특집논문 (Special Paper)
방송공학회논문지 제24권 제5호, 2019년 9월 (JBE Vol. 24, No. 5, September 2019)
https://doi.org/10.5909/JBE.2019.24.5.735
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

음소 인식을 위한 스파이크그램 기반의 음성 특성 추출 기술

한 석 현[a], 김 재 원[a], 안 순 호[a], 신 성 현[a], 박 호 종[a]‡

Speech Feature Extraction based on Spikegram for Phoneme Recognition

Seokhyeon Han[a], Jaewon Kim[a], Soonho An[a], Seonghyeon Shin[a], and Hochong Park[a]‡

요 약

본 논문에서는 스파이크그램을 기반으로 음소 인식을 위한 특성을 추출하는 방법을 제안한다. 음소 인식에 널리 사용되는 푸리에 변환 기반의 특성은 청각 기관의 동작에 부합하는 과정으로 구해지지 않으며 프레임 단위로 추출되어 높은 시간 해상도를 가지지 못한다. 따라서 음소 인식의 성능 향상을 위해 높은 시간 해상도를 가지면서 인간의 청각기관을 모델링 하는 새로운 음성 특성 추출 기술이 요구된다. 본 논문에서는 청각 기관의 특성 추출 및 전달 과정을 모델링 하는 기법인 스파이크그램을 사용하여 음성 신호를 분석하고, 이로부터 음소 인식을 위한 특성을 추출하는 방법을 제안한다. 심층 신경망 기반의 음소 인식기를 사용하여 제안한 특성의 음소 인식 성능을 측정하였고, 짧은 음소에 대해 제안 특성이 기존 푸리에 변환 기반의 특성보다 우수한 성능을 가지는 것을 확인하였다. 이 결과로부터 청각 모델을 기반으로 추출된 새로운 음성 특성을 사용하여 음소 인식이 가능함을 확인할 수 있다.

Abstract

In this paper, we propose a method of extracting speech features for phoneme recognition based on spikegram. The Fourier-transform-based features are widely used in phoneme recognition, but they are not extracted in a biologically plausible way and cannot have high temporal resolution due to the frame-based operation. For better phoneme recognition, therefore, it is desirable to have a new method of extracting speech features, which analyzes speech signal in high temporal resolution following the model of human auditory system. In this paper, we analyze speech signal based on a spikegram that models feature extraction and transmission in auditory system, and then propose a method of feature extraction from the spikegram for phoneme recognition. We evaluate the performance of proposed features by using a DNN-based phoneme recognizer and confirm that the proposed features provide better performance than the Fourier-transform-based features for short-length phonemes. From this result, we can verify the feasibility of new speech features extracted based on auditory model for phoneme recognition.

Keyword : Spikegram, Speech feature, Phoneme recognition, Deep neural network

a) 광운대학교 전자공학과(Dept. of Electronics Engineering, Kwangwoon University)
‡ Corresponding Author : 박호종(Hochong Park)
E-mail: hcpark@kw.ac.kr
Tel: +82-2-940-5104
ORCID: https://orcid.org/0000-0003-1600-6610

※ 이 논문의 연구 결과 중 일부는 한국방송·미디어공학회 “2019년 하계학술대회”에서 발표한 바 있음.
※ 본 논문은 2019년도 광운대학교 교내학술연구비 지원과 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원(NRF-2016R1D1A1B03930923)을 받아 수행된 연구임. (The present Research has been conducted by the Research Grant of Kwangwoon University in 2019 and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2016R1D1A1B03930923)).
· Manuscript received July 23, 2019; Revised September 4, 2019; Accepted September 4, 2019.

Ⅰ. 서 론

음성 인식은 인간-컴퓨터 상호작용 (human-computer interaction, HCI) 등 다양한 응용 분야에서의 중요성 때문에 최근 많은 주제에 대하여 집중적으로 연구되고 있다. 그러나 짧은 신호, 잡음, 특성 왜곡 등의 문제점으로 인하여 현재의 음성 인식 기술은 인간보다 낮은 성능을 가지며, 인간과 동등한 성능을 가지기 위해서는 기존의 수학적 방법론과 다르게 인간의 청각 기관 동작을 모방하는 새로운 접근법이 필요하다.

