인문연구 제96호
Journal of the Humanities
2021. Vol. 96, No. 1, pp.1~24
http://dx.doi.org/10.21211/JHUM.96.1

영어 발음에 대한 한국인 평가자의 성향과 일관성 연구*

박상진** · 이유리*** · 윤여신**** · 윤규철*****

〈차 례〉

1. 머리말
2. 연구방법
3. 연구결과
4. 결론

【국문초록】

본 논문의 목적은 영어 발음 평가를 수행하는 비원어민 평가자들의 음성학적 요소별 평가 성향과, 절대 평가나 상대 평가에 따른 평가 종류별 점수 분포 양상, 또 시간 차이를 둔 평가들 사이의 평가 일관성 등을 알아보는 것이다. 자동 평가 점수와 평가자

* 이 연구는 2021년도 영남대학교 학술연구조성비에 의한 것임.
** 영남대학교 영어영문학과 강사, 제1저자(주저자)
*** 영남대학교 영어영문학과 강사, 제2저자
**** 대구대학교 영어영문학과 박사, 제3저자
***** 영남대학교 영어영문학과 교수, 교신저자

들의 수동 평가 점수를 분석한 결과, 실험에 참여한 네 평가자 모두 분절음보다는 초분절음에 상대적인 비중을 더 두는 것으로 나타났으며, 절대 평가와 상대 평가 두 평가 방식에 있어서는 유의미한 차이를 보이지 않았으나, 개별 학생별로 평가 점수 차이가 매우 불규칙한 분포를 나타냈다. 또 한달 전후에 시행한 동일한 상대 평가에서도 통계적인 차이는 대체로 없었으나, 개별 학생 점수 분포는 매우 불규칙하고 일관성이 결여된 것으로 나타났다.

주제어: 영어, 한국인 평가자, 평가 성향, 평가 일관성

1. 머리말

전 세계적으로 비원어민의 영어 사용이 급증하고 있는 추세 속에서 영어 발음에 대한 평가는 그 중요성이 날로 증가하고 있다. 한국인 학습자들의 영어 발음에 대한 평가를 영어 원어민이 수행하느냐, 혹은 한국인 교수자가 수행하느냐, 아니면 컴퓨터에 의한 자동 평가를 수행하느냐는 많은 연구자들의 관심사가 된지 오래이다(Kim, 2009; Lowenberg, 2000; Yun, 2012; Kang & Ahn, 2012; Yoon, 2013ab; Ryu & Chung, 2016).

사람이 평가를 수행하는 경우에 이들 평가자들의 특성이나 경향, 일관성에 대한 연구는 매우 다양한 만큼, 의견도 분분하다. 평가자들의 평가 기준에 대한 연구를 보면, 공통적으로 채택되고 있는 기준은 문법의 정확성, 어휘 적절성, 유창성, 발음 명료성 등이다(Chung et al., 2008; Kang, 2013; Yu, 2014). 대답이 정해져 있지 않는 개방형 질문에 대한 평가시 이들 기준들이 다양하고 통합적으로 적용되어야 하지만, 대답이 확정되어 있거나 주어진 문장을 소리 내어 읽은 경우에는 발화 문장의 발음 명료성이나 유창성

평가가 상대적으로 더 중요할 것임은 자명하다.

발음 평가에 있어 평가자들의 차이를 한국어와 영어의 음소 목록 차이나, 음소배열론 차이로 설명하는 연구도 있고(Yun, 2009), 억양 범위, 휴지 구간 길이, 발화 속도 측면에서 살펴본 사례도 있다(Kang & Rhee, 2011). 발음 평가를 운율과 분절음으로 나누어 독립적으로 평가할 수 있음을 제시한 연구(Park, 2012; Yoon, 2013ab)는 비원어민 평가자들의 성향을 연구할 수 있는 새로운 방식을 제안하고 있다.

Yoon(2013ab)에 따르면, 발화 문장을 구성하고 있는 초분절음의 세 가지 요소, 즉 F0(기본주파수), 강도, 길이와 분절음 요소인 스펙트럼, 총 네 가지 요인들을 계량화하기 위하여, 우선 원어민과 학습자의 동일한 발화 문장을 서로 비교하여 그 유클리드 거리(Euclidean distance) 차이를 네 개의 축으로 수치화한 다차원 모델로 구축한 다음, 이 모델에 근거하여 미지의 학습자 발화의 평가 점수를 통계적으로 예측한다고 한다.

이 방식은 기존의 평가 기준인 발음의 명료성이나 유창성을 구성하고 있는 초분절음적 요인들과 분절음적 요인을 보다 근본적으로 수치화하여 제시하고 있는 점이 독특하다. 발음 평가에 있어서 명료성 혹은 유창성이라는 추상적인 개념도 중요하지만, 이들 개념을 구성하는 물리적 요소인 F0, 강도, 길이, 분절음 스펙트럼을 이처럼 계량화할 수 있다면, 비원어민 평가자들이 이들 중에 어느 요인에 보다 더 비중을 두는지 파악하여 해당 평가자의 평가 성향을 설정하고 훈련하는 데에 소중한 자료가 될 것이다.

또 평가자들의 평가 기준뿐 아니라, 평가 방식이 상대 평가이냐 절대 평가이냐 혹은 이들을 혼합한 방식이냐도 중요한 사안이다. 2018년 이후 수학능력시험에서 영어 평가가 절대 평가로 전환됨에 따라 영어 말하기 평가도 이러한 추세를 고려할 필요가 있다.

