한국정보전자통신기술학회논문지(jkiiect)'20-6, Vol.13 No.3
논문 20-13-03-197 http://dx.doi.org/10.17661/jkiiect.2020.13.3.197

음성특징의 거리에 기반한 한국어 발음의 시각화

복거철*

Visualization of Korean Speech Based on the Distance of Acoustic Features

Gou-Chol Pok*

요 약 한국어는 자음과 모음과 같은 음소 단위의 발음은 고정되어 있고 표기에 대응하는 발음은 변하지 않기 때문에 외국인 학습자가 쉽게 접근할 수 있다. 그러나 단어와 어구, 문장을 말할 때는 음절과 음절의 경계에서 소리의 변동이 다양하고 복잡하며 표기와 발음이 일치하지 않기 때문에 외국어로서의 한국어 표준 발음 학습은 어려운 면이 있다. 그러나 영어 같은 다른 언어와 달리 한국어의 표기와 발음의 관계는 논리적인 원리에 따라 예외 없이 규칙화 할 수 있는 장점이 있으므로 발음 오류에 대해 체계적인 분석이 가능한 것으로 여겨진다. 본 연구에서는 오류 발음과 표준 발음의 차이를 컴퓨터 화면상의 상대적 거리로 표현하여 시각화하는 모델을 제시한다. 기존 연구에서는 발음의 특징을 단지 컬러 또는 3차원 그래픽으로 표현하거나 입과 구강의 변화하는 형태를 애니메이션으로 보여 주는 방식에 머물러 있으며 추출하는 음성의 특징도 구간의 평균과 같은 점 데이터를 이용하는데 그치고 있다. 본 연구에서는 시계열로 표현되는 음성데이터의 특성 및 구조를 요약하거나 변형하지 않고 직접 이용하는 방법을 제시한다. 이를 위해서 딥러닝 기법을 토대로 자기조직화 알고리즘과 variational autoencoder(VAE) 모델 및 마코브 확률모델을 결합한 확률적 SOM-VAE 기법을 사용하여 클러스터링 성능을 향상시켰다.

Abstract Korean language has the characteristics that the pronunciation of phoneme units such as vowels and consonants are fixed and the pronunciation associated with a notation does not change, so that foreign learners can approach rather easily Korean language. However, when one pronounces words, phrases, or sentences, the pronunciation changes in a manner of a wide variation and complexity at the boundaries of syllables, and the association of notation and pronunciation does not hold any more. Consequently, it is very difficult for foreign learners to study Korean standard pronunciations. Despite these difficulties, it is believed that systematic analysis of pronunciation errors for Korean words is possible according to the advantageous observations that the relationship between Korean notations and pronunciations can be described as a set of firm rules without exceptions unlike other languages including English. In this paper, we propose a visualization framework which shows the differences between standard pronunciations and erratic ones as quantitative measures on the computer screen. Previous researches only show color representation and 3D graphics of speech properties, or an animated view of changing shapes of lips and mouth cavity. Moreover, the features used in the analysis are only point data such as the average of a speech range. In this study, we propose a method which can directly use the time-series data instead of using summary or distorted data. This was realized by using the deep learning-based technique which combines Self-organizing map, variational autoencoder model, and Markov model, and we achieved a superior performance enhancement compared to the method using the point-based data.

Key Words : Feature Clustering, Korean Pronunciation, SOM-VAE, 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 May 29, 2020 Revised June 11, 2020 Accepted June 11, 2020

1. 서론

음성의 시각화에 대한 연구는 기존에 많은 연구자들이 관심을 가지고 의미 있는 결과를 도출하였다[1,2,3,4]. Beskow[1]는 각 음성에 대해 입술 모양과 혀의 위치를 그래픽으로 표현하고 3차원 애니메이션으로 보여 주는 시각화 시스템을 제안하였다. Watanabe[2]는 Time Delay 신경망을 이용하여 음소의 특징을 구한 후 이를 컬러로 표현하는 방법을 제시하였다. Ueda[3]는 단어 및 문장에 대한 표준 발음과 사용자 발음의 스펙트럼과 컬러로 표현된 특성을 동시에 보여 주며 비교를 할 수 있도록 하였다. 이와 같은 시각화 방법들은 단지 개별 음성의 특징을 추출하고 시각화하여 보여 줄 뿐이므로 사용자의 발음이 표준발음으로부터 얼마나 멀리 떨어져 있고, 어떤 식으로 발화를 하여야 표준 발음에 가까이 근접해 나갈 수 있는지에 대한 정보를 제시할 수 없는 단점이 있다.

