창성을 독립적으로 보지 않고 발음과 묶어 "발음과 유창성"이라는 하나의 평가 루브릭으로 채점이 진행된 점은 아쉬운 부분으로 볼 수 있을 것이다. 또한, 각 측정값이 어떤 방식으로 도출되었는지에 대해서는 기술하지 않았다.

김나미·김영주(2018)는 중급 수준의 비원어민 한국어 학습자 61명의 발화를 수집하였고, 이에 대해 다섯 명의 한국어 교사가 참여하여 연구에서 선정된 유창성 루브릭별 채점을 진행했다. 이후 발화 데이터에 대해 수동 전사를 토대로 분석을 진행한 결과 (ㄱ)문법, (ㄴ)어휘, (ㄷ)발음, (ㄹ)속도 및 더듬거림/반복/휴지, (ㅁ)억양/강세/리듬, 총 다섯 가지 요인이 상관관계가 있다고 보고했다. 그러나 이 연구는 요인들 간 상관관계만을 보고할 뿐 유창성 평가에 있어 각 요인이 어느 정도의 영향력을 가지는가에 대한 구체적인 보고가 되어있지 않으며, 분석 데이터가 양적으로 매우 부족함을 언급하고 있다.

요약하자면, 비원어민의 한국어 발화에 있어서 소규모 음성자료를 주로 수동, 혹은 부분적 자동화로 추정되는 음향 분석의 결과 값을

[⑥]평가자들은 발음과 유창성을 하나의 평가 요소 채점하였다.

도출하여, 평가자의 점수와의 상관관계를 밝혀 어떤 자질들이 유창성 판단에 결정적인지를 규명하고자 하는 시도는 일부 지속되어 왔다. 그럼에도 불구하고 비원어민 한국어 유창성에 대한 정의와 상관 자질을 규정하는 데에는 어려움이 있는 상황이다(김나미·김영주, 2018; 이경, 2015). 이 같은 국내의 발화 유창성 연구 현황은 서구어를 대상으로 진행되어 왔던 발화 유창성 기초 연구에서 보이는 양상과 유사하다(제2절 참고). 또한, 비원어민 한국어 발화 유창성을 보다 명확하게 논의하기 위해서는 유창성을 발음 평가의 일부로 보는 시각(이향, 2013)에서 벗어나 독립적 요소로 분석하여 재고할 필요가 있다.

이에 더해, 대규모 비원어민 한국어 발화 데이터를 분석할 수 있는 유창성 자질을 추출하는 방식에 대해 구체적으로 보고하는 연구, 말하기 유창성 자동 평가 방안에 대한 연구, 각 자질 별 추출 결과 값에 대한 신뢰성과 정확성을 보고하는 연구 역시도 아직까지 부족한 실정이다. 신뢰성 검증에 있어서는 유창성 평가를 하는 전문 평가자 내/간의 신뢰도를 파악하는 연구에 그쳤다(김나미·김영주, 2017).

다만, 한국어 학습자 유창성 자동 평가에서 범위를 넓혀 말하기 자동 평가라는 포괄적 관점에서, 발음 자동 평가의 신뢰성 검토를 시도한 연구인 김형민·고현준(2022)를 살펴보았다. 이 연구는 발음 평가를 위해 API가 평가한 점수(자동 평가 점수)와 한국어 교육 전문가 10명의 평가 점수를 비교함으로써, 문장 읽기 발화에서 높은 상관관계를 지님을 밝혔다. 이를 통해, 발음 자동 평가 API의 성능에 대한 신뢰성을 기초로 비원어민 한국어 발음 자동 평가의 잠재성을 주장하였다. 그러나 자유 발화 음성 데이터에 대해서는 상관관계가 전혀 없었기 때문에 기술의 보완이 필요함을 지적했다. 하지만 비원어민 한국어 발화 유창성의 자동 평가 신뢰성 검토를 시도한 것은 아니라는 점에서 본 연구의 취지와는 차이가 있다. 또한, 비원어민 한국어 맥락에서의 자동 평가 연구는 주로 음성 공학 분야에서 발음 영역에 초점을 맞춰 진행되어 왔다.

한국어 학습에 대한 수요가 꾸준히 증가하고 한국어 능력 시험을 치르는 학습자들이 증가하고 있는 현 시점에서, 말하기 평가의 핵심 요소인 발화 유창성의 상관 자질에 대한 탐색을 통해 유창성 평가 준거 확립은 필수적이다. 뿐만 아니라 대규모 비원어민 한국어 발화 데이터에 대해서 효율적 평가가 이루어질 수 있는 유창성 자동 평가 시스템을 위한 기초 연구를 진행할 필요가 있다.

제 3 장. 실험 방법

제 1 절. 데이터

이 절에서는 이 연구에서 사용한 말뭉치는 AIHUB에서 다운로드 가능한 '**교육용 유럽어 모국어 사용자의 한국어 음성**' 이라는 명칭의 데이터로, 발음 및 말하기 교육과 평가를 위한 인공지능 프로그램 개발을 목적으로 구축되었다. 이 말뭉치는 총 1,540시간 분량으로, 총 3,502명의 비원어민 화자들이 구축에 참여하였다. 이들은 다양한 L1 배경을 가지고 있다. 스페인어(12.8%), 러시아어(11.3%), 터키어(11%), 프랑스어, 이탈리아어, 포르투갈어, 루마니아어, 헝가리어, 독일어는 각 10% 미만으로 구성되어 있으며, 기타 언어가 26%로 이루어져 있다.

데이터의 하위 구성은 발음 교육 및 평가용 데이터(43.3%)와 말하기 교육 및 평가용 데이터(56.7%)이다. 이 연구에서는 발음 교육 및 평가용 데이터를 사용했는데, 이는 화자가 단어 읽기, 문장 읽기, 단락 읽기라는 세 가지 유형의 과업을 수행한 발화를 담고 있다. 이 중 단락 읽기 유형의 과업에 대한 발화 데이터가 분석 대상으로 선정되었다. 그 이유는 다음과 같다. 단어 읽기 과업과는 달리, 문장 읽기와 단락 읽기 과업은 평가자가 매긴 유창성 점수가 포함되어 있다. 또한 같은 낭독 데이터라고 할지라도, 단편적인 언어 사용만 관찰 가능한 문장 읽기 유형보다는 단어와 단어 사이만이 아니라 문장 간 휴지부와 같은 다양한 맥락에서의 언어 사용을 관찰할 수 있고, 음성의 흐름을 보다 연속적으로 관찰할 수 있기 때문에 최종적으로 단락 읽기 과제[⑦]가 본 연구의 목적 달성에 가장 적합하다고 판단했다.

