한국정보전자통신기술학회논문지(jkiiect)'19-10 Vol.12 No.5
논문 19-12-05-512 http://dx.doi.org/10.17661/jkiiect.2019.12.5.512

음성특징의 거리 개념에 기반한 한국어 모음 음성의 시각화

복거철*

Speech Visualization of Korean Vowels Based on the Distances Among Acoustic Features

Gouchol Pok*

요 약 음성을 시각적으로 표현하는 것은 외국어를 습득하는 과정의 학습자나 음성을 직접 들을 수 없는 청각장애자에게 매우 유용하며 기존에 다수의 연구가 이루어졌다. 그러나 기존의 연구들은 발음의 특징을 단지 컬러로 표현한다든가 입모양을 3차원 그래픽으로 표현하거나 입과 구강의 변화하는 형태를 애니메이션으로 보여 주는 방식에 머물러 있다. 따라서 이런 방식을 사용하는 학습자들은 자신의 발음이 표준 발음과 얼마나 멀리 떨어져 있는지 알 수가 없고 더 나아가서 학습 중에 스스로 교정을 해 나가는 시스템을 개발하기가 기술적으로 어려운 단점이 있다. 이를 극복하기 위해 본 논문에서는 음성 간의 상대적 거리를 토대로 음성을 시각화하는 모델을 제시하고, 이를 한국어 모음에 적용하여 모음의 음성적 특징을 이용한 시각화의 구체적인 구현 방법을 제시한다. 음성데이터에서 F1, F2, F3의 세 개의 포먼트를 구하고 이들 특징벡터를 코호넨 자기조직화맵 알고리즘으로 2차원 화면에 사상하여 각 음성을 화면 위의 각 점에 대응하여 표현하였다. 제안하는 시스템의 실제적인 구현은 인터넷에 공개된 음성처리 공개소프트웨어를 사용하고 한국인 교사의 표준 발음과 한국어를 배우고 있는 외국인 유학생의 음성을 이용하여 음성특징의 상호간 거리를 구하였으며, 사용자 인터페이스는 자바스크립트를 이용하여 구현하였다.

Abstract It is quite useful to represent speeches visually for learners who study foreign languages as well as the hearing impaired who cannot directly hear speeches, and a number of researches have been presented in the literature. They remain, however, at the level of representing the characteristics of speeches using colors or showing the changing shape of lips and mouth using the animation-based representation. As a result of such approaches, those methods cannot tell the users how far their pronunciations are away from the standard ones, and moreover they make it technically difficult to develop such a system in which users can correct their pronunciation in an interactive manner. In order to address these kind of drawbacks, this paper proposes a speech visualization model based on the relative distance between the user's speech and the standard one, furthermore suggests actual implementation directions by applying the proposed model to the visualization of Korean vowels. The method extract three formants F1, F2, and F3 from speech signals and feed them into the Kohonen's SOM to map the results into 2-D screen and represent each speech as a pint on the screen. We have presented a real system implemented using the open source formant analysis software on the speech of a Korean instructor and several foreign students studying Korean language, in which the user interface was built using the Javascript for the screen display.

Key Words : Acoustic features, Formant, Self organizing Map, Speech processing, Speech visualization

This paper was supported by the Korea National Research Foundation with the research grant of NRF-2017R1D1A2B03028954.

*Division of Computer and IT Instruction, PaiChai University (gcpok@pcu.ac.kr)

Received October 02, 2019 Revised October 17, 2019 Accepted October 17, 2019

1. 서론

음성은 의사소통의 기본적인 매체로서 일상생활에서 아주 중요한 역할을 하므로 많은 연구의 대상이 되고 있지만[1] 정확한 발음을 반복하여 교육하거나 학습자의 발음을 교정해 나가는 시스템은 아직 완벽하지 않으며 특히 외국어 교육이나 청각 장애자들에게 정확한 발음을 교육하기 위해서 음성을 시각화하는 시스템을 위한 깊은 연구가 필요한 실정이다. 이를 위해서 음소차원의 자음과 모음에 대한 폭 넓은 연구가 이루어지고 있으나 모음의 조음에 관한 연구는 자음에 대한 연구에 비해 비교적 덜 활발하게 이루어졌는데 그 이유는 방법적인 어려움 때문이다[2].

전통적으로 모음의 조음 위치는 모음사각도로 표현 되는 모음공간을 이용하여 예측하여 왔다. 그러나 이 방법은 상대적인 모음 위치를 표시할 뿐이고 정확한 조음 위치를 반영하지 못하는 단점이 있으며, 이런 단점을 보완하기 위하여 음향학적인 관점에서 모음을 분석하려는 시도로서 포먼트(formant) 주파수 분석 기법이 제시되었다[2].