영어 말하기나 발음 평가에 대한 기존의 연구는 점수의 변별력

을 고려하기 위하여 거의 대부분 상대 평가를 전제로 한 것들이다. 같은 발화 문장들을 대상으로 동일한 평가자가 상대 평가와 절대 평가를 동시에 실시한 연구는 찾아보기 힘들다. 좀 더 면밀한 평가자 분석을 위해서는 평가자가 절대 혹은 상대 평가에 있어서 어떠한 성향을 보이는지도 주목해야 할 필요가 있다. 즉, 평가자의 일관성에 절대 혹은 상대 평가 방식에 따른 평가 특성도 포함되어야 할 것으로 보인다.

이러한 배경을 바탕으로, 본 연구에서는 비원어민 평가자들의 발음 요인별 평가 성향과 평가 종류나 기간에 따른 일관성을 알아보기 위하여 두 가지 측면을 살펴보고자 한다. 첫째는 평가자들이 발화 문장의 음성학적 구성 요소인 F0, 강도, 길이, 스펙트럼 중에서 어느 요인에 상대적으로 높은 비중을 두고 평가하는지 평가자 성향을 살펴보고, 둘째로 동일한 발화 문장에 대하여 상대 평가와 절대 평가를 동시에 실시하고, 약 한 달 후에 동일한 평가를 재실시하여 평가자의 평가 종류별, 기간별 차이와 일관성을 알아보고자 한다.

2. 연구방법

1) 평가 대상 문장

평가에 사용될 문장은 연구자의 영어 과목을 수강한 학생들로부터 녹음한 자료이다. Rhee et al.(2003)에서 사용한 문장 중에서 여덟 개의 문장을 선택하였고, 이들을 백여 명이 넘는 학생들이 녹음을 하였다. 녹음은 조용한 강의실에서 Shure SM-10 헤드셋 마이크를 쓰고 진행되었고, 컴퓨터의 프랏(Boersma, 2001)에 22kHz의 표본주파수로 녹음되었다.

녹음된 문장들 중에서 실제 평가에 사용될 문장을 선별하기 위하여 각 문장별로 연구자로부터 부여된 점수를 고려하여 문장별로 높은 점수, 중간 점수, 낮은 점수를 받은 학생을 각각 다섯 명씩 선별하여 총, 8 문장 x 5명 x 3 점수 = 120문장을 선별하였다. 미리 점수별로 분류한 것은 실험에 참여한 평가자들이 알지 못하며, 단지 고른 평가 점수가 나올 수 있도록 실험의 편의를 위한 것이다.

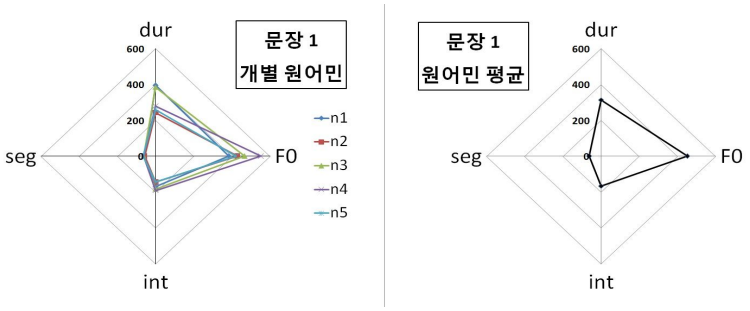
문장1: The dancing likes only the apple pies.
문장2: The players sent the cap to Susan.
문장3: People couldn't sleep well last night because of the noise.
문장4: When did he go skiing, Wednesday or Thursday?
문장5: I see the moon shining brightly all over the lake.
문장6: How long does it take to the next bus stop?
문장7: Thirteen years later, Mary met him at the same place.
문장8: If a tree could talk, what would it say?

선별된 120개의 실험 문장들에 대하여 청취 평가시 음량 레벨을 정규화하기 위하여 강도를 70dB로 모두 조정하였다.

2) 원어민과의 비교를 통한 자동 계량화

평가자들이 F0, 강도, 길이, 분절음 스펙트럼 등 네 가지의 초분절음/분절음적 요인들 중 어느 요인에 더 비중을 두는지 계량화된 분석을 하기 위하여, 윤규철(2013ab)에서 제시한 대로, 평가 대상이 되는 120개의 문장에 대하여 동일한 원어민 문장을 이용하여 네 가지 측면에서 차이를 수치화하였다.

여덟 개의 문장 각각에 대하여 모범적이라 판단되는 다섯 명의 원어민 발화 문장을 Rhee et al.(2003)에서 선별하여 이용하였다. 프랏 스크립트를 이용하여 평가 대상 한 문장 당 원어민 문장 다섯 개를 다중 비교하여 네 개의 요인별로 하나의 평균값을 구해내었다. 다섯 명의 모범 원어민 문장을 이용한 이유는 어떤 문장을 모범적으로 읽는 방법이 하나가 아니기 때문에, 모범 답안의 다양성을 확보하기 위한 것이다. 한 문장 당 네 가지 요인에 대한 수치 네 개가 출력되었다.



〈그림 1〉 계량화된 원어민 문장

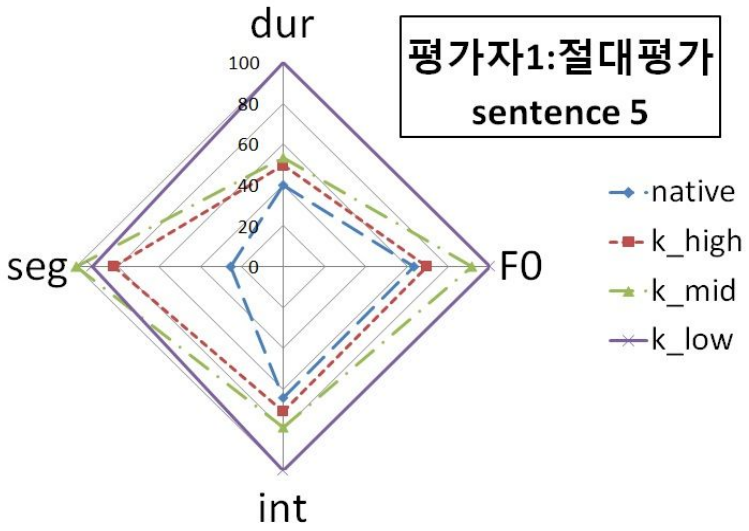
〈Figure 1〉 Praat experiment script screen