본 연구에서는 학습자의 발음이 표준 발음으로부터 얼마나 떨어져 있는지 그 차이를 상대적인 거리 개념을 도입하여 화면에 표시하는 모델을 제시한다. 학습자는 자신의 발음이 표준발음으로부터 얼마나 떨어져 있는지 확인을 할 수 있으며, 동시에 자신의 혀와 입모양이나 구강의 모양을 변화시켜가면서 교사의 지도가 없이 표준 발음을 습득할 수 있다. 그림 1에 도식화하여 나타낸 전체 발음 학습시스템의 네 모듈 중에서 본 연구에서 다루는 부분은 네 번째의 표준발음 및 오류발음 공간 시각화 모듈로서 주요 가능은 음성에서 특징을 추출하여 클러스터링을 수행한 후 이를 화면에 사상하는 방법을 개발하여 제시하는 것이다. 최종 목표 시스템은 입력 음성으로부터 지속시간, 음의 세기(intensity), 음높이에 대한 시간상의 패턴, 포먼트 주파수의 시간상 변화 패턴, 음성 포락선(envelope)의 패턴, 기식의 양 등의 다중모드 자질을 추출하고, 표준 발음과 오류 발음 사이의 거리를 이들 자질들의 차이로 표현할 수 있는 척도를 설계하며, 이 척도에 따른 발음상의 차이를 2차원 컴퓨터 화면에 상대적인 거리로서 시각적으로 표현하는 소프트웨어 시스템을 개발하는 것이다. 이 시스템의 가장 큰 특징은 한 단어 또는 어구에 대한 발음의 차이를 컴퓨터 화면에 거리로 표현하는 '발음 공간'을 생성하여 이를 이용하는 학습자는 발성기관과 조음 방법을 조정해 가면서 스스로 발음을 교정하며 표준 발음에 접근해 갈 수 있는 수월성에 있다. 본 연구에서 표준발음은 한국인 교사의 정확한 발음을 의미한다.

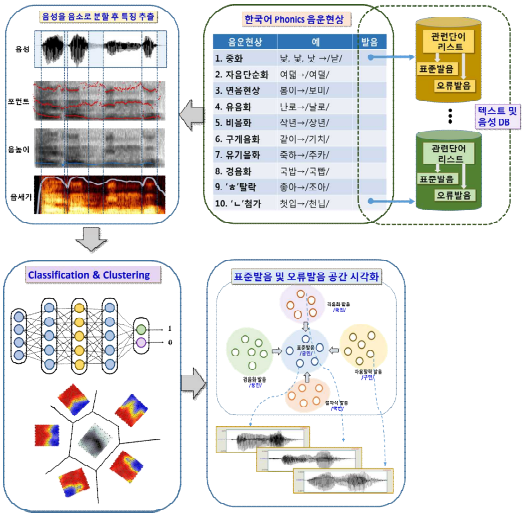

그림 1. 한국어 발음교육 시스템의 개념도
Fig. 1. Schematic overview of the instruction system for Korean pronunciation

2. 방법

2.1 음성특징 추출시 고려할 점

발음 오류의 원인은 복합적인 양상을 띠고 있기 때문에 다중모드 자질(multimodal features)을 다루어야 하고, 이에 따라 표준 발음과 오류 발음 사이의 차이를 단순 비교하기에는 어려움이 있다. 예를 들어 평음, 격음, 경음을 구분하는 중요한 자질은 후행 모음의 음높이(pitch)와 기식의 양으로서, 이들 음향적 자질을 /다/, /타/, /따/ 발음에 대해 나타내면 그림 2와 같다. 두

