스파이크그램 기반의 주파수 및 시간 특성을 이용한 음소 인식

*한석현 김재원 안순호 신성현 박호종

광운대학교

*sah0322@naver.com

Phoneme Recognition using Temporal and Spectral Features based on Spikegram

*Han, Seokhyeon Kim, Jaewon An, Soonho Shin, Seonghyeon Park, Hochong

Kwangwoon University

요약

본 논문에서는 스파이크그램 기반의 주파수 및 시간 특성을 이용한 음소 인식 방법을 제안한다. 기존의 MFCC 특성은 프레임 단위의 평균 특성이기 때문에 시간 해상도가 낮고, 짧은 음소의 특성을 반영하기에는 어려움이 있다. 반면, 스파이크그램은 청각 모델을 기반으로 샘플 단위로 계산하기 때문에 높은 시간 해상도를 가진다. 고 해상도의 스파이크그램을 분석하면 음소 인식에 특화된 특성 벡터를 추출할 수 있다. 추출된 특성으로 심층 신경망을 학습시켜 음소 인식기를 구현하였고, TIMIT 데이터 세트로 성능을 평가하였다. 성능 평가를 통하여 스파이크그램 기반의 새로운 시간-주파수 특성을 사용하여 MFCC 특성과 유사한 성능의 음소 인식이 가능한 것을 확인하였다.

1. 서론

기존 음소 인식 시스템에서는 Mel-frequency cepstral coefficient (MFCC)를 이용한 방법이 우수한 성능을 가진다[1]. MFCC 특성을 추출하는 과정은 음성의 스펙트럼을 얻기 위해서 short time fourier transform (STFT)을 요구한다. 하지만 STFT는 프레임 단위의 평균 특성이기 때문에 시간 해상도가 낮고, 파찰음 (affricate), 폐쇄음 (stops)과 같은 짧은 음소의 특성을 반영하기에는 어려움이 있다.

본 논문에서는 이러한 문제를 해결하기 위해 스파이크그램 기반의 특성 벡터를 추출한다[2]. 스파이크그램은 인간의 청각 시스템을 모델링한 감마톤의 equivalent rectangular bandwidth (ERB) 필터와 상관도가 높은 스파이크를 시간-주파수 축에 나열한다. 샘플 단위로 스파이크그램을 생성하므로 STFT보다 높은 시간 해상도를 가진다[2, 3].

스파이크그램 기반 새로운 시간-주파수 특성 추출 기술은 기존 MFCC 기반 음소 인식 시스템에 비해 시간 축에서 보다 높은 해상도를 가진다. 또한, 제안하는 특성은 기존 푸리에 변환 기반 MFCC 특성을 이용한 방법과 비슷한 성능을 제공하는 것을 확인하였다.

2. 제안하는 방법

스파이크를 추출하기 위해서는 감마톤의 ERB 필터를 커널로 사용한다. 스파이크그램을 생성하기 위한 특정 시간동안 각 채널 별로 커널과의 상관도를 구한다. 이중 가장 큰 상관도를 가지는 스파이크의 채널, 위치, 이득, 추출된 스파이크로 복원한 신호에 대한 SNR을 저장하며 스파이크를 추출한다. 스파이크를 추출하면서 감마톤 필터와 이득의 곱만큼 원 신호에서 빼서 분리해낸다. 이후 분리된 신호로부터 다시 가장 큰 상관도를 가지는 스파이크를 추출하는 것을 반복해 스파이크그램을 생성한다[2, 3]. 스파이크는 복원된 신호에 대한 PSNR이 50dB에 도달할 때까지 추출하였다. 그림 1은 음성 신호의 스펙트로그램과 스파이크그램을 비교한 것으로, 스파이크그램은 32개의 감마톤 필터뱅크를 사용하여 생성하였다.

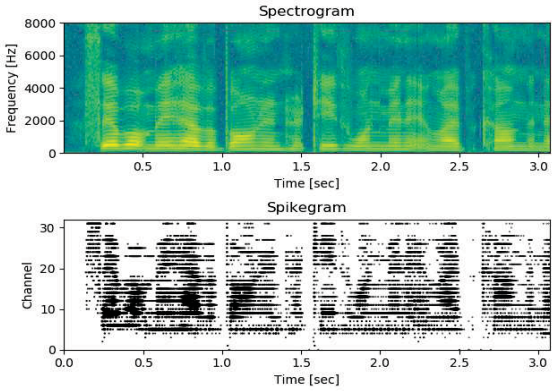

그림 1. 음성 신호의 스펙트로그램과 스파이크그램
위: 스펙트로그램, 아래: 스파이크그램

Fig. 1. Spectrogram and spikegram of speech signal
Top: spectrogram, Bottom: spikegram

그림 2는 스파이크그램으로부터 주파수 기반 특성과 시간 기반 특성을 추출하는 과정을 나타낸다. 샘플링 주파수가 16kHz인 음원에 대해, 25 ms (400 samples)의 프레임 단위로, 10 ms씩 옮기며 특성을 추출하였다. m은 채널 인덱스, i는 샘플의 시간 위치, g_i^m은 (i, m)에 위치한 스파이크의 이득을 의미한다. 채널 별로 생성된 스파이크의 이득을

합하여 32개 주파수 기반 특성 G_m을 생성한다. 다음, 2.5 ms의 서브 프레임을 10개 설정하였고, 각 서브 프레임에서 전 대역에서의 스파이크 이득을 합하여 10개 시간 기반 특성 T_n을 생성한다. 주파수 기반 특성과 시간 기반 특성을 더한 42개 특성 (static)의 1차 시간 미분 (delta)과 2차 시간 미분 (delta-delta)을 구해 총 126개 특성을 완성한다[1].

