pISSN 2005-8063
eISSN 2586-5854
2019. 9. 30.
Vol.11 No.3
pp. 39-47

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2019.11.3.039

Performance of Korean spontaneous speech recognizers based on an extended phone set derived from acoustic data*

Jeong-Uk Bang[1] · Sang-Hun Kim[2] · Oh-Wook Kwon[3,**]

[1]*Department of Control and Robot Engineering, Graduate School, Chungbuk National University, Cheongju, Korea*
[2]*Electronics and Telecommunications Research Institute, Daejeon, Korea*
[3]*School of Electronics Engineering, Chungbuk National University, Cheongju, Korea*

Abstract

We propose a method to improve the performance of spontaneous speech recognizers by extending their phone set using speech data. In the proposed method, we first extract variable-length phoneme-level segments from broadcast speech signals, and convert them to fixed-length latent vectors using an long short-term memory (LSTM) classifier. We then cluster acoustically similar latent vectors and build a new phone set by choosing the number of clusters with the lowest Davies-Bouldin index. We also update the lexicon of the speech recognizer by choosing the pronunciation sequence of each word with the highest conditional probability. In order to analyze the acoustic characteristics of the new phone set, we visualize its spectral patterns and segment duration. Through speech recognition experiments using a larger training data set than our own previous work, we confirm that the new phone set yields better performance than the conventional phoneme-based and grapheme-based units in both spontaneous speech recognition and read speech recognition.

Keywords: acoustic units, phone set, spontaneous speech recognition, broadcast data

1. 서론

음소 단위는 오랫동안 음성인식을 위한 음향 모델링 단위로 사용되어왔다. 최근에는 종단 간(end-to-end) 음성인식 시스템이 대중화됨에 따라 발음사전이 필요 없는 자소 단위가 관심을 끌고 있다. 하지만 자소 단위는 다양한 스펙트럼 패턴을 가지는 음성 신호가 하나의 자소 기호에 사상되기 때문에, 자소 단위는 음소 단위보다 낮은 음성인식 성능을 보인다. 이러한 결과는 일반적으로 기존의 음성인식 시스템과 최근의 종단 간 음성인식 시스템에서 관찰된다(Sainath et al., 2018).

대어휘 연속 음성인식에서 음소 단위는 그 수가 너무 적어서 다양한 음향적 변화를 모두 표현할 수 없다는 문제점이 있다.

* This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korea government [19ZS1140 Development of Core Conversational AI Technologies].
** owkwon@cbnu.ac.kr, Corresponding author
Received 24 July 2019; Revised 2 September 2019; Accepted 23 September 2019

기존의 음성인식기에서는 묵시적 방법(Hain, 2005; Young et al., 1994)과 명시적 방법(Lee & Chung, 2003)으로 단어의 다양한 발음변이 현상을 해결하고자 노력하였다. 여기서, 묵시적 방법은 인접한 음소를 고려하여 단위를 확장시킨 다음, 결정 트리(Young et al., 1994)를 사용하여 음향적 특성이 유사한 모델의 매개변수를 공유시킴으로써, 음소 단위보다 더 세분화된 단위를 생성하는 방법이다. 반면에, 명시적 방법은 단어의 변이된 발음을 발음사전에 다중 발음으로 직접 명시하여 변이된 발음을 반영시키는 방법이다. 위 두 방법은 보편적인 음성인식 시스템에서 표준으로 사용된다.

자유발화 음성에서 각 단어는 낭독체 음성보다 더 다양하게 변이된 발음을 가진다. 또한, 자유발화 음성에서의 음소 단위는 낭독체 음성에서의 음소 단위보다 모델 사이의 거리가 더 가까우며, 큰 분산 값을 갖는다(Nakamura et al., 2008). 이러한 모델들은 낮은 변별력을 가지기 때문에 최종적으로 낮은 인식 성능을 보인다. 이러한 상황에서, 공통된 스펙트럼 패턴을 군집시켜 변별력을 높인 새로운 음소 세트를 구축한다면 자유발화 음성인식기의 성능이 향상될 것이라 기대한다.

본 논문에서는 대량의 한국어 방송 데이터로부터 음소 세트를 확장시켜 자유발화 음성인식기의 성능을 향상시키는 방법을 제안한다. 제안된 단위는 세 단계로 추출된다. 먼저 방송 데이터로부터 가변 길이의 음소 레벨 세그먼트들을 추출한다. 다음으로 long short-term memory(LSTM) 구조를 기반으로 고정 길이 잠재 벡터(latent vector)를 추출한 다음, k-means 알고리즘을 사용하여 음향적으로 유사한 잠재 벡터를 군집화한다. 이후, Davies-Bouldin(DB) index(Davies & Bouldin, 1979)를 사용하여 군집 내 높은 유사도를 가지면서 군집 간 낮은 유사도를 보이는 최적의 군집 개수를 찾고, 군집된 벡터를 수집하여 새로운 음소 세트를 정의한다. 제안된 단위는 자유발화 음성인식 작업에서 음소 기반 단위와 자소 기반 단위보다 우수한 성능을 나타낸다.

본 논문은 선행 연구(Bang et al., 2019)를 확장시킨 것이다. 본 논문에서는 제안된 단위들을 자세하게 분석한 결과가 추가되었다. 또한, 선행 연구에서 향후 계획으로 언급되었던 자동으로 군집 개수를 선택하는 방법을 제시하고, 대량의 음성 데이터를 음향모델 학습에 모두 사용하여 제안된 단위의 음성인식 성능을 확인하였다. 이 논문은 다음과 같이 구성된다. 먼저, 2절에서는 선행 연구에서 제시된 단위 생성 방법과 발음사전을 구축하는 방법을 설명하고, 3절에서는 제안된 군집 개수 자동 선택 방법을 보인다. 4절에서는 실험에 사용된 한국어 방송 데이터와 제안된 단위 생성 결과를 보이고, 5절에서는 음성인식 결과를 보인다. 마지막으로 6절에서 결론을 제시한다.