이 연구에서는 이 데이터에서 학습자 240명이 발화한 총 463개의 음성 샘플이 선정되었다. 한 개의 음성 파일당 평균 일곱 개

[⑦] 부록 1 참조. 단락 읽기 과제 대본

문장(7x463; 약 3241개 문장)에 대한 발화가 담겨있으며, 평균 44.14초의 분량으로 총 5.7시간 분량의 데이터를 분석하였다. Fontan et al.(2018)에서 사용된 데이터 분량 총 48분, Detey et al.(2020)의 총 252개의 문장 수, Fontan et al.(2022)의 총 65개의 문장 수, <표 2-1>의 표본 수에서 볼 수 있듯, 이전 연구와 비교했을 때 상당히 큰 분량이라고 할 수 있다.

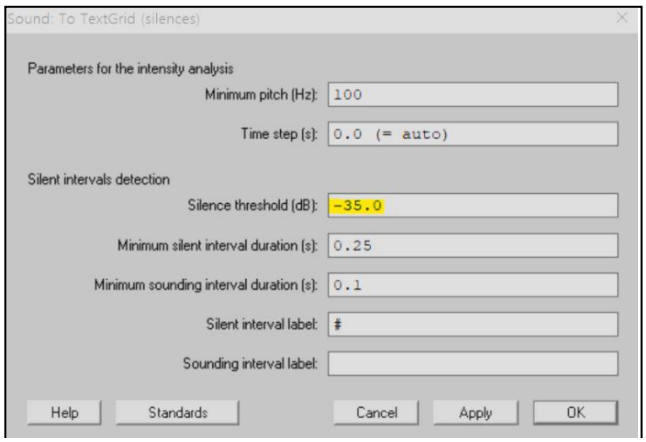
선택된 발음평가용 데이터는 발음 숙련도, 유창성, 이해 가능도라는 세 가지 항목에 대해 각 0점~5점 사이의 점수가 부여되어 있다. 데이터에 대한 설명서에 따르면, 평가자 선발 절차는 다음과 같이 설명되어 있다. 평가 항목별 교육 과정을 통해 평가 방법을 학습하고 5시간 분량의 데이터에 대해 실습 후, 실기 시험을 통과한 인원을 평가자로 선발했으며, 이들은 언어학 관련 전공 배경을 가진 교수, 한국어 강사, 대학원생, 대학생 등이다. 이 평가자들에 의해 사전 평가된 평균 점수가 기록되어 있다. 이 중에서 유창성 평가 점수는 종속변수로 활용, 즉 이 연구의 목적 중 하나인 자동으로 추출한 유창성 자질을 활용한 자동 평가의 가능성 검증을 위한 기준으로 사용되었다.

이 연구는 비원어민 한국어 맥락에서 자동으로 측정된 유창성의 시간 자질(temporal features)들의 값이 전문가의 유창성 평가를 예측하는 잠재력을 살피는 것을 목표로 한다. 한편, 억양 등의 초분절 자질과 어휘 및 문법과 같은 요인들 또한 유창성 지각에 영향을 줄 수 있다(Lennon, 1990; Riggenbach, 1991). 하지만 본 연구는 말하기 유창성 자동평가의 토대를 다지기 시작하는 연구이다. 이를 염두에 두고 자유 발화에서 나타나는 어휘/문법과 같은 변수의 영향과 최소화하고 자동 추출된 유창성의 시간 자질 요인을 집중 탐구하기 위해, 본 연구에서는 실험에 사용하는 데이터를 문단 읽기 과제 데이터로 제한하였다.

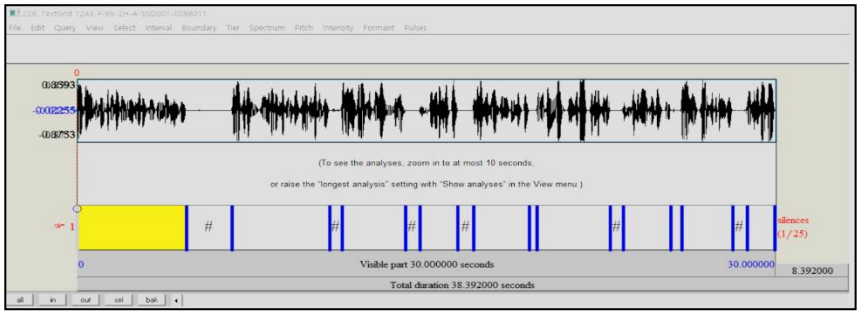
제 2 절. 유창성 자동 측정 vs. 수동 측정

이 연구에서는 먼저 L2 한국어 발화의 유창성 측정을 자동화하고 자동 측정의 신뢰성을 수동 측정과 비교하기 위해 아래 절차에 따라 실험을 진행했다.

1. 음향 분석 소프트웨어인 Praat를 활용하여 녹음된 오디오 파일에서 음향 공백 구간을 자동으로 검출했다. 음향 공백 구간을 잘 검출하기 위해 <그림 3-1>에서 확인할 수 있는 바와 같이, 임계값은 250 밀리초로 설정했다. 이 기준점(cut-off point)에 대해 학자들의 논의는 다양하나, 대부분의 연구가 200~300ms 내로 설정하는 관행이 있고, 250 밀리초는 그 중간 값에 해당한다. 위 범위 내의 설정 값이 파열음의 막음 구간(closure duration)을 포착하지 않는다고 보고한 선행 연구(Towell et al., 1996)에 근거하여 설정하였다. 그리고 <그림 3-2>에서 확인할 수 있듯, 'silence'라는 명칭의 티어가 포함된 총 463개의 Praat TextGrid 형식의 파일을 생성하였다.

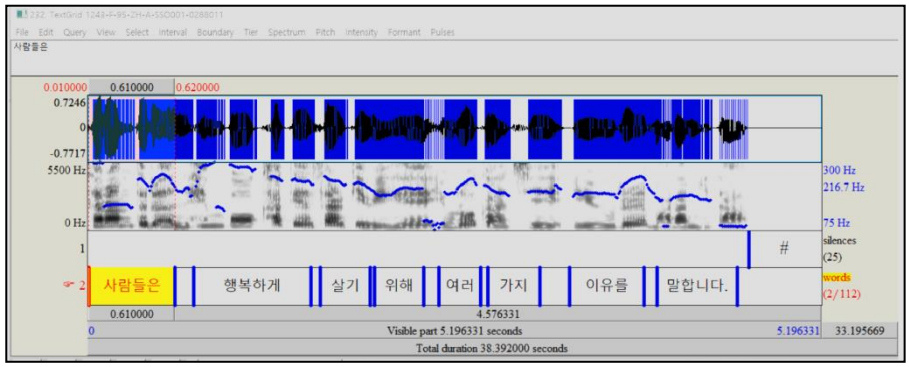


<그림 3-1> Praat의 휴지 측정 값 설정 화면



<그림 3-2> 휴지 검출 후 음성-TextGrid쌍

2. 어절 단위의 타임스탬프와 전사를 얻기 위해 네이버(Naver)에서 개발한 CLOVA Speech Recognition API를 활용했다. API를 통해 오디오 파일을 음성 인식 서버로 전송했으며, 서버에서 파일을 처리하고 텍스트 전사 형태의 인식 결과를 받았다. 해당 작업의 결과물들은 JSON 파일 형식으로 저장되었다.
3. Python 환경(버전 3.7)에서 praatIO 라이브러리를 사용하여 기존 TextGrid 파일에 CLOVA Speech API에서 얻은 ASR 결과를 통합했다. 해당 라이브러리는 TextGrid 파일을 가져오고 조작하는 함수를 제공한다. 본 연구에서는 이 라이브러리를 사용하여 'words'라는 새로운 티어를 TextGrid 파일에 추가했다 <그림3.3>.



