ISSN 2005-8063
2016. 12. 31.
Vol.8 No.4
pp. 103-113

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2016.8.4.103

조음자질을 이용한 한국인 학습자의 영어 발화 자동 발음 평가*

Automatic pronunciation assessment of English produced by Korean learners using articulatory features

류 혁 수 · 정 민 화**

Ryu, Hyuksu · Chung, Minhwa

Abstract

This paper aims to propose articulatory features as novel predictors for automatic pronunciation assessment of English produced by Korean learners. Based on the distinctive feature theory, where phonemes are represented as a set of articulatory/phonetic properties, we propose articulatory Goodness-Of-Pronunciation(aGOP) features in terms of the corresponding articulatory attributes, such as nasal, sonorant, anterior, etc. An English speech corpus spoken by Korean learners is used in the assessment modeling. In our system, learners' speech is forced aligned and recognized by using the acoustic and pronunciation models derived from the WSJ corpus (native North American speech) and the CMU pronouncing dictionary, respectively. In order to compute aGOP features, articulatory models are trained for the corresponding articulatory attributes. In addition to the proposed features, various features which are divided into four categories such as RATE, SEGMENT, SILENCE, and GOP are applied as a baseline. In order to enhance the assessment modeling performance and investigate the weights of the salient features, relevant features are extracted by using Best Subset Selection(BSS). The results show that the proposed model using aGOP features outperform the baseline. In addition, analysis of relevant features extracted by BSS reveals that the selected aGOP features represent the salient variations of Korean learners of English. The results are expected to be effective for automatic pronunciation error detection, as well.

Keywords: articulatory features, automatic pronunciation assessment, computer-assisted pronunciation training, Korean learners of English

1. 서론

최근 음성 기술이 발달함에 따라, 음성 인식을 통해 학습자의 발화를 인식하여 발음 교육에 활용하는 '컴퓨터를 이용한 발음 훈련 (CAPT; Computer-Assisted Pronunciation Training)'이 각광을 받고 있다(Eskenazi, 2009). 효과적인 CAPT를 위해서는 음성 기술을 활용한 쌍방향 교육 기능이 강화될 필요가 있다. 다시 말해, 컴퓨터가 학습자의 발화를 듣고 인식하여, 인식된 학습자의 발화에 대해 학습자의 발음 수준을 자동으로 평가하고, 발음에서 오류를 검출하고, 해당 오류에 대해 어떻게 수정하여야 하는지 교정 피드백을 제공할 수 있어야 한다.

기존의 CAPT 관련 연구에서 학습자 개인이 발화한 단어나 문장 전체에 대한 전역 점수(global score)를 계산하여 학습자의 숙련도를 평가한다(Eskenazi, 2009). 이 때, 발화 속도나 분절음/

* 이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2015R1D1A2A01061378)
** 서울대학교, mchung@snu.ac.kr, 교신저자
Received 3 November 2016; Revised 2 December 2016; Accepted 7 December 2016

휴지의 길이 등과 같은 다양한 음향 특질을 추출하고 이를 조합하여 전역 점수를 계산할 수 있다(Cucchiarini *et al.*, 2000a, 2000b, 2002; Cincarek *et al.*, 2009; Zechner *et al.*, 2009). 그 외에도 점수 계산을 위한 자질로서, 원어민 화자의 음향 모델로부터 로그 사후확률 점수와 분절음의 지속 시간 점수를 사용하기도 한다(Franco *et al.*, 1997; Neumeyer *et al.*, 2000).

기존의 영어 학습자 발음 자동 평가 시스템으로는 ETS(Education Testing Service)의 SpeechRater™(Zechner *et al.*, 2009)와 Pearson의 Versant system(Downey *et al.*, 2008) 등이 있다. SpeechRater™는 TOEFL®의 말하기 시험을 위한 자동 평가 시스템으로, 현재는 TOEFL® Practice Online(TPO)의 말하기 평가에 실제로 사용되고 있다. Zechner *et al.*(2009)의 연구에서 ETS는 비원어민 영어 학습자들의 즉흥 발화 발음 평가를 위해 29개의 자질들을 도입하였다. Zechner *et al.*(2009)에서 사용한 자질들의 대부분은 발화 단어나 휴지의 길이나 빈도와 같은 유창성과 관련이 있다. 또한, 즉흥 발화의 평가에 초점을 맞추었기 때문에, 즉흥 발화의 특성인 비유창성(disfluency)이나 반복 발화도 고려하였다. 이들 연구에서는 다중 회귀 분석과 분류 및 회귀 나무(Classification And Regression Tree; CART)를 점수 예측에 적용하여 수동 평가 점수와 0.57의 상관 계수를 보였다.

Pearson의 Versant 시스템(Downey *et al.*, 2008)은 영어 학습자의 발음을 자동으로 평가하기 위해 발화 내용, 발화 길이, 음향 특성과 관련된 자질을 개발하였다. 발화 내용에 관련된 자질은 학습자가 발화하는 어휘 종류 등을 고려하여 은닉 의미 분석(Latent Semantic Analysis; LSA)을 이용하여 계산한다. 길이 관련 자질은 원어민의 분절음 길이 분포와 학습자 발화 음소의 길이를 비교하여 계산하며, 음향 특성 관련 자질은 원어민과 학습자 음향 모델에서의 로그 우도(log-likelihood) 차이를 이용한다. Downey *et al.*(2008)의 연구에서 이들 자질을 이용하여 신경망 회귀로 발음 점수를 예측한 결과 수동평가점수와 0.826의 상관 계수를 보였다.

Ryu *et al.*(2016)의 연구에서는 외국인 한국어 학습자의 발음 평가를 위하여 기존 연구들에서 제시된 음향 특질들을 그 특성에 따라 범주화하고 선형회귀를 이용하여 점수를 예측하였다. 그 결과, 수동 평가 점수와 0.895의 상관 계수를 보였으며, 분절음이나 휴지의 길이와 관련된 자질들 보다는 발화의 속도와 관련된 자질이 평가 점수 예측에 더 큰 영향력을 가짐을 밝혔다.

하지만, 이와 같은 음향 특질에 의한 발음 자동 평가는 평가 점수만을 산출해 줄 뿐, 기존의 연구에서 평가 점수 계산에 사용하는 수치화된 음향 자질들로는 학습자들에게 교정 피드백 정보를 제공해 주기 어렵다는 한계가 있다. 예를 들어, Zechner *et al.*(2009)에서 제시한 자질 가운데 휴지 길이의 표준 편차(Silmeandev; Mean deviation of Silence duration)의 경우, 문장 전체에 대해서 전역적인 점수만을 제공하지만, 이를 통해 학습자가 어떻게 발화 또는 발음을 수정하면 좋을지에 대해서 정보를 얻기가 힘들다.

이러한 문제를 해결하기 위해서는 학습자들에게 발음에 대한 교정 피드백 정보를 제공할 수 있는 자질을 이용하여 발음을 평가하고, 오류를 검출하여 피드백을 제공하는 통합적인 CAPT 시스템이 필요하다. 학습자들에게 교정 피드백을 제공하기 위해서는 먼저 학습자의 발화에서 발음 오류를 검출할 수 있어야 한다. 이 때 효과적인 교정 피드백을 위해서는 학습자의 수준에 따라 학습자 맞춤형의 피드백을 제공할 필요가 있다. 예를 들어, 초급 학습자에게는 기본이 되는 주요 오류 사항에 대해서만 교정 피드백 정보를 제공하고, 고급 학습자에게는 주요 오류 외에도 다양한 오류 현상에 대해서도 교정 피드백 정보를 제공하는 방식으로 CAPT 시스템을 구성할 수 있다. 초급 학습자들은 주요 발음 오류의 교정이 보다 시급한 과제인 반면, 학습자의 수준이 고급에 가까울수록 발음 오류 빈도가 감소(Hong *et al.*, 2010; 홍혜진 외, 2014)하기 때문에 발음 수준 향상을 위해 주요 발음 오류 외에 다른 발음 오류들에 대한 정보를 함께 제공해 주는 것이 효과적이다. 한편, CAPT 시스템을 사용할 때에 검출된 발음 오류에 따른 교정 피드백 외에도, 발음 수준이나 발음 평가 점수와 같은 수치 정보도 함께 제공받는다면, 학습자의 현재 발음 수준을 확인할 수 있으므로 학습자 입장에서도 유용할 수 있다. 따라서 발음 오류 검출 및 그에 따른 교정 피드백 뿐 아니라, 학습자의 발음 평가 역시 CAPT에서 큰 역할을 할 수 있다.