포먼트는 폐로부터 나오는 공기의 흐름이 성대를 진동시키며 발생시킨 음원이 성도를 지나는 동안 필터링 과정을 거치면서 만들어내는 공명주파수를 의미한다[3]. 포먼트 값은 성도의 길이에 따라 다른 값을 가지기 때문에 남녀화자의 포먼트 값이 차이가 나지만 포먼트 크기를 정규화한 값은 일정한 비율로 변화하고 동일한 간격을 유지한다는 사실로부터 모든 화자들의 성도 모양은 거의 비슷하다는 것을 알 수 있다[3].

그림 1은 한국어와 영어의 모음 발음에 대한 F1 및 F2 두 개의 포먼트 값과 각 모음의 상대적 위치의 관계성을 표현한 모음 공간을 보여 주고 있다[4].

모음과 자음의 음성발화에서 혀의 역할은 매우 크게 작용하며 특히 모음의 발화에서 혀의 높이와 앞뒤 움직임, 둥근 정도에 의해서 서로 다른 모음이 생성된다. 모음과 포먼트와 관계는 첫 번째 포먼트 주파수(F1)는 혀의 높낮이와 관련이 있고, 두 번째 포먼트(F2)는 혀의 앞 뒤 움직임과 깊은 연관성이 있다[3].

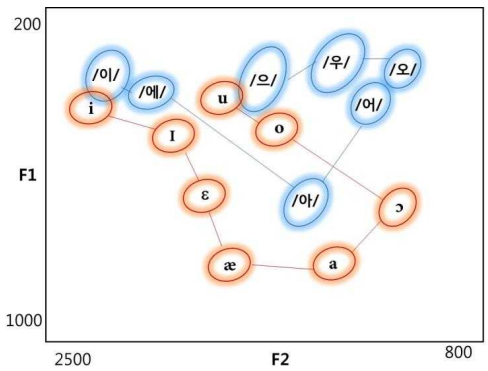

그림 1.한국어 및 영어 모음 공간 (단위: Hz)
Fig. 1. Vowel space of Korean and English vowels (Unit: Hz)

이와 같이 포먼트는 음성의 특징을 잘 표현하기 때문에 이를 음성의 시각화에 이용하면 효율적인 시스템을 구축할 것으로 예상된다.

음성의 시각화(visualization of speech)는 음성이 발화될 때 조음기관의 모양을 시각적으로 표현하는 방법이나 음성의 특징을 추출하여 화면에 색채로 변환하거나 애니메이션으로 표현하는 방법[5,6,7,8] 등이 제시되었다. 기존의 연구에서 제시하는 방법들은 모두 음성의 특징을 시각적으로 표현하는 방식에만 초점을 맞추었기 때문에 학습자/사용자의 발음이 표준 발음으로부터 얼마나 멀리 떨어져 있는지와 같은 상대적인 거리 개념이 결여되어 있어서 학습자/사용자가 조음기관의 변화를 통해 표준발음을 습득할 수 있도록 만드는 기능을 제공하지 못한다.

본 논문에서는 이런 점을 보완하기 위해 음성의 상대적 거리를 시각화하여 실시간으로 보여 주며 학습자/사용자가 표준 발음에 유사하게 발음할 때의 입모양이나 혀의 위치를 스스로 익히고 찾아 낼 수 있는 시각화 시스템을 제시하고자 한다.

2. 관련 연구

음성의 시각화에 관한 연구는 청각장애인을 위한 시스템을 개발하는 것에 초점을 두고 많은 연구가 되었다[5,6,7,8]. Watanabe[5]의 연구에서는 음성데이터의 포먼트, 피치, 멜 스펙트럼(Mel spectrum)을

Time Delay 신경망 (TDNN)의 입력데이터로 사용하여 음소의 특징을 구한 다음에 컬러로 표현하는 방법을 제시하였다. 이렇게 음성특징을 컬러로 표현하면 연속된 모음이 발화될 때 간섭현상이 발생하여 신호처리 기법으로 발견하기 어려운 특징도 비교적 쉽게 구분할 수 있는 장점이 있다고 저자들은 주장한다. Beskow[6]의 연구에서는 음성과 3차원 애니메이션을 결합하여 보여 주는 방식으로 시각화 시스템을 제안하며 하나의 발음에 대해 혀의 위치나 입술의 모양 등에 대해 그래픽으로 표현하고 있다. Ueda[7]는 Speech-ART라는 실시간 음성시각화 시스템을 소개하였는데 [5]에서 제시한 오프라인 방식을 온라인 방식으로 성능을 향상하였으며 단어/문장에 대한 표준 발음과 사용자의 발음에 대한 스펙트럼과 컬러표현을 동시에 보여 주며 비교를 할 수 있도록 하였다.