번째 행에 나타낸 음높이는 하나의 수치(주파수)로 표시되는 것이 아니라 시간상의 패턴으로 나타난다. 따라서 단어'갔**다**'가 표준 발음에서는 경음화 되어 /간**따**/로 발음되어야 하는데 학습자의 발음 오류로 인해 /간**다**/와 같이 발음된다면 /따/와 /다/에 대한 음높이 패턴 사이의 차이, 즉 오류의 정도를 나타낼 수 있는 척도 또는 방법이 필요하다. 더욱이 패턴 형태의 음높이와 더불어 세 번째 행에 나타낸 단순 수치 형태의 기식의 양도 함께 사용한 다중모드 척도를 설계하는 것은 단순한 일이 아니다. 따라서 본 연구에서는 점데이터를 기반으로 하는 클러스터링 알고리즘을 적용하지 않고 시계열 데이터를 입력으로 받아서 처리하는 방법을 사용하였으며 그에 대한 자세한 내용은 아래에 2.3에 기술하였다.

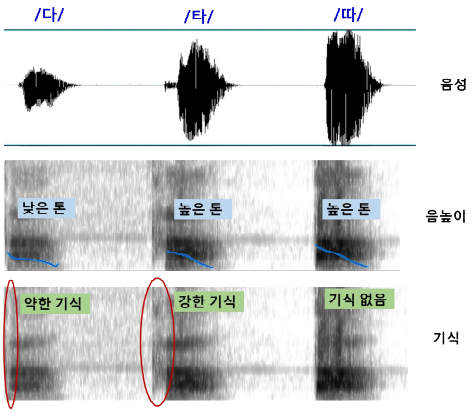

그림 2. 음성 /다/, /타/, /따/의 음높이 패턴과 기식의 특성
Fig. 2. Pitch patterns and breath features for three pronunciations

2.2 음성의 특징 추출

한국어 자음의 음향적 특성은 성대진동 시간(Voice Onset Time, VOT), 후행 모음의 기본 주파수, 파열시 나오는 기류의 세기, 폐쇄지속시간(Closure Duration, CD) 등에 의해 특징지을 수 있다. Silva[5]는 한국어 파열음의 평음, 경음, 격음을 구분하는데 있어서 VOT가 결정적인 역할을 한다고 하였다. 배재연[6]은 파열음의 조음방법에 대한 연구에서 어두 위치에서는 성대진동 시간(VOT)과 후행 모음의 길이가 파열음을 구별하는 데 중요한 역할을 하고, 어중 위치에서는 패쇄지속시간(CD)과 선행 모음의 길이 등이 중요한 역할을 한다고 하였다. 어두 위치, 즉 초성일 경우에 VOT값의 분포를 보면, 경음 /ㅃ, ㄲ, ㄸ/은 20ms 이하의 분포를 이루고, 격음 /ㅍ, ㅋ, ㅌ/는 60~80ms 구간의 분포를 가지며, 평음 /ㅂ, ㄱ, ㄷ/은 40~60ms의 분포를 보이는 것으로 관찰되었다. 어중일 경우의 분포는 경음은 20ms 이하로 줄어들고, 격음은 50~60ms 구간의 분포를 이루며, 평음은 유음화된다고 하였다.

그리고 김소현[7]의 연구에서는 후행모음의 음높이와 폐쇄지속시간은 경음, 격음, 평음을 구별해 주는 속성이 있다고 하였다. 이와 같은 연구결과를 토대로 본 연구에서는 VOT, 음높이, CD를 자음을 구분하는 특징으로 설정하였다.

모음의 특징으로는 음성데이터로부터 세 개의 포먼트 F1, F2, F3를 구하여 사용하였다. 포먼트를 구하는 방법에는 다음과 같이 LPC (linear predictive coding) 분석을 기반으로 하는 방법이 많이 사용된다[8].

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.

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

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

를 구하여 얻는다.

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

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

2.3 시계열 데이터의 클러스터링

음성데이터는 시간축 상에서 정의되는 시계열 데이터로 표현된다. 따라서 시간축 상에서 변화하는 구조를 반영할 수 있는 클러스터링 알고리즘이 요구된다. 최근에 발표된 확률적 SOM-VAE (Self-organizing map with variational auto encoder) 알고리즘은 고차원 시계열 데이터를 저차원 공간에 사상(mapping)을 할 때 부드러운 연속성을 보장하는 성능을 제공해준다[9]. 이 기법을 적용하면 시간축 상에서 변화하는 포먼트를 평균값이라는 점(point) 데이터로 표현하는 대신에 시계열 데이터를 직접 이용할 수 있으므로 정보 손실을 막을 수 있을 것으로 기대된다.