일반적으로 음성 인식의 첫 단계는 음소 (phoneme) 인식 과정이며, 음소 인식은 분석하는 정보의 종류에 따라 신호 모델과 언어 모델로 나누어진다[1]. 본 논문에서는 인간 청각 기관의 초기 동작만을 모델링하기 위해 언어 모델을 제외한 신호 모델 기반의 음소 인식만을 다룬다. 또한 음소 인식은 기능에 따라 특성 추출과 분류로 나뉘며, 음성 신호로부터 음소에 대한 핵심 정보를 표현하는 특성을 추출하고 최종적으로 음소를 인식하기 위해 특성을 분류한다.

기존의 많은 음소 인식 기술은 Mel-frequency cepstral coefficient (MFCC)를 핵심 특성으로 사용한다[1,2]. 그러나 MFCC는 음소 인식에서 두 가지 문제점을 갖는다. 첫 번째로, MFCC는 프레임 단위의 푸리에 변환을 기반으로 추출되므로 프레임 길이가 결정하는 시간 해상도 이상의 특성을 모델링하기 어렵다. 따라서 폐쇄음 (stops), 파찰음 (affricate)과 같은 길이가 짧은 음소에 대해서는 특성을 잘 반영하지 못한다. 두 번째로, MFCC는 인간 청각 기관의 동작을 모방하는 과정으로 음성 신호를 분석하지 않는다. Mel-scale을 사용하여 인간의 청각 특성을 반영하지만, 푸리에 변환은 수학적 분석 기법으로서 인간의 청각 구조와 관련이 없고, 청각 기관이 MFCC 연산을 수행하여 추출한 정보를 활용한다는 근거도 없다. 따라서 수학적 기법인 MFCC의 문제점을 해결하고 음소 인식 성능을 높이기 위해 인간의 청각 모델을 기반으로 음성 특성을 추출하는 새로운 방법이 필요하다.

본 논문의 목표는 기존 기술의 문제점을 해결하기 위해 인간의 청각 모델을 기반으로 청각 기관과 유사한 과정으로 음성 특성을 추출하는 기술을 개발하는 것이다. 이를 위해 스파이크그램 (spikegram) 기반으로 음소 인식을 위한 새로운 음성 특성을 추출하는 방법을 제안한다. 스파이크그램은 청각 기관이 음성 특성을 추출하고 전달하는 과정을 모델링한 기법으로서 음성 신호를 주파수와 시간 축 상에 청각 커널 (kernel)의 합으로 분해한 것이다[3]. 즉, 이 기법은 청각 기관이 스파이크를 통해 신호의 시간/주파수 구조를 분석하는 것을 시뮬레이션 하며, 기존 수학적 특성들과는 달리 주파수 정보를 얻기 위해 프레임 단위의 동작을 수행하지 않기 때문에 높은 시간 해상도로 음성 특성을 분석하는데 적합하다.

제안하는 방법은 대표적 청각 모델인 감마톤 (gammatone) 필터뱅크를 커널로 사용하여 입력 신호에 대한 스파이크그램을 생성하고[4,5], 생성된 스파이크그램으로부터 음소 인식을 위한 핵심 음성 특성을 시간과 주파수 영역에서 추출한다. 다음, 추출한 음성 특성을 심층 신경망 (deep neural network, DNN) 기반의 음소 인식기에 입력하여 음소 인식 성능을 측정한다. 성능 검증을 통해 제안하는 음성 특성이 장애음 (obstruent)과 같은 길이가 짧은 음소에 대해 MFCC 특성보다 더 높은 성능을 제공하는 것을 확인하였다. 이를 통해 제안 방법과 같이 인간의 청각 모델을 기반으로 추출된 새로운 음성 특성을 사용하여 음소 인식이 가능함을 확인할 수 있다.