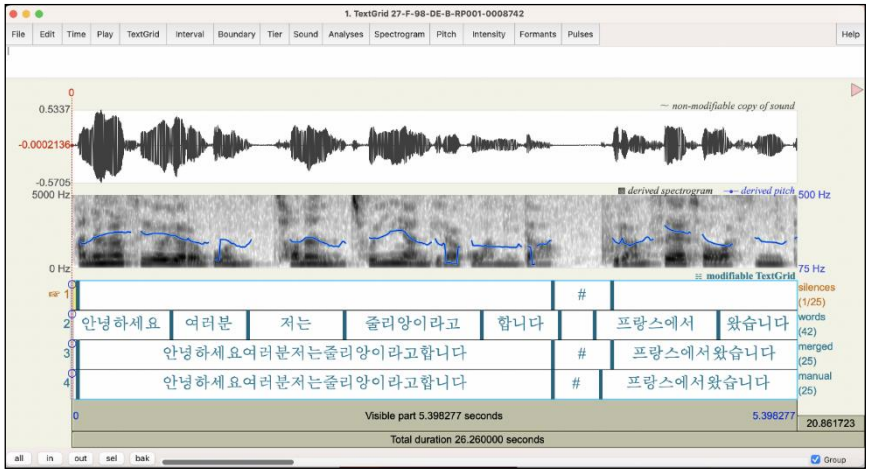
<그림 3.3> 'words' 티어 형성

4. 분석의 용이성을 위해 praatIO 라이브러리를 활용하여 'silences' 티어와 'words' 티어를 하나의 티어로 병합하여 'merged' 티어가 생성

되었으며(<그림 3-4>), 해당 티어에는 1단계 과정에서 검출된 음향 공백 구간과 함께 음성인식 결과물이 모두 포함되어 있다.

위 네 단계 과정을 통해 생성된 생성된 'merged' 티어를 사용하여 아래 표로 제시한 유창성 자질의 결과 값을 도출했다. 또한, 자동 측정의 신뢰성을 검토를 위해서는 비교 대상인 자동 측정 결과 값을 얻어야 했다. 따라서 상술한 과정으로 생성된 463개의 음성 샘플 중 80개 샘플을 추출하여 수동 분석 대상으로 선택했다. 80개 샘플의 선정 기준은 다음과 같다. 이 논문에서 사용된 데이터는 말하기 숙련도(speaking proficiency)에 대한 자기평가(fluent, advanced, intermediate, beginner; 총 4개 수준) 정보를 담고 있다. 사용 데이터 463개(5.7시간) 중 레벨당 20개 발화 샘플을 무작위로 선정하였다.

이후 praat를 활용해 해당 샘플에서 TextGrid 파일의 'merged' 티어를 복제하고 그 티어의 이름을 'manual'로 수정했다. 그리고 이 단계에서 연구자가 발견한 오류나 부정확한 타임스탬프 및 오인식 음절을 수정했다. 마지막으로 'manual' 티어에 대해서도 유창성 자질의 결과 값을 도출했다. 이 결과 값을 'merged' 티어에서 추출한 자동 측정 결과와 비교할 것이다. 분석을 위한 최종 TextGrid 형태는 다음의 <그림 3-4>와 같다.



<그림 3-4> 분석을 위한 최종 TextGrid 캡처

제 3 절. 분석 대상 유창성 자질 정의

이 연구에서 사용한 유창성 자질과 그 정의를 <표 3-1>에 나타내었다. 추출 및 계산 과정을 효율적으로 진행하기 위해 Python praatIO 라이브러리를 활용하여 TextGrid의 'merged' 티어와 'manual' 티어에 접근, 표 3.1에 제시된 유창성 지표에 관한 각 결과 값을 도출하였다. 도출된 결과 값은 이후 통계 분석에 사용되었고, 모든 통계 분석은 IBM SPSS 통계 프로그램(버전 29.0.1.0 (171))을 사용하여 이루어졌다.

<표 3-1> 실험에 사용된 유창성 자질들과 그 정의

유창성 자질	정의
발화 속도 (Speech rate)	총 음절 수를 총 응답 시간으로 나눈 후, 60을 곱하여 분당 단위로 나타낸 지표
조음 속도 (Articulation rate)	총 음절 수를 휴지 구간의 시간을 제외한 응답 시간으로 나눈 후, 60을 곱하여 분당 단위로 나타낸 지표
발성시간 비율 (Phonation-time ratio)	휴지 구간의 시간을 제외한 발화 시간이 총 응답 시간에서 차지하는 비율을 나타낸 지표
분당 연속 발화 평균 길이 (Mean length of runs)	휴지 사이의 발화 구간에서의 평균 음절 수를 나타낸 지표
분당 휴지 평균 횟수 (Mean number of pauses)	총 휴지 횟수를 총 응답 시간으로 나눈 후, 60을 곱하여 분당 단위로 나타낸 지표
분당 휴지 평균 길이 (Mean length of pauses)	총 휴지 구간의 시간을 그 횟수로 나눈 후, 60을 곱하여 분당 단위로 나타낸 지표

참고. 휴지는 250 밀리초 이상의 공백 구간만을 기준으로 한다.

제 4 장. 실험 결과

제 1 절. 자동 vs. 수동 유창성 측정 비교

먼저 이 연구의 첫 번째 연구 질문에 대한 답을 구하기 위해 유창성 자질 자동 측정의 신뢰도를 검토했다. 비교 대상으로 삼았던 수동 측정 결과는 데이터셋에서 선택된 80개의 발화 샘플에 대해 연구자가 직접 측정한 결과 값이다. 그리고 자동 측정 방식과 수동 측정 방식에 대응표본 t-검정(paired t-test)과 Pearson 상관 분석을 적용하여 두 측정 방식에 유의미한 차이가 존재하는지 확인하였다.