이와 같이 기존의 연구에서 제시한 방법들은 모두 각각의 발음에 대한 정보를 시각화하여 보여 주거나 단지 주파수영역의 정보를 보여 줄 뿐이므로 신호처리 또는 주파수 영역에 대한 지식이 없는 사용자가 쉽게 이해하지 못하는 단점이 있다. 그리고 어떤 방법으로 발화를 하여야 표준 발음에 가까이 근접해 나갈 수 있는지에 대한 교육적 제시(instructional presentation)가 결여되어 있다.

따라서 이런 점을 보완하기 위해 본 논문에서는 학습자의 발음과 표준 발음의 음성특징의 차이를 상대적인 거리 개념을 도입하여 화면에 표시하는 모델을 제시하고, 사용자가 발화를 할 때 마다 실시간으로 자신의 발음이 표준발음으로부터 얼마나 떨어져 있는지 확인을 하고 동시에 자신의 입모양이나 혀의 위치를 변화시켜가면서 표준발음을 스스로 습득하도록 도와주는 시스템 개발의 기반을 제시한다.

3. 제안 방법

3.1 음성시각화 시스템

본 논문에서 구현하려고 하는 음성의 거리기반 시각화를 위한 목표시스템(target system)의 전체 구조는 그림 2와 같다.

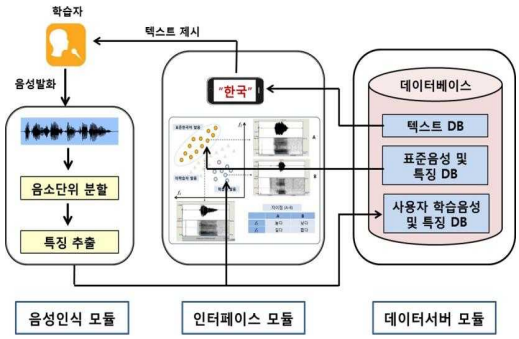

그림 2.한국어 음성 교육시스템의 구조도
Fig. 2. System structure of the instructional system for Korean language speeches

이 시스템은 크게 음성인식 모듈, 인터페이스 모듈, 데이터서버 모듈의 3부분으로 구성되었으며 본 논문에서 연구하는 부분은 인터페이스 모듈의 음성 시각화하는 부분으로서 그림 3에 자세히 나타내었다.

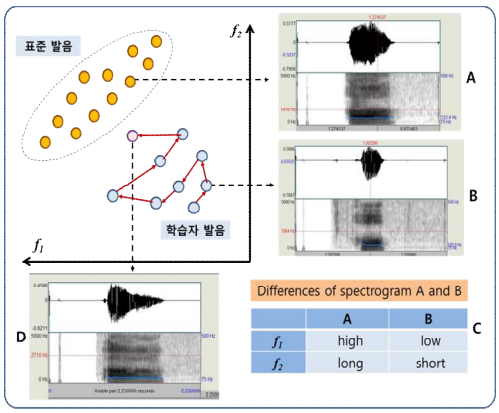

그림 3.음성특징의 상대적 거리에 기반한 시각화 화면
Fig. 3. Speech visualization on the screen based on the relative distances among acoustic features

그림 3의 왼쪽 윗부분에는 표준 발음들과 학습자의 발음들의 상대적 거리를 나타내는 점들의 집합을 나타내며, 오른쪽 윗부분에 표준발음에 해당하는 A영역과 사용자 발음에 해당하는 B영역으로 구분하여 이들 발음의 음성신호 정보(acoustic information)를 나타내며, 오른쪽 아랫부분에 이들 음성신호 정보의 차이를 명시적으로 기술해 주는 영역 C로 구성된다. 영역 D는 실제 구현하는 시스템의 필요에 의해

이전 사용자의 발음과 같은 추가정보를 나타낼 수 있는 선택적 부분이다.

위 목표 시스템의 주요 기능을 살펴보면 다음과 같다. 주어진 한 글자/단어에 대해 여러 화자의 표준 발음을 노란색 점으로 표시하고 학습자의 발음은 파란 색으로 표시하였다. 학습자는 빨간 화살표로 궤적을 표시한 것 같이 자기 발음이 표준 발음으로부터 상대적으로 얼마나 멀리 떨어져 있는지 실시간으로 관찰하면서 입의 모양과 혀의 위치를 변화시켜가며 표준 발음을 향해 교정해 나갈 수 있다.

음성신호와 음향분석도(sound spectrogram)와 같은 신호처리 기술 관련 정보를 화면에 나타낼 필요가 있을 경우 각 점들을 클릭하여 영역 A 또는 B와 같이 나타낼 수 있다. 영역 A의 표준발음 관련 정보와 영역 B의 학습자 발음 관련 특징을 명시적으로 기술하여 학습자가 어떤 부분에서 표준 발음과 차이가 나는지 인식할 수 있는 기능을 제공할 수 있다.