그림 3은 SOM-VAE모델의 개관적인 시스템플로우를 도시하였다. 입력 음성데이터 x는 신경망을 이용하여 구현한 부호기(encoder)를 통해 시간정보가 함축된 잠재적 공간(latent space)의 Z_e에 사상되고, Z_e는 다시 SOM맵의 가장 가까운 노드 Z_q에 사상되어 근사적으로 표현된다. SOM맵 공간에서는 마코브 전이 모델을 적용하여 학습을 수행하는데 이를 통해 현재 상태 Z^t_q에서 다음 상태 Z^{t+1}_q로의 전이를 확률로 예측할 수 있다.

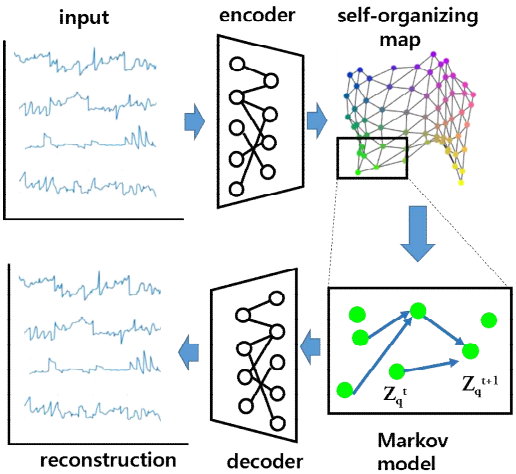

그림 3. SOM-VAE 모델의 개관적인 도해
Fig. 3. Schematic overview of SOM-VAE model

이와 같이 연속적인 시계열 데이터는 이산적인 표현으로 나타낼 수 있는데 다시 복호기(decoder)를 통하여 원래의 시계열 데이터 공간에 사상시킬 수 있다. 본 연구에서는 SOM-VAE 알고리즘에 의해서 구축되는 특징맵을 사용하여 컴퓨터 화면 위에서 음성의 상대적 거리를 나타내었다.

3. 결과

3.1 실험 데이터 및 구성

한국어 발음의 음운 변동은 다양하게 관찰되고 있으나[10] 본 연구에서는 다음과 같은 6종류의 음운 현상에 대해 고찰하였다. 즉, 경음화 (합법→/합뻡/), 비음화(작년→/장년/), 유음화(편리→/펼리/), 유기음화(파랗다→/파라타/), 니은(ㄴ)첨가(여행용→/여행뇽/), 구개음화(맏이→/마지/) 등이다. 실험 데이터는 P대학교에 유학 중인 외국인 남학생 5명과 한국인 교사 1명으로부터 수집하였다. 본 연구의 목적은 표준발음과 다양한 오류 발음 사이의 거리를 계량화하는 것이기 때문에 오류 발음을 충분히 확보하는 것이 중요하다. 따라서 화자가 자유롭게 발성하는 발음을 통해 음성을 수집하기 보다는, 제시하는 목표 발음을 하도록 유도를 하였다. 즉, 경음화가 일어나는 '합법'이란

단어를 발음할 때에 목표 발음을 /핫뻡/, /하법/, /하뻡/ 등으로 오류 발음을 제시하여 발음을 하도록 한 후 이를 수집하였다. 각 음운 현상에 대해 3개씩의 대표적인 오류발음을 설정하고 표준발음을 합쳐서 4개의 발음군을 정한 다음, 각 발음군에서 50개의 음성 표본을 수집하여 총 200개(50×4)의 음성 데이터를 확보하였다.

음성 데이터로부터 2.2에서 기술한 내용대로 VOT, 음높이, CD 및 세 개의 포먼트 F1, F2, F3를 구하였다. 음성의 발화시간은 화자마다 다르기 때문에 지속시간의 정규화가 필요하므로 시계열로 표현된 음높이와 포먼트로부터 다운 샘플링을 하여 각각 50개의 샘플을 균일한 간격에서 추출하고 VOT와 CD를 더하여 102-차원 특징벡터로 만들고 이들 특징벡터의 집합에 SOM-VAE 알고리즘을 적용하여 특징맵을 구하였다.