이에 본 연구는 발음 자동 평가를 위한 새로운 자질로서 조음 특성에 기반한 발음적합점수(articulatory Goodness-Of-Pronunciation; aGOP)를 제안하고, 이를 한국인 영어학습자의 낭독체 발음 자동 평가에 적용하여 평가 모델링 성능을 향상시키는 것을 목적으로 한다.

조음 특성 기반의 자질은 사람이 음성을 인식하는 행위의 매커니즘을 그대로 흉내낼 수 있고, 조음 특성을 이용함으로써 발화에 대한 지식을 통합할 수 있다는 장점이 있다(Lee, 2004). 또한, 동시 조음과 동화 현상을 모델링하는데 용이하며(Kirchhoff *et al*., 2002; Richardson *et al*., 2003; Lee *et al*., 2007), 언어 차이(Siniscalchi *et al*., 2008) 및 과도 조음(Metze, 2005)에 강건하다는 장점이 있다. 이러한 장점을 이용하여, 음성 인식에서 조음 특성을 이용하려는 시도들이 있었다(Kirchhoff *et al*., 2002; Richardson, 2003; Lee, 2004; Metze, 2005; Lee *et al*., 2007; Siniscalchi *et al*., 2008). 뿐만 아니라, CAPT에서도 조음 자질을 이용하려는 시도들도 있었다(Tepperman & Narayanan, 2008; Li *et al*., 2016; 류혁수 & 정민화, 2016).

Tepperman & Narayanan(2008)에서는 조음기관의 움직임을 턱의 벌어짐, 입술의 떨어짐과 같이 8가지로 나누고 이들을 수치화시켜서 조음 특성을 모델링함으로써, 비원어민 학습자의 발화 오류를 검출하고자 하였다. 하지만, 조음 기관의 움직임을 물리적으로 측정하여 값을 할당하지 않고, 음소별로 조음 기관의 위치값을 일괄적으로 부여하였다는 문제가 있다. 또한 연속적으로 움직이는 조음기관의 움직임에 대해 범주화된 값을 설정하기 때문에, 중간값의 할당이 자의적으로 결정될 수 밖에 없다는 문제가 있다.

Li *et al*.(2016)의 연구에서는 조음 자질을 조음 위치, 조음 방법, 기식성, 유성성의 4가지 범주로 나누고, 각 범주별로 해당하는 조음 속성들을 이용하여 음소들을 구분하였다. 조음 속성들

을 이용하여 음향 모델링을 시행한 후 각 분절음 별로 발음 점수를 계산하고, 이를 바탕으로 CART를 통해 음소 오류를 검출하고자 하였다. 그 결과, 조음 속성들을 자질로 이용함으로써 약 96%의 정확도로 음소 오류를 구분할 수 있었으며, 오류 검출에 사용한 조음 속성의 결과를 통해, 교정 피드백을 제공할 수 있음을 보였다. 하지만, 이 연구에서는 /m, n, l, r/과 같은 공명음들이 유성성 범주에서 유성음과 무성음 속성에 모두 포함되어 있는 등, 조음 자질 구분이 언어학적 지식과 일부 합치되지 않는 문제가 있었다. 또한, 학습자 발음 오류에 대한 언어학적 고려없이 단순히 결정나무의 뿌리에 가까울수록 해당 조음 속성이 발음 오류에서 더욱 중요한 역할을 한다고 판단하는 한계가 있었다. 예를 들어, Li *et al*.(2016)의 연구 결과에 따르면, /s/의 발음 오류에서 Fricative 자질이 Alveolar 자질보다 뿌리에 가깝게 위치한다. 하지만 , Fricative와 Alveolar 둘 다 /s/의 발음을 결정짓는 중요 자질이며, 그 둘 중 어느 하나가 더 높은 중요도를 갖는다고 말하기 어렵다.

조음자질을 이용하여 음성 인식이나 CAPT의 발음 오류 검출에 사용하고자 하는 연구들은 있었으나, 조음 자질을 이용한 자동 발음 평가에 대해서는 그동안 많은 연구가 이루어지지 않았다. 특히, 한국인의 영어 발화를 대상으로 한 자동 발음 평가와 관련해서는 현재까지 연구가 드물게 이루어졌다. 앞서 언급한 바와 같이, 학습자들에게 효과적인 피드백을 제공하기 위해서는 학습자의 수준에 따라 그 피드백 내용이 달라져야 할 필요가 있으므로, 발음 오류 검출 및 교정 피드백에 앞서 자동 발음 평가가 선행될 필요가 있다. 이와 관련하여, 류혁수 & 정민화(2016)의 연구에서는 한국인의 영어 발화에 대해 자음의 조음 자질 기반의 사후확률을 평가자질로 이용하여 평가 모델링을 수행하였다. 그 결과, 기존의 음향 특성 자질 뿐 아니라 조음 자질을 이용함으로써 평가 모델링 성능을 향상시킬 수 있음을 보인 바 있다. 본 연구는 류혁수 & 정민화(2016)의 후속 연구로서, 음성/음운론적 지식을 이용하여 조음 특성에 기반한 평가자질(aGOP)를 제안하고, 통계적인 방법론을 적용하여 평가 점수에 영향을 미치는 주요 자질들을 분석하고자 한다. 본 연구에서는 학습자의 발음을 평가하고 오류를 검출하여 피드백을 제공하는 통합적인 CAPT 프레임워크를 제안하기에 앞서, 먼저 학습자의 발음 수준을 평가하는 자동 평가로 그 범위를 한정한다.

2. 조음 자질

음소는 의미를 구분해 주는 최소 단위이다. 하지만 음소 역시 음성적 특질을 이용하여 음소를 나눌 수 있다(Hayes, 2009). 예를 들어, /p/와 /b/이라는 음소가 있을 때, 유성성(voicing)이라는 음성적 특질의 존재 유무로 두 음소를 구분할 수 있다. 유성성이 있으면([+voice]) /b/가 되고, 유성성이 없으면([-voice]) /p/가 된다. 즉, /p/와 /b/를 '변별적으로' 구분해 주는 음성적 기준은 유성성이라고 할 수 있다. 이와 같이, 한 언어에 존재하는 두 음소를 구분해 주는 최소 단위인 음성적 특질을 변별적 자질(distinctive feature)이라고 한다. 따라서 음소는 변별적 자질들의 집합(자연 부류, Natural class)으로 표현할 수 있다(전상범 2004).

이 때 음소를 표현하기 위해 사용하는 변별적 자질들은 +와 –의 2개의 값만 갖도록 한다(Chomsky & Halle, 1968). Chomsky & Halle(1968)은 이진값을 갖는 변별적 자질들을 여러 개 사용함으로서 다양한 음소들을 구분할 수 있다고 보았다. 예를 들어, /p/와 /d/는 각각 무성양순폐쇄음과 유성치경폐쇄음으로써, 공통적으로 [+consonantal, -sonorant, -delayed release]의 속성을 가지지만, labial과 voice 자질에서 차이를 갖는다. /p/는 [+labial, -voice]의 속성을 갖는 반면, /d/는 [-labial, +voice]의 속성을 갖는다. 이와 같이, 이진값을 갖는 조음/음성적 속성의 변별적 자질을 사용함으로써 음소들을 구분할 수 있다. Hayes(2009)의 연구에서는 존재 여부의 이진값 뿐 아니라, 경우에 따라서는 관계없음을 나타내는 값을 포함하는 3진값으로 음소를 분류하기도 한다. 예를 들어, /m/는 비음으로 기류가 비강을 통해 지속적으로 흘러나가기 때문에, 구강에서의 지연 폐쇄를 나타내는 delayed release라는 속성과는 관계가 멀다. 그렇기 때문에 Hayes(2009)에서는 /m/은 [0delayed release]인 것으로 본다.