3.2 제안하는 방법

위와 같은 목표시스템을 개발하기 위해서는 먼저 각 발음의 서로 다른 차이점을 잘 표현 할 수 있는 특징을 추출(extraction)해야 하고 이를 화면에 사상(mapping)하여 표현할 수 있어야 한다.

3.2.1 음성의 특징 정의 (Acoustic Features)

음성이 발화될 때 폐에서 나오는 공기의 흐름은 성도를 진동시키고 이 음원이 성도를 지나는 동안 성도의 모양에 따라 달라지는 공명 주파수를 포먼트(formant)로 정의하며 이를 이용하여 모음의 음성적 특징을 결정할 수 있다 [4, 9, 10]. 포먼트는 음성신호의 파워스펙트럼 상에서 피크를 나타내며 이는 다수의 특정 주파수 대역에 집중되는 에너지띠를 가리킨다. 보통 사람의 성도에서 최대 5개까지 형성되는 포먼트 중에서 가장 낮은 주파수 대역의 제1포먼트(F1)와 그 다음 낮은 제2포먼트(F2)가 주로 모음의 특성을 구분한다고 알려져 있으며, 제3포먼트(F3)를 포함하여 모음의 특성을 나타내기도 한다[8]. 그림 4는 /우/ 발음을 녹음한 음성신호를 윗부분에 나타내고, 그 아래에 음성을 시간, 주파수, 진폭 축에 표현한 스펙트로그램(spectrogram)을 보여준다. 스펙트로그램 위에 다섯 개의 포먼트(F1 ~ F5)를 빨간색으로 나타내고 피치를 파란색으로 나타내었다.

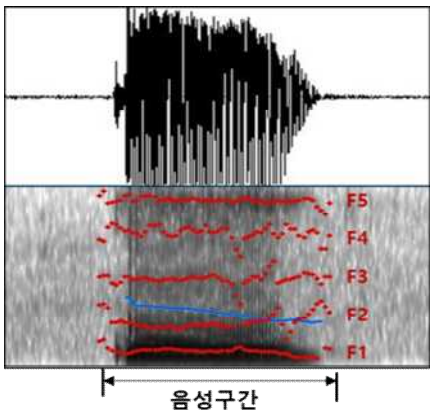

그림 4. /우/ 발음의 음성신호와 스펙트로그램상의 5개의 포먼트.
Fig. 4. Speech signal and five formants illustrated on the spectrogram for Korean /u/ pronunciation

그림 4에서 보듯이 포먼트 값은 균일하게 지속되지 않고 측정지점에 따라 값이 다르기 때문에 본 논문에서는 음성구간의 1/3 지점에서 포먼트 값을 측정값으로 정하였다 [2, 10].

3.2.2 포먼트 주파수 계산 (Prediction of Formant Frequencies)

음성데이터에서 포먼트를 구하는 방법에는 다음과 같이 LPC (linear predictive coding) 분석을 기반으로 하는 방법이 많이 사용된다[12].

1. 먼저 음성데이터를 25ms 길이의 프레임 단위로 나눈다. 전처리 작업으로 Hamming window를 곱한 후에 고주파 통과 전극 여과기 (high-pass all pole filter)를 적용한다. 각 프레임의 음성샘플 수를 N으로 표시하고, i 번째 프레임의 n 번째 샘플데이터를 $s_i(n)$으로 나타낼 때, LPC 모델은 다음과 같이 현재의 음성신호 값이 이전 p개의 음성샘플 값의

선형조합으로 근사적인 표현을 할 수 있다는 가정을 한다.

$$\hat{s}_i(n) = \sum_{k=1}^{p} a_k s_i(n-k) \quad (1)$$

계수 벡터 $\vec{a} = (a_1, ..., a_p)$는 음성신호 $s_i(n)$과 예측신호 $\hat{s}_i(n)$ 사이의 mean square error를 최소화하는 값으로서 다음 선형식의 해를 구하여 얻는다.

$$\vec{a} = \mathrm{argmin}_a \frac{1}{N} \sum_{n=1}^{N} (s_i(n) - \hat{s}_i(n))^2 \quad (2)$$

2. 위에서 구한 필터계수 $\vec{a}$를 다음과 같이 LP 켑스트럼 계수 (cepstral coefficient) $\vec{c} = (c_1, ..., c_n)$으로 변환한다.