4.1.1. 대응표본 t-검정

대응표본 t-검정은 동일한 대상 내에서 두 가지 상황(자동 측정과 수동 측정)을 비교할 때 특히 유용하며, 보다 정확하고 통제된 비교를 가능하게 한다는 이점을 지닌다. 결과는 <표 4-1>과 같으며, 모든 유창성 자질들에 대한 각 p-값이 유의수준 0.05를 초과하기 때문에 두 측정 방법 간의 평균의 차이는 없다는 귀무 가설을 채택할 수 있고, 따라서 두 방식 사이에는 통계적으로 유의미한 차이가 없다고 볼 수 있다.

분당 연속 발화 평균 길이와 분당 휴지 평균 길이에서 보이는 음수의 t-값은 자동 측정 방식이 수동 측정 방식보다 평균이 다소 높게 형성되었음을 의미한다. 한편, 분당 휴지 평균 횟수의 p-값이 유의수준에 상당히 근접한 양상을 보였으나 통계적으로 유의하지는 않다. 이러한 결과는 자동 측정 방식 결과와 수동 측정 방식 결과가 의미 있는 차이가 없다는 사실을 시사한다.

이상의 대응표본 t-검정 결과를 정리하면 <표 4-1>와 같다.

<표 4-1> 대응표본 t-검정 결과

유창성 자질	자동측정 평균값	수동측정 평균값	t	$p-$ value
발화 속도	182.28±45.83	182.50±47.74	0.78	0.44
조음 속도	248.73±44.44	249.74±42.96	0.84	0.40
발성시간 비율	72.64±11.22	72.68±10.37	0.10	0.92
분당 연속 발화 평균 길이	9.48±10.44	7.85±3.25	1.54	0.13
분당 휴지 평균 횟수	22.74±5.95	23.41±5.11	1.93	0.06
분당 휴지 평균 길이	0.71±0.25	0.71±0.26	1.12	0.27

참고. N=80, 평균±표준편차

4.1.2. Pearson 상관 분석 검정

다음으로, 유창성의 자동 측정의 신뢰성과 타당성을 검증하기 위한 두 번째 통계 처리 방법으로 Pearson 상관 분석을 시행했고, 자동 측정의 결과 값이 수동 측정의 결과 값과 얼마나 상관도가 높은 지 파악하였다. Pearson 상관 분석을 시행한 결과, 모든 유창성 자질에서 양의 상관관계가 보였으며, $p < .01$로 통계적으로 유의미한 결과가 나타났다. 이는 모든 자질들에 대해 자동 vs. 수동 측정 사이에 높은 일치도와 일관성이 존재한다는 사실을 의미한다. 단, <표 4-1>에서 보인 결과와 마찬가지로 분당 연속 발화 길이(mean length of runs), 분당 평균 휴지 횟수(mean number of pauses) 두 가지 자질은 자동 vs. 수동 측정 간 상관도에서 상대적으로 낮은 상관계수를 보였다.

이상의 모든 유창성 자질에 대해 구체적인 결과 값은 다음 <표 4-2>와 같다.

<표 4-2> Pearson 상관 분석 결과

유창성 자질	상관계수 (r)
발화 속도	.998**
조음 속도	.970**
발성시간 비율	.965**
분당 연속 발화 평균 길이	.652**
분당 휴지 평균 횟수	.815**
분당 휴지 평균 길이	.984**

참고. ** 0.01 유의수준

제 2 절. L2 한국어 말하기 평가에서의 유창성 자질의 유효성과 설명력

이 논문의 두 번째 연구 문제를 해결하기 위해, 자동 측정 방식으로 얻은 자질 별 결과 값을 이용하였다. 전문가의 유창성 평가 점수에 대한 설명력과 각각의 유창성 자질이 유창성 평가 점수에 미치는 영향력의 정도를 살펴보기 위해 단계적 다중 선형 회귀 분석(step-wise multiple linear regression)을 수행했다. 단계적 다중 선형 회귀 분석은 전진 선택(forward selection)과 후진 제거(backward elimination)의 두 단계로 진행된다. 본 통계 기법은 종속변수를 가장 잘 예측하는 변수 하위 집합을 식별하는 데에 유용할 뿐만 아니라 회귀 모델을 간소화하고 해석 가능성을 향상시키는 데에 도움이 되므로 본고의 연구 취지에 적합하다고 판단했다.

이 분석에서는 463개의 발화 샘플을 데이터로 사용했으며,

3.3절에서 기술한 여섯 가지 유창성 자질들에 대해 각각 자동 측정된 값을 독립변수로 설정했다. 또한 전문가들에 의해 사전 평가된 평균 유창성 점수를 종속변수로 설정했다. 유창성 점수 분포는 <표 4-3>와 같다.

<표 4-3> 전문가들의 유창성 평가 점수의 평균, 표준편차, 최솟값, 중앙값, 최댓값

평균값	표준 편차	최솟값	중앙값	최댓값
3.59	1.04	0	4.0	5.00

더 나아가, 이 연구에서 사용한 데이터가 발음 평가 시스템 개발을 목적으로 구축되었다는 특성을 고려하여 독립변수를 추가한 두 번째 회귀 분석도 실시하였다. 이 2차 회귀 분석에서 말뭉치 구축 과정 중에 전문가들에 의해 사전 평가된 발음 점수를 추가 독립변수로 설정하였다. 따라서, 독립변수가 자동 추출 방식으로 얻어진 유창성 자질들의 개별 수치로만 이루어진 첫 번째 회귀분석과 발음 점수까지 더해진 두 번째 회귀분석의 결과를 구분하여 차례로 살펴볼 것이다.

4.2.1. 1차 단계적 다중 선형 회귀 분석

1차 단계적 다중 선형 회귀 분석 결과 발화 속도와 분당 평균 휴지 길이, 두 개의 독립변수가 전문가의 유창성 점수를 예측하는 데에 유의미한 변수로 판정되었다.

<표 4-4> 1차 단계적 다중 선형 회귀 분석 결과

종속변수	독립변수	비표준화 계수		표준화 계수	t	*p*-value	다중공선성 통계량	
		B	표준오차	베타			공차(T)	VIF
유창성 점수	발화 속도	0.020	0.001	0.759	19.706***	< .001	0.785	1.275
	분당 평균 휴지 길이	1.810	0.300	0.232	6.033***	< .001	0.785	1.275

NOTE. R(.683), R Square(.466), Sig(<.001), Durbin-Watson(1.452)

<표 4-4>에 요약된 첫 번째 모델은 R-값이 .686으로 중간 수준의 설명력을 보였다. R-제곱 값 .466은 발화 속도와 분당 평균 휴지 길이의 결합이 유창성 점수의 분산에 대해 약 46.6%가 설명될 수 있음을 시사한다. 전체 모델에 대한 유의 수준(p 값)은 매우 유의미했다(p < .001). 그리고 Durbin-Watson 통계량은 1.452로, 잔차 간 자기상관이 없었다. 이는 해당 모델이 주어진 데이터의 패턴을 잘 포착한다는 사실을 의미한다. 공차 값인 0.785와 분산 팽창 요인(VIF)인 1.275는 모델의 독립변수 간에 상당한 다중공선성 문제가 없음을 나타낸다. 따라서 각 예측 변수가 유창성 점수를 예측하는 데 고유한 정보를 제공한다고 볼 수 있다.