3.2 실험 결과: 특징맵

각 음운 현상을 나타내는 단어의 음성데이터에 대한 특징벡터 집합에 SOM-VAE 알고리즘을 적용하여 특징맵을 구하였는데 실험 결과는 다음과 같다.

3.2.1 경음화 발음의 분포

경음화 현상은 평음이 경음으로 발음되는 현상으로 단어 '합법'이 /합뻡/으로 발음되는 현상이다. 외국인 학습자들에게 자주 발견되는 오류발음은 /하법/, /하뻡/, /핫법/ 등으로서 이들과 표준발음 /합뻡/을 포함하여 4개의 발음군을 구성하고 SOM-VAE 특징맵을 구축한 결과를 그림 4에 표시하였다. 1번으로 표시된 노드들이 표준 발음 /합뻡/을 나타내고, 2번에서 4번까지의 노드들은 각각 /하법/, /하뻡/, /핫법/을 나타낸다. 노드들의 분포를 보면 군집화가 잘 이루어져 있을을 알 수 있으며 따라서 이 특징맵을 컴퓨터 화면에 사상을 하면 학습자의 오류 발음이 표준발음으로 얼마나 떨어져 있는지 상대적인 거리를 표시할 수가 있음을 알 수 있다.

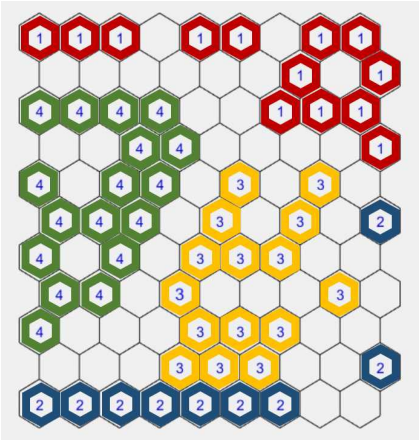

그림 4. 단어 '합법'에 대한 특징맵 상의 발음 분포
Fig. 4. Pronuciation distributions on the feature map ofr the word 'Hapbup' ('합법')

3.2.2 비음화 발음의 분포

비음화 현상은 비음이 아닌 범주에 속한 자음이 비음으로 발음되는 현상으로 단어 '작년'이 /장년/으로 발음되는 것과 같은 현상이다. 이에 관련된 오류발음은 /자년/, /자연/, /잔년/으로 정하고 표준발음을 포함하여 4개의 발음군을 구성하고 SOM-VAE 특징맵을 구축하였으며 그 결과를 그림 5에 표시하였다. 표준발음을 나타내는 1번 노드들은 2개의 그룹으로 나누어져 있음을 볼 수 있다. 이런 결과가 나온 이유는 고차원의 특징벡터들의 특성을 2차원에 사상할 때 왜곡이 일어난 결과일 수 있다. 따라서 특징맵을 화면에 사상할 때 2개의 표준발음 영역을 지정하여 나타낼 필요가 있다.

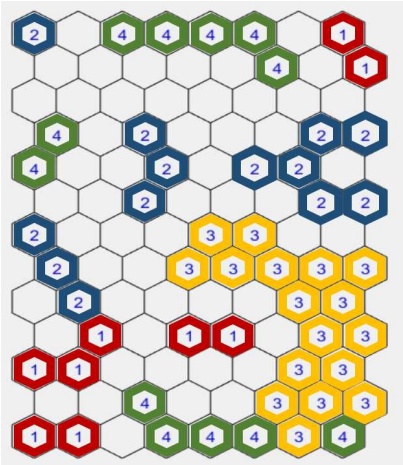

그림 5. 단어 '작년'에 대한 특징맵 상의 발음 분포
Fig. 5. Pronuciation distributions on the feature map ofr the word 'Jaknyeon' ('작년')

3.2.3 유음화 발음의 분포

한국어 자음 체계에서 유음은 /ㄹ/ 하나이다. 유음화 현상은 비음 /ㄴ/이 유음 /ㄹ/로 변동되어 실현되는 현상으로 단어 '편리'가 /펼리/로 발음되는 것 같은 현상이다. 단어 '편리'에 관련된 오류 발음은 /펴리/, /펴니/, /편니/로 정하고 표준발음을 포함하여 4개의 발음군을 구성하고 SOM-VAE 특징맵을 구축하였으며 그 결과를 그림 6에 표시하였다. 그림에서 /편니/에 해당하는 4번 노드들은 2개의 그룹으로 나누어져 있고 다른 발음들은 모두 군집화를 이루는 것을 볼 수 있다.