2. 선행 연구

본 절에서는 선행 연구(Bang et al., 2019)에서 제안된 단위 생성 방법과 발음사전 구축 방법을 설명한다. 먼저, 제안된 단위 생성 방법은 세그먼트 추출 단계, 잠재벡터 추출 단계, 세그먼트 군집화 단계로 세분화하여 설명되며, 발음사전 구축 방법은 통계적으로 음소 단위 발음사전을 제안된 단위 발음사전으로 업데이트하는 방법을 설명한다.

2.1. 세그먼트 추출

제안된 방법은 먼저 문장 단위로 구성된 방송 데이터로부터 음소 단위의 음성 세그먼트를 추출한다. 이전의 연구(Mitra et al., 2016)에서는 프레임 단위의 특징벡터로부터 새로운 단위를 탐색하였다. 하지만, 프레임 단위의 특징벡터는 25.6 ms의 좁은 음성 구간을 다루기 때문에, 자유발화 음성의 다양한 스펙트럼 패턴을 다루기에는 어려움이 있다. 이러한 이유로, 본 논문에서는 음운론적 지식을 기반으로 미리 정의된 음소 단위를 음성 세그먼트를 추출을 위한 기본 단위로 사용하였다.

음성 세그먼트는 음향모델 학습 과정에 일반적으로 사용되는 텍스트-음성 정렬(text-to-speech alignment) 방법으로 추출된다. 세그먼트 추출 실험은 선행 연구(Bang et al., 2017)에서 사용된 깊은 신경망 기반의 은닉 마르코프 모델(deep neural network-based hidden Markov model, DNN-HMM)을 텍스트-음성 정렬을 위한 기본 음향모델로 사용하며, 이들의 입력 특징벡터는 좌/우 각 7 프레임을 연결시킨 로그 멜 스케일 필터뱅크(log mel-scaled filter bank)를 사용하였다. 추출된 음성 세그먼트는 최소 3 프레임 이상의 다양한 길이를 가진다.

2.2. 잠재벡터 추출

가변 길이 세그먼트로부터 곧바로 유사한 스펙트럼 패턴을 찾는 것은 어려움이 있다. 따라서 본 단계에서는 다양한 길이를 가지는 음성 세그먼트를 고정된 길이의 잠재벡터로 변환시킨다. 잠재벡터를 추출하기 위한 방법으로 그림 1(a)의 LSTM 오토인코더 구조(Chung et al., 2016)와 그림 1(b)의 LSTM 분류기 구조를 고려할 수 있다.

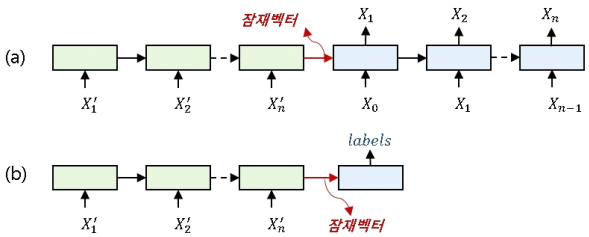

그림 1. LSTM 오토인코더 모델(a)과 분류기 모델(b)의 구조
Figure 1. Structures of LSTM auto-encoder model (a) and LSTM classifier model (b)
LSTM, long short-term memory.

그림 1에서 X_i는 길이가 n인 한 세그먼트에서 i번째 프레임의 40차 로그 멜 필터뱅크 특징벡터를 나타내며, $X_n^{'}$은 인접한 ±1 프레임을 포함한 특징벡터를 나타낸다. 그림 1(b)에서 레이블(labels)은 한국어 음성인식에서 일반적으로 사용되는 40개의 음소 기호를 나타낸다. 그림 1의 두 구조는 각 세그먼트의 특징벡터를 인코더(encoder) 단의 입력으로 동일하게 사용한다. 반면에, 디코더(decoder) 단은 오토인코더 구조에서는 각 세그먼트의

특징벡터를 출력으로 가지지만, 분류기 구조에서는 각 세그먼트의 음소 기호를 출력으로 두는 차이를 보인다. 여기서, 고정된 길이의 잠재벡터는 인코더 단에서 디코더 단으로 넘어가는 80차 잠재벡터를 추출함으로써 얻어진다.

선행 연구(Bang et al., 2019)에서 오토인코더 구조와 분류기 구조에서 얻어진 잠재벡터들을 비교하였다. 그 결과로 오토인코더 구조에서 얻은 잠재벡터는 음성 세그먼트의 길이 정보를 가장 중요한 요인으로 저장시키기 때문에, 분류기 구조에서 얻은 잠재벡터가 단위 생성을 위한 특징벡터로 더 적합함을 확인하였다. 따라서 본 논문에서는 세그먼트 추출 단계에서 얻은 다양한 길이의 세그먼트들을 분류기 구조를 사용하여 고정된 길이의 잠재벡터로 변환하였다.

실험에 사용된 LSTM 분류기는 활성화 함수로 tanh를 사용하면서 노드 개수 80개인 메모리 셀(memory cell)과 은닉층(hidden layer)으로 구성된다. 전체 데이터는 0.01의 학습률(learning rate)로 학습에 10번 사용되며, β_1=0.9, β_2=0.999, ϵ=10^{-8}의 파라미터를 가지는 아담(Adam) 최적화 알고리즘으로 학습된다. 분류기 구조에서 80차원의 잠재벡터를 사용하였지만, 차원의 크기와 인코더 단, 디코더 단의 레이어 수는 모두 최적화되지 않았다.