발화 속도 변수의 계수(B)는 0.020이며, 표준화된 계수(Beta)는 0.759로, 발화 속도가 유창성 점수에 상대적으로 강한 양의 영향을 미친다는 사실을 확인하였다. t-값은 19.706으로 매우 유의미했다(p < .001). 즉, 발화 속도와 유창성 점수 사이의 관계가 견고하다는 사실을 추론할 수 있다.

분당 평균 휴지 길이의 계수(B)는 1.810이며, 표준화된 계수(Beta)는 0.232로, 분당 평균 휴지 길이가 유창성 점수에 발화

속도에 비하면 상대적으로 적지만 여전히 유의미한 양의 영향을 미친다는 것을 나타낸다. 그리고 t-값은 6.033으로 매우 유의미했다 (p<.001), 즉, 분당 평균 휴지 길이와 유창성 점수 사이의 관계가 중요하다고 볼 수 있다.

요약하면, 상술한 결과는 발화 속도와 분당 평균 휴지 길이가 첫 회귀 모델에서 유창성 점수의 중요한 예측 변수임을 보여준다.

4.2.2. 2차 단계적 다중 선형 회귀 분석

데이터의 특성을 고려하여, 모델의 예측 능력을 향상시키기 위해 두 번째 단계적 다중 선형 회귀 분석을 실시하였다. 자동 측정된 유창성 자질 항목 이외에도 전문가들에 의해 사전 채점된 발음 점수가 독립변수로 포함되었을 때의 회귀 모델 결과를 <표 4.2.2>에 제시하였다.

<표 4-5> 2차 단계적 다중 선형 회귀 분석 결과

종속변수	독립변수	비표준화 계수		표준화 계수	t	*p*-value	다중공선성 통계량	
		B	표준오차	베타			공차 (T)	VIF
유창성 점수	발음 점수	0.601	0.034	0.565	17.937***	< .001	0.690	1.448
	발화 속도	0.011	0.001	0.408	11.533***	< .001	0.546	1.832
	분당 평균 휴지 길이	0.929	0.235	0.119	3.947***	< .001	0.750	1.333

참고. R(.828), R-Square(.686), p(< .001), Durbin-Watson(1.534)

두 번째 회귀 분석 모델은 첫 번째 회귀 분석 모델보다 더 강한 설명력을 가진다. 발음 점수를 독립변수로 추가함으로써 가치 있는 정보가 제공되었고, 유창성 점수의 전체 예측력이 향상되었다. 이 분석에서 선택된 강력한 예측 지표는 발음 점수, 발화 속도, 분당 평균 휴지 길이였다. 해당 회귀 모델은 높은 R-제곱 값인 0.686을 달성했으며, 이는 유창성 점수의 약 68.6%가 독립변수의 결합으로 설명될 수 있음을 나타낸다. 그리고 이러한 관계가 통계적으로 유의하다는 사실을 확인할 수 있었다($p < .001$). 또한, Durbin-Watson 값이 1.534로, 모델의 잔차에는 유의한 자기상관이 없었다.

발음 점수의 변수의 계수(B)는 0.601이며, 이 변수의 표준화된 계수가 0.565인 것으로 미루어 볼 때 발음 점수가 유창성에 높은 영향력을 행사한다고 볼 수 있다. 발음 점수와 유창성 점수 사이의 관계는 매우 유의미했다($p < .001$).

발화 속도 또한 그 다음으로 유창성 점수에 상당한 기여를 한다. 발화 속도의 변수의 계수(B)는 0.011이며, 0.408이라는 표준화 계수는 유창성 점수에 발음 점수 다음으로 기여하고 있음을 보여준다($p < .001$).

세 번째 중요도를 보여주는 평균 휴지 길이의 변수의 계수(B) 0.929와 0.119의 표준화된 계수는 상대적으로 낮은 영향력을 나타낸다($p < .001$).

종합하자면, 두 회귀 분석 모델 중 후자의 자동적으로 추출된 유창성 자질들의 개별 값 집합에 발음 평가 점수를 독립변수로 포함한 2차 회귀 모델이 전자의 유창성 자질 항목만으로 구성된 모델보다 우수한 설명력을 가지는 것을 확인할 수 있다.

제 5 장. 논　　의

제 1 절. 유창성 자동 vs. 수동 측정 비교를 통한 신뢰성 검증

이 연구는 비교적 간단한 절차를 거쳐 이용할 수 있는 음성인식 API와 음향 분석 프로그램 Praat를 활용하여 유창성 자질 자동 측정 방법을 마련했고, 이 방법론을 통해 측정한 각 유창성 자질의 결과 값에 대하여 신뢰성과 타당성을 서로 다른 두 가지 통계 처리 기법을 적용해 조사하였다. 먼저, 대응표본 t-검정 결과 모든 자질들에 대한 각 p-값이 유의수준 0.05를 초과하기 때문에 두 측정 방법 간의 통계적 평균 차이가 없다는 결과를 도출해내었다. 뿐만 아니라 Pearson 상관 분석을 통해 자질 별 두 측정 방법이 상당히 높은 수준의 양의 상관관계를 보임을 확인했다. 즉, 이 연구는 성능 향상에 주된 초점을 맞추는 이전 연구들과 달리, 각 자질 별 자동 추출 방식의 측정 평균 값과 수동 측정 평균과의 비교를 통해 자동 측정의 결과 값의 정확성을 검증했다.

2.2절에서는 음성학 분야의 선행 연구를 살펴보았고, 이들이 유창성을 과학적으로 엄밀하게 정의하고자 하는 기초 연구로서 의의를 지님을 확인했다. 또한 이들은 이후 진행된 자동화 연구들이 설정한 연구가설의 토대를 제공했다. 다만, 주로 인간의 유창성에 대한 지각을 포착하기 위해 다양한 음향 자질들을 도입하여 소량의 음성 표본에 대해 수동 분석한 후, 높은 상관관계를 보이는 자질들을 보고하는 데에 집중되어 있었다. 2.3절에서는 융합 분야에서 진행되고 있는 유창성 평가의 자동화 관련 연구를 확인했다. 이들은 평가의 객관성과 일관성을 갖춘 평가 도구의 마련을 위해 진행된 연구였다. 이들은 공통적으로 평가 모델의 단일 성능을 보고하는 데에 그치고 있었다. 이 연구는 자질별 평균값으로 대표되는 각 측정 결과 값을 명확히 제시하였다(<표4.1.1> 참고)는 데에 의의를 지닌다. 이는 학습자별 개인화된

학습도구로서 잠재성을 갖는다. 예를 들어, 개별 학습자마다 유창성 향상을 위해 평균 이하의 측정 값을 보이는 자질들에 대해 구체적 피드백을 제공할 수 있다.