$$c_m = \begin{cases} a_m + \sum_{k=1}^{m-1} (1 - \frac{k}{m}) a_k c_{m-k} & 1 \le m \le p \\ \sum_{k=1}^{p} (1 - \frac{k}{m}) a_k c_{m-k} & p < m \le n \end{cases} \quad (3)$$

3. LP 스펙트럼에서 낮은 주파수로부터 피크가 되는 지점을 차례로 구하면 F1, F2, F3를 얻을 수 있다.

3.2.3 화면 사상 (Mapping to Screen)

각 음성을 표현하는 특징을 F1, F2, F3의 세 개의 포먼트 값을 원소로 하는 3차원 벡터로 정의하기 때문에 이를 2차원 화면에 표현하려면 차원 감소를 하는 과정이 필요하다. 이를 위해서 SOM 알고리즘 [13]을 이용하면 3차원 공간에 분포한 음성 데이터들의 상호 거리에 기반한 내재적인 구조를 유지하면서 2차원 공간에 사상시킬 있다. SOM의 구조는 그림 5에 나타내었듯이 2차원 그리드(grid)로 구성된 뉴런 노드와 연결 가중치로 이루어진다.

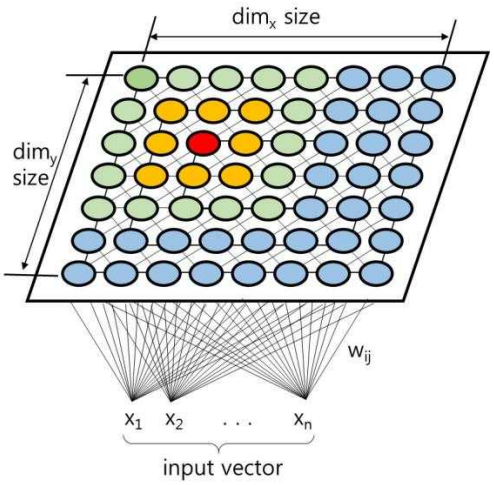

그림 5. SOM의 구조
Fig. 5. Structure of SOM

즉, 입력 벡터 $\vec{x} = (x_1, ..., x_n)$의 j 번째 요소 x_j와 i 번째 노드는 연결가중치 w_{ij}로 연결되며 다음과 같은 학습 과정을 통해 최적화된 w_{ij}의 값을 구한다.

1. 연결가중치 w_{ij}를 0과 1 사이의 난수로 초기화 한다.
2. 모든 입력벡터에 대해 하나씩 i 번째 노드 사이의 거리를 계산한다.

$$D_{ij} = \sum_{j=1}^{N} (w_{ij} - x_j)^2 \quad (4)$$

3. 위에서 계산한 D_{ij}값 중에서 제일 작은 값을 가지는 노드 i_0의 연결가중치와 이웃 노드의 연결가중치를 갱신한다.

$$w_{i_0}^{t+1} = w_{i_0}^{t} + \mu_t \lambda (x - w_{i_0}^{t}) \quad (5)$$

여기서 μ_t는 시간에 따라 값이 변하는 학습률 (learning rate)이고 λ함수는 가우시안 분포함수와 같이 중심에서 멀어질수록 낮은 가중치를 부여하는

함수이다.

4. 현재 입력 벡터가 마지막 벡터이거나 최대 반복 횟수에 도달했으면 학습을 종료하고, 그렇지 않은 경우 위의 2번으로 가서 학습을 계속한다.

위와 같은 학습과정을 마치면 그 결과로 코드벡터의 집합을 얻게 된다. 이를 토대로 화면 사상을 하는 과정은 다음 장에서 기술한다.

4. 실험 및 고찰

제안한 방법을 수집한 음성 데이터를 사용하여 실제로 구현하기 위해 다음과 같이 실험을 진행하였다. 음성특징 추출은 공개 소프트웨어인 Deep Phonetic Tools[14] 사용하였다.

4.1 음성 녹음 환경 및 자료 수집

실험을 위한 음성자료는 한국인 남성 교사와 P대학교에 유학 중인 외국인 남학생 10명의 발음을 녹음하여 수집하였다. 한국인 교사의 표준발음 녹음은 비교적 방음이 잘되어 있는 방에서 헤드셋 마이크를 이용하여 녹음을 하였고, 학습자 발음은 잡음 환경하의 발음을 수집하려는 목적으로 유학생들의 개인 스마트폰을 이용하여 녹음하였다.

한국어 모음 중에 /아, 에, 이, 오, 우/의 다섯 모음을 각각 5회씩 발성하여 녹음을 하고, 이 중에서 발음이 명확하지 않은 것을 제외하고 실험 데이터를 구성하였다. 실험데이터의 구성은 각 모음에 대해 세 개의 그룹으로 나누어 표준 발음 10개 (표준발음군), 학습자 발음 중 표준발음에 가까운 발음 10개 (발음군 1), 표준 발음에서 비교적 멀리 떨어진 발음 10개(발음군 2)로 선택하여 구성하였다.