2.3. 세그먼트 군집화

가장 일반적인 군집화 알고리즘인 k-means 알고리즘(MacQueen, 1967)을 사용하여 이전 단계에서 추출된 잠재벡터를 군집시킨다. k-means 알고리즘에 사용되는 유클리드 거리(Euclidean distance)는 매니폴드 공간(manifold space)에서 유의미한 거리 값을 가진다. 약 5,000만 개의 세그먼트로 구성된 전체 방송 데이터를 모두 군집화 실험에 사용하기에는 물리적으로 어려움이 있었다. 따라서 각 음소 별로 10,000개의 세그먼트를 임의로 선택한 다음, 선택된 400,000개(=40×10,000)의 샘플을 대상으로 k-means 알고리즘을 수행하였다.

2.4. 발음사전 업데이트

발음사전 업데이트 단계에서는 음소 단위의 발음사전(lexicon)을 제안된 단위의 발음사전으로 업데이트시킨다. 이를 위해서, 먼저 앞서 학습된 k-means 모델을 사용하여 전체 방송 데이터의 모든 세그먼트에 대한 군집번호를 계산한다. 그림 2는 한글 단어 '말'의 음소 시퀀스[m a r]에 대한 각 군집번호의 히스토그램이다. 본 단계에서는 제안된 방법을 설명하기 위해서 120개의 군집(k=120)을 사용하였다.

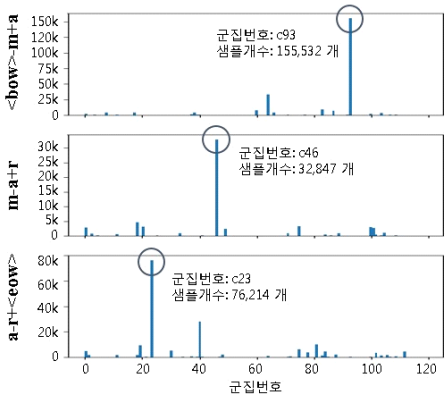

그림 2. 음소 시퀀스 [m a r]에 대한 각 군집번호의 히스토그램
Figure 2. Histogram of each cluster index for the phoneme sequence [m a r]

여기에서 기호 '<bow>'와 기호 '<eow>'는 단어의 시작과 끝을 나타내는 더미 기호(dummy symbol)이다. 단어 '말'의 첫 번째 음소 /m/는 '<bow>-m+a'와 같이 인접한 음소를 고려시켰을 때 군집번호 93에서 가장 높은 빈도 값을 보인다. 유사하게, 음소 /a/와 음소 /r/는 인접한 음소를 고려하였을 때 군집번호 46, 군집번호 23에서 가장 높은 빈도 값을 보인다.

이후, 수식 (1)과 같이 음소 시퀀스(P)가 주어졌을 때 가장 높은 조건부 확률(Bang et al., 2019)을 가지는 군집번호 시퀀스(C)를 찾는다. 여기서 낮은 빈도를 가지는 음소 기호로부터 조건부 확률을 추정하는 것을 피하기 위해, 수식 (2)와 같이 문맥 길이에 제약($\widetilde{P}_i$)을 두어 조건부 확률을 계산하였다.

$$\log P(\boldsymbol{C}|\boldsymbol{P}) = \sum_{i=1}^{n} \log P(C_{i-1}, C_i|\widetilde{P}_i) - \sum_{i=1}^{n} \log P(C_{i-1}|\widetilde{P}_i) \tag{1}$$

$$\widetilde{P}_i = \begin{cases} P_i^{tri} & \text{if} \quad \#\{P_i^{tri}\} > \alpha \times \#\{P_i^{bi}\} \\ P_i^{bi} & elif\ \#\{P_i^{bi}\} > \alpha \times \#\{P_i^{uni}\} \\ P_i^{uni} & else \end{cases} \tag{2}$$

수식 (1)에서 $P(C_{i-1}, C_i|\widetilde{P}_i)$는 n의 길이를 가지는 단어의 i번째 문맥 의존 음소($\widetilde{P}_i$)가 주어졌을 때, $i-1$번째와 i번째에 군집번호 C_{i-1}와 C_i가 나타날 확률을 의미한다. 수식 (2)에서, #$\{P_i^{tri}\}$는 i번째 음소의 좌/우에 위치한 음소를 고려한 문맥 의존 단위의 세그먼트 개수를 의미하며, #$\{P_i^{bi}\}$는 좌측 음소만을 고려한 문맥 의존 단위의 세그먼트 개수, #$\{P_i^{uni}\}$는 단일 음소 단위의 세그먼트 개수를 의미한다. 본 논문에서는 수식 (2)의 임계값(α)으로 예비 실험을 통해 설정된 0.1을 사용하였다.

그림 3은 자소 단위, 음소 단위 그리고 제안된 단위로 구성된 발음사전의 예시이다. 음소 단위 기반의 발음사전은 모든 단어에서 음소 기호 /m/(파란색), /a/(녹색), /h/(빨간색)를 동일하게 사용한다. 반면에, 제안된 단위 기반의 발음사전은 단어의 문맥에 따라 확장된 단위를 사용한다. 아래의 예제에서 음소 기호 /m/은 93번째 군집인 c93과 41번째 군집인 c41로 확장되며, 음소 기호 /a/는 101번째 군집인 c101과 18번째 군집인 c18로 확장된

다. 또한, 음소 기호 /h/는 71번째 군집인 c71과 100번째 군집인 c100으로 확장되어 새로운 발음 단위로 사용된다.