이와 같은 분석을 종합했을 때, 본고에서 시도한 유창성 측정의 자동화 방안은 신뢰성을 바탕으로 다양한 유창성 자질들을 평가할 때 수동 측정법의 대안, 즉 말하기 자동 평가 시스템의 구성 모듈로 발전할 수 있는 기초 연구로서의 가능성을 보여준다. 그리고 이 연구는 양적 변수가 ASR 시스템의 음향 모델(Cucchiarini et al., 2000), 음향 분석 알고리즘(Fontan et al., 2018), 종단간 (end-to-end) 모델(Liu et al., 2023)과 같은 계산 모델을 활용하지 않고, 음성인식 API와 음향분석 프로그램 praat를 활용하여 일련의 과정을 자동화하여 용이한 분석 절차를 통해 일련의 과정의 신뢰성을 보여주었다. 이러한 시도는 말하기 유창성 관련 유사 연구를 진행하는 연구자들에게도 이점으로 작용할 것으로 기대된다.

제 2 절. L2 한국어 말하기 자동 평가에서 유창성 자질들의 유효성 및 설명력

이 논문의 두 번째 연구 질문은 자동화 방식으로 얻은, 기존의 L2 영어를 바탕으로 연구되어 온 유창성 자질들이 비원어민의 한국어 발화에 대해 원어민 전문가의 유창성 점수를 유의미하게 설명해낼 수 있는지를 살피는 것이었다. 이를 위해 이 연구에서는 약 5.7시간에 이르는 3,241개의 문장으로 이루어진 발화 데이터를 분석함으로써 이전 응용 음성학 분야의 연구와는 차별화되는 대규모 분석을 진행하였고, 이를 통해 충분한 신뢰성을 확보했다.

이 질문에 대한 답을 구하기 위해 두 차례에 걸친 단계적 다중 선형 회귀 분석을 사용했다. 1차 회귀 분석 모형에서 발화 속도(speech rate) > 분당 휴지 평균 길이(mean length of pauses) 순의 영향력을

보였으며, 이 두 변수가 핵심적인 예측 변수로 선정되었다. 발화 속도는 2장에서 살펴본 모든 선행 연구가 공통적으로 가장 예측력이 높으며 범언어적으로 신뢰할 만한 척도로 보고하고 있는 자질이다. 이 연구에서도 동일한 결과를 도출하였다. 이는 발화 속도가 비원어민 한국어 유창성 평가(읽기과제)에서도 유의미한 자질로 기능하고 있다는 점을 시사하며 즉, 범언어적 유창성 자질로 보기에 적절하다. 한편, 분당 휴지 평균 길이의 경우 L2 영어를 분석 대상으로 삼았던 Tavakoli et al.(2020)의 연구 결과와 일치한다. 반면, L2 네덜란드어 맥락에서는 분당 휴지 평균 길이에 있어 상대적으로 높지 않은 영향력이 있다고 보고한 Cucchiarini et al.(2000)과는 상반되는 결과이다.

이와 같은 결과는 유창성 자질을 범언어적으로 적용하는 데에는 주의를 기울여야 한다고 주장한 Baker-Smemoe et al.(2014)를 뒷받침한다. 다만, 상관계수 R값은 .686이므로 상관관계가 있다고 할 수 있지만 결정계수 R-제곱 값은 .466에 그친다. 따라서, 적어도 이 연구에서 사용한 발음 평가용 데이터의 한국어 읽기 과제 맥락에서는 유창성 점수를 예측하기 위해서는 추가적인 유창성 자질의 도입이 필요한 것으로 보인다.

한편, 사용된 음성 데이터의 특성, 즉 발음 평가용 데이터인 점을 고려하여 발음 평가 점수를 독립변수 목록에 추가한 2차 회귀 분석 모델은 평가자의 유창성 점수에 대한 설명력이 향상되는 양상을 관찰했다. 영향력 정도는 발음 점수 > 발화 속도 > 분당 휴지 평균 길이 순으로 나타났다. 이러한 결과는 비원어민 한국어 발화의 유창성 평가 시, 발음의 질이 유창성 평가에 큰 영향을 미친 것으로 해석할 수 있다. 즉, 평가자가 유창성을 평가할 때 유창성에 관계없는 발음 정확도에 많이 영향을 받았음을 의미한다. 그러나 평가자들이 해당 데이터를 평가할 때, 이미 발화자가 주어진 프롬프트를 읽었다는 사실을 알고 있으며, 어휘와 문법과 같은 다른 말하기 요소에 어떤 편차도 존재하지 않는다는 사실을 인지한 상태로 채점을 진행했다는 점을 인지해야 한다. 다시 말해, 이와 같이 통제된 조건에서는 평가자로 하여금 발화자의

정확한 발음을 기대하게 만드는 효과를 일으킬 수 있기 때문에 해석에 주의를 기울여야 할 필요가 있다.

제 3 절. 한계와 후속 연구

이 연구는 L2 한국어 맥락에서의 유창성 자동 평가를 위한 초기 시도로서, 실험 데이터를 낭독 발화 수행 발화로 제한하였다. 자유 발화 과제에서 나타나는 화자간 어휘 사용 범위나 문법 정확성 등과 같은 기타 언어 현상의 차이는 최소화하고 우선적으로 탐구 범위를 시간적 특성(temporal features)으로 제한하여 그 유효성을 살피고, 측정의 자동화에 초점을 맞추었기 때문이다. 따라서 후속 연구에서는 탐구 범위를 자유 발화 과제 음성 데이터로 확장해 볼 것이다. 이는 실제 발화와 유사한 상황에서의 발화도 다량의 데이터로 입력 받아 고속 자동 처리하는 강건한 시스템이 필요하기 때문이다. 또한 L2 한국어의 유창성에 대한 설명력을 높이기 위해 추가 음향 자질을 고안해 보는 것도 한국어 유창성에 대한 이해를 높이는 데에 일조할 수 있을 것이다.

제 4 절. 결 론

이 논문에서는 자동 유창성 측정 방법의 신뢰성과 타당성을 조사하고, 비원어민 서구권 언어(주로, 영어) 유창성 평가에 사용된 유창성 자질들이 비원어민 한국어 발화 유창성 평가에서도 유의미한 설명력을 가지는지 살펴보았다.