우선 다섯 명의 원어민들(n1~n5) 자체 내의 차이를 네 요인별로 구해 레이더 차트로 표현하면 〈그림 1〉과 같다. 그림에서 네 가지 요인 F0, 강도, 길이, 분절음 스펙트럼은 각각 F0, int, dur, seg로 표시되었다. 문장 1에 대하여 원어민 레이더 차트가 나타내는 것은 네 가지 요인별로 원어민들이 서로 얼마만큼의 차이를 나타내느냐이다.

〈그림 1〉에서 왼쪽 차트는 각 원어민에 대하여 나머지 네 명의 원어민 발화를 비교하여 그 차이를 네 가지 요인 측면에서 숫자로 표현한 것이다. 각 축을 나타내는 네 가지 요인들은 단위가 다르

므로 구체적인 수치보다는 최대값과의 비율이 의미 있게 된다. 그림의 오른쪽 차트는 원어민들 수치를 하나의 평균값으로 나타낸 것이다.

각 학생의 계량화된 수치는 다섯 명의 모범 원어민 문장들과의 다중 비교로 구해지지만, 레이더 차트로 표현할 때는 시각적 편의상 〈그림 2〉와 같이 원어민 평균치(native)와 함께 표시하도록 하겠다. 또한 각 축의 최대값 대비 백분율로 표시하여 단위와 무관하게 집단간 비율의 차이을 비교할 수 있도록 하였다.



〈그림 2〉 평가된 학생들의 레이더 차트 샘플

〈Figure 2〉 A sample radar chart for evaluated students

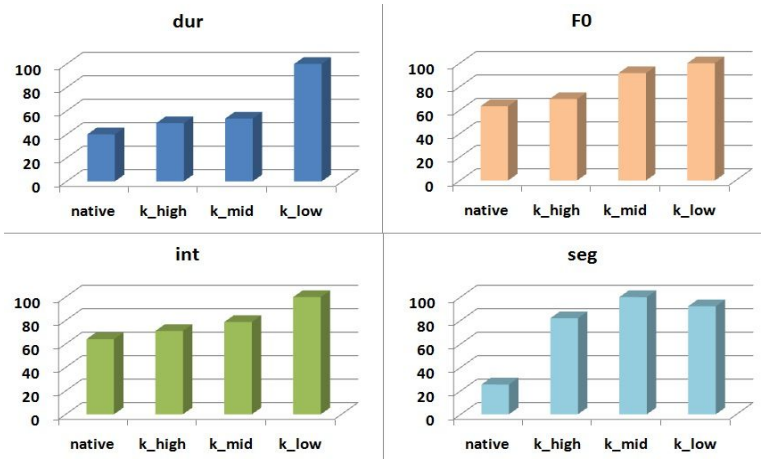
이처럼 원어민과의 자동화된 비교를 통해 계량화된 학생들의 수치 모음은 평가자들이 수동으로 평가하기 이전에 마련된 값들로서 컴퓨터로 계산된 일종의 자동 평가 점수(computerized scores)로

볼 수 있다. 이 수치들을 사람이 수행한 수동 평가 점수(evaluators' scores)과 비교 분석하면 평가자별로 네 개의 요인에 따른 평가 성향을 유추해 낼 수 있다는 것이 본 논문이 주장하는 바이다.

〈그림 2〉는 실제로 1번 평가자가 문장 5에 대하여 절대 평가를 시행한 것을 나타낸 것이다. 높은 점수(5점), 중간 점수(3점), 낮은 점수(1점)을 받은 학생들만의 수치를 각각 k_high, k_mid, k_low로 표시한 것이다. 그림에서 보듯, 제일 안쪽에 있는 원어민(native) 수치보다 학생들의 수치는 모두 다 밖에 있는데, 이는 원어민 문장들끼리의 비교값보다 해당 학생들과 원어민들과의 비교값들이 더 크다는 것을 나타낸다. 즉, 네 요소 측면에서 원어민 문장보다 벗어난 정도가 더 크다는 것이며 발음이 좋지 않을 수 있다는 것을 나타내는 것이다.

점수별 학생들을 비교해 보면, 낮은 점수를 받은 학생들의 수치가 길이, F0, 강도 등 세 가지 요인 축에서 보면 가장 밖에 있는 것을 볼 수 있는데, 이는 기준이 된 원어민 문장보다 많이 벗어나 있다는 것을 의미한다. 수치를 비교해 보면, 높은 점수의 학생들이 가장 안쪽에, 중간 점수의 학생들이 그 밖에, 낮은 점수 학생들이 분절음 스펙트럼을 제외하면 가장 밖에 있는데, 이것을 보면, 컴퓨터로 계산된 자동 평가 점수와 평가자1의 수동 평가 점수가 대체로 같은 성향을 보이는 것으로 해석할 수 있다.

레이더 차트의 네 가지 요소 축을 시각적 편의를 위해 네 개의 막대그래프로 나타내면 〈그림 3〉과 같다. 그래프를 비교하여 살펴보면, 이 문장의 경우 평가자 1은 분절음 스펙트럼(seg 축)보다는 초분절음 요인들(F0, dur, int 축)을 평가 등급 구별에 이용한 것으로 보인다. 다시 말하면, 평가자1은 적어도 이 문장에 있어서는 분절음보다는 초분절음에 있어서 평가 가중치를 두는 성향이 있는 것으로 판단할 수도 있을 것이다.