3.2.4 유기음화 발음의 분포

유기음화는 격음화라고 불리기도 하며 자음 /ㄱ, ㄷ, ㅂ, ㅈ/이 'ㅎ'을 만나 / ㅋ, ㅌ, ㅍ, ㅊ/으로 발음되는 발음되는 현상으로 단어 '파랗다'가 /파라타/로 발음되는 것과 같은 현상이다. 이에 관련된 오류 발음은 /파랏다/, /파라다/, /파라따/로 설정하여 표준발음을 포함하여 4개의 발음군을 구성하고 SOM-VAE 특징맵을 구축하였으며 그 결과를 그림 7에 표시하였다. 표준발음을 나타내는 1번 노드들은 군집화가 잘 이루어진 것을 볼 수 있다. 따라서 화면 사상을 실행하기에 적당함을 알 수 있다. 다만 /파랏다/에 해당하는 2번 노드들은 3개의 그룹으로 나누어져 있음을 볼 수 있다.

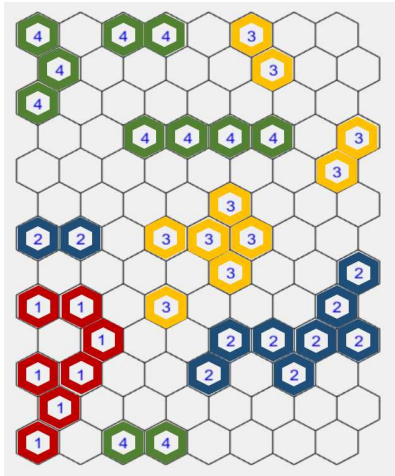

그림 6. 단어 '편리'에 대한 특징맵 상의 발음 분포
Fig. 6. Pronuciation distributions on the feature map ofr the word 'Pyunli' ('편리')

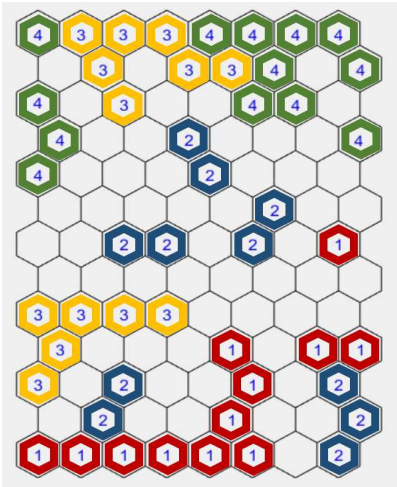

그림 7. 단어 '파랗다'에 대한 특징맵 상의 발음 분포
Fig. 7. Pronuciation distributions on the feature map ofr the word 'Paratta' ('파랗다')

3.2.5 니은('ㄴ') 첨가 발음의 분포

니은('ㄴ') 첨가는 발음의 원래 형태에 관계가 없는 니은('ㄴ')음이 추가되는 현상을 의미하며 단어 '여행

용'이 /여행뇽/으로 발음되는 것과 같은 현상이다. 오류 발음은 /여해용/, /여해뇽/, /여핸뇽/로 설정하여 표준발음을 포함하여 4개의 발음군을 구성하고 SOM-VAE 특징맵을 구축하였으며 그 결과를 그림 8에 표시하였다. 그림에서 표준발음을 나타내는 1번 노드들이 2개의 그룹으로 나누어진 것을 볼 수 있다. 앞서 언급을 하였듯이 표준발음을 두 개의 그룹으로 설정하여 화면사상을 하면 이런 구조에서도 발음 사이의 상대적 거리를 잘 표현할 수가 있다.