Ⅱ. 제안하는 음성 특성 추출 방법

1. 스파이크그램

인간의 청각 기관인 달팽이관 내부에는 긴 섬모 모양의 청세포가 모여 있고, 소리의 주파수에 따라 다른 위치의 청세포가 반응한다[6]. 반응한 청세포는 해당 정보를 청신경 (auditory nerve)을 통해 인간의 뇌로 전달하고, 이때 전해지는 신호를 스파이크라 한다. 음성 신호 $x(t)$를 식 (1)과 같이 커널을 기반으로 스파이크의 합으로 분해할 수 있고[3], 이렇게 구한 스파이크 위치와 크기를 시간/주파수 영역에 표시하여 스파이크그램을 생성한다.

$$x(t)=\sum_{m=1}^{M}\sum_{i=1}^{n_m} g_i^m \phi_m\left(t-\tau_i^m\right)+\epsilon(t) \tag{1}$$

여기서 $\phi_m(t)$는 밴드별 커널 함수, m은 커널 인덱스, M은 커널의 개수, n_m은 커널 당 스파이크의 활성 횟수, g_i^m은 스파이크의 크기, τ_i^m는 스파이크의 위치, $\epsilon(t)$는 모델링 오차를 의미한다. 청각 이론과 신경망 이론에 의하면 각 시간과 주파수 밴드별 스파이크 크기는 해당 시간과 밴드에서의 청신경의 발화율 (firing rate)을 나타낸다[3].

본 논문에서는 스파이크를 추출하기 위해서 식 (2)와 같은 감마톤의 equivalent rectangular bandwidth (ERB) 필터를 커널로 사용한다[7].

$$\phi(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft) \qquad (2)$$

여기서 a는 필터 진폭, n은 필터 차수, f는 Hz 단위의 필터 중간 주파수, b는 Hz 단위의 필터 대역폭이다. Glasberg과 Moore의 감마톤 모델에 따라 감마톤을 생성하였고, 32개 감마톤 ERB 필터를 각각 에너지 1로 정규화 하여 사용하였다[7]. 그림 1은 식 (2)에 의해 생성된 감마톤 필터의 일부 파형이고, 그림 2는 32개 감마톤 필터의 주파수 응답 그래프이다. 청각의 장소 이론 (place theory)에 따르면 달팽이관 기저막은 비선형적인 주파수 분포를 가지며[6], 그림 2의 감마톤 필터는 비선형적인 청각 시스템의 동작과 동일하게 높은 밴드일수록 넓은 대역폭을 가진다.

본 논문에서는 스파이크그램을 생성하기 위해 matching pursuit (MP) 알고리즘을 사용한다[8]. 먼저, 식 (1)과 같이 음성 신호를 커널의 합으로 분해하기 위해 특정 시간 위치에서 신호와 32 밴드 커널 $\phi_m(t)$과의 상관도를 구한다. 이중 가장 큰 상관도를 가지는 커널의 밴드 m과 위치 τ_i^m를 저장하고, 신호와 커널로부터 구한 상관도의 크기 g_i^m을 저장한다. 이 과정을 통하여 하나의 스파이크 (τ_i^m, g_i^m)가 정의된다. 다음, 추출된 스파이크에 해당하는 감마톤 필터와 크기의 곱만큼을 원 신호 $x(t)$에서 빼서 잔여 (residual) 신호를 구한다. 이후 잔여 신호로부터 다시 가장 큰 상관도를 가지는 스파이크를 추출하는 것을 반복해 스파이크그램을 생성한다. 추출된 스파이크로부터 식 (1)에 따라 신호를 복원할 수 있으며, 복원 신호의 peak signal-to-noise ratio (PSNR)이 50dB에 도달할 때까지 스파이크를 추출한다. 그림 3은 음성 신호의 스펙트로그램과 스파이크그램을 비교한 것이고, 스파이크의 크기 정보는 점의 크기로 표현하였다.

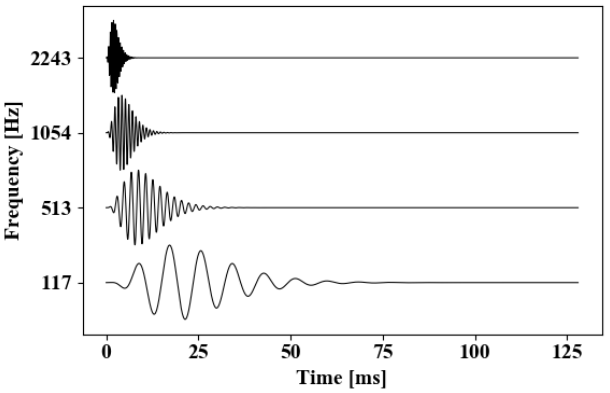


그림 1. 제안 방법에서 사용하는 감마톤 필터의 파형
Fig. 1. Waveform of gammatone filter used in the proposed method