〈그림 3〉 레이더 차트를 막대그래프로 표현

〈Figure 3〉 A bar chart from the sample radar chart

3) 평가 실험

평가 실험은 프랏의 청취 실험 기능인 ExperimentMFC 오브젝트를 활용하였다. 프랏 스크립트를 이용하여 발음 평가를 1(worst)부터 5(best)까지의 스케일로 절대 평가 1회, 상대 평가 1회를 연이어 실시하도록 작성하였다(〈그림 4〉 참조). 원하면 반복하여 청취할 수 있게 하였고, 직전의 평가로 되돌아가서 다시 할 수도 있게 하였다. 40개의 문장마다 휴식 시간을 부여하였으며, 각 평가는 20분 정도, 평가자당 총 40여분이 소요되었다.

평가 방식인 절대 평가와 상대 평가와 관련하여, 절대 평가시에는 원하면 좋은 점수를 학생 수와 관계 없이 줄 수 있고, 상대 평가시에는 고른 점수 분포를 위하여 높은 점수와 낮은 점수를 좀 더 엄격하게 주어야 한다고 설명해 주었다. 구체적으로 점수별로 몇 명씩 주어야 한다고 제한을 두지 않은 이유는, 각 평가자들이

갖고 있는 절대 혹은 상대 평가의 주관적 개념을 파악하고자 하였기 때문이다.



〈그림 4〉 평가 실험 스크립트 실행 화면
〈Figure 4〉 Praat experiment script screen

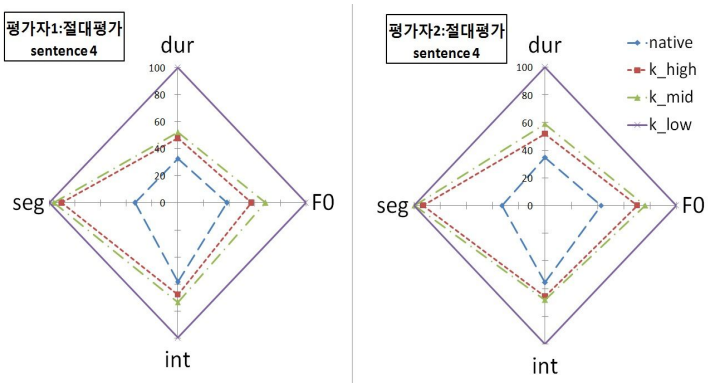
실험을 비대면으로 실시하기 위하여 실험 방법을 동영상으로 녹화하여 실험 자료와 함께 이메일로 전송하여 평가자들이 자택에서 평가를 실시하도록 하였고 결과 파일을 연구자에게 재전송하여 분석하였다. 연구자를 포함하여 영어를 10년 이상 가르쳐온 교수자 총 네 명의 평가자가 실험에 참여하였고, 평가 점수는 4 명 평가자 x (절대평가 120개 + 상대평가 120개) = 960개였다.

약 한 달 후에 같은 실험을 다시 시행하도록 부탁하여 추가로 960개의 평가 점수를 취합하였다. 수거한 평가 점수들은 절대 평가와 상대 평가의 점수 분포 및 한 달 전후의 평가 일관성 검토를 위하여 히스토그램 등의 기술 통계를 활용하여 분석하였다.

3. 연구결과

1) 평가자들의 요인별 수동 평가 성향

수동 평가 실험의 결과를 레이더 차트 분석을 통해 알아보았다. 우선 평가 결과를 정성적으로 살펴보기 위하여, 문장4에 대한 평가자1과 평가자2의 절대 평가 결과를 〈그림 5〉에 나타내었다.



〈그림 5〉 문장4에 대한 절대 평가 레이더 차트

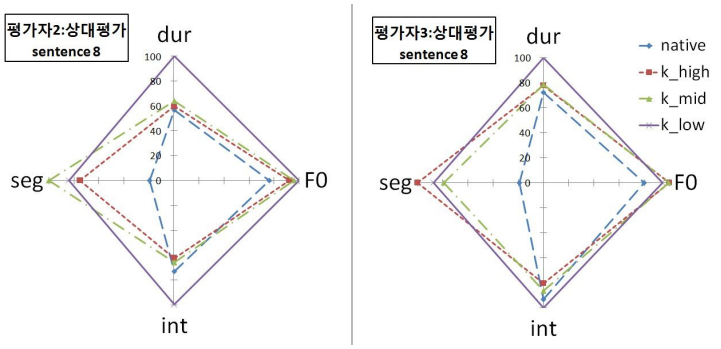
〈Figure 5〉 Radar charts of absolute evaluation for sentence 4

두 평가자 모두 공통적으로 높은 등급 점수를 받은 학생들이 원어민들의 수치에 가장 가까웠고, 그 다음으로 정도의 차이는 적으나 중간 등급 점수를 받은 학생들이었고, 낮은 등급 점수를 받은 학생들이 큰 차이로 가장 멀게 나타났다. 이를 통해, 컴퓨터로 미리 계산된 자동 평가 점수와 평가자들이 부여한 수동 평가 점수가 서로 동일한 성향을 갖는다는 것을 볼 수 있다.

또한 네 가지 음성학적 요인들에 있어서 분절음보다는 초분절음적 요인들에서 등급별 차이가 더 크다는 것을 볼 수 있어, 평가자

들이 초분절음 요인들에 상대적으로 더 가중치를 둔다는 것을 유추해 볼 수 있다. 특히, 높은 점수와 중간 점수를 받은 학생들과 낮은 점수를 받은 학생들 사이의 자동 평가 점수 수치 차이가 그 비율에 있어 매우 크다는 것도 알 수 있다.