4.2 포먼트 계산

음성 자료의 샘플링 주파수는 11,025Hz로 설정하고 모음구간은 STE(short time energy)와 STZCR(short time zero crossing rate)를 이용하여 구하였다. STE는 유한 구간에서 신호의 에너지를 나타나며 다음과 같이 정의된다.

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \qquad (6)$$

여기서 $h(\bullet)$는 윈도우 함수이다.

모음과 같은 유성음의 에너지는 무성음이나 잡음보다 에너지가 크다는 사실을 이용하면 모음구간을 추출할 수 있다. STZCR은 신호의 부호가 바뀌는 비율을 의미하며 음성 신호는 대체적으로 잡음보다 낮은 주파수를 가지는 성질이 있으므로 음성 신호는 낮은 STZCR 값을 가지는 반면 잡음은 높게 나타나며 다음과 같이 정의된다.

$$Z_n = |s[x(m)] - s[x(m-1)]|\, w(n-m) \qquad (7)$$

여기서 $s[\bullet]$는 부호를 나타내는 sign 함수이다. 이들 기법을 종합하여 음성구간은 STE가 높고 STZCR이 낮은 구간을 찾아서 지정한다.

음성구간을 지정한 후 오픈소프트웨어 Deep Phonetic Tools의 모듈을 사용하여 각 음성데이터에 대해 F1, F2, F3의 세 개의 포먼트를 구하였다. 즉, 각 모음별로 수집한 각 발음군의 10개 음성에서 포먼트값을 구했으며 이들 10개 포먼트값에 대한 평균은 표 1에 정리하였다.

4.3 화면 사상

각 발음에 대해 30개의 음성데이터에서 추출한 특징벡터를 가지고 SOM 특징맵을 구하였다. 그림 6(a)는 /아/ 발음에 대한 특징맵을 보여준다. N으로 표시한 노드는 표준발음에 대한 코드벡터를 나타내고, S 노드는 발음군 1을 나타내며, T 노드는 발음군 2를 나타낸다.

표 1. 실험 음성의 포먼트 값 평균(Hz)
Table 1. Formant means for speech data(Hz)

발음	그룹	F1	F2	F3
/아/	표준발음	651.3	1156.1	2515.3
	발음군 1	639.1	1280.8	2528.1
	발음군 2	763.4	1446.0	2734.3
/에/	표준발음	391.1	1876.5	2593.7
	발음군 1	341.4	1784.2	2453.9
	발음군 2	424.0	1947.8	2534.2
/오/	표준발음	320.6	587.9	2583.2
	발음군 1	342.8	750.5	2681.4
	발음군 2	390.3	811.1	2970.0
/우/	표준발음	324.6	595.5	2508.3
	발음군 1	409.4	962.0	2600.3
	발음군 2	390.7	1090.7	2790.5
/이/	표준발음	236.5	1156.0	2515.8
	발음군 1	422.1	1695.3	2960.2
	발음군 2	301.9	2234.5	2960.8

그림 6(a)의 특징맵에서 학습자가 이르러야 할 표준발음 영역을 노란색으로 나타내었다. 아무런 표시가 없는 4개의 노란색 노드는 현재의 학습데이터(training data)가 부족함을 보여 주며 추후 표준발음 음성을 수집하여 보완해 갈 수 있다. 노란색 노드 중에서 S로 표시된 노드는 학습자 발음 중에서 표준발음과 거의 유사한 예를 보여 주며 이 노드에는 표준발음도 포함되었지만 표시가 다만 S로 되었을 뿐이다. 4 번째 행 첫 째 열의 회색의 N노드는 비록 표준발음이지만 클러스터 중심에서 떨어져 있으므로 학습자가 따라야 할 표준발음에 속하지 않는다는 의미로 노란색으로 표시한 목표 발음에서 제외할 수 있음을 보여준다. 그림 6(b)는 위의 특징맵을 실제로 컴퓨터 화면에 사상한 예를 보여 준다. 표준발음 노드들을 빨간색의 점으로 표시하고 발음군 1은 녹색으로 표시했으며 발음군 2은 파란색으로 표시하였다. 각 점들은 음성 데이터들의 상대적 위치를 보여 주고 있다. 따라서 본 시스템을 구현하였을 경우, 사용자가 발화를 하면 발음의 특징에 따른 위치가 화면의 점에 대응되어 표시가 된다. 사용자가 표준발음이나 기존의 발음군 1과 발음군 2의 점을 클릭하여 그 발음과 자신의 발음에 대한 음향적인 특성을 비교할 수 있는데 그림 7은 이를 보여준다.