이 논문은 다음과 같은 연구 현황에 기인하여 진행되었다. 말하기 영역 시험의 평가는 일반적으로 전문가들에 의해 이루어진다. 그러나 이는 많은 시간과 비용을 요구할 뿐만 아니라, 평가자 개개인의 경험과 직관에 근거한 주관이 개입될 가능성이 다분하다. 특히, 대규모로

이루어지는 공인 외국어 말하기 평가의 경우, 평가자 내 및 평가자 간 신뢰도를 확보하기 어려워 객관적이고 일관성 있는 평가 결과를 기대하기 어렵다는 한계가 존재한다.

이를 극복하기 위해 응용 언어학 분야에서 수행한 여러 선행 연구에서는 전문가 또는 일반인의 청각 인상을 반영하는 유창성 지표를 설정하고 수치화하여 객관적인 평가 기준을 마련하고자 했다. 그리고 말하기 자동 평가 기술은 평가 항목을 자동으로 추출하여 평가에 들어가는 시간과 노력을 크게 단축하고 보다 객관적인 평가 결과를 제공할 수 있다는 점에서 대규모 공인 외국어 발화 평가에 도움을 줄 수 있다.

그러나 응용 언어학 분야의 연구들은 대개 L2 영어 학습자의 소규모 발화 데이터를 대상으로 하고 있으므로 대규모 L2 한국어 학습자의 발화에 대한 유창성을 평가하는 데에도 유의미한 지표로 기능하는지에 관한 검증이 요구된다. 또한, 유창성 특징 자동 추출과 관련하여 각 특징 수치들이 수동 추출과 비교하여 얼마나 정확하게 추출되는가에 관한 논의는 충분히 이루어지지 않았다.

이 연구는 두 가지 측면에서 의미 있는 결과를 도출해 내었다. 첫째, 자동화된 유창성 측정의 신뢰성을 수동 측정과 비교하여 제시하여 검증함으로써 그간의 자동평가 시스템 관련 연구들이 성능 향상에 초점을 맞추었던 것과 달리, 자질 별 구체적인 결과 값을 투명하게 제공하여 신뢰성을 검증하였다. 이는 자동 측정 방법이 유창성 관련 변수를 평가하는 데 신뢰할 수 있고, 타당한 대안으로 고려될 수 있다는 것을 시사한다. 추후의 연구에서는 이 논문에서 개발한 유창성 자동 측정 방법의 결과물을 AI 모델링에 활용할 수 있을지 확인해 보는 기회를 가지고자 한다. 간결한 구조를 가지지만 피드백을 제공할 수 없는 AI 모델이 가지는 한계를 극복하여 학습자의 유창성에 대한 구체적인 관리가 이루어질 수 있을 것으로 기대된다. 자동화된 기술의 사용은 객관성, 표준화 및 효율성과 같은 이점을 제공하여 L2 언어 학습 및 평가 분야에서 특히 가치가 있다. 또한, 자동화된 측정 방법의

정확성과 효율성에 대한 근거를 제공하여 대규모 데이터 수집과 분석을 용이하게 한다.

둘째, 비원어민 한국어 발화 유창성 점수에 대한 회귀 분석 결과를 통해 전통적 유창성 자질인 발화 속도와 분당 평균 휴지 길이가 유창성 점수의 가장 중요한 예측 요인으로 나타났다. 이는 대부분의 선행 연구와 일치하는 결과로써 범언어적 유창성 자질이라고 볼 만하다. 한편, L2 한국어 말하기 자동 평가에서 발음 정확도를 고려하는 것의 중요성을 확인하였다. 구체적으로는 발음 정확도가 모델의 예측 능력을 향상시키는 것으로 나타났으며, 이는 전반적인 유창성 인식에 큰 영향을 미침을 의미한다.

종합하면, 이 논문은 자동 유창성 측정 방법의 신뢰성과 타당성에 대한 근거를 제시하고 이를 통해 얻은 자질 별 결과 값을 활용해 L2 한국어 말하기 유창성을 평가하는 데에 유의미한 설명력을 지님을 조사하였다. 이러한 결과는 언어 학습 및 평가 분야에 기여하여 자동화된 기술이 객관적이고 표준화되며 효율적인 유창성 측정을 제공할 수 있다는 잠재력을 보여준다. 향후 연구는 이러한 결과를 바탕으로 다양한 맥락 특히, 자유 발화 환경에서의 유창성 평가에 적용함으로써 보다 더 나아간 응용 가능성을 탐구할 수 있으며, 설명력을 높이기 위한 유창성 자질의 모색을 추구해볼 수 있을 것이다.

참고 문헌

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals, 47*(4), 707-728.

Chandel, A., Parate, A., Madathingal, M., Pant, H., Rajput, N., Ikbal, S., Deshmukh, O., & Verma, A. (2007). Sensei: Spoken language assessment for call center agents. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU).

Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535-544.

Cucchiarini, 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.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223-243.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557.

Deshmukh, O. D., Kandhway, K., Verma, A., & Audhkhasi, K. (2009). *Automatic evaluation of spoken English fluency.* Paper presented at the 2009 IEEE International Conference on

Acoustics, Speech and Signal Processing.

Detey, S., Fontan, L., Le Coz, M., & Jmel, S. (2020). Computer-assisted assessment of phonetic fluency in a second language: a longitudinal study of Japanese learners of French. *Speech Communication*, *125*, 69-79.

Fillmore, C. J. (1979). On fluency. In *Individual differences in language ability and language behavior* (pp. 85-101): Elsevier.

Fontan, L., Le Coz, M., & Detey, S. (2018). *Automatically Measuring L2 Speech Fluency without the Need of ASR: A Proof-of-concept Study with Japanese Learners of French.* Paper presented at the INTERSPEECH.

Fontan, L., Kim, S., De Fino, V., & Detey, S. (2022). Predicting speech fluency in children using automatic acoustic features. 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

Foster, P. (2020). Oral fluency in a second language: A research agenda for the next ten years. *Language Teaching, 53*(4), 446-461.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299-323.

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.

Hilton, H. (2014). Oral fluency and spoken proficiency: Considerations for research and testing. *Measuring L2 proficiency: Perspectives from SLA, 27*, 53.

Hong, H., Ryu, H., & Chung, M. (2014). The relationship between segmental production by Japanese learners of Korean and pronunciation evaluation. *Phonetics and Speech Sciences, 6*(4), 101-108.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA, 32*, 1-20.

Iwashita, N., Brown, A., McNamara, T., & O' Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied linguistics, 29*(1), 24-49.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language learning, 49*(2), 303-342.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145-164.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning, 40*(3), 387-417.

Liu, W., Fu, K., Tian, X., Shi, S., Li, W., Ma, Z., & Lee, T. (2023). An ASR-free Fluency Scoring Approach with Self-Supervised Learning. *arXiv preprint arXiv:2302.09928*.