표 1. 범주 별 조음 속성 및 해당 음소 목록
Table 1. Articulatory attributes and the corresponding phonemes in terms of categories

Category	Attribute	Phonemes
Manner (9)	Consonantal	p, b, m, f, v, th, dh, t, d, s, z, n, l, ch, jh, sh, zh, r, y, k, g, ng, hh, w
	Sonorant	m, n, l, r, y, ng, w, iy, uw, ih, uh, eh, ow, ah, ao, ae, aa, aw, ay, ey, oy, er
	Continuant	f, v, th, dh, s, z, l, sh, zh, r, y, hh, w, iy, uw, ih, uh, eh, ow, ah, ao, ae, aa, aw, ay, ey, oy, er
	Delayed release	f, v, th, dh, s, z, ch, jh, sh, zh
	Approximant	l, r, y, w, iy, uw, ih, uh, eh, ow, ah, ao, ae, aa, aw, ay, ey, oy, er
	Nasal	m, n, ng
	Stop	p, b, t, d, k, g
	Fricative	f, v, th, dh, s, z, sh, zh
	Affricate	ch, jh
Place (14)	Labial	p, b, m, f, v, uw, uh, ow, ao, aw, oy
	Round	w, uw, uh, ow, ao, aw, oy
	Labiodental	f, v
	Coronal	th, dh, t, d, s, z, n, l, ch, jh, sh, zh, r, er
	Anterior	th, dh, s, z, n, l, er
	Distributed	th, dh, ch, jh, sh zh, r, er
	Strident	s, z, ch, jh, sh, zh
	Lateral	l
	Dorsal	y, k, g, ng, w
	High	y, k, g, ng, w, iy, uw, ih, uh, ay, ey, oy
	Low	ae, aa, aw, ay
	Front	y, iy, ih, eh, ae, ay, ey, oy
	Back	w, uw, uh, ow, ah, ao, aa, ay, aw, oy
	Tense	y, w, iy, uw, eh, ow, oy, ey, er
Laryngeal (1)	Voice	b, m, v, dh, d, z, n, l, jh, zh, r, y, g, ng, w, aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw,

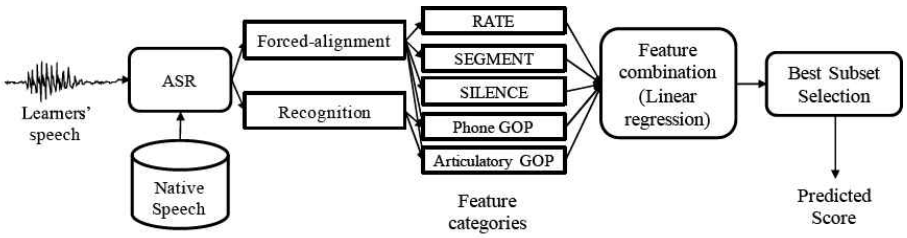

그림 1. 자동 발음 평가 모델링 구조도
Figure 1. Pronunciation assessment modeling framework

비원어민인 학습자의 발음과 원어민이 발화하는 발음의 차이는 조음 방식의 차이에서 기인한다. 따라서 컴퓨터를 이용한 발음 교육에서 발음 평가를 시행할 경우, 학습자의 조음이 원어민의 조음과 작은 차이를 보이면, 발음이 유사하다고 판단할 수 있으므로 평가 점수가 높아지고, 반대로 학습자와 원어민의 조음이 상이할수록 평가 점수가 낮아질 것으로 예상할 수 있다. 이와 같은 점에 착안하여, 본 연구에서는 Chomky & Halle(1968)와 Hayes(2009)의 연구를 기반으로 변별적 자질로 나타나는 조음 특성을 이용하여 자동 발음 평가 모델링을 시행하고자 한다. Hayes(2009)에서는 조음 방법, 조음 위치, 후두부에서의 성문 움직임을 기준으로 하여, 조음 특성을 크게 3가지 범주(조음 방법, 조음 위치, 후두 특질)로 구분한다. 각각의 범주에 따른 조음 속성 및 해당 속성이 +로 나타나는 음소들의 목록은 <표 1>과 같다. <표 1>에서의 음소 집합은 CMU pronouncing dictionary v0.7b(Weide, 2014)의 CMU 39 유사 음소 단위(Phoneme-Like Unit; 이하 PLU)의 표기를 따랐다.

Witt & Young(2000)에서는 외국어 학습자의 발음 오류 검출을 위해, 발음적합점수(Goodness-Of-Pronunciation; GOP)의 개념을 제안한 바 있다. GOP는 음소에 대한 정규화된 사후 확률(normalized posterior probability)로 정의하며(Witt & Young, 2000), 학습자가 발화한 음소가 원어민 음향모델에서 떨어져 있는 정도를 계산할 수 있다. 따라서 학습자의 발음이 원어민의 기준 발음과 차이가 많이 날수록 GOP가 커지는 특성을 보인다. 본래 GOP는 개별 발음 오류를 탐지하기 위해 제안된 자질이나(Witt & Young, 2000), 전체 문장에 대해 전역 범위에서 점수를 계산할 경우, GOP를 발음 평가를 위한 자질로 사용할 수 있다(Neumeyer *et al*., 2000; Shi *et al*., 2016). 일반적으로 GOP가 클수록 평가 점수가 하락하는 음의 상관관계를 보인다. GOP는 다음의 식 (1)과 같이 계산할 수 있다.

$$\begin{aligned} GOP &\equiv \frac{|\log\ p(q_i|O_i)|}{N_i} \\ &\approx |\frac{\log\ p(O_i|q_i)}{N_i} - \frac{\log\ \max_{j=1}^{J} p(O_i|q_j)}{N_i}| \\ &= |p_{q_i(\text{forced})} - p_{q_i(recognition)}| \end{aligned} \quad (1)$$

이 때, N_i와 $p(O_i|q_i)$는 각각 관측값인 O_i를 구성하는 프레임의 개수와 음소 q_i가 주어졌을 때 관측값 O_i가 나타나는 우도를 의미한다. $\frac{\log(p(O_i|q_i)}{N_i}$가 음소 q_i가 주어졌을 때 관측값 O_i가 나타날 평균 로그 우도를 의미하므로, 이는 음소 q_i에 대한 강제 정렬 확률 $p_{q_i(\text{forced})}$에 해당한다. 또한, $\frac{\log\ \max_{j=1}^{J} p(O_i|q_j)}{N_i}$는 모든 가능한 J개의 음소에 대해서, 음소 q_j가 주어졌을 때 관측값 O_i이 나타나는 우도 가운데 가장 큰 값의 평균 로그 우도를 의미한다. 이는 자유 음소 인식에 의한 확률 $p_{q_i(recognition)}$로 해석할 수 있다. 따라서 GOP는 특정 구간의 프레임에 대해 강제 정렬에 의한 우도와 음소 인식에 의한 우도의 차이의 절대값으로 계산된다.

본 연구에서는 자동 발음 평가를 위한 자질로 조음 특성에 기반을 둔 조음 기반 GOP(articulatory GOP; aGOP[1])를 제안한다. aGOP는 원어민과 학습자의 조음을 비교하기 위해서 위의 <표 1>에서 구분한 조음 속성에 대해, 각 조음 속성 별로 학습자의 발화에서 해당 조음 속성이 나타나는지의 확률을 다음의 식 (2)와 같이 계산할 수 있다.

$$\begin{aligned} aGOP_i^k &\equiv \frac{|\log\ p(q_i^k|O_i)|}{N_i} \\ &\approx |\frac{\log\ p(O_i|q_i^k)}{N_i} - \frac{\log\ \max_{j=1}^{J} p(O_i|q_j^k)}{N_i}| \\ &= |p_{q_i^k(\text{forced})} - p_{q_i^k(recognition)}| \end{aligned} \quad (2)$$

이 때, k는 <표 1>에서 제시된 조음 속성의 종류를 나타내고, q_i^k는 i번째 분절음 위치에서의 k번째 조음 속성의 값을 의미하며, + 또는 -의 이진값을 갖는다. N_i와 $p(O_i|q_i^k)$는 각각 관측값인 O_i를 구성하는 프레임의 개수와 q_i^k가 주어졌을 때 관측값 O_i가 나타나는 우도를 의미한다. 또한, 식 (1)에서 강제 정렬과 음소 인식의 우도 차이로 GOP를 계산한 것과 마찬가지로, aGOP 역시 q_i^k에 대한 강제 정렬 우도 $p_{q_i^k(\text{forced})}$와 자유 음소 인식에 의한

1 본 연구에서는 음소 단위의 GOP와 조음 자질 기반의 GOP를 구분하기 위하여, 음소 단위의 GOP는 GOP로, 조음 자질 기반의 GOP는 aGOP로 명명하기로 한다.