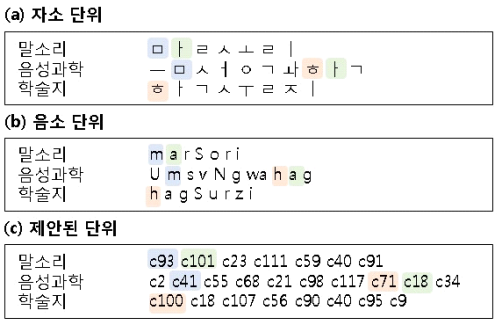

그림 3. 자소 단위(a), 음소 단위(b), 제안된 단위(c)로 구성된 발음사전의 예시
Figure 3. Example of lexicons constructed based on (a) the grapheme unit, (b) the phoneme unit, and (c) the proposed unit

3. 군집 개수 선택 방법

본 절에서는 선행 연구(Bang et al., 2019)에서 향후 계획으로 언급되었던 자동으로 군집 개수를 선택하는 방법을 제시한다. 세그먼트 군집화 단계에서 k-means 알고리즘의 최적의 군집 개수(k)는 DB 지수(Davies & Bouldin, 1979)를 사용하여 선택된다. 군집화 평가 척도인 DB 지수는 군집 간 거리 값과 군집 내 잠재벡터들의 분산 값의 비율로 나타낸다. 이때, DB 지수는 아래의 수식으로 계산된다.

$$DB = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\frac{S_i + S_j}{M_{ij}} \tag{3}$$

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j - A_i|^2\right)^{1/2} \tag{4}$$

$$M_{ij} = \|A_i - A_j\|_2 \tag{5}$$

수식 (3)에서 S_i은 i번째 군집 내 유사도를 의미하고, M_{ij}는 i번째 군집과 j번째 군집 간 유사도를 나타낸다. 여기서 군집 내 유사도(S_i)는 수식 (4)와 같이 표현되며, T_i개의 잠재벡터를 가지는 i번째 군집에서 j번째 잠재벡터 X_j와 중심점 A_i 사이의 평균 거리를 의미한다. 군집 간 유사도(M_{ij})는 수식 (5)와 같이 표현되며, i번째 군집의 중심점 A_i와 j번째 군집의 중심점 A_j 사이의 거리를 의미한다. 마지막으로, DB 지수는 전체 N개의 군집을 대상으로 군집 내 유사도와 군집 간 유사도의 평균 비율을 계산함으로써 얻어진다.

최적의 군집 개수는 DB 지수가 가장 작은 것으로 선택된다. 여기서 DB 지수는 군집 내 분산 값이 작으면서 군집 간 거리가 클수록 낮은 수치를 보인다. 이는 음성인식 결과를 통해서 군집 개수를 선택하는 선행 연구(Bang et al., 2019)와의 차이점이다. 아래의 표 1은 전체 세그먼트에 대해서 군집 개수를 40개에서 160개로 40개씩 증가시켰을 때 측정되는 DB 지수를 보인다.

표 1. 군집 개수에 따른 DB 지수 비교
Table 1. Comparison of the Davies-Bouldin (DB) index according to the number of clusters

군집개수(개)	40	80	120	160
DB 지수	2.68	2.67	**2.65**	2.67

군집 개수에 따른 DB 지수 비교 실험에서는 120개의 군집을 가질 때, DB 지수 2.65의 가장 우수한 성능을 보였다. 이 결과로부터 이후 음성인식 실험에서 120개로 확장된 단위에서 가장 좋은 성능을 보일 것을 기대할 수 있다.

4. 단위 생성 실험 및 결과

본 절에서는 다양한 길이를 가지는 음성 세그먼트로부터 고정된 길이의 잠재벡터를 추출하는 다양한 방법들을 비교하고, 군집화 결과로 얻어진 120개의 군집으로부터 각 세그먼트들의 스펙트럼 패턴, 세그먼트 길이 등을 분석한다.

4.1. 한국어 방송 데이터

음소 세트 확장 실험을 위해서 약 1,000시간의 한국어 방송 음성 데이터(Bang et al., 2017)를 사용한다. 방송 데이터는 미리 수집된 방송 오디오와 자막 텍스트로부터 가볍게 감독된 접근법(lightly supervised approach; Lamel et al., 2002)을 사용하여 자동으로 구축되었다. 정제된 방송 데이터는 배경 잡음과 배경 음악이 혼합되어 있으며, 일부 음성 신호와 일치하지 않는 잘못 전사된 텍스트를 가진다.

방송 데이터는 크게 뉴스, 다큐멘터리, 시사, 교양, 드라마, 예능, 어린이의 7가지 장르로 구성된다. 수집된 데이터에서 낭독체 음성을 많이 포함하는 뉴스, 다큐멘터리 그리고 시사 장르는 각각 25%, 12% 및 5%로 구성되며, 전체 데이터의 42%를 차지하였다. 반면에 자유발화를 많이 포함하는 교양, 드라마, 예능, 어린이 장르는 각각 22%, 16%, 12% 및 3%로 구성되며, 전체 데이터의 54%를 차지하였다. 나머지 4%는 스포츠 방송과 음악 장르의 프로그램으로 구성되었다. 아래의 그림 4는 한국어 방송 데이터의 장르별 분포를 나타낸다.

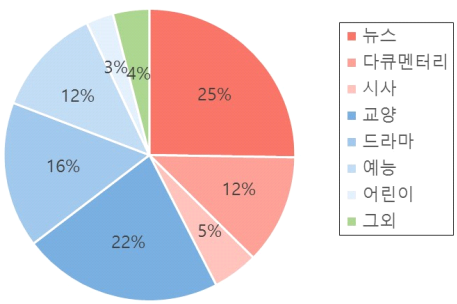

그림 4. 한국어 방송 데이터의 장르별 분포
Figure 4. Distribution for each genre of Korean broadcast data

낭독체 음성은 주로 아나운서나 사회자가 정확한 발음으로 발성하는 음성으로 구성된다. 반면에, 자유발화 음성은 원하지 않는 일시 중지, 발화 끊김, 장음화, 간투사, 자기 교정 및 반복된 단어로 채워진다. 다양한 자유발화 현상을 가지는 방송 데이터로부터 공통적으로 나타나는 스펙트럼 패턴을 찾는다면 자유발화 음성에 적합한 새로운 단위를 생성할 수 있을 것이다.