자동 평가 점수와 수동 평가 점수가 모든 요소면에서 일치하지 않는 경우도 물론 있다. 문장8에 대한 평가자2와 평가자3의 결과를 〈그림 6〉에 나타내었다. 평가자2와 평가자3 모두 분절음 요소에 있어서 수동과 자동 평가 점수 순위가 일치하지 않았는데, 구체적으로 평가자2의 경우는 중간 등급이, 평가자3의 경우는 높은 등급이 원점에서 가장 멀었다.



〈그림 6〉 문장8에 대한 상대 평가 레이더 차트

〈Figure 6〉 Radar charts of relative evaluation for sentence 8

이러한 결과는 평가자2가 해당 문장을 평가할 때 분절음보다는 초분절음에 비중을 더 두었다는 것을 반영한 것이라고 볼 수 있을 것이다. 특히, 평가자3의 경우에는 초분절음 요인들 중에서 F0보다는 길이와 강도에 상대적으로 비중을 더 두었다고 말할 수 있다.

〈표 1〉 평가자의 요인별 가중치 부여 여부

〈Table 1〉 Weighting of factors by evaluators

해당여부 O X		평가자1		평가자2		평가자3		평가자4	
		절대	상대	절대	상대	절대	상대	절대	상대
문장1	seg	X	X	X	X	O	X	X	X
	F0	O	O	O	O	O	O	X	O
	dur	O	O	O	O	O	O	X	O
	int	O	O	O	O	O	O	X	O
문장2	seg	X	X	X	X	X	X	X	X
	F0	O	O	O	O	X	O	O	O
	dur	O	O	O	O	X	O	O	O
	int	O	O	X	X	X	O	O	O
문장3	seg	X	X	X	X	X	X	X	X
	F0	O	O	X	O	O	O	O	O
	dur	O	O	O	O	O	O	O	O
	int	X	X	X	X	O	O	X	O
문장4	seg	O	O	O	O	O	O	O	O
	F0	O	O	O	O	O	O	O	O
	dur	O	O	O	O	O	O	O	O
	int	O	O	O	O	O	O	O	O
문장5	seg	X	O	X	X	X	X	O	O
	F0	O	O	X	O	O	O	O	O
	dur	O	O	O	O	O	O	O	O
	int	O	O	O	O	O	O	O	O

문장6	seg	O	X	X	O	X	X	X	O
	F0	O	X	O	O	O	X	O	X
	dur	O	O	O	O	O	O	O	O
	int	O	O	X	O	O	O	O	O

문장7	seg	X	O	O	O	X	O	X	X
	F0	X	X	X	X	X	X	X	X
	dur	O	O	O	O	O	O	O	O
	int	O	O	O	O	O	O	O	O

문장8	seg	O	X	O	X	O	X	X	O
	F0	O	O	O	O	X	O	O	O
	dur	O	O	O	O	O	O	O	O
	int	O	O	O	O	O	O	O	O

평가자별로 어느 요인에 비중을 더 두었는지 레이더 차트를 기반으로 표로 나타내면 〈표 1〉과 같다. 대체로 평가자들은 문장4를 제외하고, 상대 평가 혹은 절대 평가에 있어서 분절음 요소보다는 초분절음적인 요소들 즉, F0, 강도, 길이에 상대적으로 비중을 더 두고 평가를 했음을 알 수 있다. 특히, 문장7의 경우 F0보다는 길이나 강도의 비중이 높음을 알 수 있다. 따라서 자동 평가 점수와 수동 평가 점수를 비교함으로써 해당 평가자의 평가 성향을 유추해 볼 수 있고, 이는 평가자의 훈련이나 검증에 유용한 자료로 쓰일 수도 있을 것이다.

2) 평가 유형에 따른 차이

평가자들이 절대 평가 혹은 상대 평가에 따라 어떠한 차이를 보이는지 살펴보기 위하여, 지면 관계 문장1에서 문장4까지 평가자별로 각 학생들에게 부여한 수동 점수를 〈표 2〉에 나타내었다.

〈표 2〉 평가자의 학생별 절대 평가 점수

〈Table 2〉 Scores for students from absolute evaluation

	절대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 1	평가자1	3	2	3	2	3	3	4	4	3	3	5	5	5	4	4
	평가자2	3	2	4	2	3	3	3	4	4	2	4	3	4	3	3
	평가자3	2	2	2	1	3	3	3	3	3	3	5	3	4	4	5
	평가자4	1	1	1	1	2	2	2	2	1	2	4	5	5	4	3

	절대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 2	평가자1	1	2	3	2	1	2	2	4	3	2	5	4	5	4	5
	평가자2	2	3	3	3	3	4	3	4	2	4	5	5	4	4	4
	평가자3	2	2	2	1	3	3	2	2	2	3	3	5	4	5	4
	평가자4	1	1	1	1	1	2	2	2	3	2	4	4	4	4	5

	절대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 3	평가자1	2	2	1	1	2	2	3	3	3	4	4	4	5	4	5
	평가자2	3	2	2	3	3	3	3	4	3	3	4	4	4	3	4
	평가자3	2	1	3	1	2	3	4	3	3	3	5	4	4	4	5
	평가자4	1	1	1	1	1	3	2	2	1	2	4	2	4	3	3

	절대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 4	평가자1	2	3	3	1	1	4	2	4	3	3	5	5	4	5	5
	평가자2	2	3	3	2	2	5	3	5	3	4	4	4	4	5	5
	평가자3	3	1	2	1	1	2	2	2	3	3	5	5	4	5	4
	평가자4	1	2	2	1	1	2	1	3	2	3	4	4	5	5	3

표의 15개 열은 평가 대상이었던 수준별(high, mid, low) 학생들 15명을 나타낸다. 등급 1, 3, 5는 밝기에 따른 회색으로 구분하였다. 〈표 2〉에서 보듯, 네 명의 평가자들은 절대 평가에 있어서 같은 학생한테도 한 등급 혹은 두 등급 이상의 차이를 보여 매우 많은 차이를 보였다. 전체적으로는 등급이 나누어져 있으나, 세부

적으로는 매우 불규칙한 차이를 보이는 것을 알 수 있다. 평균값을 살펴보기 전에 자료 전체를 보는 이유는 평가자별 불규칙성 혹은 일관성의 결여가 학생 한 명에게는 점수 한 등급의 차이와 당락을 결정할 수도 있기 때문이다.