우도 $p_{q_i^k(recognition)}$의 차이의 절대값으로 계산할 수 있다.

앞서 1장에서 언급한 바와 같이 Li *et al*.(2016)의 연구에서도 발음 오류 자동 검출을 위해 조음 자질을 이용하고 있다. 하지만, Li *et al*.(2016)의 연구는 두자음(onset)으로 연구 범위를 한정하고 있으며, 1장에서 지적한 바와 같이 속성 분류에 있어서도 일부 언어학적 오류가 있었다. 본 연구는 먼저 모든 자/모음의 음소를 대상으로, 조음 자질을 자동 발음 평가에 적용하고 있다는 점에서 차별점을 갖는다. 이는 향후 조음 자질을 이용하여 발음 오류 검출 및 교정 피드백을 포함하는 통합적인 CAPT 프레임워크 연구를 위한 초기 연구로서의 성격을 갖는다. 또한, Hayes(2009)의 기준에 따라 음성학/음운론 지식에 의거하여 조음 속성을 보다 상세히 구분함으로써, 자동 발음 평가 모델의 자질로 반영한다. 이를 통해, 학습자의 발화로부터 보다 다양한 조음 정보를 도출하여 자동 발음 평가에 활용할 수 있다.

3. 실험 방법

3.1. 자동 발음 평가 모델링 구조도

한국인 영어학습자의 발음 자동 평가를 위해, 자동 발음 평가의 전체적인 구조를 <그림 1>과 같이 나타내었다. <그림 1>에서 보는 바와 같이, 북미 영어 모국어 화자의 음향 모델을 이용하여, 한국인 영어학습자의 발화에 대한 강제 정렬 및 음성 인식을 시행한다. 그 결과를 통해, 발음 평가를 위한 자질들을 계산한다. 본 연구에서 사용하는 자질들의 종류와 그 특성에 대해서는 3.2에서 상세하게 설명할 예정이다. 자질들을 추출한 후 자질 결합(Feature combination) 단계를 거치게 된다. 이 단계에서는 다중 선형 회귀와 같은 통계적인 방법을 통해서, 학습자의 발음 점수를 예측한다. 하지만, 선형회귀의 경우, 설명 변수들 사이의 독립이 가정되지 않을 경우 잉여성이 발생할 수 있다. 따라서 잉여성을 최소화하고, 관련성이 높은 주요 자질들을 추출하기 위해서 최량 부분 집합 선택(Best subset selection; 이하 BSS)과 같이 주요 자질을 선택하는 과정을 거쳐서 최종적으로 발음 점수를 예측하게 된다.

3.2. 평가 자질

자동 발음 평가를 위해, 본 연구에서 제안하는 조음 자질 기반의 aGOP 외에도 다양한 평가 자질들이 여러 연구에서 사용되어 왔다(Cucchiarini *et al*,, 2000a, 2000b, 2002; Zechner *et al*., 2009). 본 연구에서는 통계 모델링과 결과 분석을 위해, 기존 연구들에서 제안된 자질들을 그 특성에 따라 발화 속도와 관련된 자질(RATE), 분절음의 길이와 관련된 자질(SEGMENT), 휴지의 길이와 관련된 자질(SILENCE), 음소 기반의 GOP(GOP)의 4개 범주로 구분하였다. 여기에 본 연구에서 제안하는 조음 기반의 GOP인 aGOP 자질까지 포함하면 총 5개 범주가 된다. 각 범주별 포함된 자질의 종류와 그에 따른 상세 설명은 다음의 <표 2>와 같다.

표 2. 자동 발음 평가에 사용되는 자질 종류 세부 설명
Table 2. Descriptions of features for automatic pronunciation assessment

Categ.	Feature	Description
aGOP (24)	aGOP	조음 자질 기반 Goodness-Of-Pronunciation
GOP (1)	GOP	음소 기반 Goodness-Of-Pronunciation
(5) RATE	ROS	발화 속도 (음소) Rate-Of-Speech
	AR	조음 속도 (음소) Articulation Rate
	PTR	음소-시간 비 Phonation-Time Ratio
	Wpsec	조음 속도 (단어) Words Per SECond
	Wpsecutt	발화 속도 (단어) Words Per SECond in UTTerance
(6) SEGMENT	Globsegdur	휴지를 포함한 모든 분절음 길이 GLOBal SEGment DURation
	Segdur	휴지를 제외한 분절음의 길이 SEGment DURation
	Wdpchk	chunk당 평균 단어 개수 Words Per Chunk
	Secpchk	chunk당 평균 발화 길이 Seconds Per Chunk
	Secpchkmeandev	chunk당 평균 발화길이 평균 절대편차 SEConds Per Chunk MEAN DEViation
	Wdpchkmeandev	chunk당 평균 단어개수 평균절대 편차 Words Per Chunk MEAN DEViation
(11) SILENCE	Numsil	휴지 개수 NUMbers of SILence
	Silpwd	단어당 평균 휴지 개수 SILences Per Word
	Silpsec	초당 평균 휴지 개수 SILences Per SECond
	Silmean	평균 휴지 길이 SILence MEAN
	Silmeandev	휴지 길이 평균 편차 SILence MEAN DEViation
	Longpfreq	긴 휴지 (0.5s)빈도 LONG Pause FREQuency
	Longpmn	긴 휴지 평균 길이 LONG Pause Mean
	Longpwd	단어당 평균 긴 휴지 개수 LONG pauses Per Word
	Longpmeandev	긴 휴지 평균 편차 LONG Pause MEAN DEViation
	Silstddev	휴지 길이 표준 편차 SILence Standard DEViation
	Longpstdev	긴 휴지 표준 편차 LONG Pause Standard DEViation

RATE 범주는 학습자가 발화하는 발화 속도와 관련된 자질들을 다룬다. RATE 범주에서 ROS(Rate of Speech)는 휴지를 포함하는 전체 음소 개수 N_{phone}와 발화 시간 T_{total}의 비로 다음의 식 (3)과 같이 정의된다(Cucchiarini *et al*., 2000a).

$$ROS = \frac{N_{phone}}{T_{total}} \quad (3)$$

Cucciarini *et al*.(2000a)의 연구에서는 ROS가 수동 평가 발음 점수와의 상관계수가 약 0.81로 나타났으며, 전체 발화길이와의 상관계수인 0.79보다 다소 높은 상관성을 나타냄을 보인 바 있다. 한편, <표 2>에서의 Wpsecutt 역시 발화 속도를 나타내는 자질이지만, 발화 음소의 개수를 사용하는 ROS와는 달리, Wpsecutt는 발화된 단어의 개수(Zechner *et al*., 2009)를 사용한다는 점에서 차이가 있다.

AR (Articulation Rate)은 문장 내 휴지를 제외한 발화의 길이 $T_{NoPause}$와 전체 음소 개수 N_{phone}의 비로 식 (4)와 같이 계산할 수 있다(Cucchiarini *et al*., 2000b).

$$AR = \frac{N_{phone}}{T_{NoPause}} \quad (4)$$

Cucciarini *et al*.(2002)의 연구에 따르면, 낭독체의 경우 AR은 수동 평가 점수와 0.83의 상관성을 보였으나, 즉흥 발화의 경우에는 0.07로 상관성이 매우 낮아짐을 보인 바 있다. <표 2>에서의 Wpsec 역시 조음 속도를 의미하지만, ROS와 Wpsecutt의 관계와 마찬가지로, 음소의 개수가 아닌 발화된 단어의 개수(Zechner *et al*., 2009)를 사용한다는 점에서 AR과 차이를 갖는다.