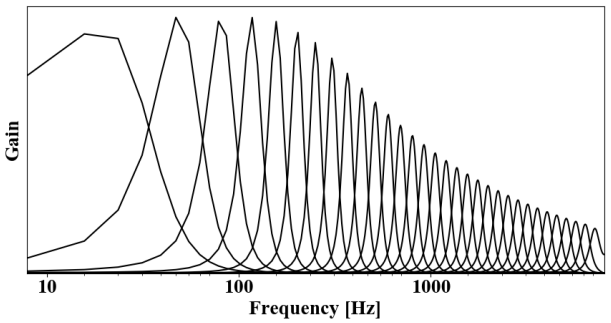


그림 2. 제안 방법에서 사용하는 32 밴드 감마톤 필터뱅크의 주파수 응답
Fig. 2. Frequency response of 32-band gammatone filterbank used in the proposed method

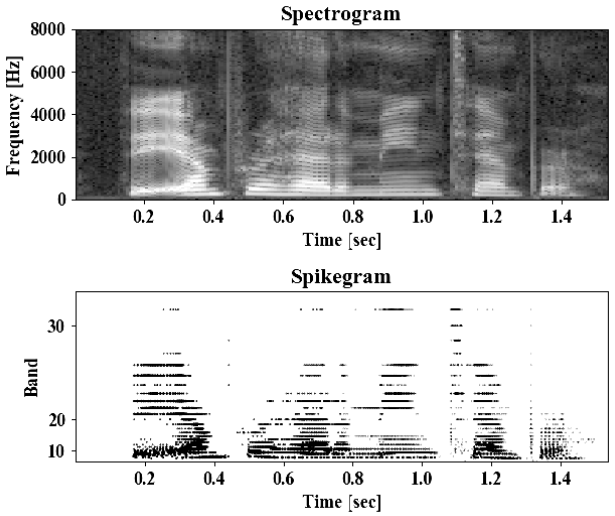


그림 3. 음성 신호의 스펙트로그램(위)과 스파이크그램(아래)의 예
Fig. 3. Spectrogram (top) and spikegram (bottom) of speech signal

감마톤 필터의 파형은 그림 1과 같이 주파수가 낮을수록 필터의 최대점이 원점과 멀어지고, 따라서 위상을 조정하지 않은 감마톤 필터는 시간 지연된 스파이크를 추출한다. 이는 시간 해상도가 높은 스파이크그램에서 잘못된 시간 기반 특성을 추출하게 한다. 이러한 문제점을 해결하기 위해 스파이크그램을 생성한 다음 각 밴드별로 지연된 시간만큼 스파이크의 위치를 조정한다. 그림 4는 위상을 조정하기 전 스파이크그램과 위상을 조정한 후 스파이크그램을 보여준다. 그림 4에서 세로 점선은 음성이 시작되는 지점을 의미한다. 위상을 조정한 후 스파이크의 시간 지연이 사라진 것을 확인할 수 있다.

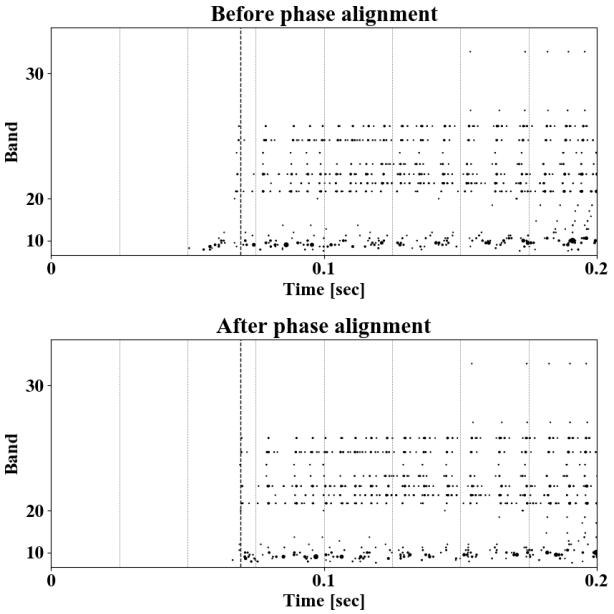


그림 4. 위상 조정 전(위)과 후(아래)의 스파이크그램
Fig. 4. Spikegram before (top) and after (bottom) phase alignment