같은 문장들에 대한 상대 평가 결과를 〈표 3〉에 나타내었다. 정도의 차이는 있지만 절대 평가에서와 마찬가지로 매우 불규칙한 점수 분포를 보이고 있다. 평가자에 따라 위로 혹은 아래로 한 등급 차이를 보이는 학생이 다수 보이고 있다.

〈표 3〉 평가자의 학생별 상대 평가 점수

〈Table 3〉 Scores for students from relative evaluation

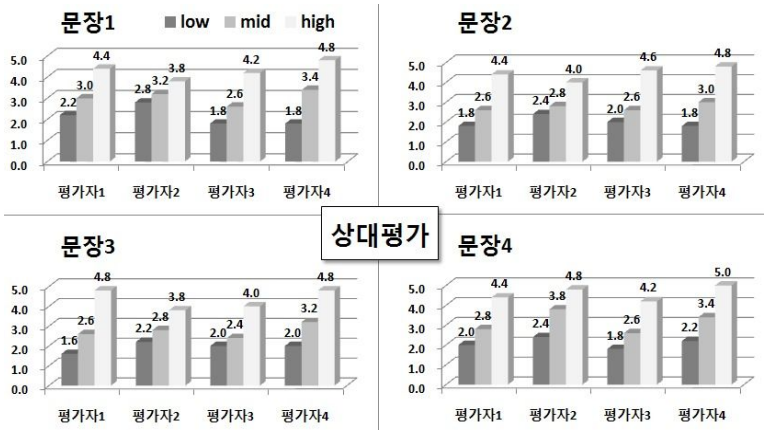
	상대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 1	평가자1	3	2	2	2	2	3	3	4	2	3	5	5	5	3	4
	평가자2	3	2	4	2	3	3	3	4	3	3	5	3	4	4	3
	평가자3	2	1	2	2	2	2	3	3	2	3	5	4	4	4	4
	평가자4	2	2	2	1	2	3	3	4	3	4	5	5	5	5	4

	상대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 2	평가자1	2	2	2	2	1	3	3	3	2	2	5	4	5	4	4
	평가자2	2	2	2	3	3	3	3	3	2	3	4	4	4	4	4
	평가자3	2	2	2	2	2	2	3	3	3	2	4	5	5	4	5
	평가자4	2	2	2	1	2	3	2	3	4	3	5	5	5	4	5

	상대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 3	평가자1	1	1	3	1	2	2	3	3	2	3	5	4	5	5	5
	평가자2	3	2	2	2	2	2	3	3	3	3	4	3	4	4	4
	평가자3	2	2	2	2	2	2	2	3	3	2	5	3	4	4	4
	평가자4	3	3	1	1	2	3	4	3	3	3	5	4	5	5	5

	상대평가	low	low	low	low	low	mid	mid	mid	mid	mid	high	high	high	high	high
문장 4	평가자1	2	3	3	1	1	3	2	3	3	3	5	4	4	5	4
	평가자2	2	3	3	2	2	4	3	5	4	3	5	5	4	5	5
	평가자3	2	3	2	1	1	2	2	3	3	3	5	5	3	4	4
	평가자4	2	4	3	1	1	5	2	4	3	3	5	5	5	5	5

상대 평가 점수에 대한 평가자별 평균값을 막대그래프로 나타내면 〈그림 6〉과 같다. 개별 점수 표에서와 마찬가지로 평균값에 있어서도 같은 등급이라도 평가자별로 대부분 차이가 있음을 볼 수 있다.



〈그림 6〉 상대 평가 등급별 평균값

〈Figure 6〉 Mean scores for relative evaluation

평가자들의 문장별 절대 평가와 상대 평가 점수 분포 사이의 차이를 쌍체 t-검정으로 확인한 결과, 평가자1은 문장1($t(14) = 2.646$, $p < 0.05$))과 문장4($t(14) = 2.256$, $p < 0.05$)에 있어 유의미한 차이가 있었고, 평가자2는 문장2($t(14) = 3.500$, $p < 0.05$)에서만 차이가 있었고, 평가자3은 모든 문장에 있어서 차이가 없었으나, 평가자4는 문장7을 제외하고 모든 문장에 있어서 유의미한 차이를 보였다(문장1: $t(14) = -5.137$, $p < 0.05$; 문장2: $t(14) = -6.205$, $p < 0.05$; 문장3: $t(14) = -6.141$, $p < 0.05$; 문장4: $t(14) = -4.090$, $p < 0.05$; 문장5: $t(14) = -4.026$, $p < 0.05$; 문장6: $t(14) = -3.500$, $p < 0.05$; 문장8: $t(14) = -3.568$, $p < 0.05$). 특히, 평가자4는 거의 모든 문장에서 유의미한 차이를 보이는 가운데, 상대 평가 평균 점수가 절대 평가 평균 점수보다 높은 특징도 나타내고 있다. 나머지 세 평가자들은 몇 가지 경우를 제외하고는 상대 평가와 절대 평가에 있어서 큰 차이를 보이지 않았던 것으로 나타났다.

또한 절대 평가와 상대 평가에 있어서 문장별로 네 평가자들의

평가 점수 분포 차이를 비모수 분산분석인 Kruskal-Wallis 검정을 수행한 결과, 문장3의 절대 평가에 있어서만 $F(3,55) = 3.124$, $p < 0.05$로 평가자들 사이에 유의미한 차이를 나타내었고 나머지 전체에서는 차이를 보이지 않았다.