4.2. 잠재벡터 추출 방법의 비교

음성 세그먼트로부터 고정된 길이의 잠재벡터를 추출하는 방법은 다양하다. 본 절에서는 1) 선형보간법, 2) LSTM 오토인코더, 3) LSTM 분류기의 3가지 방법으로 추출된 잠재벡터를 비교하고 LSTM 분류기 구조를 선택한 이유를 설명한다.

여기서, 선형보간법은 다양한 길이를 가지는 세그먼트에서 가운데와 마지막에 위치하는 40차 로그 멜 필터뱅크 특징벡터 2개를 선택하여 총 80차 잠재벡터를 얻는다. 반면에 LSTM 오토인코더와 분류기는 2절에서 설명된 그림 1의 구조를 사용하며, 인코더 단에서 디코더 단으로 넘어가는 80차 잠재벡터를 고정된 길이의 특징벡터로 사용한다. 그림 5는 3가지 방법으로 추출된 잠재벡터들의 분포 및 길이 정보를 보인다.

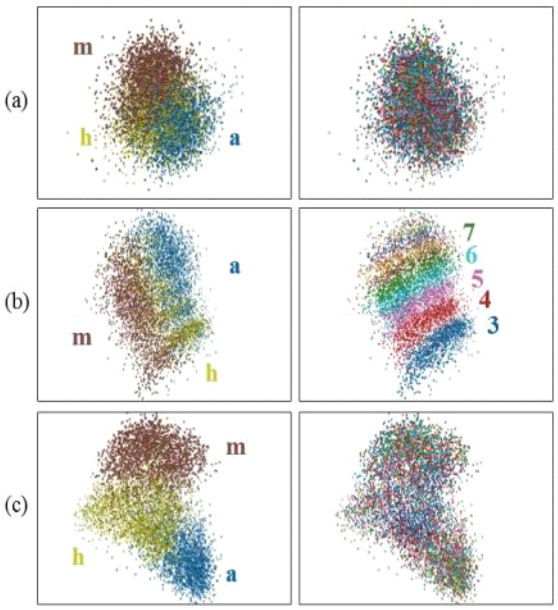

그림 5. 선형변환(a), 오토인코더(b), 분류기(c)에서 얻어진 잠재벡터의 음소(왼쪽) 및 길이(오른쪽) 정보에 따른 산점도
Figure 5. Scattering plots of latent vectors according to phoneme class (left) and duration (right) obtained by (a) linear interpolation, (b) auto-encoder, and (c) classifier

3가지 방법으로 추출된 잠재벡터를 시각화하여 분석하였다. 시각화 실험은 그림 3의 발음사전 예시에서 보인 /m/, /a/, /h/ 음소를 대상으로, 주성분분석법을 이용하여 80차 잠재벡터를 3 차원으로 축소시켜 수행한다. /m/, /a/, /h/ 이외의 음소에서도 유사한 경향의 산점도를 보이는 것으로 나타났다.

선형보간법으로 얻어진 잠재벡터(그림 5(a))는 음소 /a/와 /h/가 서로 비슷한 위치에 존재한다. 이들은 각 음소들이 가진 음향적 특징 차이를 잘 표현하지 못하기 때문에, 이후 군집화를 통해서 공통된 스펙트럼 패턴을 찾는데 부적합하다.

반면에, LSTM 오토인코더로 얻어진 잠재벡터(그림 5(b))는 비선형 함수를 통과하여 얻어진다. 오토인코더 구조는 먼저 인코더 단에서 각 세그먼트를 대표하는 잠재벡터를 추출한 다음, 디코더 단에서 잠재벡터를 입력받아 다시 원래의 세그먼트를 추정하도록 설계된다. 그 과정에서, 잠재벡터는 디코더 단에서 다시 원래의 세그먼트로 복원할 수 있도록 각 세그먼트의 정보를 잘 압축시킨 정보를 갖는다. 하지만, 그림 5(b)에서 얻어진 잠재벡터가 세그먼트의 프레임 길이에 군집이 생성되었음을 관찰했다. 이는 프레임 길이 정보가 원래의 세그먼트를 다시 복원하는데 가장 중요한 요인이기 때문이다.

마지막으로, 제안된 단위 생성 방법에서 사용된 LSTM 분류기로 얻어진 잠재벡터(그림 5(c))는 세그먼트 길이 정보를 강조하지 않으면서 각 음소 기호를 잘 표현하였다. 이는 원래의 세그먼트를 복원시키는 오토인코더 구조와 다르게 40개의 음소 기호를 추정하도록 학습되었기 때문이다. 결과적으로, 이 방법은 다른 방법으로 추출된 잠재벡터보다 각 음소의 음향 특성을 더 잘 나타내는 것을 시각적으로 확인하였다.

3가지 방법으로 얻어진 잠재벡터는 각각 12.2, 9.7, 5.0의 DB 지수를 얻었다. 여기서 DB 지수는 낮을수록 좋은 품질을 나타낸다. 결과적으로, LSTM 분류기 모델로부터 얻은 잠재벡터에서 DB 지수 5.0의 가장 우수한 성능을 보였다(표 2 참조).