PTR (Phone-Time Ratio)은 문장 내 휴지를 제외한 발화의 길이 $T_{NoPause}$와 전체 발화의 길이 T_{total}의 비로 식 (5) 와 같이 정의된다(Cucchiarini *et al*., 2000b).

$$PTR = \frac{T_{NoPause}}{T_{total}} \quad (5)$$

Cucchiarini *et al*.(2002)의 연구에 따르면, PTR 자질은 낭독체의 수동 평가 점수와의 상관계수가 0.86으로 강한 상관성이 있음을 보인 바 있다.

SEGMENT 범주는 평균 및 표준/절대 편차를 포함한 분절음 빈도 및 길이와 관련된 자질들을 포함하고 있다. 반면, SILENCE 범주는 휴지의 빈도 및 길이와 관련된 자질들을 다룬다. SEGMENT와 SILENCE 범주들에서 다루어지는 모든 자질들은 Zechner *et al*.(2009)의 연구에서 제시된 바 있다. Zechner *et al*.(2009)는 1장에서 언급한 바와 같이 ETS의 TOEFL 말하기 자동 평가에 관한 연구로서, 비원어민 영어 학습자 발음 자동 평가를 위해 다양한 자질들을 제안하였다. 대부분은 유창성과 관련된 자질들로서, 휴지 및 발화 단어의 길이와 빈도에 관련된 특성을 의미한다. Zechner *et al*.(2009)의 연구는 즉흥 발화의 발음 평가에 주안점을 두고, 비유창성(disfluency)이나 반복 빈도, 언어 모델 점수와 같은 자질도 고려하였다. 반면, 본 연구에서는 주어진 문장을 발화하는 낭독체 환경에서의 발음 평가에 초점을 두고 있으므로, 비유창성, 반복 빈도, 언어모델 점수 등의 즉흥 발화와 관련된 자질들은 고려에서 제외하였다.

또한, 발음 평가를 위해 사용하는 자질로 음소 단위의 GOP(Goodness-Of-Pronunciation)가 있다. 본래 GOP는 앞서 2장에서 언급한 바와 같이 개별 발음 오류를 탐지하기 위해 제안된 자질이나(Witt & Young, 2000), 전체 문장에 대해 전역 범위에서 점수를 계산할 경우 GOP를 발음 평가를 위한 자질로 사용할 수도 있다(Neumeyer *et al*., 2000; Shi *et al*., 2016). GOP 자질은 2장의 식 (1)과 같이 계산된다. Neumeyer *et al*.(2000)의 연구에서는 전체 문장에서의 GOP가 수동 평가와 높은 상관성을 가짐을 보인 바 있으며, Shi *et al*.(2016)는 그림자처럼 따라 말하기(shadowing) 과업에서도 GOP가 자동 평가를 위한 자질로 사용될 수 있음을 보였다.

또한, 앞서 2장에서 언급한 바와 같이, 음소 레벨에서의 GOP 외에도 본 연구에서 제안한 조음 자질 기반의 aGOP를 발음 평가 모델을 위한 자질로 포함하여 자동 발음 평가를 수행한다. 또한, 평가를 위해서는 문장 단위의 전역적 범위에서 aGOP를 계산하여야 한다. 이를 위해, 식 (2)에서 분절음 단위에서 계산한 aGOP를 다음의 식 (6)과 같이 문장 단위의 전역 자질로 바꾸어 준다.

$$aGOP^k = \frac{\sum_{m=1}^{M} aGOP_m^k}{M} \quad (6)$$

이 때, M은 문장을 구성하는 분절음의 전체 개수를 의미한다.

이를 통해, 음성학/음운론적 지식에 기반한 조음 자질을 이용한 평가 모델의 평가 점수 예측 성능을 관찰하고자 한다.

3.3. 코퍼스 및 음성 인식

한국인 영어 학습자의 자동 평가를 위해서, 한국인 영어 발화 데이터로는 K-SEC(Korean-Spoken English Corpus) 코퍼스(이하 K-SEC 코퍼스)와 한국전자통신연구원(ETRI)에서 제공한 한국인 영어 학습자 낭독체 발화 코퍼스(이하 ETRI 코퍼스)를 사용하였다. K-SEC 코퍼스는 한국인 영어 학습자 336명의 발화를 포함하고 있으며, 한국어 기본단어, 한국어 이야기, 영어 자모음, 영어 어휘, 영어 문장, 영어 이야기 등 6개 세트로 구성되어 있다(이석재 외, 2003). 그 중 본 연구의 평가 모델링을 위해서 12명의 학습자가 발화한 431개의 영어 문장을 선택하였다. ETRI 코퍼스는 한국인 영어 학습자 151명이 발화한 21,110문장으로 구성되어 있으며, 그 중 본 연구를 위해서 140명이 발화한 800 문장을 사용하였다. 평가 모델에 사용한 전체 1,231문장(K-SEC 코퍼스 431문장 + ETRI 코퍼스 800문장) 가운데, 평가 모델의 학습 및 테스트를 위해, 각각 119명이 발화한 1,001 문장(학습 집합)과 33명이 발화한 230 문장(테스트 집합)으로 구분하였다. 학습 발화와 테스트 발화의 화자는 서로 겹치지 않도록 하였다.

학습자 발화 음성 인식 및 강제 정렬을 위해서, PLU 집합으로는 CMU pronouncing dictionary v0.7b(Weide, 2014)의 CMU39

PLU 집합을 기반으로, WSJ(Wall Street Journal) 코퍼스(Garofolo *et al*., 2007)의 북미 영어 모국어 화자의 약 37,000 문장으로 학습된 영어 원어민 음향 모델을 사용하였다. 음향모델은 Kaldi(Povey *et al*., 2011)의 HMM-DNN을 이용하여, PLU 단위로 학습하였다. aGOP 계산을 위해서, PLU 단위의 음향 모델 뿐 아니라, 조음 속성 별로도 음향모델을 학습하였다.

3.4. 수동 평가

한국인 영어 학습자의 낭독체 문장 발화에 대해서 수동 평가를 시행하였다. 한국인 영어 교육 전문가 2명이 문장 발화 평가에 참여하였으며, 라이커트 척도(Likert scale)를 이용하여 이루어졌다. 라이커트 척도란 응답자가 서열성을 갖는 응답 범주 가운데 하나를 고르도록 하는 형식을 말한다. 본 연구에서는 5점 척도(1=very poor, 5=excellent)를 사용하였다.

평가 문장에 대해서 평가자는 각 문장별로 학습자의 발음에 대한 전체적인 인상을 평가하였다. 전체적인 인상이란 평가자가 학습자의 발화를 들었을 때, 특정 분절음이나 음운 규칙 등과 같이 발음의 특정 측면에 주의를 기울이지 않고, 평가자들이 발음에 대해 느끼는 종합적인 판단을 말한다. 이 때, 평가자가 필요에 따라 동일한 문장을 반복하여 들을 수 있도록 하였다.

수동 평가자간 신뢰도는 Pearson 평가자간 상관계수(Pearson's inter-rater correlation coefficients)로 측정하였다. 상관계수 계산 결과, 전체 데이터에 대해서 0.766이었고, 각 데이터 별로 K-SEC 코퍼스에서는 0.862, ETRI 코퍼스에서는 0.701로 나타났다. 이러한 상관 계수 결과는 강한 신뢰성을 갖는 것으로 평가할 수 있다(Evans, 1996). 평가자들의 평가 결과가 완전히 동일하기를 기대하기는 어렵다는 점을 감안할 때, 위의 평가자간 신뢰도는 충분히 높다고 볼 수 있다.