각 평가자의 문장별 상대 평가와 절대 평가 점수 분포가 통계적으로는 불규칙한 차이를 나타내었고, 각 문장 평가에 있어서는 네 평가자들이 통계적으로 거의 차이를 나타내지 않았으나, 이미 지적한 바와 같이 통계적 경향과 각 학생들의 등급 점수는 본질적인 차이가 있다. 동일한 학생에 대하여 평가자들이 하나 혹은 두 등급 차이로 점수를 부여한다면 일관성이라는 신뢰도에 큰 타격이 불가피하기 때문이다. 단기간 혹은 장기간의 면밀한 평가 훈련을 거치게 되는 경우라도 평가자들이 평가 때마다 매번 동일한 점수를 부여하게 될는지도 의문의 여지가 있다.

3) 평가간 기간에 따른 차이

평가자들이 시간이 지난 후에도 동일한 성향의 평가를 수행하는지 파악하기 위하여, 약 한 달 후에 동일한 학생들 평가 자료에 대하여 상대 평가를 수행하도록 하였다.

한 달 기간 전후의 평가 점수 차이를 평가자별 문장별로 쌍체 t-검정으로 살펴보았더니, 대체로 유사한 평가를 하였으나, 평가자 2는 문장3($t(14) = 3.500$, $p < 0.05$)과 문장5($t(14) = 2.256$, $p < 0.05$)에 있어서 차이가 있었으며, 평가자 3은 문장1($t(14) = 2.432$, $p < 0.05$)과 문장3($t(14) = 2.249$, $p < 0.05$)에 있어서 평가에 차이를 보였다. 나머지 평가자들은 차이가 없었다. 한 달 전후 두 집단 사이의 차이가 없는 경우에도 세부 평가 점수에 있어서는 한 등급 차이가 나는 경우가 매우 많았고, 두 등급 차이가 나는 경우도 종종 보였다.

또한 문장별로 네 평가자들의 평가 점수들을 Kruskal-Wallis 검정으로 살펴보았더니, 전체적으로 평가자들 사이에 유의미한 차이는 없었다. 통계적으로는 네 평가자들이 모두 유사한 평가를 하였다고 볼 수 있으나, 전술한 바와 같이 개별적인 점수에서는 매우 불규칙하고 등급의 차이가 많아, 평가를 받는 학생들의 입장에서는 신뢰성 있는 평가라고 볼 수 없을 수도 있을 것이다.

4. 결론

본 논문에서는 비원어민 평가자들이 영어 발음을 평가하는데 있어 어떠한 평가 성향을 지니고 있는지를 두 가지 측면에서 살펴보았다. 우선 Yoon(2013ab)에 따라 평가 대상이 되는 한국인 학생들의 모든 문장을 모범 답안 격인 원어민 문장들과 함께 분절음과 초분절음 요인별로 자동 평가 점수를 컴퓨터로 미리 계산하였고, 이를 수동 평가 점수와 비교하여 평가자들이 각각 어느 요인들에 비중을 두는지 유추해낼 수 있었다. 그 다음은 평가자들이 절대 평가와 상대 평가를 수행한 수동 평가 점수들을 이용하여 평가자들의 평가 일관성이나 평가 점수 분포의 불규칙성 등의 양상을 알아냈다.

발화된 영어 문장의 분절음을 원어민과 한국인과의 스펙트럼 비교를 통해 유클리드 거리로 계량화하고, 초분절음은 원어민과 한국인의 F0, 길이, 강도를 각각 비교하여 유클리드 거리로 계량화한 다음, 평가자들이 부여한 수동 평가 점수별로 이 네 가지 음성학적 요소를 레이더 차트 위에 집단별 표시를 하게 되면, 각 평가자들이 어떠한 요소에 비중을 더 두어 발음 평가를 수행했는지 밝혀낼 수 있다는 점은 이 연구의 매우 유용한 성과 중의 하나라고 말할 수 있다. 그 결과, 정도의 차이는 다소 있지만, 실험에 참여

한 평가자들은 대체로 분절음보다는 초분절음에 비중을 더 두는 평가를 수행했다고 볼 수 있었다.

절대 혹은 상대 평가에 의한 평가자들의 수동 평가 점수를 살펴본 결과, 평가자들은 대체로 두 평가 방식에 있어 유의미한 차이를 보이지 않았다. 이는 등급별 구체적인 학생 수나 백분율과 같은 제한이 주어지지 않으면 평가자들은 절대 혹은 상대 평가에 대한 막연한 개념은 있으나, 실제 평가에 있어 구체적이고 명확한 구분을 하지 않는다는 사실은 반영하는 것이라 볼 수 있을 것이다.

절대 혹은 상대 평가별로 평가자들의 수동 평가 점수 분포를 비교 분석한 결과, 통계적인 경향은 볼 수 있으나 구체적인 점수에 있어서는 매우 불규칙한 분포를 보이고 있었다. 같은 학생이라도 한두 등급이 높거나 낮게 평가되는 경우가 매우 많아 평가자들에 의한 수동 평가 신뢰성이나 일관성에 의문이 제기되는 것으로 나타났다. 학생들이 자신의 평가 등급에 이의를 제기할 경우 난감한 상황이 벌어질 가능성이 매우 높다. 같은 문장을 여러 평가자가 평가하여도 평가자마다 학생별 평가 등급 점수가 조금씩 다르게 나타났으며, 이는 평균값에도 반영되었다.

한 달의 기간이 지난 후에 평가자들에게 동일한 자료에 대한 상대 평가를 수행한 경우에도 유사한 결과를 얻을 수 있었다. 전체적으로는 통계적으로 유사한 평가를 했다고 볼 수 있지만, 개별 문장이나 학생별로는 불규칙하고 일관적이지 못한 점수 평가가 이루어지는 것으로 나타났다.