표 2. 다양한 방법으로 얻어진 잠재벡터의 DB 지수 비교
Table 2. Davies-Bouldin (DB) indices of latent vectors obtained by various methods

추출 방법	(a) 선형변환	(b) 오토인코더	(c) 분류기
DB 지수	12.2	9.7	**5.0**

4.3. 군집 분석

군집화 결과로 얻어진 120개의 군집과 40개의 기본 음소 사이의 관계를 시각적으로 확인하였다. 실험은 음소 별로 10,000개의 세그먼트를 추출하여 각 군집에 할당되는 세그먼트 개수와 그들의 대표 스펙트럼 패턴 그리고 프레임 길이 정보를 비교한다. 여기서, 스펙트럼 패턴은 길이가 30 프레임이 되도록 모든 세그먼트를 선형보간 처리한 다음, 각 군집에 속한 세그먼트들을 모두 누적시켜 출력하였다. 그림 6은 그림 3의 발음사전 예시에 사용된 음소 /m/, /a/, /h/의 세그먼트 정보를 보인다.

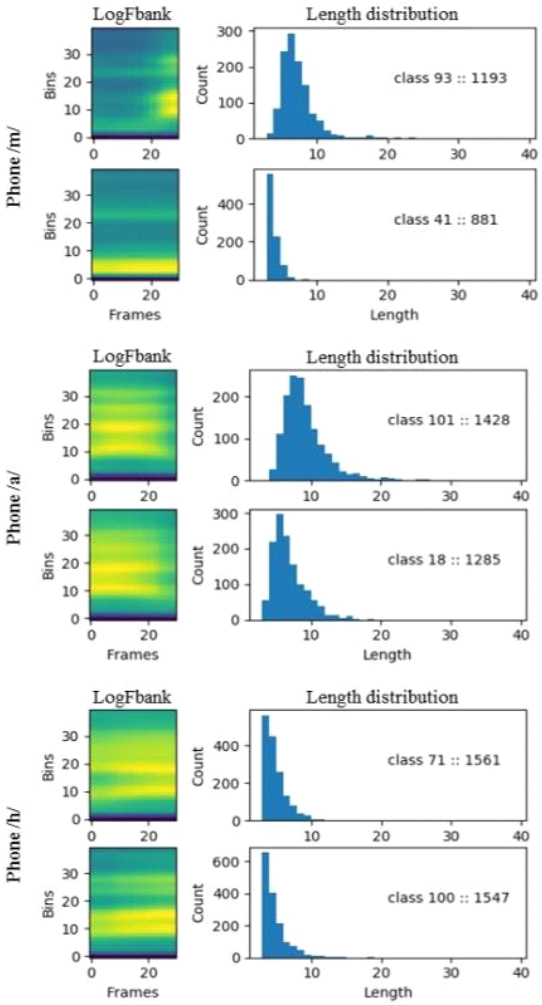

그림 6. 음소 /m/, /a/, /h/의 세그먼트 정보
Figure 6. Segment information of the phoneme /m/, /a/, and /h/

음소 /m/에 속한 10,000개의 세그먼트는 93번째 군집에서 1,193개의 세그먼트가 가지며, 41번째 군집에서 881개의 세그먼트를 가졌다. 93번째 군집의 스펙트럼 패턴은 시작 부분에 묵음 구간을 가지는 것으로 볼 때, 단어의 시작 지점에 위치하는 것으로 판단된다. 반면에, 41번째 군집의 세그먼트들은 대부분 짧은 지속시간을 가지며, 저주파 대역에서 연속적인 스펙트럼 패턴을 보였다. 그림 6에서 음소 /a/의 경우에는 세그먼트의 길이, 음소 /h/의 경우에는 인접한 음소에 따라서 각 세그먼트는 서로 다른 군집으로 할당되었다.

유사한 스펙트럼 패턴을 가지는 세그먼트들이 각 군집에 모이는 것을 관찰하였다. 동일한 음소를 가지는 세그먼트임에도 불구하고, 각 세그먼트는 길이 분포나 스펙트럼 패턴에 따라서로 다른 군집으로 할당되었다. 결과적으로, 각 군집은 주변의 음소 기호나, 단어에서의 음소 위치 그리고 음소의 길이에 따라 구분된다는 것을 발견하였다.

5. 음성인식 실험 및 결과

5.1. 실험 환경

모든 음성인식 실험은 Kaldi 도구(Povey et al., 2011)를 사용하여 수행하였다. 음성인식기에 사용된 특징벡터는 발화 단위로 평균과 분산을 정규화한 40차 로그 멜 필터뱅크를 사용하였으며, 문맥을 고려하기 위해서 좌/우의 인접한 프레임을 2개씩 연결하여 전체 5개 프레임을 음향모델의 입력으로 사용하였다.

음향모델은 LSTM 기반의 HMM을 사용하였다. HMM은 비묵음 기호와 묵음 기호에 대해서 각각 3개와 5개의 상태 열을 가지는 left-to-right HMM을 사용하였다. LSTM은 1,024개의 메모리 셀을 가지는 3개의 층을 사용하였으며, 256개의 은닉 노드를 가지는 투영 층(projection layer; Sak et al., 2014)은 추가하였다. 출력 층은 softmax 활성함수를 가진 약 8,000개의 노드를 사용하였다. 나머지 매개 변수는 Kaldi 도구(Povey et al., 2011)의 LSTM 예제 스크립트가 제공하는 기본 설정 값을 사용하였다.

언어모델은 SRILM 도구(Stolcke, 2002)를 이용하여 미리 수집된 방송 자막 텍스트로 학습되었으며, 생성된 모델은 unigram에 약 176만 개, bigram 약 3,158만 개, trigram 약 4,294만 개를 가졌다. 디코더에서 음향모델 가중치는 0.125로 설정하고, 빔 크기(beam size)와 격자 빔 크기(lattice beam size)는 각각 10.0과 5.0으로 설정하였다. 인식성능은 의사형태소 단위를 사용하여 단어 오류율(word error rate, WER)을 계산하여 확인하였다.