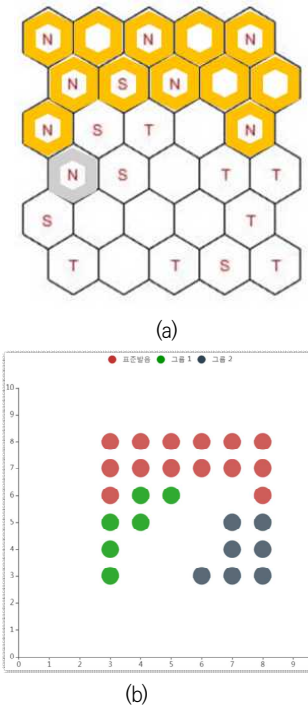

그림 6. (a) /아/ 발음의 SOM 특징맵 (b) 화면 사상 예
Fig. 6. (a) Feature map of /아/ speeches (b) Screen mapping example

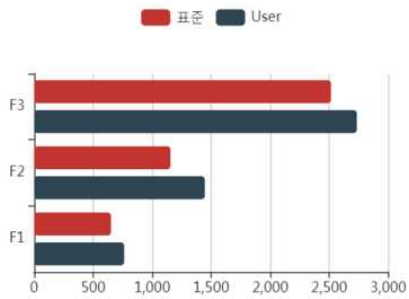

그림 7. 표준음성과 사용자 음성의 특성비교
Fig. 7. Comparison of acoustic features between standard pronunciation and the user's pronunciation

그림 8은 /에/ 발음의 SOM 특징맵과 그에 대응하는 화면 사상의 예를 보여 준다. 가운데를 중심으로 분포된 노란색의 표준발음 주위로 윗부분에는 주로 발음군 1이 분포되어 있고 아랫부분에는 주로 발음군 2가 분포되어 있다. 다만 청각적으로 표준발음에서 더 멀리 떨어진 발음군 2가 표준발음 그룹에서

더 멀리 떨어져 있는 것을 알 수 있다.

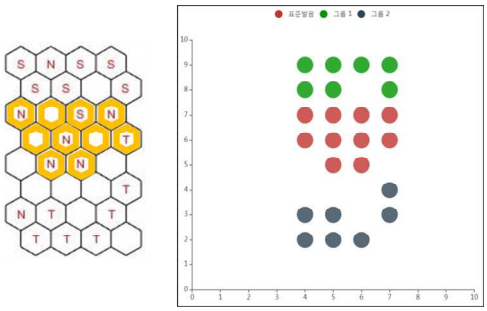

그림 8. /에/ 음성의 SOM 특징맵과 화면 사상 예
Fig. 8. Feature map of /에/ speeches and their screen mapping example

/오/ 발음의 SOM 특징맵과 그에 대응하는 화면 사상의 예를 그림 9에 나타내었다. /오/ 발음의 경우에는 표준발음의 클러스터는 대체적으로 조밀한 양태(compact shape)를 보이지만 발음군 1과 발음군 2의 음성들은 청각적인 분류와는 다르게 서로 겹치고 있고 각각 조밀한 모습을 보이지 않고 있다. 그리고 5번 째 행에 회색으로 표시된 표준발음 노드가 학습자의 발음들과 섞여 있은 경우에는 앞서 언급하였듯이 표준발음으로 분류하지 않는 방향으로 시스템을 구성할 수 있다. 이와 같이 클러스터의 구성에서 일관되지 않은 모습을 보이는 경우에 화면 맵 구성은 새로운 음성 데이터를 수집하여 특징을 추출한 후에 학습을 하여 구성하는 것이 바람직하다. 따라서 그림 9의 /오/ 발음 같이 학습자 음성들이 조밀한 클러스터를 형성하지 않는 경우를 대비하여 학습자의 샘플데이터를 충분하게 수집하는 대책이 마련되어야 한다. 이와 같이 한국어 모음에 대해 F1, F2, F3의 3개 포먼트를 사용하여 구현한 시스템은 대체적으로 조밀한 클러스터를 형성하고 이를 이용하여 학습자의 발음이 표준발음으로부터 얼마나 떨어져 있는지 스스로 학습하면서 인지할 수 있는 시스템을 구현해 보였다.