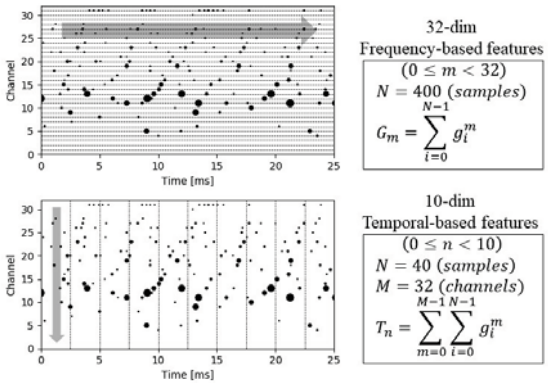

그림 2. 스파이크그램 기반 특성 추출 방법. 위. 32개 채널 이득 합 특성, 아래. 10개 서브 프레임 이득 합 특성.

Fig. 2. Features extraction based on spikegram. Top: 32 channel features of gain sum, Bottom: 10 sub-frame features of gain sum.

3. 성능 평가

음성 신호의 프레임 (25ms)마다 제안한 스파이크그램 기반의 특성을 추출하여 하나의 음소를 인식하였다. 모든 특성 값은 각각 평균 0과 분산 1로 정규화 하였다[1].

음소 인식 심층 신경망의 성능 평가를 위해 TIMIT 데이터 세트를 사용하였다. 61개의 음소들은 39개로 재구성하여 사용하였다[4]. 630명의 화자가 다양한 사투리로 녹음한 문장들은 결과의 편향을 발생시키므로 사용하지 않았다. 462명의 화자가 녹음한 training set을 학습 데이터로, 50명의 화자가 녹음한 development set을 검증 데이터로 사용했다. 최종 성능은 development set과 중복되지 않으며, 24명의 화자가 녹음한 core test set을 사용하여 평가하였다[1].

음소를 인식하기 위한 심층 신경망은 3개의 은닉층을 가진 DNN이며, 은닉 뉴런의 수는 2000, 1000, 1000이다. 활성 함수는 ReLU이며, 출력층에는 softmax 함수를 적용하였다. Adam을 사용해 심층 신경망을 학습하였다[5].

표 1은 MFCC와 제안하는 특성을 사용한 경우의 정확도를 나타낸다. 40개 MFCC 특성의 1차 시간 미분과 2차 시간 미분을 더하여 총 120개 특성을 사용하였다. MFCC의 정확도는 67.74%이다. 제안하는 특성은 시간적으로 짧은 음소에 대해 상대적으로 잘 분류하며, 정확도는 65.06%이다.

표 1. MFCC와 제안하는 특성의 음소 인식 정확도

Table 1. Phoneme recognition accuracy for MFCC and proposed features

Features	Dimension	Accuracy (%)
MFCC	120	67.74
Proposed features	126	65.06

표 2는 MFCC와 제안하는 특성의 음소 class 인식 정확도를 나타낸다. 파찰음 (affricate)과 폐쇄음 (stops)은 다른 음소 class에 비해 평균 음소 길이가 짧다[6]. 높은 시간 해상도의 제안하는 특성은 프레임 단위의 평균 특성인 MFCC보다 짧은 신호를 분석하기에 유리하다. 제안하는 특성은 파찰음과 폐쇄음에 대해서 우수한 성능을 제공함을 확인할 수 있다.

표 2. MFCC와 제안하는 특성의 음소 class 인식 정확도

Table 2. Recognition accuracy of each phoneme class for MFCC and proposed features

Phoneme class	Features	
	MFCC	Proposed features
Affricate	**40.71**	**41.11**
Fricative	70.14	70.05
Nasals	64.14	58.99
Semi-vowels and Glides	56.68	56.50
Vowels	55.49	52.80
Stops	**56.65**	**57.50**
Others	92.77	92.42

4. 결론

본 논문에서는 스파이크그램과 심층 신경망 기반의 음소 인식 방법을 제안하였다. 스파이크그램에서 주파수 특성과 시간 특성을 추출하고 1차 시간 미분과 2차 시간 미분을 더하여 심층 신경망의 입력으로 사용하였다. 본 논문은 스파이크그램 기반 새로운 시간-주파수 특성 추출 기술을 제안하였고, 평균 길이가 짧은 음소 class에 대해 기존 MFCC 특성보다 높은 성능을 제공하는 것을 확인하였다.

감사의 글

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016R1D1A1B03930923).

참고문헌

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533-1545, Oct. 2014.

[2] 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.

[3] E. Smith and M. Lewicki, "Efficient Auditory Coding," *Nature*, vol.439, no.7079, pp. 978-982, Feb. 2006.

[4] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 37, no. 11, pp. 1641 - 1648, Nov. 1989.

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

[6] 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.