5.2. 음성 데이터베이스

단위 생성 실험에 사용되었던 약 1,000시간의 한국어 방송 데이터를 음향모델 학습에 모두 사용하였다. 이전의 연구(Bang et al., 2019)에서는 200시간으로 구성된 소량의 데이터를 사용하지만, 본 연구에서는 대량의 음성 데이터를 사용하여 제안된 단위를 사용한 인식 실험의 신뢰성을 높였다.

평가 데이터는 방송 뉴스 장르로 구성된 약 3.5시간의 낭독체 음성 데이터(‘낭독체’)와 조용한 실험실 환경에 직접 녹음된 약 7시간의 자유발화 음성 데이터(‘자유발화’)를 사용하였다. 여기서, 자유발화 데이터는 일상생활에서 사용되는 일반적인 대화 내용들로 구성되며, 뉴스 내용으로 구성된 낭독체 데이터와 다른 도메인을 가진다. 모든 평가 데이터는 음향모델 학습 및 새로운 단위 구축 실험에 중복되어 사용되지 않는다.

5.3. 자소 및 음소 단위 음성인식 실험

먼저 자소 단위와 음소 단위의 성능을 비교한다. 이들의 성능을 비교하기 위해서 한국어 음성인식기에서 일반적으로 사용되는 40개의 음소 단위와 52개의 자소 단위로 구성된 발음사전을 생성하였다. 여기서, 자소 단위는 그림 3(a)와 같이 각 음절을 초성, 중성, 종성으로 변환하여 얻어진다. 아래의 표 3은 낭독체 데이터와 자유발화 데이터에서의 자소 및 음소 단위의 음성인식 성능을 보인다.

표 3. 자소 단위와 음소 단위의 단어 오류율(%)
Table 3. WER(%) of the phoneme unit and the grapheme unit

WER(%)	자소 단위	음소 단위
낭독체	22.1	**20.3**
자유발화	53.6	**52.6**

WER, word error rate.

낭독체 데이터에서 음소 단위는 자소 단위보다 상대적으로 8.1%의 더 좋은 성능을 보인다. 다른 언어(Killer et al., 2003)와 성능 차이를 비교해볼 때, 한국어는 다소 미미한 음성인식 성능 차이를 보인다. 이것은 한국어가 사람의 말소리를 기호로 나타내는 표음문자에 가깝기 때문이며, 이러한 결과는 한국어와 유사한 특성을 가진 바스크어에서도 비슷하게 나타난다. 반면에, 자유발화 데이터에서는 1.9%의 미미한 차이를 보인다. 이러한 현상은 음소 단위가 자유발화 음성에서 높은 단위 내 분산 값을 가지고, 단위 사이의 거리는 짧아지기 때문으로 판단된다. 결과적으로, 낭독체 음성에서 음소 단위는 자소 단위보다 더 좋은 성능을 보였지만, 자유발화에서는 단위 사이의 변별력이 낮아졌기 때문에 미미한 성능 차이를 보였다.

본 논문의 결과는 더 많은 양의 음성 데이터를 사용하였음에도 선행 연구(Bang et al., 2019)에서 제시된 결과보다 낮은 성능을 보인다. 이러한 차이는 두 실험의 인식 어휘 개수 차이에 의한 것이다. 선행 연구에서는 음성 데이터의 전사 텍스트를 사용하여 약 8 만 개의 unigram 개수를 가지는 작은 언어모델을 사용하였다. 반면에, 본 논문에서는 다양한 말뭉치를 추가적으로 사용하여 약 22 배 증가된 176만 개의 unigram 개수를 가지는 대량의 언어모델을 사용하였다.

5.4. 제안된 단위 음성인식 실험

본 절에서는 제안된 단위생성 방법으로 구축된 새로운 음소 세트의 음성인식 성능을 확인한다. 군집 개수를 120개로 설정한 경우에 낭독체 데이터에서 18.5%, 자유발화 데이터에서 48.9%의 가장 낮은 단어 오류율을 얻었다. 이 결과는 앞서 수행된 군집 개수에 따른 DB 지수 비교 실험에서 기대한 결과와 동일하다. 아래의 표 4는 다양한 군집 개수에 따른 제안된 단위의 음성인식 성능을 보인다.

표 4. 군집 개수에 따른 제안된 단위의 단어 오류율(%)
Table 4. WER (%) of the proposed unit according to the number of clusters

WER(%)	40	80	120	160
낭독체	20.2	19.0	**18.5**	18.9
자유발화	52.9	51.0	**48.9**	50.0

WER, word error rate.

제안된 단위는 음소 단위에 비해서 낭독체 음성에서 8.9%의 상대적 단어 오류율을 감소시키는 효과를 보였다. 이는 유사한 스펙트럼 패턴을 가지는 세그먼트들을 군집시킴으로써, 인접한 음소나 음소 위치, 음소 길이에 따라 세분화된 단위를 사용하여 단위 사이의 변별력이 높아졌기 때문이다. 또한, 다른 도메인으로 구성된 자유발화 데이터에서도 약 7.0%의 상대적 단어 오류율을 감소시키는 효과를 보임을 확인하였다.

6. 결론

본 논문에서는 자유발화 음성인식기의 성능 향상을 위해서 대량의 한국어 방송 데이터로부터 새로운 음소 세트를 구축하는 방법을 제안하였다. 그 과정에서 가변 길이의 음성 세그먼트로부터 다양한 방법으로 얻어진 고정된 길이의 잠재벡터를 비교하였다. 또한, 자동으로 새로운 음소 세트의 개수를 선택하는 방법을 설명하고, 제안된 단위에 속한 음성 세그먼트의 스펙트럼 패턴과 길이 분포를 분석하였다. 결과적으로, 제안된 단위는 DB 지수가 가장 낮은 군집 개수에서 최적의 군집 개수를 보임을 확인하고, 각 군집은 인접한 음소나 음소의 위치, 음소 길이에 따라서 변별력을 가짐을 확인하였다.