3.5. 평가 모델링

한국인 영어 낭독체 발화의 발음 자동 평가를 위해서, <그림 1>에서 기술한 바와 같이 다중선형회귀 분석을 시행하였다. 다중선형회귀에서 <표 2>와 3.2에서 기술한 평가 자질들을 설명 변수로, 그리고 수동 평가자들의 점수를 반응 변수로 설정하였으며, 이를 통해 발음 평가 점수를 예측하였다. 통계 패키지 R 3.2.4 버전(R Core Team, 2016)를 이용하여 다중선형회귀 분석을 수행하였다.

선형회귀분석에서 사용한 평가 자질들 가운데, 평가 점수 예측에 영향을 많이 미치는 자질들의 부분집합을 구하고 이를 통해, 발음 점수 예측 성능 향상이 이루어지는지를 알아보기 위해, 최량 부분 집합 선택(Best Subset Selection; BSS)을 적용하였다. BSS는 모든 가능한 설명 변수들의 조합에 대해 최소 제곱법을 이용하여 회귀분석을 시행하고, 베이즈 정보기준(Bayesian Information Criterion; BIC)을 최소화하는 자질들의 최적 부분집합을 찾는 방법을 말한다(James *et al*., 2013). BSS의 최적 파라미터를 찾기 위해, 훈련 집합에 대해 10회 교차 검증(10-fold cross validation)을 시행하였다. 통계 프로그램 R의 'leaps' 패키지(Lumley & Miller, 2009)를 이용하여 BSS를 수행하였다.

<표 2>에서와 같이, 평가 모델링을 위해 aGOP, GOP, RATE, SEGMENT, SILENCE의 다섯 종류의 자질 범주를 사용하였다. 본 연구의 목적이 aGOP가 자동 발음 평가에 미치는 영향을 알아 보는 것이므로, aGOP를 제외한 나머지 자질들을 이용한 모델들을 베이스라인으로 삼고, aGOP 자질을 추가한 모델들을 제안 모델로 삼아 성능을 비교하였다. 이 때, 평가 모델링에 사용한 모든 자질들은 z값 정규화(z-score normalization)하였다. 베이스라인과 제안 모델의 자질 조합은 다음의 <표 3>과 같다.

표 3. 평가 모델링 세부 구성
Table 3. Details of assessment modeling

No.	Baseline	Proposed
(1)	SILENCE	aGOP + SILENCE
(2)	SEGMENT	aGOP + SEGMENT
(3)	RATE	aGOP + RATE
(4)	SILENCE + SEGMENT + RATE	aGOP + SILENCE + SEGMENT + RATE
(5)	GOP + SILENCE + SEGMENT + RATE	aGOP + GOP + SILENCE + SEGMENT + RATE (FULL)

<표 3>에서 보는 바와 같이, 먼저 베이스라인에서 모델 (1) - 모델 (3)까지 평가 자질 범주를 하나씩만 포함하였을 때의 평가 모델링 성능을 알아본다. 모델 (4)에서는 기존 연구들에서 제안하는 길이 관련 자질들을 모두 포함하였을 때의 평가 모델 성능을, 그리고 모델 (5)에서는 GOP까지 포함하는 모든 자질을 포함한 모델의 성능을 알아보는 것을 목적으로 한다. 제안 모델에서는 모델 (1) - 모델 (5)까지의 베이스라인 조합에 aGOP 범주를 추가하여, 베이스라인과 그 성능을 비교한다. 제안 모델의 (5)는 모든 자질 범주를 포함하고 있기 때문에 FULL 모델로 표시하였다.

위의 <표 3>에서 나타나는 자질 조합들에 대해 다중회귀분석을 시행하고, 가장 높은 성능을 보이는 모델에 대해 BSS를 적용한다. 이를 통해 주요 자질을 추출함으로써, 평가 점수 예측에 영향력을 미치는 자질들의 특성을 살펴본다.

4. 실험 결과 및 논의

4.1. 수동 평가 결과

한국인 영어 발화에 대한 발음 자동 평가를 위해, 정답 점수로 사용될 평가자에 의한 수동 평가를 실시한 결과, 평균 발음 점수는 <표 4>와 같이 2.98점 (표준 편차 = 1.05)으로 나타났다.

표 4. 수동 평가 발음 점수의 평균, 표준편차, 최소값, 중간값, 최대값
Table 4. Mean, standard deviation, and minimum, median and maximum value for manual pronunciation score

Mean	Std. dev.	Min	Median	Max
2.98	1.05	1.00	3.00	5.00

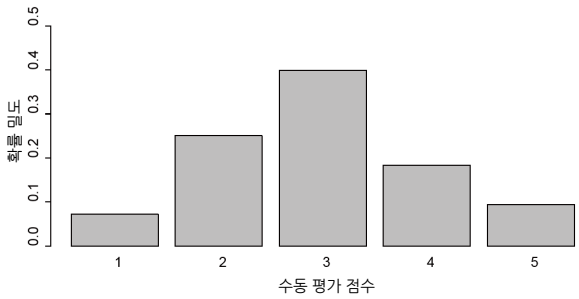

그림 2. 수동 발음 평가 점수의 확률밀도 그래프
Figure 2. Bar plot of probability densities of manual pronunciation score

또한 <그림 2>에서 볼 수 있는 바와 같이 학습자의 수동 발음 평가 점수 분포를 살펴볼 수 있다. 최대 5점인 라이커트 척도에서 중간 점수에 해당하는 3점을 받은 학습자의 발화가 전체 발화 가운데 약 40%를 차지하고 있으며, 최고점인 5점과 최저점인 1점을 받은 문장이 각각 9.5%와 7.2%를 차지하는 것으로 나타났다.

4.2. 평가 모델링 성능 비교

앞서 3.5에서 언급한 바와 같이, 발음 평가를 위한 자질들의 범주를 조합하여, 베이스라인과 aGOP를 추가한 제안 모델의 성능을 비교하였다. 평가 모델을 통해 예측한 점수와 실제 수동 평가 점수의 Pearson 상관계수로 평가 모델의 성능을 측정하였다. 베이스라인과 제안 모델의 성능은 다음의 <표 5>와 같다.

표 5. 평가 자질 범주 결합에 따른 베이스라인과 제안 모델의 성능 비교
Table 5. Performance by comparison between the baseline and the proposed model in terms of combination of feature categories

No.	Baseline		Proposed	
	Model	Corr.	Model	Corr.
(1)	SILENCE	0.608	aGOP + SILENCE	0.610
(2)	SEGMENT	0.651	aGOP + SEGMENT	0.653
(3)	RATE	0.691	aGOP + RATE	0.702
(4)	SILENCE + SEGMENT + RATE	0.692	aGOP + SILENCE + SEGMENT + RATE	0.705
(5)	GOP + SILENCE + SEGMENT + RATE	0.697	FULL	**0.715**
Corr. b/w human raters				0.766

먼저 베이스라인에서, 하나의 자질 범주만 사용한 모델 (1) - (3)을 비교하면, RATE 범주의 자질만 사용한 모델 (3)의 성능이 0.691로 가장 높게 나타나는 것을 알 수 있다. 모델 (3)은 RATE 범주의 자질 5종을 이용한 반면, 모델 (1)은 SILENCE 범주의 11종 자질을, 그리고 모델 (2)는 SEGMENT 범주의 자질 6종을 사용한다. 모델 (3)이 더 적은 수의 자질을 사용하고 있음에도 불구하고, 더 많은 수의 자질을 사용하고 있는 모델 (1)에 비해 0.1, 그리고 모델 (2)에 비해 0.05 가량 더 높은 성능을 보여주고 있다. 또한, SILENCE, SEGMENT, RATE 범주를 모두 사용하고 있는 모델 (4) 및 GOP 까지 포함하는 모델 (5)와 비교해서도 성능의 차이가 크게 나타나지 않는다. 모델 (4)와 모델 (5)과 비교하였을 때, 사용하는 자질이 총 22종, 23종임에 비해, 모델 (3)은 발화 속도와 관련된 자질 5종만을 사용하고 있다. 그럼에도 불구하고, 모델 (3)의 성능(0.691)과 모델 (4), (5)의 성능(0.692, 0.697) 차이가 각각 0.001과 0.006으로, 거의 차이가 나지 않는다. 이러한 점에서 발화 속도와 관련된 자질들이 발음 자동 평가에서 효과적으로 기능하고 있음을 알 수 있다. 이러한 결과는 Cucchiarini *et al*.(2000a, 2000b, 2002)의 연구 결과와도 합치한다.