이러한 모든 연구 결과를 고려하여 볼 때, 영어 발음 평가를 비원어민인 교수자가 수동으로 수행할 경우에, 미리 면밀하게 마련된 평가 지침에 근거한 집중적인 훈련이 평가자들에게 필요할 것으로 생각된다. 특히, 평가 훈련에 참여할 비원어민 영어 교수자들의 평가 성향을 미리 파악하여 같은 성향을 지닌 평가자들만을 선별하거나, 혹은 미리 파악된 평가자 평가 성향을 평가 지침에 맞

도록 변화시키는 추가적인 교육이나 훈련이 필요할 것이다. 궁극적으로는 사람 평가자의 평가 행위를 거의 완벽하게 학습하여 수행할 수 있는 인공 지능 개발이 필요할 것으로 사료된다.

참고문헌

1. 논문

Boersma, P. 「Praat, a system for doing phonetics by computer」. Glot International, 5(9/10), 341–345. Byrd, D. (1993). 54,000 American stops. UCLA Working Papers in Phonetics, 83, 2001, pp. 97–115.

Chung, H., Jang, T. Y., Yun, W., Yun, I., & Sa, J.. 「A study on automatic measurement of pronunciation accuracy of English speech produced by Korean learners of English」. Language and Linguistics 42, 2008, pp. 165–196.

Kang, S. 「The study on Korean raters' characteristics for Korean English oral performance」. Studies in Linguistics 26, 2013, pp. 1–21.

Kang, S. & Rhee, S–C. 「A study on the suprasegmental parameters exerting an effect on the judgment of goodness or badness on Korean–spoken English」. Phonetics and Speech Sciences 3(2), 2011, pp. 3–10.

Kang, S–H. & Ahn, H–K. 「A comparative study on criteria and tasks in Korean English speaking assessment by native and non–native raters」. Language Research 48(2), 2012, pp. 1–25.

Kim, Y–H. 「An investigation into native and non–native teachers' judgments of oral English performance: A mixed methods approach」. Language Testing, 26(2), 2009, pp. 187–217.

Lowenberg, P. H. 「Assessing English proficiency in the global context: The significance of non–native norms. In Kam, H. W., editors. Language in the global context: Implications for the language classroom」. Singapore: SEAMEO Regional Language Center,

2000, pp. 207–229.

Park, H. 「A study of an independent evaluation of prosody and segmentals: with reference to the difference in the foreign accent of Korean, Chinese, and Japanese learners of English」. Phonetics and Speech Sciences 4(4), 2012, pp. 37–43.

Rhee, S–C., Lee, S–H., Kang, S–K., & Lee, Y–J. 「Design and construction of Korean–spoken English corpus (K–SEC)」. Malsori 46, 2003, pp. 159–174.

Ryu, H. & Chung, M. 「Automatic pronunciation assessment of English spoken by Korean learners using phone–level articulatory posterior probability」. Proceedings of the 2016 Spring Conference of the Korean Society of Speech Sciences, 2016, pp. 101–102.

Yoo., E–M. 「A comparative study on assessing oral tests between native and non–native English teachers in Korean secondary schools」. M.A. thesis. Korea National University of Education, Chung–buk, Korea. 2014.

Yoon, K. 「A study on human evaluators using the evaluation model of English pronunciation」. Phonetics and Speech Sciences 5(4), 2013a, pp. 109–119.

Yoon, K. 「A study on the relationship between manual and automatic segmental evaluation through spectral comparisons」. The New Studies of English Language & Literature 55, 2013b, pp. 241–261.

Yun, W. 「An analysis of the Korean inter–rater difference in evaluating English pronunciations of Korean speakers」. Studies in Foreign Language Education 23(2), 2009, pp. 85–103.

Yun, W. 「The objectives of English pronunciation evaluations and the usability of machine scoring」. The Journal of Linguistic Science 61, 2012, pp. 167–184.

(투고일: 2021. 6. 23 심사완료일: 2021. 9. 16 게재확정일: 2021. 9. 24)

성명: 박상진, 제1저자(주저자)
소속: 영남대학교 영어영문학과
주소: 경북 경산시 대동 214-1 영남대학교
전자우편: parkesl@hanmail.net

성명: 이유리, 제2저자
소속: 영남대학교 영어영문학과
주소: 경북 경산시 대동 214-1 영남대학교
전자우편: yuri2077@hanmail.net

성명: 윤여신, 제3저자
소속: 대구대학교 영어영문학과
주소: 경북 경산시 대구대로 201 대구대학교
전자우편: yys5752@hanmail.net

성명: 윤규철, 교신저자
소속: 영남대학교 영어영문학과
주소: 경북 경산시 대동 214-1 영남대학교
전자우편: kyoon@ynu.ac.kr

[Abstract]

A Study on the Korean Evaluators' Tendency and Consistency for Learners' English Pronunciation

Park, Sang-jin · Lee, Yu-ri · Yoon, Yeo-shin · Yoon, Kyu-chul

The purpose of this work was to examine the evaluation tendency of the non-native evaluators with respect to the phonetic factors, the distribution pattern of evaluation scores from the absolute and relative evaluation, and the evaluation consistency of the evaluators between the two evaluations that were performed one month apart. From the comparison of computerized automatic scores and manual scores, we found that the evaluators tend to put more emphasis on the suprasegmental factors than on the segmental factor. Contrary to the statistical analyses, the actual distribution of the manual scores by the evaluators was irregular and inconsistent. The same type of relative evaluation over a one-month period produced manual scores that were also irregular and inconsistent.

Key words: English, Korean evaluators, evaluation tendency, evaluation consistency