2. 핵심 음성 특성 추출

스파이크그램은 스파이크의 시간 위치를 샘플 단위로 정확히 표현하는 장점을 가지지만, 이와 같은 높은 시간 해상도의 스파이크그램은 활성화된 스파이크 이외에 많은 빈 공간을 가진다. 따라서 희박한 (sparse) 데이터의 스파이크그램을 그대로 특성으로 사용하지 않고, 대신 스파이크그램으로부터 의미 있는 정보를 얻기 위해 핵심 특성을 추출한다. 그림 5는 본 논문에서 제안하는 스파이크그램으로부터 주파수 기반 특성과 시간 기반 특성을 추출하는 과정을 나타낸다. 샘플링 주파수가 16 kHz인 음원에 대해 25 ms (400 샘플) 구간별로 특성을 추출하고 25 ms 마다 하나의 음소를 인식한다. 그림 5에서 m은 밴드 인덱스, n은 서브 프레임 인덱스, i는 샘플의 시간 위치, g_i^m은 (i, m)에 위치한 스파이크의 크기를 의미한다. 구간 길이는 400 샘플이므로 시간 위치 i는 0~399의 범위를 가진다. 밴드별로 생성된 스파이크의 크기를 그림 5의 가로 화살표와 같이 시간 축으로 합하여 32개 주파수 기반 특성 G_m을 생성한다. 다음, 각 구간을 K개의 서브 프레임으로 분할하고 각 서브 프레임에서 모든 스파이크 크기를 그림 5의 세로 화살표와 같이 시간과 밴드 축으로 합하여 K개의 시간 기반 특성 T_n을 생성한다.

이와 같이 25 ms 구간별로 주파수 기반 특성과 시간 기반 특성을 더한 $(32 + K)$개 정적 상태 (static) 특성을 구하고, 모든 특성에 대하여 구간에 대한 1차 시간 미분 (delta)과 2차 시간 미분 (delta-delta)을 구해 총 $(96 + 3K)$개 특성을 완성한다. 시간 미분 d_t를 구하는 방법은 식 (3)과 같다[9]. 이때, t는 구간 인덱스, c는 특성 값이다. 1차 시간 미분과 2차 시간 미분을 구할 때 R은 동일하게 2를 사용하였다.

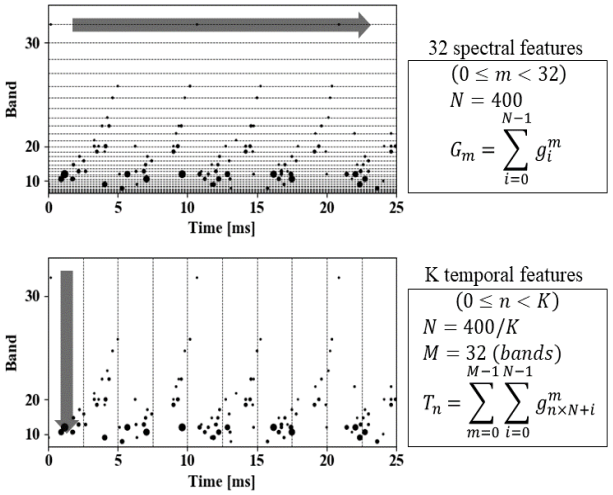


그림 5. 스파이크그램으로부터 32개의 주파수 기반 특성(위)과 K개의 시간 기반 특성(아래)을 구하는 과정
Fig. 5. Procedure of extracting 32 spectral features (top) and K temporal features (bottom) from spectrogram

$$d_t = \frac{\sum_{n=1}^{R} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{R} n^2} \quad (3)$$

Ⅲ. 성능 평가

제안하는 특성의 음소 인식 성능을 기존 음소 인식에서 주로 사용하는 스펙트로그램과 MFCC 특성의 성능과 비교하여 제안 특성의 음소 인식 가능성을 검증하였다. 각 특성을 비슷한 차원으로 맞추고 동일한 분류기로 음소 인식 성능을 측정하여 공정한 성능 비교가 되도록 하였다. 음소 인식을 위한 분류기는 3개의 은닉층을 가지는 DNN이며, 은닉 뉴런의 수는 2000, 1000, 1000이다. 은닉층에는 ReLU, 출력층에는 softmax 함수를 적용하였고, drop-out의 keep probability를 0.8로 설정하였으며, Adam을 사용해 DNN을 학습하였다[10].