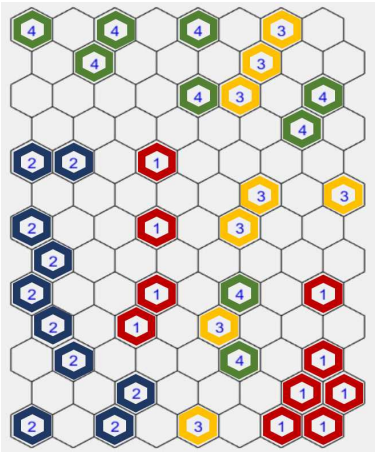

그림 8. 단어 '여행용'에 대한 특징맵 상의 발음 분포
Fig. 8. Pronuciation distributions on the feature map ofr the word 'Yeohangyong' ('여행용')

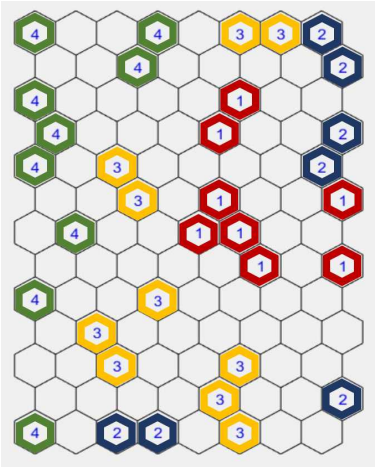

그림 9. 단어 '맏이'에 대한 특징맵 상의 발음 분포
Fig. 9. Pronuciation distributions on the feature map ofr the word 'Maji' ('맏이')

3.2.6 구개음화 발음의 분포

구개음화는 구개음이 아닌 음운이 경구개음으로 발음되는 현상으로 단어 '맏이'가 /마지/로 발음되는 것과 같은 현상이다. 이에 관련된 오류 발음은 /마이/, /마디/, /마시/로 설정하여 표준발음을 포함하여 4개의 발음군을 구성하고 SOM-VAE 특징맵을 구축하였으며 그 결과를 그림 9에 표시하였다. 그림에서 표준발음을 나타내는 1번 노드들은 군집화가 잘 이루어진 것을 볼 수 있다. 다만 /마디/에 해당하는 2번 노드들과 /마이/에 해당하는 3번 노드들은 2개의 그룹으로 나누어져 있음을 볼 수 있다.

위에서 고찰한바와 같이 본 연구에서 제안하는 방법에 의해서 발음 데이터들은 포먼트 특징에 따라서 화면에 적절하게 사상시킬 수 있음을 보였다.

3.3 시계열 기반 특징과 점기반 특징의 비교

앞 절에서 도식화한 특징맵은 시각적인 판단을 위한 것으로서 군집화의 성능을 객관적이고 정량적으로 판단하기 위해서 군집 타당성 지표(Cluster Validation Index, CVI)를 구하여 비교할 수가 있다. 본 연구에서 제안하는 시계열 데이터를 이용하는 방법으로 구하는 특징맵 클러스터와 시계열의 평균을 구하고 이를 대표값으로 하는 기존의 점 데이터(point data) 기반 방법으로 구하는 특징맵 클러스터의 군집화 성능을 비교하기 위해서 Davies Bouldin index (DBI)[11]와 Silhouette index[12]를 계산하였다.

DBI는 클러스터내의 모든 그룹을 두 개의 쌍으로 연결 짓고 이 쌍에 대해서 각 그룹의 크기의 합을 각 그룹의 중심 사이의 거리로 나눈 값으로서 나타낸다. DBI 값이 유효할수록 같은 그룹 내의 데이터의 분포는 밀집화를 이루게 되고, 다른 그룹 사이의 거리는 커지므로 DBI 값이 작을수록 클러스터는 더욱 유효하다고 판단할 수 있다. DBI는 다음과 같이 정의된다.

$$DBI = \frac{1}{N}\sum_{n=1}^{N}\frac{S_i + S_j}{M_{ij}} \quad (4)$$

여기에서 S_i는 i번째 그룹을 나타내며 M_{ij}는 S_i와 S_j의 거리를 나타낸다.

$$S_i = \left\{\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j - A_i|^q\right\}^{\frac{1}{q}} \quad (5)$$

여기에서 T_i는 i번째 클러스터의 개체 X_j의 개수를 나타내고 A_i는 i번째 클러스터의 중심점을 나타낸다.