향후 연구로 자소 단위로부터 제안된 단위를 생성한 결과를 확인하고, 발음사전을 요구하지 않는 종단 간 음성인식 시스템에서 자소 단위와 제안된 단위 사이의 음성인식 성능을 비교할 계획이다. 또한, 초성, 중성, 종성 별로 음소를 확장시킨 단위와 제안된 단위 사이의 성능을 비교할 계획이다. 마지막으로, 다양한 차원의 잠재벡터를 비교하여 음성 세그먼트를 표현하는데 최적의 차원을 찾을 계획이다.

References

Bang, J. U., Choi, M. Y., Kim, S. H., & Kwon, O. W. (2017, August). Improving speech recognizers by refining broadcast data with inaccurate subtitle timestamps. *Proceedings of the Interspeech 2017* (pp. 2929-2933). Stockholm, Sweden.

Bang, J. U., Choi, M. Y., Kim, S. H., & Kwon, O. W. (2019, September). Extending an acoustic data-driven phone set for spontaneous speech recognition. *Proceedings of the Interspeech 2019* (pp. 4405-4409). Graz, Austria.

Chung, Y. A., Wu, C. C., Shen, C. H., Lee, H. Y., & Lee, L. S. (2016, September). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *Proceedings of the Interspeech 2016* (pp. 410-415). San Francisco, CA.

Hain, T. (2005). Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Communication*, *46*(2), 171-188.

Killer, M., Stuker, S., & Schultz, T. (2003). Grapheme based speech recognition. *Proceedings of the Eurospeech 2003* (pp. 3141-3144). Geneva, Switzerland.

Lamel, L., Gauvain, J. L., & Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, *16*(1), 115-129.

Lee, K. N., & Chung, M. (2003, January). Modeling cross-morpheme

pronunciation variations for Korean large vocabulary continuous speech recognition. *Proceedings of the Eurospeech 2003* (pp. 261-264). Geneva, Switzerland.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). Berkeley, CA.

Mitra, V., Vergyri, D., & Franco, H. (2016, September). Unsupervised learning of acoustic units using autoencoders and Kohonen nets. *Proceedings of the Interspeech 2016* (pp. 1300-1304). San Francisco, CA.

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, *22*(2), 171-184.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., ... Vesely, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Hawaii.

Sainath, T. N., Prabhavalkar, R., Kumar, S., Lee, S., Kannan, A., Rybach, D., Schoglo, V., ... Chiu, C. C. (2018, April). No need for a lexicon? Evaluating the value of the pronunciation lexica in end-to-end models. *Proceedings of the International Conference on Acoustics, Speech, Signal Processing* (pp. 5859-5863). Calgary, Canada.

Sak, H., Senior, A., & Beaufays, F. (2014, September). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of the Interspeech 2014* (pp. 338-342). Singapore.

Stolcke, A. (2002, September). SRILM-an extensible language modeling toolkit. *Proceedings of the Interspeech 2002* (pp. 901-904). Denver, CO.

Young, S. J., Odell, J. J., & Woodland, P. C. (1994, March). Tree-based state tying for high accuracy acoustic modelling. *Proceedings of the Workshop on Human Language Technology* (pp. 307-312). Plainsboro, NJ.

• **방정욱 (Jeong-Uk Bang)**
충북대학교 제어로봇공학전공 박사과정
충북 청주시 서원구 충대로 1(개신동)
Tel: 043-261-3374
Email: jubang@cbnu.ac.kr
관심분야: 음성인식, 음성정렬, 음성 데이터 정제

• **김상훈 (Sang-Hun Kim)**
한국전자통신연구원 책임연구원
대전 유성구 가정로 218
Tel: 042-860-5141
Email: ksh@etri.re.kr
관심분야: 음성인식, 자동통역

• **권오욱 (Oh-Wook Kwon)** 교신저자
충북대학교 전자공학부 교수
충북 청주시 서원구 충대로 1(개신동)
Tel: 043-261-3374
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 음성신호처리, 오디오신호처리

음향 데이터로부터 얻은 확장된 음소 단위를 이용한 한국어 자유발화 음성인식기의 성능*

방 정 욱[1] · 김 상 훈[2] · 권 오 욱[1]

[1]충북대학교 일반대학원 제어로봇공학전공, [2]한국전자통신연구원, [3]충북대학교 전자공학부

국문초록

본 논문에서는 대량의 음성 데이터를 이용하여 기존의 음소 세트를 확장하여 자유발화 음성인식기의 성능을 향상시키는 방법을 제안한다. 제안된 방법은 먼저 방송 데이터에서 가변 길이의 음소 세그먼트를 추출한 다음 LSTM 구조를 기반으로 고정 길이의 잠복벡터를 얻는다. 그런 다음, k-means 군집화 알고리즘을 사용하여 음향적으로 유사한 세그먼트를 군집시키고, Davies-Bouldin 지수가 가장 낮은 군집 수를 선택하여 새로운 음소 세트를 구축한다. 이후, 음성인식기의 발음사전은 가장 높은 조건부 확률을 가지는 각 단어의 발음 시퀀스를 선택함으로써 업데이트된다. 새로운 음소 세트의 음향적 특성을 분석하기 위하여, 확장된 음소 세트의 스펙트럼 패턴과 세그먼트 지속 시간을 시각화하여 비교한다. 제안된 단위는 자유발화뿐만 아니라, 낭독체 음성인식 작업에서 음소 단위 및 자소 단위보다 더 우수한 성능을 보였다.

핵심어: 음향 단위, 음소 세트, 자유발화 음성인식, 방송 데이터

* 본 논문은 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음(19ZS1140, Conversational AI 공통핵심기술 연구과제).