한편, 제안 모델과 베이스라인의 성능을 비교하면, aGOP 범주의 자질을 포함함으로써 (1) - (5)의 모든 모델에서 성능이 향상되고 있음을 관찰할 수 있다. 특히, 모델에 사용하는 자질의 종류가 늘어날수록 베이스라인과 제안 모델의 절대적인 성능도 좋아질 뿐 아니라, 베이스라인 대비 제안 모델의 성능 향상 정도도 커지고 있음을 알 수 있다. 본 연구에서 제안한 모든 자질들을 사용하는 모델 (5)의 경우, 수동 평가 점수와의 상관계수가 0.715로써 수동 평가자간 상관 계수인 0.766에 가장 가까운 값을 보여주고 있다. 또한, 베이스라인의 모델 (5)와 비교했을 때, 상관계수가 0.697에서 0.715로 증가하여, 상대적인 성능 향상이 약 6% 가까이 나타남을 알 수 있다. 이를 통해, 본 연구에서 제안한 조음 특성에 기반한 평가 자질(aGOP)이 발음 평가 모델링의 성능 향상에 유의미한 역할을 하고 있다고 볼 수 있다.

4.3. 주요 자질 선택

4.2에서 살펴본 바와 같이 제안 모델 중에서 FULL 모델이 가장 높은 성능을 보이고 있으므로, FULL 모델에 대해 3.5에서 기술한 BSS를 적용하여 주요 자질을 추출하였다. 그 결과 선택된 주요 자질 및 주요 자질이 속한 범주, 그리고 해당 주요 자질의 가중치는 다음의 <표 6>과 같다. <표 6>에서는 선택된 주요자질들을 범주 별로 분류하고, 이를 가중치의 크기에 따라 내림차순으로 정렬하였다.

표 6. 추출된 주요 자질의 종류, 범주 및 해당 자질의 가중치
Table 6. Selected salient features, the corresponding categories and the corresponding weights

Category	Feature	Weight
aGOP	Delayed Release	0.491
	Lateral	0.360
	Nasal	0.254
	Stop	0.143
	Front	0.123
	Tense	0.105
	Distributed	0.098
	Voice	0.046
	Labial	0.034
RATE	Wpsecutt	0.647
	Wpsec	0.481
	AR	0.377
SILENCE	Silpsec	0.154
	Silpwd	0.039
	Silmean	0.025
SEGMENT	Segdur	0.104
	Wdpchk	0.034

FULL 모델에서 사용된 전체 자질은 총 47종이며, BSS를 적용하여 수동 평가 점수를 가장 잘 설명해주는 자질들의 부분집합을 찾은 결과, <표 6>에서와 같이 17종의 주요 자질이 선택되었다. 이들 17종의 주요 자질을 이용하여 선형회귀분석을 시행한 결과, 수동평가점수와의 상관계수는 0.720으로 <표 5>의 FULL 모델의 0.715에 비해, 약간의 성능 향상을 관찰할 수 있다.

추출된 주요 자질을 <표 6>에서 살펴보면, 전체 17종의 주요 자질 가운데 조음 자질과 관련된 aGOP 범주의 자질이 9종으로 가장 많은 비중을 차지하고 있었다. 선택된 자질의 개수 뿐 아니라 가중치 측면에서 살펴볼 때, aGOP 범주의 주요 자질들이 상당 부분 상위에 위치하고 있음을 알 수 있다. aGOP 범주에서 Delayed Release, Lateral, Nasal과 같은 자질들이 상위에 위치하고 있으며, 이들의 가중치는 0.2-0.5 정도이다. 이들의 가중치는 SILENCE나 SEGMENT 범주에 속하는 주요 자질들의 가중치보다 높으며, RATE 범주에 속하는 주요 자질들의 가중치와 유사한 값을 보인다. 이를 통해, 조음 특성에 기반한 자질들이 자동 발음 평가 모델링에서 비교적 큰 영향을 미치고 있음을 미루어 판단할 수 있다.

한편, RATE, SILENCE, SEGMENT 범주에서도 각각 2-3종의 자질이 주요 자질 목록에 포함되어 있는데, RATE의 자질들이 SILENCE나 SEGMENT 범주들에 비해, 높은 가중치를 가지고 있음을 알 수 있다. 다시 말해, 길이나 속도와 같은 유창성과 관련된 자질 가운데, 발화 속도와 관련된 RATE 범주의 자질들이 길이와 관련된 SILENCE나 SEGMENT 자질들보다 평가 점수에 미치는 영향이 더 크게 나타남을 알 수 있다. 이와 같은 결과는 앞서 4.2에서 <표 5>의 베이스라인 성능을 분석한 결과와도 일치한다.

마지막으로, aGOP 범주에서 선택된 주요 자질들과 한국인의 영어 발화에서 주로 나타나는 발음 변이들과의 상관성을 살펴본다. Hong *et al*.(2014)에 따르면, 한국인의 영어 발화에서 나타나는 자음의 주요 발음변이로 /f, v, θ, z/가 각각 /p, b, d, s/ 로 발화되는 변이를 지적한 바 있다. 이들은 모두 한국어에 없는 음소로 인해 나타나는 변이 현상으로, 이 중 /z/ →/s/를 제외한 나머지 세 가지의 변이는 마찰음이 파열음으로 실현되었다는 공통점이 있다. 이를 조음 특성으로 나타내면, [+delayed release, -stop] → [-delayed release, +stop]의 변이로 표현할 수 있다. <표 6>에서 보는 바와 같이, Delayed release(0.491)와 Stop(0.143)이 비교적 큰 가중치를 갖는 주요 자질로 선택되고 있으므로 이와 같은 주요 발음 변이들이 평가 모델링에서도 조음 특성 차원에서 반영되고 있음을 알 수 있다. /z/ →/s/의 변이 역시 유성성 (Voice)의 변화([+voice] → [-voice])로 설명할 수 있는데, Voice 역시 <표 6>에서 주요 자질로 선택되고 있다. 또한, Jang(2005)에서는 한국인 영어학습자의 주요 발음 변이로 Tense/lax 모음의 미구분, 유음화, 비음화와 같은 한국어 음소 변동 규칙의 적용 등을 지적한 바가 있다. 조음 자질 측면에서 이와 관련된 자질들로는 Tense, Lateral, Nasal을 들 수가 있으며, 이들 역시 평가 모델에서의 주요 자질로 선택되고 있다. 이상에서 보는 바와 같이, <표 6>에서 주요 자질로 선택된 aGOP 자질들과 한국인 영어 학습자들이 보이는 주요 발음 변이들의 관련성을 살펴 볼 수 있었다. 하지만, 평가라는 행위는 문장 전체에 대해서 전역 범위(global scale)로 이루어질 뿐 개별 분절음 단위로 판단하지는 않는다. 따라서, 실제로 이와 같은 주요 발음 변이들을 학습자 발화에서 aGOP 범주의 자질들로 얼마나 정확하게 검출할 수 있는지는 오류 검출 실험을 통해서 검증할 필요가 있다.

5. 결론

본 연구에서는 한국인 영어 학습자의 자동 발음 평가 모델링을 위해 조음 특성 기반의 GOP(aGOP)를 평가 자질로 제안하였다. 그리고, 기존의 발음 자동 평가 연구에서 사용된 자질들을 그 특성에 따라 범주화하고, 본 연구에서 제안한 자질들을 포함하여 평가 모델의 발음 점수 예측 성능을 선형회귀분석을 통해 살펴보았다. 또한, 최량 부분 집합 선택(Best Subset Selection) 기법을 이용하여, 평가 점수에 영향력을 많이 미치는 주요 평가 자질들을 추출하고, 이들의 특성을 분석하였다. 분석 결과, 본 연구에서 제안한 aGOP 자질들을 포함할 때, 포함하지 않은 베이스라인보다 평가 점수 예측 성능이 향상됨을 관찰할 수 있었다. 따라서, 본 연구에서 새롭게 제안한 조음 특성을 반영한 자질들이 자동 평가 모델링에 유용하게 사용될 수 있음을 알 수 있다. 또한, 주요 평가 자질들을 추출한 결과, aGOP 자질들이 추출된 주요 자질 가운데 절반 가량을 차지하고 있었으며, 가중치 측면에서도 상당한 비중을 차지함을 알 수 있었다. 또한, 언어학적 분석을 통해 선택된 주요 자질들이 기존의 다른 연구들에서 분석한 한국인 영어학습자들의 발음 변이와도 관련성이 있음을 보였다.