성능 평가는 TIMIT 데이터 세트를 사용하여 진행하였다. 음운학에 따라 지정된 기존 61개의 음소는 혼동되기 쉬우며 불필요한 분류가 존재한다. 따라서 61개의 음소들을 효율적으로 39개로 재구성하였다[11]. 462명의 화자가 녹음한 training set을 학습 데이터로, 50명의 화자가 녹음한 development set을 검증 데이터로 사용하였다. 최종 성능은 development set과 중복되지 않으며, 24명의 화자가 녹음한 core test set을 사용하여 평가하였다[2]. 화자와 상관없이 인식하는 화자독립 (speaker-independent) 조건으로 성능을 평가하였다.

표 1은 서브 프레임 개수 K에 따른 제안하는 특성의 음소 인식 정확도이다. 서브 프레임의 길이를 줄여갈수록 성능이 높아지다가 일정 길이 이하로 줄이면 성능이 다시 하락하는 것을 확인하였다. 이는 적당한 길이의 서브 프레임이 음성 신호의 시간적 구조를 잘 모델링하는 것을 의미한다. 표 1의 결과를 바탕으로 제안 특성에서 K = 10으로 설정하고 126개 특성에 대하여 성능을 평가한다.

표 1. 서브 프레임 개수 K에 따른 음소 인식 정확도
Table 1. Recognition accuracy as a function of the number of sub-frames, K

K	Sub-frame length (ms)	Number of features	Accuracy (%)
1	25	99	63.91
2	12.5	102	65.05
4	6.25	108	65.06
8	3.125	120	65.07
10	**2.5**	**126**	**65.26**
20	1.25	156	64.89
40	0.625	216	64.84

성능 비교를 위한 스펙트로그램 특성과 MFCC 특성도 제안 특성과 동일하게 25 ms 단위로 추출하고 25 ms마다 음소를 인식한다. 먼저, 200 bin의 주파수 해상도를 가진 스펙트로그램을 얻은 후, 5 bin 단위로 에너지를 구하여 40 밴드 스펙트로그램 기반의 에너지 특성을 얻는다. 이후 40 밴드 스펙트로그램 특성의 1차 시간 미분과 2차 시간 미분을 더하여 총 120개 스펙트로그램 특성을 구한다. 다음, 푸리에 변환을 통해 25 ms 단위의 파워 스펙트럼을 구하고 40 밴드의 Mel-filterbank 에너지를 기반으로 discrete cosine transform (DCT)를 실행하여 40개의 MFCC를 구하고, 40개 MFCC의 1차 시간 미분과 2차 시간 미분을 더하여 총 120개 MFCC 특성을 구한다.

표 2는 스펙트로그램 특성, MFCC 특성, 제안하는 특성의 음소 class별 인식 정확도를 나타낸다. 표 2의 음소 class는 평균 음소 길이가 짧은 순으로 표기하였다[12]. 파찰음 (affricate), 폐쇄음 (stops), 마찰음 (fricative)으로 구성된 장애음 (obstruent)은 구강 통로가 폐쇄되거나 마찰이 생겨서 나는 소리를 뜻하며, 다른 음소 class에 비해 평균 음소 길이가 짧다[12,13]. 반면, 비음 (nasals), 반모음 (glides), 모음

표 2. 스펙트로그램, MFCC, 제안하는 특성의 음소 class별 인식 정확도
Table 2. Recognition accuracy of spectrogram, MFCC and proposed features for phoneme class

Phoneme class		Phonemic length (ms)	Accuracy (%)		
			Spectro-gram	MFCC	Proposed
Obstruent	**Stops**	**20.25**	**49.83**	**56.65**	**57.74**
	Affricate	**28.50**	**34.29**	**40.71**	**41.11**
	Fricative	**34.63**	**66.16**	**70.14**	**70.78**
Sonorant	**Glides**	**41.57**	**54.69**	**56.68**	**56.73**
	Nasals	41.83	57.95	64.14	59.44
	Vowels	44.75	54.37	55.49	52.34
Others		-	92.90	92.77	92.23
Total		-	65.56	67.74	65.26