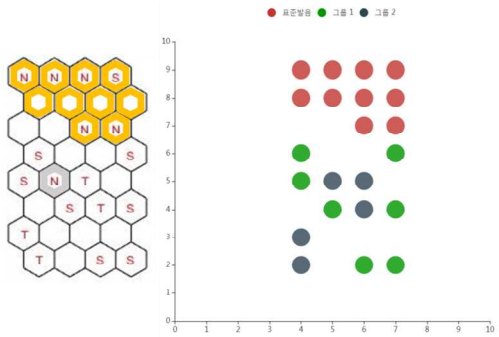

그림 9. /오/ 발음의 SOM 특징맵과 화면 사상 예
Fig. 9. Feature map of /오/ speeches and their screen mapping example

4. 결론

본 논문에서는 기존의 음성시각화 방법과 달리 음성의 특징 사이의 상대적 거리를 기반으로 표준발음으로부터 학습자의 발음이 얼마나 많이 떨어져 있는지를 확인할 수 있는 시각화 모델을 제시하였고 이를 한국어 모음에 대해 구체적인 구현을 시행하였다. 3개의 포먼트를 이용한 모음의 특징에 코호넨 자기조직화맵 알고리즘을 적용하여 특징맵을 구한 후 이를 컴퓨터 화면에 사상하는 시스템을 구현하였다. 실험을 통해 3개의 포먼트를 이용하면 각 모음의 조밀한 클러스터를 얻을 수 있음을 확인하였다. 이를 그림 2에서 제시한 한국어 발음 교육시스템 상의 사용자인터페이스에 적용하면 외국인을 위한 발음교정이나 장애인을 위한 발음교육에 유용하게 사용할 수 있을 것으로 기대되며, 특히 본 연구의 후속으로 이어질 연구에서 학습자 스스로 발성을 연습하면서 실시간으로 교정이 가능한 발음 교육시스템을 개발하는 토대가 된다.

REFERENCES

[1] G. J. Borden, K. S. Harris, and L. J. Raphael, Speech science primer: physiology, acoustics, and perception of speech (Kim et al., Trans.), Seoul: Hankookmunhwasa, 2000.

[2] H.-Y. Sim , C.-H. Choi and S. H. Choi,

"Characteristics of Vowel Formants, Vowel Space, and Speech Intelligibility Produced by Children Aged 3-6 Years," Audiology and Speech Research, vol.12, no. 4, pp. 260-269, 2016.

[3] B. Yang, "A study on vowel formant variation by vocal tract modification," Phonetics and Speech Sciences vol. 3, no. 4, pp. 83-92, 1998.

[4] G. C, Yoon "A Comparative Study on the Male and Female Vowel Formants of the Korean Corpus of Spontaneous Speech," Phonetics and Speech Sciences vol. 7, no. 2, pp. 131~138, 2015.

[5] A. Watanabe, S. Tomishige, and M. Nakatake "Speech Visualization by Integrating Features for the Hearing Impaired", IEEE Trans. Speech Audio Proc., vol 8, no 4, pp. 454-466, 2000.

[6] J. Beskow, O. Engwall, B. Granstrom, P. Nordqvist, and P. Wik, "Visualization of Speech and Audio for Hearing Impaired Persons," Technology and Disability, vol 20, pp. 97-107, 2008.

[7] Y. Ueda, T. Sakada, and A. Watanabe, "Real-time Speech Visualization System for Speech Training and Diagnosis," Audio Engineering Society Convention Paper 8184, 2010 November 4, San Fransico, USA.

[8]D. S. Kim, T. H. Lee, and D. M. Lee, "An ambient display for hearing impaired people," Proc. Human Computer Interface Korea (HCI2006), pp.46 - 51, 2006.

[9] J.-H. Lee and H.-J. Chung, "A Study on Frequency Characteristics of Korean Phonemes," Audiology, 제1권, pp. 59-66, 2005.

[10]P. Denes and E. Pinson, The Speech Chain, W. H. Freeman and Company, (Ko et al., Trans.) 1995.

[11] B. Yang, "Formant Measurements of Complex Waves and Vowels Produced by Students," Phonetics and Speech Sciences vol. 15, no. 3, pp. 39-52, 2008.

[12] Y. Dissen, J. Goldberg, and J. Keshet, "Formant Estimation and Tracking: A Deep Learning Approach", J. Acoustic Society, vol.145, no.2, pp.1-11, 2019.

[13] Kohonen, "Clustering, taxonomy, and topological maps of patterns," Proc. 6th Int. Conf. on Pattern Recognition, pp. 114-128, Washington, DC. IEEE Computer Soc. Press.

[14] State-of-the-art accurate phonetic tools based on machine-learning, https://mlspeech.github.io/index.html.

저자약력

복 거 철(Gouchol Pok) **[회원]**

1981년 8월: 연세대학교 수학과 이학사

1999년 8월: Texas A&M Unversity, 컴퓨터공학과 공학박사

2001년 3월 ~ 2010년 2월: 연변과학기술대학교 교수

2016년 3월 ~ 현재: 배재대학교

〈관심분야〉 신경망, 패턴인식, 영상처리