본 연구는 음성/음운론적 지식을 반영한 조음 특성에 기반한 평가 자질을 제안하였다. 이를 통해 기존의 발음 평가 연구에서 주로 길이, 속도와 같은 시간 관련 자질들을 주로 사용한 것을 넘어, 음성/음운론적 지식을 이용하여 자동 발음 평가 성능 향상에 기여 할 수 있음을 보였다는 점에서 의의를 갖는다. 하지만, 4.3에서 언급한 바와 같이 조음 특성 기반의 자질들이 주요 발음 변이를 정확하게 탐지할 수 있는지에 대해서는 분절음 단위의 발음 오류 검출까지 확장할 필요가 있다. 조음 특성을 이용한 발음 오류 검출 모델링을 통해, 조음 자질을 이용한 발음 평가, 오류 검출, 피드백에 이르는 통합적인 CAPT 프레임워크의 가능성을 기대할 수 있다. 추후 연구에서는 조음 자질을 이용한 발음 오류 검출 실험에 대한 추가 연구를 진행하며, 발음 평가 모델링에서 선형회귀 이외에 CART나 DNN과 같은 다른 평가 모델링 방법을 적용하고, 다른 L1-L2 쌍에 발음 평가를 적용함으로써, 본 연구에서 제안하는 조음 기반의 자질들이 언어에 강건하게 나타나는지를 검증하고자 한다.

참고문헌

Alderson, C. J., Wall, D., & Claphaim, C. (1996). *Language Test Construction and Evaluation.* Cambridge: Cambridge University

Press.

Cheun, S. (2004). *Phonology*. Seoul: Seoul National University Press. (전상범 (2004). *음운론*. 서울: 서울대학교 출판부.)

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Cincarek, T., Gruhn, R., Hacker, C., Nöth, E., & Nakamura, S. (2009). Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Computer Speech & Language*, 23(1), 65-88.

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30(2-3), 109-119.

Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2), 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6), 2862-2873.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the Usefulness of the Versant for English Test: A Response. *Language Assessment Quarterly*, 5(2), 160-167.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51(10), 832-844.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Franco, H., Neumeyer, L., Yoon, K., & Ronen, O. (1997). Automatic pronunciation scoring for language instruction. *Proceedings of IEEE International Conference on the Acoustics, Speech, and Signal Processing(ICASSP) 1997* (pp. 1471-1474). München, Germany. 21-24 April, 1997.

Garofalo, J., Graff, D., Paul, D., & Pallett, D. (2007). *CSR-1 (WSJ0) complete*. Philadelphia: Linguistic Data Consortium.

Hong, H., Kim, S., & Chung, M. (2011). How Korean learner's English proficiency level affects English speech production variations. *Phonetics and Speech Sciences*, 3(3), 115-121.

Hong, H., Kim, S., & Chung, M. (2014). A corpus-based analysis of English segments produced by Korean learners. *Journal of Phonetics*, 46, 52-67.

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. (홍혜진·류혁수·정민화 (2014). 일본인 한국어 학습자의 분절음 실현과 발음 평가의 상관성. *말소리와 음성과학*, 6(4), 101-108.)

James, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with application in R*. New York: Springer.

Jang, T. (2005). Construction of an English speech database for Korean learners of English, *Language and Linguistics*, 35, 292-309. (장태엽 (2005). 한국인 영어학습자의 영어음성 데이터베이스 구축에 관한 연구. *언어와 언어학*, 35, 292-309)

Kirchhoff, K., Fink, G. A., & Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3-4), 303-319.

Lee, C.-H. (2004). From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition. *Proceedings of INTERSPEECH 2004* (pp. 109-112). Jeju Island, Korea. 4-8 October, 2004.

Lee, C.-H., Clements, M. A., Dusan, S., Fosler-Lussier, E., Johnson, K., Juang, B.-H., & Rabiner, L. R. (2007). An overview on automatic speech attribute transcription (ASAT). *Proceedings of INTERSPEECH 2007* (pp. 1825-1828). Antwerp, Belgium. 27-31 August, 2007.

Li, W., Li, K., Siniscalchi, S. M., Chen, N. F., & Lee, C.-H. (2016). Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-guided and Data-driven Decision Trees. *Proceedings of INTERSPEECH 2016* (pp. 3127-3131). San Francisco, CA. 8-12 September, 2016.

Lumley, T., & Miller, A. (2009). leaps: regression subset selection. Retrieved from https://cran.r-project.org/package=leaps on October 20, 2016.

Metze, F. (2005). *Articulatory features for conversational speech recognition*. Ph.D. Dissertation, Universität Fridericiana zu Karlsruhe, München, Germany.

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30(2-3), 83-93.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011)*.

R Core Team (2016). R: language and environment for statistical computing. Retrieved from http://www.r-project.org on October 20, 2016.

Rhee, S., Lee, S., Kang, S., & Lee, Y. (2003). Design and Construction of Korea-Spoken English Corpus (K-SEC). *Malsori*, 46, 159-174. (이석재·이숙향·강석근·이용주 (2003). 한국인의 영어 음성 코퍼스 설계 및 구축. *말소리*, 46, 159-174.)

Richardson, M., Bilmes, J., & Diorio, C. (2003). Hidden-articulator Markov models for speech recognition. *Speech Communication*, 41(2-3), 511-529.

Ryu, H., & Chung, M. (2016). 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* (pp. 101-102). (류혁수·정민화 (2016). 조음 기반의 음소 레벨 사후 확률을 이용한 한국인 영어 학습자 유창성 자동 평가. *한국음성학회 봄 학술대회 발표논문집*, 101-102.)

Ryu, H., Hong, H., Kim, S., & Chung, M. (2016). Automatic Pronunciation Assessment of Korean Spoken by L2 Learners Using Best Feature Set Selection. *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA ASC) 2016*, accepted.

Shi, S., Kashiwagi, Y., Toyama, S., Yue, J., Yamauchi, Y., Saito, D., & Minematsu, N. (2016) Automatic assessment and error detection of shadowing speech: case of English spoken by Japanese learners. *Proceedings of INTERSPEECH 2016* (pp. 3142-3146). San Francisco, CA. 8-12 Sep, 2016.

Siniscalchi, S. M., Svendsen, T., & Lee, C.-H. (2008). Toward a detector-based universal phone recognizer. *Proceedings of IEEE International Conference on the Acoustics, Speech, and Signal Processing(ICASSP) 2008* (pp. 4261-4264). Las Vegas, NV. 31 March - 04 April, 2008.

Tepperman, J., & Narayanan, S. (2008). Using articulatory representations to detect segmental errors in nonnative pronunciation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), 8-22.

Weide, R. L. (2014). The CMU pronouncing dictionary 0.7b. Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict on October 20, 2016.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3), 95-108.

Zechner, K., Higgins, D., Xi, X. M., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10), 883-895.

• **류혁수 (Ryu, Hyuksu)**
서울대학교 언어학과
서울시 관악구 관악로 1
Tel: 02-880-9039
Email: oster01@snu.ac.kr
관심분야: 음성인식, 음성학, 컴퓨터 기반 언어교육

• **정민화 (Chung, Minhwa)** 교신저자
서울대학교 언어학과
서울시 관악구 관악로 1
Tel: 02-880-9195 Fax: 02-882-2451
Email: mchung@snu.ac.kr
관심분야: 음성인식, 음성언어처리, 컴퓨터 기반 언어교육