Silhouette index는 다음과 같이 정의 되는데

$$SI(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \quad (6)$$

여기에서 $a(i)$는 i번째 개체와 같은 클러스터에 속한 개체들 사이 거리들의 평균이다. $b(i)$는 i번째 개체와 다른 클러스터에 속한 요소들 사이 거리들의 평균을 클러스터마다 각각 구한 뒤, 이 중에서 최소값을 나타낸다.

Silhouette index 값에서 가장 이상적인 경우는 $a(i)$가 0인 경우로서 한 클러스터의 모든 개체가 떨어져 있지 않고 붙어있는 상태를 나타나며 Silhouette index 값은 1이 된다. 최악의 경우는 서로 다른 클러스터가 구분되지 않으면 $b(i)$값은 0이 된다.

표 1은 본 연구에서 제안하는 시계열 기반 특징벡터를 이용하는 방법과 기존의 점데이터 기반 특징벡터를 이용한 방법의 군집타당성 지표값을 비교한 것이다.

표 1. 군집타당성지표 값의 비교
Table 1. Comparison of Cluster Validity Index

	DB index		Silhouette index	
	제안 방법	비교 방법	제안 방법	비교 방법
‘합법’	**0.67**	0.89	**0.75**	0.66
‘작년’	**0.78**	0.91	**0.61**	0.54
‘편리’	**0.71**	0.87	**0.73**	0.67
‘파랗다’	**0.80**	0.90	**0.59**	0.49
‘여행용’	**0.82**	0.95	**0.56**	0.47
‘맏이’	**0.79**	0.92	**0.62**	0.57

표 1에서 DBI는 작을수록, Silhouette index는 클수록 클러스터가 더욱 유효하다. 표 1이 보여주듯이 본 연구에서 제안하는 방법이 더 유효하게 클러스터를 구성하는 것을 알 수 있다.

4. 결론

한국어 발음은 음절과 음절의 경계에서 음운의 변동이 다양하고 복잡한 언어에 속하여서 언어의 표기와 발음이 일치하지 않는 특징이 있다. 따라서 외국인 학습자가 한국어의 표준 발음을 학습하기가 어려운 면이 있다. 본 연구에서는 발음의 차이를 컴퓨터 화면상의 상대적 거리로 표현하는 시각화하는 모델을 개발하기 위하여 시계열로 표현되는 음성데이터 패턴을 직접 이용하는 방법을 제시하였다. 즉, 딥러닝 기법이 적용된 자기조직화 알고리즘과 VAE 모델 및 마코브 확률모델을 결합한 확률적 SOM-VAE 기법을 사용하여 클러스터링 성능을 향상시켰다. 제안하는 방법을 6개의 음운 현상을 나타내는 발음에 적용시켜 본 결과 화면상의 위치에 적절히 배치할 수 있음을 실험으로 확인하였다.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] D. Silva, "Variation in Voice Onset Time for Korean Stops: A Case for Recent Sound Change", Korean Linguistics, vol. 13, 2006.

[6] J. Y. Bae, "Acoustic Characteristics of Korean Stop Sounds According to Phonetic Environment: Focusing on Features on the Time Line", Phonectics and Speech Sciences, vol. 5.2, pp.139-159, 1999.

[7] S. H. Kim, "A Study on Korean Affricates Produced by Vietnamese Speakers", vol. 59, pp. 145-168, Korean Linguistics, 2013.

[8] 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

[9] V. Fortuin, M. Huser, F. Locatello, H. Stratman, and G. Ratsch, "Deep Self-Organization: Interpretable Discreate Representation Learning on Time Series", arXiv:1806.02199, 2018.

[10] H. S. Shin, "Phonological Information in Korean Language", Prunsasang, 2016.

[11] D. Davies and D. Bouldin. "A Cluster Separation Measure". IEEE Trans on Pattern Analysis and Machine Intelligence. vol.2, pp.224-227, 1979.

[12] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. vol. 20, pp.53-65, 1987.

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