(vowel)으로 구성된 공명음 (sonorant)은 성대를 떨게 한 공기가 비강이나 구강으로 흘러 나갈 때 덜 막혀 울리는 소리로, 평균 음소 길이가 길다[12,13]. 표 2에서 보듯이 높은 시간 해상도를 가지는 제안 특성은 길이가 짧은 장애음에 대해 프레임 기반의 스펙트로그램과 MFCC 특성보다 우수한 성능을 제공함을 확인할 수 있다. 제안하는 특성이 기존 특성보다 95개의 장애음을 잘 분류하였고, 높은 성능을 가지는 것을 확인하였다. 반면, 평균 음소 길이가 긴 공명음에 대해서는 제안하는 특성이 기존 특성보다 낮은 성능을 가진다. TIMIT 데이터 세트는 묵음 (silence)을 39개 음소 중 'etc'로 정의하며, 음소 class로는 others로 분류한다. 제안하는 특성은 others에 대하여 스펙트로그램과 MFCC 특성에 비해 낮은 성능을 가지고, 다른 음소 class에 비해 others의 빈도가 높으므로 others에서의 낮은 성능이 전체 인식 성능을 하락시키는 요인으로 작용한다. 결론적으로, 청각 기관을 모델링하는 새로운 접근법에 따라 구한 특성을 사용하여 음소 인식이 가능하고 특히 짧은 음소에 대하여 성능이 우수한 것을 확인할 수 있다.

그림 6은 제안하는 특성의 음소 인식 정확도에 대한 혼동행렬이다. 5% 미만은 표기하지 않았으며, 39개로 재구성한 음소별로 나타내었다. 음소 'uh'에 대해서 정확도 7%로 낮은 성능을 보이고 있으며, 이는 음소 'uh'가 데이터 세트에서 낮은 비중을 차지하여 다른 음소들에 비해 심층 신경망에 충분히 학습되지 않았기 때문이다. 또한 많은 음소들을 묵음이 포함되어 있는 'etc'로 잘못 분류한 것을 확인할 수 있으며, 이는 'etc'가 데이터 세트에서 높은 비중을 차지해 학습의 불균형을 발생시켰기 때문이다.