Mao, S., Wu, Z., Jiang, J., Liu, P., & Soong, F. K. (2019). NN-based ordinal regression for assessing fluency of ESL speech. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Ryu, H., Hong, H., Kim, S., & Chung, M. (2016). *Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection.* Paper presented at the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

Segalowitz, N. (2010). *Cognitive bases of second language fluency*: 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.

Tavakoli, P., Nakatsuhara, F., & Hunter, A. m. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169-191.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. *Planning and task performance in a second language, 239273*.

Towell, R. (2012). Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 47-69.

Yan, X., Kim, H. R., & Kim, J. Y. (2021). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing, 38*(4), 485-510.

Chandel, A., Parate, A., Madathingal, M., Pant, H., Rajput, N., Ikbal, S., Deshmukh, O., & Verma, A. (2007). Sensei: Spoken language assessment for call center agents. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU),

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: 다국면 라쉬 모형 활용. *언어학 연구* (45), 483-524.

김나미, & 김영주. (2018). 말하기 유창성 평가에서의 평가 구인 간 상관관계-한국어 중급 학습자의 발화를 중심으로. *한국어 의미학, 59*, 87-108.

김태경, & 박초롱. (2015). 외국어로서의 한국어 발화 유창성 변화 연구: 중국어 모어 화자를 대상으로: 중국어 모어 화자를 대상으로. *언어과학연구, 75*, 129-150.

김형민, & 고현준. (2022). 한국어 학습자 발화에 대한 발음 평가 API와 한국어교육 전문가의 평가 비교. 이중언어학, 90, 29-52.

이경. (2015). 한국어 어휘 유창성과 평가 구인에 대한 인식 연구-채점자 논평을 중심으로. *이중언어학*(58), 59-86.

이향. (2013). 발음 평가에 있어서 정확성, 유창성, 이해명료성, 이해가능성 기준 간의 영향 관계 연구. *언어와 문화, 9*(3), 221-243.

양승희, & 정민화. (2017). 비원어민 한국어 말하기 숙련도 평가와 평가항목의 상관관계. *말소리와 음성과학, 9*(3), 49-56.

부 록

<부록 1: 단락 읽기 과제 대본>

1. RP001

안녕하세요 여러분, 저는 줄리앙이라고 합니다.
프랑스에서 왔습니다.
고향에서는 한국어를 혼자 공부했습니다.
아직 한국어를 잘 못 하지만 한국생활이 재미있을 것 같습니다.
여러분과 같이 공부하게 되어 정말 기쁩니다.

2. RP002

오늘은 한국어 수업이 있어요.
수업은 오전 아홉 시에 시작해요.
집에서 학교까지 삼십 분이 걸려요.
일곱 시에 일어나서 아침을 먹었어요.
여덟 시에 집에서 나갔어요.
지하철역에 도착했어요.
그런데 지갑이 없었어요.
책상 위에 지갑이 있었어요.
지하철역까지 뛰어서 갔어요.
학교에 여덟 오십 오 분에 도착했어요.
수업에 안 늦었지만 정말 힘들었어요.

3. RP003

의사소통 과정 중에서 말하기와 쓰기 같은 표현 기능이 중요할까?
듣기와 읽기 같은 기능이 더 중요할까?
나는 이해 영역에 더 중요하고, 특히 듣기가 기본이라고 생각한다.
다른 이의 말을 듣고 그 말을 이해해야 의사소통을 계속 할 수 있기 때문이다.

따라서, 진정한 의사소통은 듣기에서부터 시작한다고 주장하고 싶다.

4. RP004

나는 아름다운 색을 사랑한다.
예전 우리 유치원 선생님이 주신 색종이 같은 빨간색, 보라색,
주색, 녹색, 이런 색깔을 나는 좋아한다.
나는 우리나라 가을 하늘을 사랑한다.
나는 오래된 가구의 색을 좋아한다.
늙어가는 학자의 희끗희끗한 머리카락을 좋아한다.
나는 이른 아침의 새소리를 좋아하며 봄 시냇물 흐르는
소리를 즐긴다.
갈대에 부는 바람 소리를 좋아하며,
바다의 파도 소리를 들으면 아직도 가슴이 뛴다.

Abstract

A Basic Study for Evaluating and Providing Feedback on the Korean Fluency of Non-Native Speakers

Mikyoung Kim
The Department of Linguistics
The Graduate School
Seoul National University

This study serves as a foundational research aiming to automate the evaluation of speech fluency — a critical assessment criterion in spoken language evaluation — in the context of non-native Korean learners, and to provide feedback.

Among the various fluency-related features previously studied for L2 English, we selected several key fluency features identified in prior research and included them as subjects of analysis. To determine whether these features are also applicable to the evaluation of non-native Korean fluency, we propose a method for automatically extracting the values of each feature from non-native learners' Korean utterances. By examining the explanatory power of the combinations of fluency feature values which are automatically extracted in predicting and explaining the fluency scores assigned by human evaluators, we explored the importance and utility of these attributes in the automatic evaluation and feedback of non-native Korean fluency, thereby laying the groundwork for further research

in this area.

For the automatic extraction method of fluency feature values proposed in this paper, we validated its reliability by comparing the automatically extracted values with the results of manual acoustic analysis. This method is deemed highly useful for providing valuable fluency feedback to Korean language learners.

In this study, we analyzed the speech data of non-native Korean learners, which is available from AIHUB. The most important features in fluency assessment, such as speech rate, articulation rate, phonation-time ratio, mean length of runs, mean number of silent pauses per minute, and mean length of silent pauses, were automatically extracted. Using paired t-tests and Pearson's correlation analysis, we verified the reliability of our automated method by comparing it with manual acoustic analysis. Moreover, in two consecutive multiple linear regression analyses to examine which fluency features function as significant indicators in the evaluation of non-native Korean fluency, the explanatory power of the first multiple regression model, which used only fluency features to predict and explain the fluency scores of human evaluators, was R2=.466. However, the explanatory power increased to R2=.686 when the pronunciation evaluation score was added as an independent variable in the second multiple regression model. Furthermore, we examined the features selected by the stepwise regression model and their influence on the regression model. The results of the first model identified speech rate (Beta=0.759) and mean length of silent pauses per minute (Beta=0.232) as significant predictors of fluency scores, which aligns with prior studies demonstrating a high correlation between the results of our automated fluency assessment method and human auditory

perception of fluency. In the second model, including the pronunciation score as an independent variable, we found that pronunciation score (Beta=0.565), speech rate (Beta=0.408), and mean length of silent pauses per minute (Beta=0.235) impacted the regression model in that order.

Keywords : utterance fluency, non-native Korean fluency, fluency features, automatic fluency measurement, reliability validation of automatized evaluation, fluency feedback
Student Number : 2019-29853