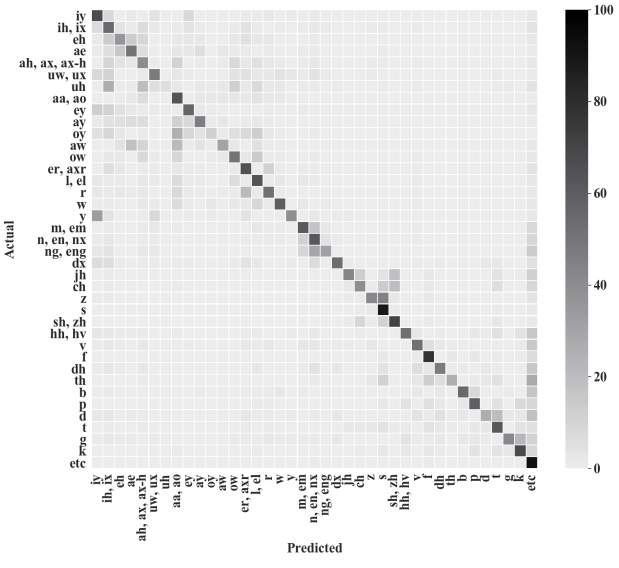


그림 6. 제안하는 특성의 음소 인식 혼동행렬
Fig. 6. Confusion matrix of proposed features

Ⅳ. 결 론

본 논문에서는 청각 기관의 동작을 모델링 하는 스파이크그램을 구하고 이를 기반으로 음소 인식을 위한 새로운 특성을 추출하는 방법을 제안하였다. 음성 신호의 스파이크그램을 주파수와 시간 축으로 분석하여 주파수 기반 특성과 시간 기반 특성을 추출하고, 이에 대한 1차 시간 미분과 2차 시간 미분을 추가하여 최종 126개 음성 특성을 구한다. 제안한 특성은 샘플 단위의 시간 해상도를 가지는 스펙트로그램으로부터 구하므로, 프레임 단위로 구하는 기존의 MFCC 특성에 비해 높은 시간 해상도를 가진다. 제안한 특성을 사용하면 짧은 음소에 대하여 MFCC 특성보다 우수한 성능을 가지는 것을 확인하였고, 이를 통해 청각 동작을 기반으로 구한 새로운 음성 특성을 사용하여 음소 인식이 가능한 것을 확인하였다. 추가 연구를 통하여 스파이크그램과 타이밍 기반의 심층 신경망을 이용한 음소 인식기에 대하여 연구하고 성능을 향상시킬 계획이다.

참 고 문 헌 (References)

[1] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer Publishing Company, Incorporated, 2014.

[2] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, Vol. 22, No. 10, pp. 1533-1545, Oct. 2014, doi:10.1109/TASLP.2014. 2339736.

[3] E. Smith and M. Lewicki, "Efficient Auditory Coding," *Nature*, Vol. 439, No. 7079, pp. 978-982, Feb. 2006, doi:10.1038/nature04485.

[4] W.-J. Jang, H.-W. Yun, S.-H. Shin and H. Park, "Music genre classification using spikegram and deep neural network," *J. of Broadcast Engineering*, Vol. 22, No. 6, pp. 693-701, Nov. 2017, doi:10.5909/JBE. 2017.22.6.693.

[5] S.-H. Shin, H.-W. Yun, W.-J. Jang and H. Park, "Extraction of acoustic

features based on auditory spike code and its application to music genre classification," *IET Signal Processing*, Vol. 13, No. 2, pp. 230-234, Apr. 2019, doi:10.1049/iet-spr.2018.5158.

[6] G. Mather, *Foundations of Perception*, Psychology Press, 2006.

[7] M. Slaney, "*An Efficient Implementation of the Patterson - Holdsworth Auditory Filter Bank,*" Apple Computer Technical Report #35, 1993.

[8] J. Tropp and A. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Trans. on Information Theory*, Vol. 53, No. 12, Dec. 2007, doi:10.1109/TIT. 2007.909108.

[9] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall, 2001.

[10] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, The MIT Press, Cambridge and London, 2016.

[11] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Audio, Speech, Lang. Process.*, Vol. 37, No. 11, pp. 1641 – 1648, Nov. 1989, doi:10.1109/29. 46546.

[12] N. Faraji, S. M. Ahadi and H. Sheikhzadeh, "Sequential method for speech segmentation based on Random Matrix Theory," *IET Signal Processing*, Vol. 7, No. 7, pp. 625-633, Sept. 2013, doi:10.1049/iet-spr. 2011.0471.

[13] P. Ladefoged and I. Maddieson. *The Sounds of the World's Languages*. Oxford, OX, UK: Blackwell Publishers, 1996.

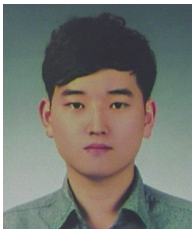
저 자 소 개

한 석 현

- 2019년 2월 : 광운대학교 전자공학과 공학사
- 2019년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0001-8871-5403
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

김 재 원

- 2019년 2월 : 광운대학교 전자공학과 학사
- 2019년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0002-6496-842X
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝



안 순 호

- 2019년 2월 : 광운대학교 전자공학과 학사
- 2019년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0001-9482-3478
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

저 자 소 개

신 성 현

- 2016년 2월 : 광운대학교 전자공학과 공학사
- 2016년 3월 ~ 현재 : 광운대학교 전자공학과 석박사통합과정
- ORCID : https://orcid.org/0000-0002-2343-8983
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

박 호 종

- 1986년 2월 : 서울대학교 전자공학과 공학사
- 1987년 12월 : Univ. of Wisconsin-Madison 공학석사
- 1993년 5월 : Univ. of Wisconsin-Madison 공학박사
- 1993년 9월 ~ 1997년 8월 : 삼성전자 선임연구원
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : https://orcid.org/0000-0003-1600-6610
- 주관심분야 : 오디오/음성 신호처리, 3D 오디오, 음악정보처리